

ENCYCLOPEDIA OF

SMART MATERIALS

VOLUME 1 and VOLUME 2

Mel Schwartz

The *Encyclopedia of Smart Materials* is available Online at
www.interscience.wiley.com/reference/esm



A Wiley-Interscience Publication
John Wiley & Sons, Inc.

This book is printed on acid-free paper. ☺

Copyright © 2002 by John Wiley and Sons, Inc., New York. All rights reserved.

Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4744. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 605 Third Avenue, New York, NY 10158-0012, (212) 850-6011, fax (212) 850-6008, E-Mail: PERMREQ @ WILEY.COM.

For ordering and customer service, call 1-800-CALL WILEY.

Library of Congress Cataloging in Publication Data

Encyclopedia of smart materials / Mel Schwartz, editor-in-chief.

p. cm.

“A Wiley-Interscience publication.”

Includes index.

ISBN 0-471-17780-6 (cloth : alk. paper)

1. Smart materials—Encyclopedias. I. Schwartz, Mel M.

TA418 9.S62 E63 2002

620.1'1—dc21

2001056795

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

CONTRIBUTORS

- D. Michelle Addington**, *Harvard University, Cambridge, MA*, Architecture
- Yasuyuki Agari**, *Osaka Municipal Technical Research Institute, Jotoku, Osaka, Japan*, Polymer Blends, Functionally Graded
- U.O. Akpan**, *Martec Limited, Halifax, NS, Canada*, Vibration Control in Ship Structures
- Samuel M. Allen**, *Massachusetts Institute of Technology, Cambridge, MA*, Shape-Memory Alloys, Magnetically Activated Ferromagnetic Shape-Memory Materials
- J.M. Bell**, *Queensland University of Technology, Brisbane Qld*, Windows
- Yves Bellouard**, *Institut de Systèmes Robotiques Ecole Polytechnique Fédérale de Lausanne Switzerland*, Microrobotics, Microdevices Based on Shape-Memory Alloys
- Davide Bernardini**, *Università di Roma "La Sapienza", Rome, Italy*, Shape-Memory Materials, Modeling
- A. Berry**, *GAUS, University de Sherbrooke, Sherbrooke, Quebec, Canada*, Vibration Control in Ship Structures
- O. Besslin**, *GAUS, University de Sherbrooke, Sherbrooke, Quebec, Canada*, Vibration Control in Ship Structures
- Mahesh C. Bhardwaj**, *Second Wave Systems, Boalsburg, PA*, Nondestructive Evaluation
- Vivek Bharti**, *Pennsylvania State University, University Park, PA*, Poly(Vinylidene Fluoride) (PVDF) and Its Copolymers
- Rafael Bravo**, *Universidad del Zulia, Maracaibo, Venezuela*, Truss Structures with Piezoelectric Actuators and Sensors
- Christopher S. Brazel**, *University of Alabama, Tuscaloosa, Alabama*, Biomedical Sensing
- W.A. Bullough**, *University of Sheffield, Sheffield, UK*, Fluid Machines
- J. David Carlson**, *Lord Corporation, Cary, NC*, Magnetorheological Fluids
- Aditi Chattopadhyay**, *Arizona State University, Tempe, AZ*, Adaptive Systems, Rotary Wing Applications
- Peter C. Chen**, *Alexandria, VA*, Ship Health Monitoring
- Seung-Bok Choi**, *Inha University, Incheon, Korea*, Vibration Control
- D.D.L. Chung**, *State University of New York at Buffalo, Buffalo, NY*, Composites, Intrinsically Smart Structures
- Juan L. Cormenzana**, *ETSII | Polytechnic University of Madrid, Madrid, Spain*, Computational Techniques For Smart Materials
- Marcelo J. Dapino**, *Ohio State University, Columbus, OH*, Magnetostrictive Materials
- Jerry A. Darsey**, *University of Arkansas at Little Rock, Little Rock, AR*, Neural Networks
- Kambiz Dianatkah**, *Lennox Industries, Carrollton, TX*, Highways
- Mohamed Dokainish**, *McMaster University, Hamilton, Ontario, Canada*, Truss Structures with Piezoelectric Actuators and Sensors
- Sherry Draisey**, *Good Vibrations Engineering, Ltd, Nobleton, Ontario, Canada*, Pest Control Applications
- Michael Drake**, *University of Dayton Research, Dayton, OH*, Vibrational Damping, Design Considerations
- Thomas D. Dziubla**, *Drexel University, Philadelphia, PA*, Gels
- Hiroshi Eda**, *IBARAKI University, Nakanarusawa, Japan*, Giant Magnetostrictive Materials
- Shigenori Egusa** (Deceased), *Japan Atomic Energy Research Institute, Takasaki-shi, Gunma, Japan*, Paints
- Harold D. Eidson**, *Southwestern University, Georgetown, TX USA*, Fish Aquatic Studies
- Arthur J. Epstein**, *The Ohio State University, Columbus, OH*, Magnets, Organic/Polymer
- John S.O. Evans**, *University of Durham, Durham, UK*, Thermoresponsive Inorganic Materials
- Frank Filisko**, *University of Michigan, Ann Arbor, MI*, Electrorheological Materials
- Koji Fujita**, *Kyoto University, Sakyo-ku, Kyoto, Japan*, Triboluminescence, Applications in Sensors
- Takehito Fukuda**, *Osaka City University, Sumiyoshi-ku, Osaka, Japan*, Cure and Health Monitoring
- C.R. Fuller**, *Virginia Polytechnic Institute and State University, Blacksburg, VA*, Sound Control with Smart Skins
- I. Yu. Galaev**, *Lund University, Lund, Sweden*, Polymers, Biotechnology and Medical Applications
- David W. Galipeau**, *South Dakota State University, Brookings, SD*, Sensors, Surface Acoustic Wave Sensors
- L.B. Glebov**, *University of Central Florida, Orlando, FL*, Photochromic and Photo-Thermo-Refractive Glasses
- J.A. Güemes**, *Univ. Politecnica, Madrid, Spain*, Intelligent Processing of Materials (IPM)
- Andrew D. Hamilton**, *Yale University, New Haven, CT*, Gelators, Organic
- Tian Hao**, *Rutgers—The State University of New Jersey, Piscataway, NJ*, Electrorheological Fluids
- J.S. Harrison**, *NASA Langley Research Center, Hampton, VA*, Polymers, Piezoelectric
- Bradley R. Hart**, *University of California, Irvine, CA*, Molecularly Imprinted Polymers
- Alisa J. Millar Henrie**, *Brigham Young University, Provo, UT*, Magnetorheological Fluids
- Kazuyuki Hirao**, *Kyoto University, Sakyo-ku, Kyoto, Japan*, Triboluminescence, Applications in Sensors
- Wesley P. Hoffman**, *Air Force Research Laboratory, AFRL / PRSM, Edwards AFB, CA*, Microtubes
- J. Van Humbeeck**, *K.U. Leuven-MTM, Katholieke Universiteit Leuven, Heverlee, Belgium*, Shape Memory Alloys, Types and Functionalities
- Emile H. Ishida**, *INAX Corporation, Minatomachi, Tokoname, Aichi, Japan*, Soil-Ceramics (Earth), Self-Adjustment of Humidity and Temperature
- Tsuguo Ishihara**, *Hyogo, Prefectural Institute of Industrial Research Suma-ku, Kobe, Japan*, Triboluminescence, Applications in Sensors
- Yukio Ito**, *The Pennsylvania State University, University Park, PA*, Ceramics, Transducers
- Bahram Javidian**, *Rutgers University, Piscataway, NJ*, Ceramics, Piezoelectric and Electrostrictive
- Andreas Janshoff**, *Johannes-Gutenberg-Universität, Mainz, Germany*, Biosensors, Porous Silicon
- T.L. Jordan**, *NASA Langley Research Center, Hampton, VA*, Characterization of Piezoelectric Ceramic Materials
- George Kavarnos**, *Pennsylvania State University, University Park, PA*, Poly(Vinylidene Fluoride) (PVDF) and Its Copolymers
- Andrei Kholkin**, *Rutgers University, Piscataway, NJ*, Ceramics, Piezoelectric and Electrostrictive
- Jason S. Kiddy**, *Alexandria, VA*, Ship Health Monitoring
- L.C. Klein**, *Rutgers—The State University of New Jersey, Piscataway, NJ*, Electrochromic Sol-Gel Coatings
- T.S. Koko**, *Martec Limited, Halifax, NS, Canada*, Vibration Control in Ship Structures
- Tatsuro Kosaka**, *Osaka City University, Sumiyoshi-ku, Osaka, Japan*, Cure and Health Monitoring
- Joseph Kost**, *Ben-Gurion University of the Negev, Beer Sheva, ISRAEL*, Drug Delivery Systems
- D. Kranbuehl**, *College of William and Mary, Williamsburg, Virginia*, Frequency Dependent Electromagnetic Sensing (FDEMS)
- Smadar A. Lapidot**, *Ben-Gurion University of the Negev, Beer Sheva, Israel*, Drug Delivery Systems
- Manuel Laso**, *ETSII | Polytechnic University of Madrid, Madrid, Spain*, Computational Techniques For Smart Materials
- Christine M. Lee**, *Unilever Research US Edgewater, NJ*, Langmuir-Blodgett Films

- F. Rodriguez-Lence**, *EADS-CASA Getafe, Madrid, Spain*, Intelligent Processing of Materials (IPM)
- Malgorzata M. Lencka**, *OLI Systems, Inc. Morris Plains, NJ*, Intelligent Synthesis of Smart Ceramic Materials
- T.W. Lewis**, *University of Wollongong, Wollongong, Australia*, Conductive Polymers
- Fang Li**, *Tianjin University, Tianjin, China*, Chitosan-Based Gels
- Anthony M. Lowman**, *Drexel University, Philadelphia, PA*, Gels
- Daoqiang Lu**, *Institute of Technology, Atlanta, GA*, Electrically Conductive Adhesives for Electronic Applications
- Shijian Luo**, *Georgia Institute of Technology, Atlanta, GA*, Conductive Polymer Composites with Large Positive Temperature Coefficients
- L.A.P. Kane-Maguire**, *University of Wollongong, Wollongong, Australia*, Conductive Polymers
- A. Maignan**, *Laboratoire CRISMAT, ISMRA, CAEN Cedex, FRANCE*, Colossal Magneto-resistive Materials
- Arumugam Manthiram**, *The University of Texas at Austin, Austin, TX*, Battery Applications
- P. Masson**, *GAUS, University de Sherbrooke, Sherbrooke, Quebec, Canada*, Vibration Control in Ship Structures
- Hideaki Matsubara**, *Atsuta-ku, Nagoya, Japan*, Self-diagnosing of Damage in Ceramics and Large-Scale Structures
- J.P. Matthews**, *Queensland University of Technology, Brisbane Qld, Windows*
- B. Mattiasson**, *Lund University, Lund, Sweden*, Polymers, Biotechnology and Medical Applications
- Raymond M. Measures**, *Ontario, Canada*, Fiber Optics, Bragg Grating Sensors
- Rosa E. Meléndez**, *Yale University, New Haven, CT*, Gelators, Organic
- J.M. Menendez**, *EADS-CASA Getafe, Madrid, Spain*, Intelligent Processing of Materials (IPM)
- Zhongyan Meng**, *Shanghai University, Shanghai, People's Republic of China*, Actuators, Piezoelectric Ceramic, Functional Gradient
- Joel S. Miller**, *University of Utah, Salt Lake City, UT*, Magnets, Organic/Polymer; Spin-Crossover Materials
- Nezih Mrad**, *Institute for Aerospace Research, Ottawa, Ontario, Canada*, Optical Fiber Sensor Technology: Introduction and Evaluation and Application
- Rajesh R. Naik**, *Wright-Patterson Air Force Base, Dayton, Ohio*, Biomimetic Electromagnetic Devices
- R.C. O'Handley**, *Massachusetts Institute of Technology, Cambridge, MA*, Shape-Memory Alloys, Magnetically Activated Ferromagnetic Shape-Memory Materials
- Yoshiki Okuhara**, *Atsuta-ku, Nagoya, Japan*, Self-diagnosing of Damage in Ceramics and Large-scale Structures
- Christopher O. Oriakhi**, *Hewlett-Packard Company, Corvallis, OR*, Chemical Indicating Devices
- Z. Ounaies**, *ICASE/NASA Langley Research Center, Hampton, VA*, Characterization of Piezoelectric Ceramic Materials; Polymers, Piezoelectric
- Thomas J. Pence**, *Michigan State University, East Lansing, MI*, Shape-Memory Materials, Modeling
- Darryll J. Pines**, *University of Maryland, College Park, MD*, Health Monitoring (Structural) Using Wave Dynamics
- Jesse E. Purdy**, *Southwestern University, Georgetown, TX*, Fish Aquatic Studies
- Jinhao Qiu**, *Tohoku University Sendai, Japan*, Biomedical Applications
- John Rajadas**, *Arizona State University, Tempe, AZ*, Adaptive Systems, Rotary Wing Applications
- Carolyn Rice**, *Cordis-NDC, Fremont, CA*, Shape Memory Alloys, Applications
- R. H. Richman**, *Daedalus Associates, Mountain View, CA*, Power Industry Applications
- Richard E. Riman**, *Rutgers University, Piscataway, NJ*, Intelligent Synthesis of Smart Ceramic Materials
- Paul Ross**, *Alexandria, VA*, Ship Health Monitoring
- Ahmad Safari**, *Rutgers University, Piscataway, NJ*, Ceramics, Piezoelectric and Electrostrictive
- Daniel S. Schodek**, *Harvard University, Cambridge, MA*, Architecture
- Jeffrey Schoess**, *Honeywell Technology Center, Minneapolis, MN*, Sensor Array Technology, Army
- Johannes Schweiger**, *European Aeronautic Defense and Space Company, Military Aircraft Business Unit, Muenchen, Germany*, Aircraft Control, Applications of Smart Structures
- K.H. Searles**, *Oregon Graduate Institute of Science and Technology, Beaverton, OR*, Composites, Survey
- Kenneth J. Shea**, *University of California, Irvine, CA*, Molecularly Imprinted Polymers
- Songhua Shi**, *Institute of Technology, Atlanta, GA*, Flip-Chip Applications, Underfill Materials
- I.L. Skryabin**, *Queensland University of Technology, Brisbane Qld, Windows*
- N. Sponagle**, *DREA, Dartmouth, NS, Canada*, Vibration Control in Ship Structures
- R. Stalmans**, *Flexmet, Aarschot, Belgium*, Shape Memory Alloys, Types and Functionalities
- Dave S. Steinberg**, *Westlake Village, CA*, Vibrational Analysis
- Claudia Steinem**, *Universität Regensburg, Regensburg, Germany*, Biosensors, Porous Silicon
- Morley O. Stone**, *Wright-Patterson Air Force Base, Dayton, Ohio*, Biomimetic Electromagnetic Devices
- J. Stringer**, *EPRI, Palo Alto, CA*, Power Industry Applications
- A. Suleman**, *Instituto Superior Técnico, Lisbon, Portugal*, Adaptive Composite Systems: Modeling and Applications
- J. Szabo**, *DREA, Dartmouth, NS, Canada*, Vibration Control in Ship Structures
- Daniel R. Talham**, *University of Florida, Gainesville, FL*, Langmuir-Blodgett Films
- Katsuhisa Tanaka**, *Kyoto Institute of Technology, Sakyo-ku, Kyoto, Japan*, Triboluminescence, Applications in Sensors
- Mami Tanaka**, *Tohoku University Sendai, Japan*, Biomedical Applications
- Brian S. Thompson**, *Michigan State University, East Lansing, MI*, Composites, Future Concepts
- Harry Tuller**, *Massachusetts Institute of Technology, Cambridge, MA*, Electroceramics
- Kenji Uchino**, *The Pennsylvania State University, University Park, PA*, Ceramics, Transducers
- Eric Udd**, *Blue Road Research, Fairview, Oregon*, Fiber optics, Theory and Applications
- Anthony Faria Vaz**, *Applied Computing Enterprises Inc., Mississauga, Ontario, Canada & University of Waterloo, Waterloo, Ontario, Canada*, Truss Structures with Piezoelectric Actuators and Sensors
- A.G. Vedeshwar**, *University of Delhi, Delhi, India*, Optical Storage Films, Chalcogenide Compound Films
- Aleksandra Vinogradov**, *Montana State University, Bozeman, MT*, Piezoelectricity in Polymers
- G.G. Wallace**, *University of Wollongong, Wollongong, Australia*, Conductive Polymers
- Lejun Wang**, *Institute of Technology, Atlanta, GA*, Flip-Chip Applications, Underfill Materials
- Zhong L. Wang**, *Georgia Institute of Technology, Atlanta, GA*, Smart Perovskites
- Phillip G. Wapner**, *ERC Inc., Edwards AFB, CA*, Microtubes
- Zhongguo Wei**, *Dalian University of Technology, Dalian, China*, Hybrid Composites
- Michael O. Wolf**, *The University of British Columbia, Vancouver, British Columbia, Canada*, Poly(P-Phenylenevinylene)
- C.P. Wong**, *Georgia Institute of Technology, Atlanta, GA*, Conductive Polymer Composites with Large Positive Temperature Coefficients; Electrically Conductive Adhesives for Electronic Applications
- C.P. Wong**, *Georgia Institute of Technology, Atlanta, GA*, Flip-Chip Applications, Underfill Materials
- Chao-Nan Xu**, *National Institute of Advanced Industrial Science and Technology (AIST), Tosu, Saga, Japan*, Coatings

Hiroaki Yanagida, *University of Tokyo, Mutuno, Atsuta-ku, Nagoya, Japan*, Environmental and People Applications; Ken-Materials; Self-diagnosing of Damage in Ceramics and Large-scale Structures

Dazhi Yang, *Dalian University of Technology, Dalian, China*, Hybrid Composites

Kang De Yao, *Tianjin University, Tianjin, China*, Chitosan-Based Gels

Yu Ji Yin, *Tianjin University, Tianjin, China*, Chitosan-Based Gels

Rudolf Zentel, *University of Mainz, Mainz, Germany*, Polymers, Ferroelectric liquid Crystalline Elastomers

Q.M. Zhang, *Pennsylvania State University, University Park, PA*, Poly(Vinylidene Fluoride) (PVDF) and Its Copolymers

Feng Zhao, *Tianjin University, Tianjin, China*, Chitosan-Based Gels

Libo Zhou, *IBARAKI University, Nakanarusawa, Japan*, Giant Magnetostrictive Materials

Xinhua Zhu, *Nanjing University, Nanjing, People's Republic of China*, Actuators, Piezoelectric Ceramic, Functional Gradient

ENCYCLOPEDIA OF SMART MATERIALS

Editor-in-Chief

Mel Schwartz

Editorial Board

Alok Das

Air Force Research Laboratory/VSD

US Air Force

Michael L. Drake

University of Dayton Research Institute

Caroline Dry

Natural Process Design

School of Architecture

University of Illinois

Lisa C. Klein

Rutgers—The State University of New Jersey

S. Eswar Prasad

Sensor Technology Limited

Buddy D. Ratner

University of Washington

Craig A. Rogers

James Sirkis

CiDRA Corporation

Junji Tani

Tohoku University

C.P. Wong

Georgia Institute of Technology

Editorial Staff

Vice-President, STM Books: **Janet Bailey**

Vice-President and Publisher: **Paula Kepos**

Executive Editor: **Jacqueline I. Kroschwitz**

Director, Book Production and Manufacturing:

Camille P. Carter

Managing Editor: **Shirley Thomas**

Editorial Assistant: **Surlan Murrell**

PREFACE

The Encyclopedia of Smart Materials (ESM) contains the writings, thoughts, and work of many of the world's foremost people (scientists, educators, chemists, engineers, laboratory and innovative practitioners) who work in the field of smart materials. The authors discuss theory, fundamentals, fabrication, processing, application, applications and uses of these very special, and in some instances rare, materials.

The term "smart structure" and "smart materials" are much used and abused.

Consideration of the lexicology of the English language should provide some guidelines, although engineers often forget the dictionary and evolve a language of their own. Here is what the abbreviated Oxford English Dictionary says:

- *Smart*: severe enough to cause pain, sharp, vigorous, lively, brisk . . . clever, ingenious, showing quick wit or ingenuity . . . selfishly clever to the verge of dishonesty;
- *Material*: matter from which a thing is made;
- *Structure*: material configured to do mechanical work . . . a thing constructed, complex whole.

The concept of "smart" or "intelligent" materials, systems, and structures has been around for many years. A great deal of progress has been made recently in the development of structures that continuously and actively monitor and optimize themselves and their performance through emulating biological systems with their adaptive capabilities and integrated designs. The field of smart materials is multidisciplinary and interdisciplinary, and there are a number of enabling technologies—materials, control, information processing, sensing, actuation, and damping—and system integration across a wide range of industrial applications.

The diverse technologies that make up the field of smart materials and structures are at varying stages of commercialization. Piezoelectric and electrostrictive ceramics, piezoelectric polymers, and fiber-optic sensor systems are well-established commercial technologies, whereas micro-machined electromechanical systems (MEMS), magnetostrictive materials, shape memory alloys (SMA) and polymers, and conductive polymers are in the early stages of commercialization. The next wave of smart technologies will likely see the wider introduction of chromogenic materials and systems, electro- and magneto-rheological fluids, and biometric polymers and gels.

Piezoelectric transducers are widely used in automotive, aerospace, and other industries to measure vibration and shock, including monitoring of machinery such as pumps and turbomachinery, and noise and vibration control. MEMS sensors are starting to be used where they offer advantages over current technologies, particularly for static or low frequency measurements. Fiber-optic systems are increasingly being used in hazardous or difficult

environments, such as at high temperatures or in corrosive atmospheres.

Automotive companies are investigating the use of smart materials to control vehicles in panels, such as damping vibration in roof panels, engine mounts, etc. Aerospace applications include the testing of aircraft and satellites for the strenuous environments in which they are used, both in the design phase and in use, as well as for actuators or devices to react to or control vibrations, or to change the shape of structures.

In civil engineering, especially in earthquake-prone areas, a number of projects are under way to investigate the use of materials such as active composites to allow support systems of bridges (and the like) to handle such shocks without catastrophic failure. These materials can be used in many structures that have to withstand severe stresses, such as offshore oil rigs, bridges, flyovers, and many types of buildings.

The ESM will serve the rapidly expanding demand for information on technological developments of smart materials and devices. In addition to information for manufacturers and assemblers of smart materials, components, systems, and structures, ESM is aimed at managers responsible for technology development, research projects, R&D programs, business development, and strategic planning in the various industries that are considering these technologies. These industries, as well as aerospace and automotive industries, include mass transit, marine, computer-related and other electronic equipment, as well as industrial equipment (including rotating machinery, consumer goods, civil engineering, and medical applications).

Smart material and system developments are diversified and have covered many fields, from medical and biological to electronic and mechanical. For example, a manufacturer of spinal implants and prosthetic components has produced a prosthetic device that dramatically improves the mobility of leg amputees by closely recreating a natural gait.

Scientists and doctors have engineered for amputees a solution with controllable magneto-rheological (MR) technology to significantly improve stability, gait balance, and energy efficiency for amputees. Combining electronics and software, the MR-enabled responsiveness of the device is 20 times faster than that of the prior state-of-the-art devices, and therefore allows the closest neural human reaction time of movement for the user. The newly designed prosthetic device therefore more closely mimics the process of natural thought and locomotion than earlier prosthetic designs.

Another example is the single-axis accelerometer/sensor technology, now available in the very low-profile, surface-mount LCC-8 package. This ceramic package allows users to surface-mount the state-of-the-art MEMS-based sensors. Through utilization of this standard packaging profile, one is now able to use the lowest

profile, smallest surface-mountable accelerometer/sensor currently available. This sensor/accelerometer product technology offers on-chip mixed signal processing, MEMS sensor, and full flexibility in circuit integration on a single chip. Features of the sensor itself include continuous self-test as well as both ratiometric and absolute output. Other sensor attributes include high long-term reliability resulting from no moving parts, which eliminates striction and tap-sensitive/sticky quality issues.

Application areas include automotive, computer devices, gaming, industrial control, event detection, as well as medical and home appliances. In high-speed trains traveling at 200 km/h, a droning or rumbling is often heard by passengers. Tiny imperfections in the roundness of the wheels generate vibrations in the train that are the source of this noise. In addition to increasing the noise level, these imperfect wheels lead to accelerated material fatigue. An effective countermeasure is the use of actively controlled dampers. Here a mechanical concept—a specific counterweight combined with an adjustable spring and a powerful force-actuator—is coupled with electronic components. Simulations show what weights should be applied at which points on the wheel to optimally offset the vibrations. Sensors detect the degree of vibration, which varies with the train's speed. The electronic regulator then adjusts the tension in the springs and precisely synchronizes the timing and the location of the counter-vibration as needed. Undesirable vibration energy is diffused, and the wheel rolls quietly and smoothly. In this way, wear on the wheels is considerably reduced.

The prospects of minimized material fatigue, a higher level of travel comfort for passengers, and lower noise emissions are compelling reasons for continuing this development.

Novel composite materials discovered by researchers exhibit dramatically high levels of magneto-resistance, and have the potential to significantly increase the performance of magnetic sensors used in a wide variety of important technologies, as well as dramatically increase data storage in magnetic disk drives. The newly developed extraordinary magnetoresistance (EMR) materials can be applied in the read heads of disk drives, which, together with the write heads and disk materials, determine the overall capacity, speed, and efficiency of magnetic recording and storage devices. EMR composite materials will be able to respond up to 1000 times faster than the materials used in conventional read heads, thus significantly advancing magnetic storage technology and bringing the industry closer to its long-range target of a disk drive that will store a terabit (1000 gigabits) of data per square inch.

The new materials are composites of nonmagnetic, semiconducting, and metallic components, and exhibit an EMR at room temperature of the order of 1,000,000% at high fields. More importantly, the new materials give high values of room-temperature magnetoresistance at low and moderate fields. Embedding a highly conducting metal, such as gold, into a thin disc of a nonmagnetic semiconductor, such as indium antimonide, boosts the magnetoresistance, and offers a number of other advantages. These include very high thermal stability, the potential for much

lower manufacturing costs, and operation at speeds up to 1000 times higher than sensors fabricated from magnetic materials.

Envisioned are numerous other applications of EMR sensors in areas such as consumer electronics, wireless telephones, and automobiles, which utilize magnetic sensors in their products. Future EMR sensors will deliver dramatically greater sensitivity, and will be considerably less expensive to produce.

Another recent development is an infrared (IR) gas sensor based on MEMS manufacturing techniques. The MEMS IR gas SensorChip will be sensitive enough to compete with larger, more complex gas sensors, but inexpensive enough to penetrate mass-market applications. MEMS technology should simplify the construction of IR gas sensors by integrating all the active functions onto a single integrated circuit.

Tiny electronic devices called “smart dust,” which are designed to capture large amounts of data about their surroundings while floating in the air, have been developed. The project could lead to wide array of applications, from following enemy troop movements and detecting missiles before launch to detecting toxic chemicals in the environments and monitoring weather patterns. The “Smart Dust” project aims to create massively distributed sensor networks, consisting of hundreds to many thousands of sensor nodes, and one or more interrogators to query the network and read out sensor data. The sensor nodes will be completely autonomous, and quite small. Each node will contain a sensor, electronics, power supply, and communication hardware, all in a volume of 1 mm³.

The idea behind “smart dust” is to pack sophisticated sensors, tiny computers, and wireless communications onto minuscule “motes” of silicon that are light enough to remain suspended in air for hours at a time. As the motes drift on the wind, they can monitor the environment for light, sound, temperature, chemical composition, and a wide range of other information, and transmit the data back to a distant base station. Each mote of smart dust is composed of a number of MEMS, wired together to form a simple computer. Each mote contains a solar cell to generate power, sensors that can be programmed to look for specific information, a tiny computer that can store the information and sort out which data are worth reporting, and a communicator that enables the mote to be interrogated by the base unit. The goals are to explore the fundamental limits to the size of autonomous sensor platforms, and the new applications which become possible when autonomous sensors can be made on a millimeter scale.

Laser light can quickly and accurately flex fluid-swollen plastics called polymer gels. These potential polymer muscles could be used to power robot arms, because they expand and contract when stimulated by heat or certain chemicals. Gel/laser combinations could find applications ranging from actuators to sensors, and precisely targeted laser light could allow very specific shape changes. Polymer gels have been made to shrink and swell in a fraction of a second. Targeting laser light at the center of a cylinder made of *N*-isopropylacrylamide pinches together the tube's edges to form a dumb-bell shape. The cylinder

returns to its original shape when the laser is switched off. This movement is possible because in polymer gels, the attractive and repulsive forces between neighboring molecules are finely balanced. Small chemical and physical changes can disrupt this balance, making the whole polymer to violently expand or collapse. Also it has been shown that radiation forces from focused laser light disturb this delicate equilibrium, and induce a reversible phase transition. Repeated cycling did not change the thresholds of shrinkage and expansion; also, the shrinking is not caused by temperature increases accompanying the laser radiation.

The field of smart materials offers enormous potential for rapid introduction and implementation in a wide range

of end-user sectors industries. Not only are the organizations involved in research and preliminary development keen to grow their markets in order to capitalize on their R&D investment, but other technologically aware companies are alerted to new business opportunities for their own products and skillsets.

The readers of this ESM will appreciate the efforts of a multitude of researchers, academia, and industry people who have contributed to this endeavor. The editor is thankful to Dr. James Harvey and Mr. Arthur Biderman for their initial efforts in getting the project off the ground and moving the program.

Mel Schwartz

Table of Contents

Preface	vii
Actuators to Architecture	
Actuators, Piezoelectric Ceramic, Functional Gradient	1
Introduction	1
Actuators	1
Piezoelectric Ceramics	2
Functionally Graded Materials	7
Summary	1
Acknowledgments	14
Bibliography	15
Adaptive Composite Systems: Modeling and Applications	16
Introduction	16
Actuators and Sensors	16
Adaptive Composite Modeling	18
Applications	20
Concluding Remarks	25
Bibliography	25
Adaptive Systems, Rotary Wing Applications	28
Introduction	28
Active / Passive Control of Structural Response	29
Passive / Active Control of Damping	30
Trailing Edge Flaps	32
Servoflap	34
Active Twist	35
Modeling	37
Future Directions	39
Bibliography	40
Aircraft Control, Applications of Smart Structures	42
Introduction	42
Smart Structures for Flight in Nature	43
General Remarks on Aspects of Aircraft Design	44
Traditional Active or Adaptive Aircraft Control Concepts	44
The Range of Active Structures and Materials Applications in Aeronautics	45
Aircraft Structures	45
Smart Materials for Active Structures	47
The Role of Aeroelasticity	47
Overview of Smart Structural Concepts for Aircraft Control	50
Quality of the Deformations	54
Achievable Amount of Deformation and Effectiveness of Different Active Aeroelastic Concepts	55
Need for Analyzing and Optimizing the Design of Active Structural Concepts	56
Summary, Conclusions, and Predictions	57
Bibliography	58

Architecture	59
Introduction	59
Material Considerations in Architecture	59
Traditional Material Classification in Architecture	59
Traditional Technology Classification in Architecture	59
Proposed Classification System for Smart Materials	60
Taxonomy of Smart Materials	60
Categories of Applications	61
Future Design Approaches in Architecture	65
Bibliography	67
 Battery to Biosensors	
Battery Applications	68
Introduction	68
Electrochemical Concepts	68
Batteries	70
Smart Batteries	72
Concluding Remarks	81
Acknowledgments	81
Bibliography	81
 Biomedical Applications	82
Introduction	82
Properties of SMAs for Biomedical Applications	83
Examples of Biomedical Applications	84
Current Biomedical Applications of SMA	88
Current Biomedical Applications of Piezoelectric Materials	91
Bibliography	94
 Biomedical Sensing	95
Introduction	95
Medical, Therapeutic, and Diagnostic Applications of Biosensors	98
Polymers as Electrode Coatings and Biosensor Mediators	99
Immobilization Techniques and Materials	100
Smart Polymers for Immobilization and Bioconjugate Materials	103
Biosensor Operation	103
Glucose Sensors	104
Other Analytes for Biological Sensing	107
Modes of Response in Smart Polymers	107
Molecular Imprinting	109
Possibilities for Future Development	109
Bibliography	109
 Biomimetic Electromagnetic Devices	112
Introduction	112
Biological Ultraviolet and Visible Systems	112
Biological Infrared Detection	115
Electromagnetic Applications of Biomimetic Research	119
Acknowledgments	120
Bibliography	120

Biosensors, Porous Silicon	121
Introduction	121
Historical Overview	122
Porous Silicon Formation	122
Characterization of Porous Silicon	122
Optical Properties of Porous Silicon	124
Functionalization of Porous Silicon Surfaces	127
Porous Silicon Chemosensors	129
Biosensor Applications of Porous Silicon	130
Conclusions	137
Bibliography	137
 Ceramics to Coatings	
Ceramics, Piezoelectric and Electrostrictive	139
Introduction	139
Piezoelectric and Electrostrictive Effects in Ceramic Materials	139
Measurements of Piezoelectric and Electrostrictive Effects	141
Common Piezoelectric and Electrostrictive Materials	143
Piezoelectric Composites	144
Applications of Piezoelectric / Electrostrictive Ceramics	146
Future Trends	147
Bibliography	147
 Ceramics, Transducers	148
Introduction	148
Piezoelectricity	149
Piezoelectric Materials	151
Applications of Piezoelectricity	154
Bibliography	161
 Characterization of Piezoelectric Ceramic Materials	162
Introduction	162
Piezoelectric Materials: History and Processing	162
Piezoelectric Constitutive Relationships	165
Piezoelectric Parameters: Definitions and Characterization	166
Conclusion	172
Acknowledgment	172
Bibliography	172
 Chemical Indicating Devices	173
Introduction	173
Chemical Indicating Devices Are Smart	174
Classification	174
General Operating Principles	174
Choice of Indicators	174
Ph Indicators	174
Indicator Materials	175
Temperature and Time-Temperature Indicators (TTI)	175
Anticounterfeiting and Tamper Indicator Devices	179
Conclusion	181
Bibliography	182

Chitosan-Based Gels	182
Introduction	182
Supramolecular Interactions and Gel Formation	182
Applications	185
Bibliography	188
Coatings	190
Introduction	190
Nondestructive ML from Alkaline Aluminates Doped with Rare-Earth Ions	191
Repeatable ML of Transition Metal Ions Doped Zinc Sulfide	195
Application of Smart Coating with ML Monitoring in Dynamic Stress and Impact Materials for Novel Stress Display	198
Bibliography	201
Colossal Magnetoiresistive to Cure and Health	
Colossal Magnetoiresistive Materials	202
Introduction	202
CMR in Hole-Doped $\text{Ln}_{0.7}\text{AE}_{0.3}\text{MnO}_3$ Perovskites	202
Origin of the CMR Effect: Manganese Mixed Valency and Double Exchange	203
Chemical Factors Governing CMR Properties	204
Charge Ordering in Perovskite Manganites	207
Other CMR Manganites	211
Conclusion	212
Bibliography	213
Composites, Future Concepts	214
Introduction	214
Historical Prologue	214
Biologically-Inspired Creativity in Engineering	217
Smart Materials and Structures: Current Noncommercial Technologies	217
Smart Materials and Structures: The Future	219
Concluding Comments	223
Composites, Intrinsically Smart Structures	223
Introduction	223
Cement-Matrix Composites for Smart Structures	223
Polymer-Matrix Composites for Smart Structures	233
Conclusion	240
Bibliography	240
Composites, Survey	243
Introduction	243
Engineering Materials	243
Composite Materials	245
Microscale Behavior	246
Mesoscale Behavior	250
Macroscale Behavior	258
Other Considerations	264
Bibliography	264

Computational Techniques For Smart Materials	265
Introduction	265
Smart Materials, Memory Effects, and Molecular Complexity	266
Continuum and Molecular Descriptions of Smart Materials	268
Smart Materials and Nonequilibrium Thermodynamics	271
Outlook	273
Notation	273
Bibliography	273
Conductive Polymer Composites with Large Positive Temperature Coefficients	274
Introduction	274
Basic Theory of Conductive Polymer Composites and PTC Behavior	275
Effect of Conductive Fillers on PTC Conductive Polymer	276
Effect of Polymer Matrix on PTC Behavior	277
Effect of Processing Condition and Additives	278
Application of PTC Conductive Polymer Composite	278
Summary	278
Bibliography	278
Conductive Polymers	279
Introduction	279
Synthesis and Properties	279
Chemical and Physical Stimuli	281
Actuators	285
Information Processing	285
Energy Conversion / Storage	286
Polymer Processing	286
Device Fabrication	287
Conclusions	288
Bibliography	288
Cure and Health Monitoring	291
Smart Monitoring System	291
Cure Monitoring	292
Health Monitoring	301
Bibliography	316
Drug Delivery to Environmental and People Applications	
Drug Delivery Systems	319
Introduction	319
Development of Controlled Drug Delivery	319
Pulsatile Systems	320
Self-Regulated Systems Mechanisms	322
Concluding Remarks	328
Bibliography	328
Electrically Conductive Adhesives for Electronic Applications	331
Introduction	331
Electrically Conductive Adhesives (ECAs)	331
Improvement of Electrical Conductivity of ICAs	332
Improvement of Contact Resistance Stability	332
Improvement of Impact Performance	334
High-Performance Conductive Adhesives	335
Bibliography	335

Electroceramics	337
Introduction	337
Electroceramics and Smart Systems	337
Electromechanical Actuators	338
Actuator Materials	339
Basic Relationships and Phenomena	339
Electrostrictive Behavior in Materials	342
Materials Systems	343
Thermal Systems	344
Smart Chemical Systems Applications	347
Optical Materials	351
Conclusions	355
Acknowledgments	355
Bibliography	355
Electrochromic Sol-Gel Coatings	356
Definition of Electrochromism	356
Conclusions	360
Acknowledgment	361
Bibliography	361
Electrorheological Fluids	362
Introduction	362
Dielectric Properties of Heterogeneous Systems	363
Experimental Facts	364
Theoretical Treatment	369
The Mechanism of the ER Effect	371
The Yield Stress Equation	372
Conclusion	375
Bibliography	375
Electrorheological Materials	376
Introduction	376
Background	377
Materials	379
Mechanical (Rheological) Properties of ER Materials	382
Mechanical Models	385
Theories of ER	385
Applications	388
Bibliography	390
General References	391
Environmental and People Applications	392
Introduction	392
Forum for Intelligent Materials	392
Frontier Ceramics Project	392
Bibliography	394
Fiber Optics to Frequency Dependent Electromagnetic Sensing	
Fiber Optics, Bragg Grating Sensors	395
Introduction	395
Bragg Grating Reflection	395
Fiber Bragg Gratings	398
Sensor Multiplexing	401
Optothermo-Mechanical Equations	402
Strain and Temperature Sensitivity	403
Measurements of Strain and Temperature	404

Fiber Optics, Bragg Grating Sensors	
Fiber Bragg Grating Sensor Demodulation	408
Fiber Bragg Grating Sensor Applications	412
Conclusions	414
Bibliography	414
Fiber Optics, Theory and Applications	415
Introduction	415
Fiber Optic Sensors for Smart Structures	415
Fiber Optic Smart Structure Applications	419
Summary	422
Acknowledgment	422
Bibliography	422
Fish Aquatic Studies	423
Introduction and Overview	423
Applications of Smart Materials and Smart Structures in Fish Aquatic Studies	424
Conclusions	436
Bibliography	437
Flip-Chip Applications, Underfill Materials	438
Introduction	438
Development of No-Flow Flip-Chip Underfills	441
Development of Novel Reworkable Underfills	444
Development of Molded (Tablet) Flip-Chip Underfills	446
Development of Wafer-Level-Applied Flip-Chip Underfills	446
Bibliography	447
Fluid Machines	448
Introduction	448
Intelligent Hydraulics	449
Smart Fluids	450
Flexible Machine Operation	451
ESF Controllers	451
Typical Application	453
Future Trends and Limitations	455
Bibliography	456
Frequency Dependent Electromagnetic Sensing (FDEMS)	456
Introduction	456
Background	456
Instrumentation	457
Theory	457
Relationship to Macroscopic Performance and Processing Properties	458
Experimental In Situ Monitoring of Molecular Mobility and Relationship to Changes in Processing Properties	459
Application to Autoclave Cure Monitoring of Viscosity In Situ in the Mold and Model Verification	460
Monitoring Resin Position During RTM	462
Smart Automated Cure Control	463
Smart Automated Cure of a Polyimide	464
Conclusions: Processing	466
Life Monitoring, a Smart Material	467
Life Monitoring - Results	467
Conclusions: Life Monitoring	469
Acknowledgments	469
Bibliography	470

Gelators to Giant Magnetostrictive Materials	
Gelators, Organic	471
Molecular Recognition and Supramolecular Materials	471
Intermolecular Interactions	472
Organogelation	474
Conclusion	489
Acknowledgement	489
Bibliography	489
Gels	490
Introduction	490
Structure and Properties of Hydrogels	491
Classifications	494
Applications	497
Bibliography	500
Giant Magnetostrictive Materials	503
Introduction	503
Giant Magnetostrictive Materials	504
GMM Manufacturing Process	508
Applications	510
Bibliography	519
Health Monitoring to Hybrid Composites	
Health Monitoring (Structural) Using Wave Dynamics	520
Introduction	520
Dereverberated Transfer Functions of Structural Elements	520
DTF Responses of Nonuniform Structures	525
Damage Detection Approach Based on DTF Response	533
Damage Detection in a Building Structure Using DTF	539
Summary and Conclusions	545
Acknowledgments	545
Bibliography	545
Highways	545
Introduction	545
Smart Materials	545
Objectives of Smart Highways	546
Smart Highways	548
Advanced Automobiles	548
Update on Smart Highway Projects Under Construction	550
Smart Highways in Japan	550
Summary	551
Acknowledgments	551
Bibliography	551
Hybrid Composites	551
Introduction	551
Shape Memory Alloy Fiber / Metal Matrix Composites	552
Shape Memory Alloy Fiber / Polymer Matrix Composites	552
SMA Particulate / Aluminum Matrix Composites	554
Ceramic Particulate / SMA Matrix Composites	555
Magnetic Particulate / SMA Matrix Composites	555
SMA / Si Heterostructures	556
SMA / Piezoelectric Heterostructures	556
SMA / Terfenol-D Heterostructures	557

Intelligent Processing to Langmuir-Blodgett Films	
Intelligent Processing of Materials (IPM)	559
The Concept of Intelligent Materials Processing	559
Intelligent Processing of Composite Materials	560
Intelligent Processing of Metallic Materials	567
Conclusions	568
Bibliography	568
Intelligent Synthesis of Smart Ceramic Materials	568
Introduction	568
Thermodynamic Model	569
Stability and Yield Diagrams	571
Validation and Applications of Thermodynamic Modeling	578
Conclusions	579
Bibliography	579
Ken-Materials	581
Introduction	581
Ken Materials Research Consortium	581
Significance of R & D in the New Millennium	581
Directions of Technology: Miniaturization, Enlargement, Integration, and Brevity	582
Examples of Ken Materials	582
Environmentally Related Materials	582
Avoiding the Spaghetti Syndrome of Technology	583
Bibliography	583
Langmuir-Blodgett Films	584
Introduction	584
History	584
Equipment	584
Langmuir Monolayers	585
Langmuir-Blodgett Films	587
Smart Materials Applications	589
Conclusion	590
Bibliography	590
Magnets to Microrobotics	
Magnets, Organic / Polymer	591
Introduction	591
V(TCNE) _x Y(Solvent) Room Temperature Magnets	594
M(TCNE) _{2.x} (CH ₂ Cl ₂) (M = Mn, Fe, Co, Ni) High Room Temperature Magnets	595
Hexacyanometallate Magnets	595
Uses of Organic / Polymeric Magnets	596
Acknowledgment	596
Bibliography	596
Magnetorheological Fluids	597
Introduction	597
Definition	597
History	598
Common Nomenclature	598
Material Choice	598
Basic Composition	598

Magnetorheological Fluids	
Theory	598
MR Properties	599
Applications	600
Bibliography	600
Magnetostrictive Materials	600
Introduction	600
Materials Overview	602
Physical Origin of Magnetostriction	603
Material Behavior	603
Linear Magnetostriction	606
Other Magnetostrictive Effects	608
Magnetostrictive Transducers	609
Concluding Remarks	618
Bibliography	618
Microrobotics, Microdevices Based on Shape-Memory Alloys	620
Introduction	620
Shape-Memory Alloys (SMA): A Summary of Their Properties	620
Designing SMA Actuators: General Principles	623
Shape-Memory Alloys for Microapplications	630
A Concept of Smart SMA Microdevices	637
Future Trends	641
Conclusion	643
Bibliography	643
Microtubes to Molecularly Imprinted Polymers	
Microtubes	644
Introduction	644
AFRL Microtube Technology	647
Microtube Devices Based on Surface Tension and Wettability	654
Conclusions	666
Acknowledgments	666
Bibliography	666
Molecularly Imprinted Polymers	667
Introduction	667
Polymer Chemistry	668
Applications	672
Conclusions	680
Bibliography	680
Neural Networks to Nondestructive Evaluation	
Neural Networks	682
A Short Tutorial on Artificial Neural Networks	682
Introduction	682
Selecting Input Data	683
Preparation of Data	684
Number of Neurons in Hidden Layer(s)	684
Supervised Versus Unsupervised Neural Networks	684
Artificial Neural Network Extrapolations of Heat Capacities of Polymeric Materials to Very Low Temperatures	684
Results	686
Conclusions	686

Artificial Neural Network Modeling of Monte Carlo Simulated Properties of Polymers	686
Introduction	687
Calculations	687
Results	688
Conclusions	689
Summary	689
Acknowledgement	690
Bibliography	690
Nondestructive Evaluation	690
Introduction	690
Reality that Defies Noncontact Ultrasound	692
Pursuit of Noncontact Ultrasonic Transducers	693
Piezoelectric Transducers for Unlimited Noncontact Ultrasonic Testing	694
Noncontact Ultrasonic Analyzer	698
Reflection and Transmission in Noncontact Mode	698
Very High Frequency NCU Propagation in Materials	701
Noncontact Ultrasonic Measurements	701
Applications of Noncontact Ultrasound	706
Conclusions	713
Acknowledgements	714
Bibliography	714
Optical Fiber Sensor to Optical Storage Films	
Optical Fiber Sensor Technology: Introduction and Evaluation and Application	715
Introduction	715
Fiber Optics	715
Classification of Optical Fiber Sensors	718
Optical Fiber Sensing Mechanisms	720
Fiber Optic Sensor Applications	730
Conclusions	735
Acknowledgments	736
Bibliography	737
Optical Storage Films, Chalcogenide Compound Films	738
Introduction	738
Essentials of Optical Recording Media	738
Properties of Chalcogenide Compound Films	739
Optical Storage in Chalcogenide Films	749
Future Directions	751
Bibliography	751
Paints to Poly(Vinylidene Fluoride)	
Paints	754
Introduction	754
Basic Concepts of Smart Paints	754
Piezoelectric Composites	754
Composition of Smart Paints	755
Formation of Smart Paint Films	755
Evaluation of Smart Paint Films	756
Factors Determining Poling Behavior of Smart Paint Films	758
Techniques for Applying Smart Paint Films	759
Future Directions	759
Acknowledgments	760
Bibliography	760

Pest Control Applications	761
Introduction	761
Sound as a Pest Deterrent	761
Cavitation as a Destructor	762
Notation	770
Bibliography	770
Photochromic and Photo-Thermo-Refractive Glasses	770
Introduction	770
Physical Principles of Photosensitivity in Glasses	770
Induced Coloration of Reversible Photochromic Glasses	772
Heterogeneous Photochromic Glasses	774
Optical Waveguides in Photochromic Glasses	775
Induced Refraction Through Irreversible Photoinduced Crystallization	776
Photo-Thermorefractive Glass	777
Bragg Gratings in PTR Glass	778
Summary	780
Bibliography	780
Piezoelectricity in Polymers	780
Introduction	780
Piezoelectricity: An Overview	781
Synthetic Piezoelectric Polymers	781
Electromechanical Properties of PVDF	782
Nonlinear and Time-Dependent Effects	785
Applications	788
Concluding Remarks	790
Bibliography	791
Poly(<i>P</i>-Phenylenevinylene)	793
Introduction	793
Methods of Preparation	793
Properties	797
Applications	799
Future Considerations	804
Bibliography	805
Poly(Vinylidene Fluoride) (PVDF) and Its Copolymers	807
Introduction	807
Synthetic Pathways and Molecular and Crystal Structures	808
Processing and Fabrication	811
Electromechanical Properties in Normal Ferroelectric PVDF and Its Copolymers	815
Relaxor Ferroelectric Behavior and Electrostrictive Response in P(VDF-TrFE) Based Copolymers	819
Concluding Remarks	824
Bibliography	824
Polymer Blends to Power Industry	
Polymer Blends, Functionally Graded	826
Introduction	826
Mechanism of Diffusion-Dissolution Method	827
Preparation and Characterization of Several Types of Functionally Graded Polymer Blends	829
Functional and Smart Performances and the Prospect for Application	831
Bibliography	834

Polymers, Biotechnology and Medical Applications	835
Introduction	835
Smart Polymers Used in Biotechnology and Medicine	835
Applications	838
Conclusion	847
Bibliography	847
Polymers, Ferroelectric Liquid Crystalline Elastomers	850
Introduction	850
Synthesis of Ferroelectric LC-Elastomers	850
Conclusion	858
Acknowledgment	859
Bibliography	859
Polymers, Piezoelectric	860
Introduction	860
Semicrystalline Polymers	861
Amorphous Polymers	867
Characterization and Modeling	870
Acknowledgments	872
Bibliography	872
Power Industry Applications	873
Introduction	873
Overview of the Electric Power Industry	874
Smart Sensors	876
Smart Sensor-Actuators	882
Challenges Awaiting Smart Materials Solutions	884
Bibliography	889
Self-Diagnosing to Shape Memory Alloys, Applications	
Self-Diagnosing of Damage in Ceramics and Large-Scale Structures	891
Introduction	891
Self-Diagnosis Function of Fiber-Reinforced Composite with	
Conductive Particles	891
Mortar Block Tests Application of the Self-Diagnosis Composite	
to Concrete Structures	897
Bibliography	903
Sensor Array Technology, Army	903
Introduction	903
Background	903
Technical Approach	904
Sensors, Surface Acoustic Wave Sensors	910
Introduction	910
Background	910
Experimental Procedures for Saw Sensing	912
Discription of Applications and Experimental Results	913
Conclusions	920
Bibliography	920

Shape Memory Alloys, Applications	921
Introduction	921
Types of Shape-Memory Alloys	921
Designing with Shape Memory Alloys	923
SMA Applications	928
Future Trends	935
Bibliography	935
Additional Reading	936
Shape-Memory Alloys, Magnetically Activated to Ship Health Monitoring	
Shape-Memory Alloys, Magnetically Activated Ferromagnetic Shape-Memory Materials	936
Introduction	936
Field-Induced Strain in FSMAs	938
Quantitative Models of Twin-Boundary Motion	941
Field-Induced Strain Under Load	945
Discussion	947
Summary	950
Acknowledgments	950
Bibliography	950
Shape Memory Alloys, Types and Functionalities	951
Shape-Memory Alloy Systems	951
Functional Properties of Shape-Memory Alloys	957
Bibliography	962
Shape-Memory Materials, Modeling	964
Introduction	964
Basic Material Behavior and Modeling Issues	964
State of the Art and Historical Developments	966
A Comprehensive Model for Uniaxial Stress	973
Summary and Conclusions	979
Bibliography	980
Ship Health Monitoring	981
Introduction	981
Overview of Ship Health Monitoring	981
Environmental Issues	984
Sensor Technology	985
Data	989
Commercial Systems	990
Bibliography	991
Smart Perovskites to Spin-Crossover Materials	
Smart Perovskites	992
The Family of Perovskite-Structured Materials	992
Structures and Properties	993
The Fundamental Structural Characteristics of ABO ₃ Perovskite	1001
Anion-Deficient Perovskite Structural Units - The Fundamental Building Blocks for New Structures	1003
Structural Evolution in the Family of Perovskites	1006
Quantification of Mixed Valences by EELS	1010
High-Spatial-Resolution Mapping of Valence States	1012
Summary	1012
Acknowledgment	1013
Bibliography	1013

Soil-Ceramics (Earth), Self-Adjustment of Humidity and Temperature	1014
Introduction	1014
A New Definition on Materials with Consideration for Humans and the Earth	1014
Smart Materials for the Living Environment	1016
Climate Control by Porous Bodies	1016
Using the Greatness of Nature Wisely	1017
Performance of the Hydrothermally Solidified Soil Bodies	1024
New Functional Materials	1027
Conclusions	1028
Acknowledgment	1028
Bibliography	1028
Sound Control with Smart Skins	1029
Introduction	1029
Smart Foam Skin	1030
Piezoelectric Double Amplifier Smart Skin	1032
Smart Skins for Sound Reflection Control	1034
Advanced Control Approaches for Smart Skins	1035
Conclusion	1036
Bibliography	1037
Spin-Crossover Materials	1037
Bibliography	1039
Acknowledgments	1039
Thermoresponsive to Truss Structures	
Thermoresponsive Inorganic Materials	1040
Introduction	1040
Origins of Negative Thermal Expansion	1041
Materials that Display Negative Thermal Expansion	1045
Bibliography	1053
Triboluminescence, Applications in Sensors	1054
Introduction	1054
Classification of TL	1055
TL of Oxide Crystals Doped with Rare Earths	1057
Applications of TL	1064
Bibliography	1065
Truss Structures with Piezoelectric Actuators and Sensors	1066
Introduction	1066
Truss Structure Configuration	1067
Finite Element and Modal Analysis of the Truss Structure	1068
Actuator and Sensor Construction	1072
Formulation of the State Space Dynamic Model	1072
Control Design and Simulations	1075
Experimental Results	1078
Conclusion	1080
Acknowledgments	1083
Bibliography	1083

Vibration Control to Windows

Vibration Control	1085
Introduction	1085
Vibrational Control of Smart Structures	1085
Vibrational Control of Smart Systems	1092
Bibliography	1097
Vibration Control in Ship Structures	1098
Introduction	1098
Fundamental Concepts of Ship Noise Control	1098
Sensors and Actuators for Active Noise and Vibration Control (ANVC)	1100
Applications of Noise Control in Ship Structures	1107
Recommendations on Sensors and Actuators for ANVC of Marine Structures	1111
Summary and Conclusions	1114
Bibliography	1114
Vibrational Analysis	1115
Introduction	1115
Vibrational Representation	1115
Degrees of Freedom	1115
Vibrations of Simple Structures	1115
Natural Frequencies of Uniform Beam Structures	1116
Natural Frequencies of Uniform Plates and Circuit Boards	1116
Methods of Vibrational Analysis	1117
Problems of Vibrational Analysis	1117
Problems of Material Properties	1117
Relation of Displacement to Acceleration and Frequency	1118
Effects of Vibration on Structures	1119
Estimating the Transmissibility Q in Different Structures	1119
Methods for Evaluating Vibrational Failures	1120
Determining Dynamic Forces and Stresses in Structures Due to Sine Vibration	1121
Determining the Fatigue Life in a Sine Vibrational Environment	1122
Effects of High Vibrational Acceleration Levels	1123
Making Structural Elements Work Smarter in Vibration	1123
How Structures Respond to Random Vibration	1127
Miner's Cumulative Damage for Estimating Fatigue Life	1128
Bibliography	1129
Vibrational Damping, Design Considerations	1129
Introduction	1129
Dynamic Problem Identification	1129
Dynamic Characteristics	1130
Environmental Definition	1130
Required Damping Increase	1131
Damping Concept Selection and Application Design	1131
Prototype Fabrication and Laboratory Verification	1132
Production Tooling and Field Validation	1132
Summary	1133
Bibliography	1133

Windows	1134
Introduction	1134
Architectural Glazing Applications for Smart Windows	1134
Survey of Smart Windows	1137
Electrochromic Smart Windows	1142
Future Directions	1145
Bibliography	1084

Index

Index - Absorption to Myoglobin	1147
Index - Nafion to Zirconium	1194

ACTUATORS, PIEZOELECTRIC CERAMIC, FUNCTIONAL GRADIENT

XINHUA ZHU

Nanjing University
Nanjing, People's Republic of China

ZHONGYAN MENG

Shanghai University
Shanghai, People's Republic of China

INTRODUCTION

Actuators and materials play a key role in developing advanced precision engineering. The breakthroughs in this field are closely related to the development of various types of actuators and related materials. The successes of piezoelectric ceramics and ceramic actuators have. For instance, the propagating-wave type ultrasonic motor that produces precise rotational displacements has been used in autofocusing movie cameras and VCRs (1). Multimorph ceramic actuators prepared from electrostrictive $\text{Pb}(\text{Mg}_{1/3}\text{Nb}_{2/3})\text{O}_3$ (PMN) ceramics are used as deformable mirrors to correct image distortions from atmospheric effects (2). The likelihood that the range of applications and demand for actuators will grow actively and has stimulated intensive research on piezoelectric ceramics. Functionally graded materials (FGMs) are a new class of composites that contain a continuous, or discontinuous, gradient in composition and microstructure. Such gradients can be tailored to meet specific needs while providing the best use of composite components. Furthermore, FGM technology is also a novel interfacial technology for solving the problems of the sharp interface between two dissimilar materials. In recent years, significant advances in developing FGMs have been achieved. In this paper, we introduce and summarize the recent progress in piezoelectric ceramic actuators and review recent applications of FGMs in piezoelectric ceramic devices.

ACTUATORS

Background

Microelectromechanical and intelligent materials systems have received much attention because of their great scientific significance and promising potential applications in automation, micromanipulation, and medical technology. However, most applications require a source of mechanical power, microscale motors and actuators, that provide the effect of "muscles" to make things happen. To date, several different physical mechanisms such as electrostatic and magnetostatic forces, phase changes (shape memory alloys and electro-rheological fluids), piezoelectric/electrostrictive strains, magnetostriction, and thermal stresses have been explored as potential actuation sources.

Each kind of actuator has its own advantages and drawbacks, so their selection and optimization should be determined by the requirements of the application.

Ceramic Actuators

Piezoelectric/electrostrictive ceramics are widely used in many different types of sensing-actuating devices. This is particularly true for the whole family of micro- and macro-piezoelectric ceramic actuators. Various types of ceramic actuators have been developed for different applications. From a structural point of view, ceramic actuators are classified as unimorph, bimorph, Moonie, Cymbal, and Rainbow monomorph benders (3–7). The bimorph bender consists of two thin ceramic elements (poled in opposite directions) that sandwich a thin metal shim, whereas the unimorph is simply on a thin ceramic element bonded to a thin metal plate. Although unimorph and bimorph structures have been successfully applied to many devices during the past forty years, their inability to extend the force-displacement envelope of performance has generated a search for new actuator technologies. The Moonie bender has crescent-shaped, shallow cavities on the interior surface of end caps bonded to a conventionally electroded piezoelectric ceramic disk (5). The metal end caps are mechanical transformers for converting and amplifying the lateral motion of the ceramic into a large axial displacement normal to the end caps. Both the $d_{31}(=d_{32})$ and d_{33} coefficients of a piezoelectric ceramic contribute to the axial displacement of the composite. The advantages of the Moonie include (1) a factor of 10 enhancement of the longitudinal displacement, (2) an unusually large d_{33} coefficient that exceeds 2500 pC/N, and (3) an enhanced hydrostatic response. Recent improvements in the basic Moonie design have resulted in an element named "Cymbal" (6), a device that possesses more flexible end caps that result in higher displacement. Moonie and Cymbal composites have several promising applications, such as transceivers for fish finders, positional actuators, and highly sensitive accelerometers.

Another device developed to increase the force-displacement performance of a piezoelectric actuator is the rainbow (7), a monolithic monomorph that is produced from a conventional, high-lead piezoelectric ceramic disk that has one surface reduced to a nonpiezoelectric phase by a high-temperature, chemical reduction reaction. Thermal stresses induced on cooling from the reduction temperature, due to the different thermal expansion coefficients of the reduced and unreduced layers, cause the ceramic disk to deform with high axial displacements and sustain moderate pressure. The axial displacements achieved when driven by an external electric field can be as high as 0.25 mm (for a 32-mm diameter \times 0.5-mm thick wafer), while sustaining loads of 1 kg. Displacements larger than 1 mm can be achieved by using wafers (32 mm in diameter) thinner than 0.25 mm when operating in a saddle

mode. Prototypes of rainbow pumps, sensors, actuator arrays, and optical deflectors have been demonstrated, but commercial products have not yet been produced (8).

The photostrictive actuator is another type of bimorph application. The photostrictive behavior is results from a combined photovoltaic effect and a piezoelectric effect. PZT ceramics doped with slight additives of niobium and tungsten exhibit a large photostrictive effect when irradiated by violet light. A photostrictive PLZT bimorph has been used to demonstrate the prototype of a photodriven relay for a remote microwalking device, and a photophone for the future is also envisioned (9).

There is also increasing interest in electrostrictive ceramic actuators (10) because electrostrictive ceramics do not contain ferroelectric domains, so that they can return to their original dimensions immediately, when the external electric field is reduced to zero. Therefore, the advantages of an electrostrictive actuator are the near absence of hysteresis and lack of aging behavior. However, because the electrostrictive effect is a second-order phenomenon of electromechanical coupling, it is usually necessary to apply a high voltage to achieve moderate deformation. However, a particularly large longitudinal electrostrictive strain as high as 0.1% has been achieved in PMN-based relaxor materials (11–13) just above their Curie temperature. To decrease the applied voltage, a considerable effort is now being devoted to developing multilayered actuators, which have the advantages of low operating voltage, large generated forces and displacements, quick response, and low energy consumption (10).

Composite Actuators

Piezoelectric/electrostrictive ceramics, magnetostrictive materials, and ferroelastic shape memory alloys (SMA) are all used for actuators. However, different classes of ceramic actuators require somewhat different materials. In general, an ideal actuating material should exhibit a large stroke, high recovery force, and superior dynamic response. Shape memory alloys display large strokes and forces but have an inferior dynamic response and low efficiency. Ferroelectric ceramics exhibit excellent dynamic response (of the order of microseconds), but their displacements are quite small (of the order of a few micrometers) due to their small strain magnitude ($<10^{-3}$). By combining ferroelastic shape memory alloys with piezoelectric, or magnetostrictive, materials, hybrid smart heterostructures can be fabricated, which may have the optimum characteristics of both materials. Recently much work has been done to explore the technical feasibility of smart thin film heterostructures by depositing piezoelectric thin films (PZT) on SMA substrates using techniques such as sol-gel, spin-on coating, and pulse laser deposition (14,15). However, cracking of ferroelectric thin films and the low interfacial bonding and dynamic coupling of dissimilar components are crucial problems to be solved before hybrid composites can be employed as actuators in for smart structures. To feature the diverse characteristics, various kinds of smart hybrid composites will be designed for actuating applications by incorporating these smart materials appropriately together.

It is clear that the application field of ceramic actuators is remarkably wide. However, some issues of reliability and

durability still need to be solved before ceramic actuators can become general-purpose commercialized products, especially for ferroelectric ceramic actuators. It is very desirable to develop new fatigue methodology and fracture mechanics to predict actuator dynamic response and lifetime, to assess actuator reliability, and to implement them in control methodology. The main issues in constitutive behavior are coupled field effects, modeling of phase transformations, twinning, and domain switching (in the finite strain regime). Good understanding of fatigue and fracture behavior at the microlevel and understanding interfacial failure mechanisms (active–passive and active–active) and the mechanics of hybrid active materials is the rational approach to designing actuators. As the relentless drive of electronic devices toward miniaturization, multifunction, and integration continues, the issues of optimizing the size and location of actuators based on control theory, structural response, and desired adaptability should be addressed. Thus, ceramic actuators will need to be smarter and smarter as the applications demand.

PIEZOELECTRIC CERAMICS

Background

Piezoelectricity in solids is based on the internal structures of materials. For simplicity, here, consider only a single crystal that has a defined chemical composition and consists of ions (atoms that have positive or negative charges) that are constrained to occupy positions in a specific repeating relationship to each other, thus building up the structure of the crystal lattice. The smallest repeating unit of the lattice is called the unit cell, and the specific symmetry possessed by the unit cell determines whether piezoelectricity can exist in the crystal. Among the 32 point groups, 21 classes are noncentrosymmetric (a necessary condition for piezoelectricity), and only 20 of these are piezoelectric. One class, although it lacks a center of symmetry, is not piezoelectric because of other combined symmetry elements. Furthermore, for those materials that are piezoelectric but not ferroelectric (i.e., they do not possess spontaneous polarization), the stress itself is the only means by which dipoles are generated. The piezoelectric effect is linear and reversible, and the magnitude of the polarization depends on the magnitude of the stress, the sign of the charge produced depends on the types of stresses such as tensile or compressive (16). A ferroelectric material is one that undergoes a phase transition from a high-temperature phase that behaves as ordinary dielectrics (so that an applied electric field induces an electric polarization, which goes to zero when the field is removed) to a low-temperature phase that has spontaneous polarization whose direction can be switched by an applied field. Therefore, all ferroelectric materials possess piezoelectricity. The piezoelectric effect in ferroelectric ceramics is realized by a poling process, in which an external electric field can orient the ferroelectric domains within the grains and thus produce a ceramic material that acts similarly to a single crystal that possesses both ferroelectric and piezoelectric properties. Before poling, ferroelectric ceramics do not possess any piezoelectric properties due to the random

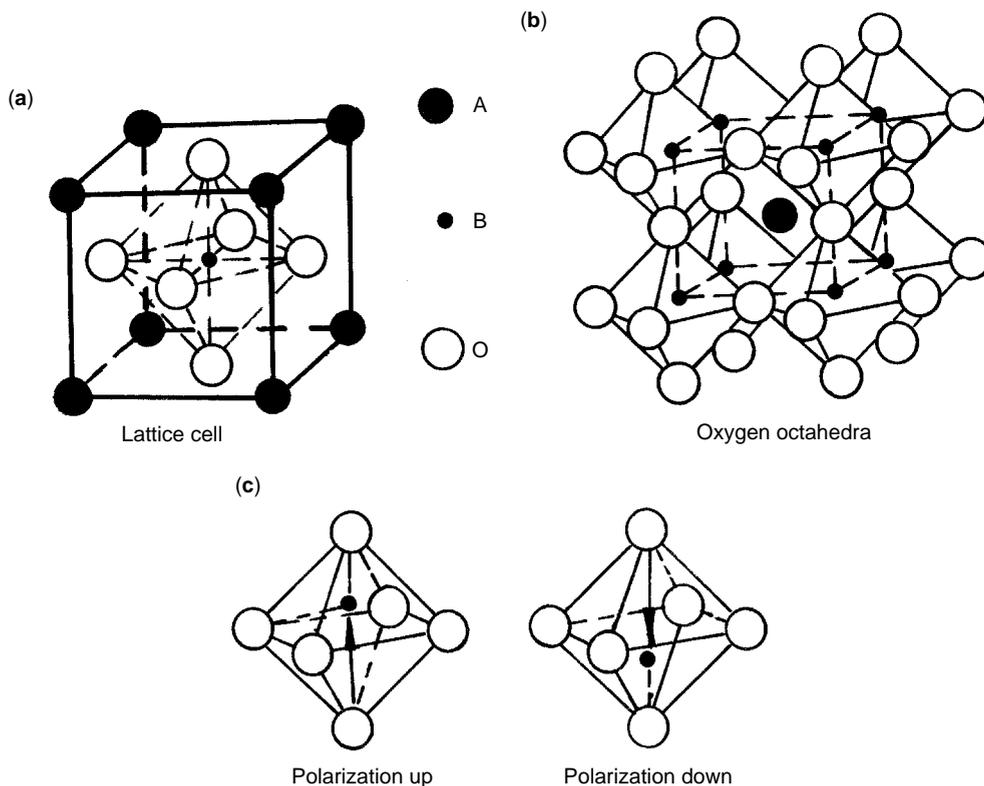


Figure 1. (a) Unit cell of ABO₃ perovskite, (b) oxygen octahedra, and (c) 180° polarization reversal for two of the six possible polarization states produced by displacement of the central cation in the tetragonal plane.

orientations of the ferroelectric domains in the ceramics. The first piezoelectric ceramic was BaTiO₃ developed commercially in the 1940s; it has an unusually high dielectric constant due to its ferroelectric (permanent internal dipole moment) nature and thus ushered in a new class of ferroelectrics of the ABO₃ perovskite structure. A typical ABO₃ unit cell is shown in Fig. 1a. As an example, the BaTiO₃ unit cell consists of a corner-linked network of oxygen octahedra where Ti⁴⁺ ions occupy B sites within the octahedral cage and the Ba²⁺ ions are situated in the interstices (A site) created by the linked octahedra, as shown in Fig. 1b. Below the Curie temperature, there is a structural distortion to a lower symmetry phase accompanied by a shift off-center of the small cation (Ti⁴⁺) along the c axis, as shown in Fig. 1c. The spontaneous polarization derives largely from the electric dipole moment created by this shift. Displacement of Ti⁴⁺ occurs along the c axis in a tetragonal structure, although it should be understood that it can also occur along the orthogonal a or b axes as well. The views of “polarization up” and “polarization down” (representing 180° polarization reversal) show two of the six possible permanent polarization positions. New ferroelectric ceramic materials were surveyed and led to the development of Pb(Zr,Ti)O₃ in the 1950s (17), it became the main industrial product in piezoelectric ceramic materials in the following 10 years. The phase diagram of the PZT pseudobinary is shown in Fig. 2, where the T_c line is the boundary between the cubic paraelectric phase and the ferroelectric phases. A significant feature in Fig. 2 is the morphotropic phase boundary (MPB), which divides the

region of the ferroelectric phase into two parts: a tetragonal phase region (on the Ti-rich side) and a rhombohedral phase region (on the Zr-rich side). In the PZT system at room temperature, the MPB occurs close to Zr/Ti = 53/47. An MPB represents an abrupt structural change (involving fractional atom shifts, but no change in near neighbors) in composition at a constant temperature within a solid solution. Usually, it occurs because of the instability of one phase (such as the ferroelectric tetragonal phase) to the other (ferroelectric rhombohedral phase) at a critical composition where two phases are energetically very similar but elastically different. Note that the dielectric constant and the piezoelectric and electromechanical behavior attain maxima, whereas the elastic constants tend to be softer in the vicinity of the MPB composition, as shown in Fig. 3. Similar phenomena are also observed in ternary solid solutions such as PLZT, and this feature is exploited in many commercial compositions because of the high property coefficients and unique structural characteristics of MPB compositions. Many have speculated concerning the reasons for the maximum in coupling at the MPB (18). Polycrystalline materials that have random grain orientation belong to the $\infty \infty m$ Curie group and show no piezoelectricity before poling because of the random orientations of the dipoles. The poled ferroelectric ceramics belong to the Curie group of ∞m , and the net residual spontaneous polarization has a component in the direction of the poling field, which is optimized in terms of the poling electric field and the poling temperature. Although piezoelectricity was first observed in single crystals, polycrystalline

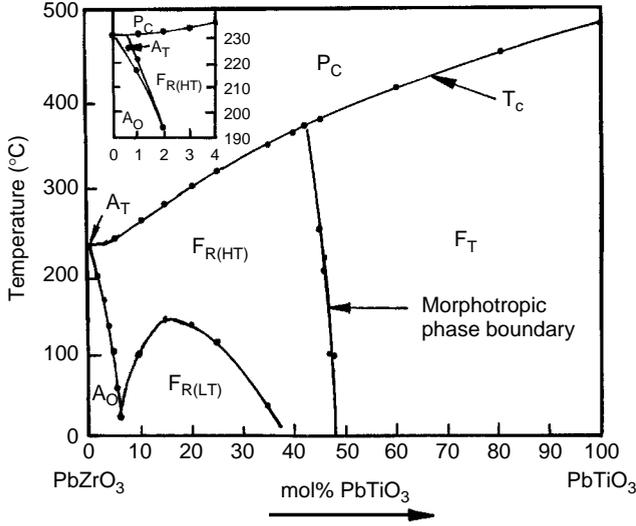


Figure 2. Phase diagram of $\text{PbZrO}_3\text{-PbTiO}_3$ system. Pc: paraelectric cubic, F_T : ferroelectric tetragonal, $F_{R(HT)}$: ferroelectric rhombohedral (high temperature form), $F_{R(LT)}$: ferroelectric rhombohedral (low temperature form), A_O : antiferroelectric orthorhombic, A_T : antiferroelectric tetragonal, T_C : curie temperature .

piezoelectric ceramics now represent the primary commercial piezoelectric material for actuators and sensors in intelligent systems and smart structures.

Dielectric and Piezoelectric Parameters

Piezoelectric ceramics are evaluated by the piezoelectric coupling factor κ (e.g., κ_{33} , κ_{31} , and κ_p), mechanical quality

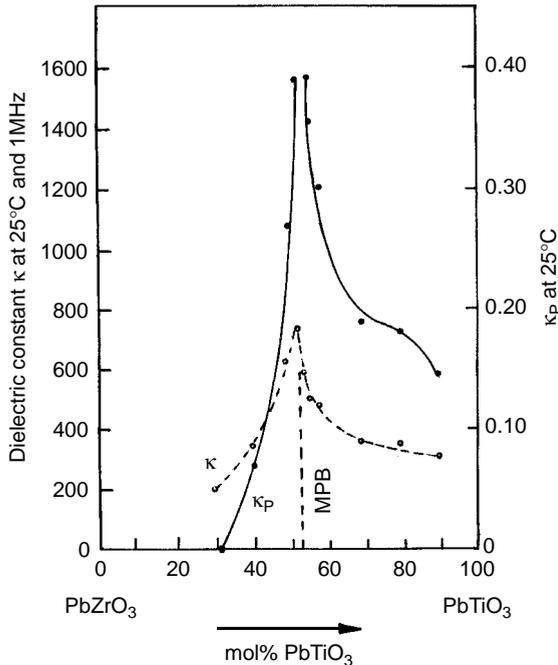


Figure 3. Dielectric constant κ and electromechanical coupling factor κ_p for the $\text{PbZrO}_3\text{-PbTiO}_3$ piezoelectric ceramic system.

factor (Q_m), frequency constant (N_l), and piezoelectric coefficients, such as the d and g coefficients that describe the interaction between the mechanical and electrical behavior of piezoelectric ceramics. The effective electromechanical coupling coefficient κ_{eff} describes the ability of the ceramic transducer to convert one form of energy to another, as defined by the equations

$$\kappa_{\text{eff}}^2 = \frac{\text{mechanical energy converted to electrical energy}}{\text{input mechanical energy}}, \quad (1)$$

or

$$\kappa_{\text{eff}}^2 = \frac{\text{electrical energy converted to mechanical energy}}{\text{input electrical energy}}. \quad (2)$$

This parameter is a function in equations for electrical/mechanical energy conversion efficiency in actuators, in bandwidth and insertion loss in transducers and signal processing devices, and in the location and spacing of critical frequencies of resonators.

The effective coupling coefficient κ_{eff} is related to the values of f_m and f_n and can be described as

$$\kappa_{\text{eff}}^2 \approx \frac{f_n^2 - f_m^2}{f_n^2}. \quad (3)$$

Values for f_m and f_n are readily measured by using a suitable bridge. The approximations in Eq. (3) are good, provided that the Q_m value for the resonator is high enough, for example, greater than 100.

The planar coupling coefficient κ_p is related to the parallel and series resonant frequency by

$$\frac{\kappa_p^2}{1 - \kappa_p^2} = f \left(J_0, J_1, \nu \frac{f_p - f_s}{f_s} \right), \quad (4)$$

where J_0 and J_1 are Bessel functions and ν is Poisson's ratio. κ_{31} can also be calculated from

$$k_{31}^2 = \frac{1 - \nu}{2} k_p^2. \quad (5)$$

The mechanical quality factor Q_m that represents the degree of mechanical loss of a piezoelectric resonator at resonance is defined as

$$Q_m = 2\pi \frac{\text{stored mechanical energy at resonance}}{\text{mechanical dissipated energy per resonant cycle}}. \quad (6)$$

Q_m can be obtained from the following equation:

$$Q_m = \frac{f_p^2}{2\pi f_s |Z_m| (C_0 + C_1) (f_p^2 - f_s^2)}, \quad (7)$$

where $|Z_m|$ is the minimum impedance at resonance and C_0 and C_1 are the capacitance shown in Fig. 4a, respectively.

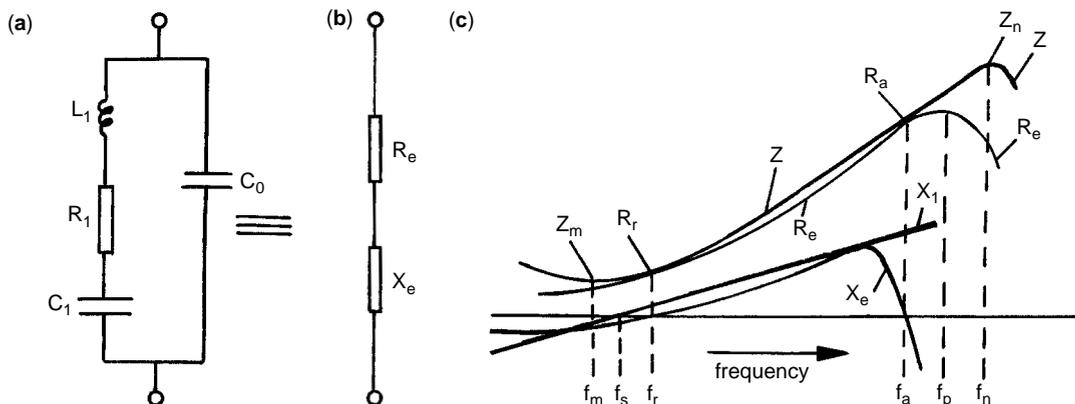


Figure 4. (a) Equivalent circuit for a piezoelectric specimen vibrating close to its fundamental resonance; (b) the equivalent series components of the impedance of (a); (c) characteristic frequencies of the equivalent circuit, the differences between f_m , f_s , and f_r , and between f_a , f_p , and f_n are exaggerated.

The frequency constant N_l is defined by the following equation:

$$N_l = l \times f_r = \frac{1}{2} \sqrt{\frac{Y}{\rho}} \quad (8)$$

where l is the length of a piezoelectric ceramic thin plate, f_r is the resonant frequency in the length direction, Y is Young's modulus, and ρ is the density.

These values of the piezoelectric properties of a material can be derived from the resonant behavior of suitably shaped specimens subjected to a sinusoidally varying electric field. To a first approximation, the behavior of the piezoelectric specimen close to its fundamental resonance can be represented by an equivalent circuit, as shown in Fig. 4a, b. The frequency response of the circuit is shown in Fig. 4c, in which various characteristic frequencies are identified. The functions f_r and f_a are the resonant and antiresonant frequencies when the reactance of the circuit is zero ($X_e = 0$); f_s is the frequency at which the series arm has zero reactance ($X_1 = 0$); f_p is the frequency when the resistive component R_e is at a maximum; f_m and f_n are, respectively, the frequencies for the minimum and maximum impedance Z of the circuit as a whole. Piezoelectric vibrators that have electrodes covering their two flat faces are used to measure the properties of piezoelectric ceramics. A more common geometry is a thin disk of diameter d electroded over both faces and poled perpendicularly to the faces. The resonance in these disk-shaped specimens is focused on a radial mode excited through the piezoelectric effect across the thickness of the disk. The details for determining piezoelectric coefficients can be found in IRE standards on piezoelectric crystals: measurements of piezoelectric ceramics [Proc. IRE 49(7); 1161–1169 (1961)].

Compositions and Properties

To meet stringent requirements for specific applications, piezoelectric ceramics under different doping conditions and hence, possessing different characteristics, have been

developed for various applications. The techniques for modifying piezoelectric ceramics include element substitution and doping. In general, the term “element substitution” implies that cations in the perovskite lattice, for example, Pb^{2+} , Zr^{4+} , and Ti^{4+} , are replaced partially by other cations that have the same chemical valence and ionic radii similar to those of the replaced ions. The new substituent cation usually occupies the same position as the replaced cation in the perovskite lattice, and thus a substitutional solid solution is formed; the term “doping” implies that some ions whose chemical valences differ from those of the original ions in the lattice, or some compounds that have the chemical formulas, $A^+B^5+O_3$ and $A^{3+}B^{3+}O_3$ are added to PZT ceramics. From a global perspective, there are essentially four types of compositional modifiers (19). The first type comprises higher covalent substitutions (donor dopants) on A and/or B sites (such as La^{3+} replacing Pb^{2+} or Nb^{5+} replacing Zr^{4+} or Ti^{4+}) to counteract the natural p-type conductivity of PZT and, thus, increase the electrical resistivity of the materials by at least three orders of magnitude. The donors are usually compensated for by the formation of A-site vacancies. The donor-doped PZT piezoelectric ceramics are usually called “soft” piezoelectric ceramics, meaning that they are easily depoled and driven nonlinear. Their main features include square hysteresis loops, low coercive fields, high remanent polarization, high dielectric constants and dielectric loss, maximum coupling factors, high mechanical compliance, and reduced aging.

The second modified type consists of lower valent substitutions (acceptor dopants) on A and/or B sites (such as Fe^{3+} replacing Zr^{4+} or Ti^{4+}). These are compensated for by the formation of oxygen vacancies. Acceptor-doped PZT piezoelectric ceramics are usually called “hard” piezoelectric ceramics because of their much enhanced linearity and high drive. Their main features are poorly developed hysteresis loops, lower dielectric constants and dielectric loss, lower compliances, and higher aging rates.

The third modified type is composed of isovalent substitutions on A and/or B sites (such as Ba^{2+} or Sr^{2+} replacing Pb^{2+} or Sn^{4+} replacing Zr^{4+} or Ti^{4+}). Such isovalent

substitutions usually broaden the temperature-dependent properties, increase the dielectric permittivity, and reduce the Curie temperature, but cause no significant change in coupling coefficient, aging rate, volume resistivity, or low amplitude mechanical or dielectric loss.

The last types of compositional modifiers are more difficult to classify and are called “thermally variable”. They can exist in more than a single valence state and at more than one type of ionic site. The feature of piezoelectric ceramics doped with variable valence additives is much improved temperature and time stability of the resonant frequency. These materials are typically used in electric wave filters or resonators where high temperature and time stability of the resonant frequency are required.

Fabrication Processes

The fabrication processes for piezoelectric ceramics are similar to those used for electronic ceramics. Sequential processes involve weighing the starting materials, mixing, preliminary calcination, milling, shape forming, removing all organic constituents, sintering, formulating electrodes, poling, measuring properties, testing, and packing. Poling is the most critical step in the total fabrication process. The poling process is necessary to induce piezoelectric properties in polycrystalline ferroelectric ceramics because ceramic bodies are generally macroscopically isotropic in the “as-sintered” condition. The poling process can be carried out by immersing the specimens in transformer oil at a temperature of 100–150 °C, while applying a static electric field of 2.5–4.5 MV/m in a desired direction for a period of 10–20 minutes to align the ferroelectric domains. In the poled condition, the ferroelectric ceramics exhibit spontaneous polarization and have a component in the direction of the applied field. To obtain the best piezoelectric properties, the temperature and applied static voltage must be optimized in the poling process. The poling temperature is limited by the leakage current, which can cause an increase in the internal temperature that leads to thermal breakdown, whereas the electric field is limited by the breakdown strength of the ceramic. Higher fields can be used if they are applied as a succession of short pulses. In another poling method, called corona poling, high voltages of the order of 10^4 V are applied either to a single needle or an array of needles; their tips are located a few millimeters from the ceramic surface, and the opposite surface of the ceramic is grounded to develop a high electric field in the ceramics. The corona poling method has many advantages over conventional poling, such as the capability of continuous poling for mass production and the use of samples that have larger surface areas. Furthermore, it diminishes the risk of electrical breakdown because the poling charge cannot be quickly channeled to a “weak spot,” as it could be when using metallic electrodes. This method has been successfully used for poling piezoelectric ceramic–polymer composites such as PZT-epoxy (20). The alignment of the ferroelectric domains in the direction of the poled field is never complete in poled ceramics. However, depending on the type of crystal structure involved, the degree of poling can be quite high and ranges from 83% for the tetragonal phase, to 86% for the rhombohedral phase, and to 91% for the orthorhombic phase. The degree of poling also increases in

ascending order from polycrystalline ferroelectric ceramics, to poled ferroelectric ceramics, to single-crystal ferroelectrics, and to single-domain single crystals.

Applications of Piezoelectric Ceramics

Both direct and inverse piezoelectric effects can be used for applications of piezoelectric ceramics. In general, the use of the direct piezoelectric effect can generate a high voltage by applying a compressive stresses, whereas using the converse piezoelectric effect, small displacements can be generated by applying an electric field to a ceramic piece. Similarly, vibrations can be produced by applying an alternating field to a ceramic piece and can be detected by amplifying the field generated by vibrations incident on the ceramic. The flexor transducer, which consists of two piezoelectric ceramic thin plates poled in opposite directions, is used in gramophone pick-ups and ultrasonic accelerometers. Filters and other devices can be made to generate surface waves at frequencies that exceed 1GHz, and ultrasonic motors provide an opportunity to illustrate many important concepts.

The ultrasonic motor is based on the concept of driving a rotor by mechanical vibration excited on a stator via the piezoelectric effect. The rotor is in contact with the stator, and the driving force is the frictional force between the rotor and stator, in contrast to conventional electric motors that are based on electromagnetic conversion. The elliptical trajectory of the surface points of a stator is used to generate a rotational, or translational, motion in the rotor or slider. These elliptical motions may be generated either by exciting both longitudinal and bending vibrations of a beam or longitudinal and torsional vibrations. The unique features of ultrasonic motors are high output torque, quick response, large holding torque without energy dissipation, and operation free of a magnetic field. Many types of ultrasonic motors have been proposed that can be classified by the type of elliptical particle motion created in the stator. Here, we introduce only two types of ultrasonic motors: the standing-wave type and the traveling-wave type.

The standing-wave type is sometimes referred to as a vibratory-coupler type, or a “woodpecker” type; a vibratory piece is connected to a piezoelectric driver, and the tip portion generates flat elliptical movement. The vibratory piece is attached to a rotor, or slider, and provides intermittent rotational torque, or thrust. In general, the standing-wave type is highly efficient, but lack of control in both clockwise and counterclockwise directions is still a problem. In comparison, a wave in the traveling-wave type motor is excited by piezoelectric elements bonded to the stator. The traveling wave is obtained by superimposing two standing natural flexural waves equal in amplitude, but differ in phase by 90°, both spatially and temporally. The temporal difference is obtained by using one wave generated by the voltage $U_0 \sin(\omega t)$ and the other by $U_0 \cos(\omega t)$; the spatial phase difference results from the $3\lambda/4$ and $\lambda/4$ gaps between the two poled segments, as shown in Fig. 5. By propagating the traveling elastic wave induced by the thin piezoelectric ring, a ring-type slider in contact with the “rippled” surface of the elastic body bonded onto the piezoelectric elements is driven in both rotational directions by exchanging the sine and cosine voltage inputs. Another advantage of this motor

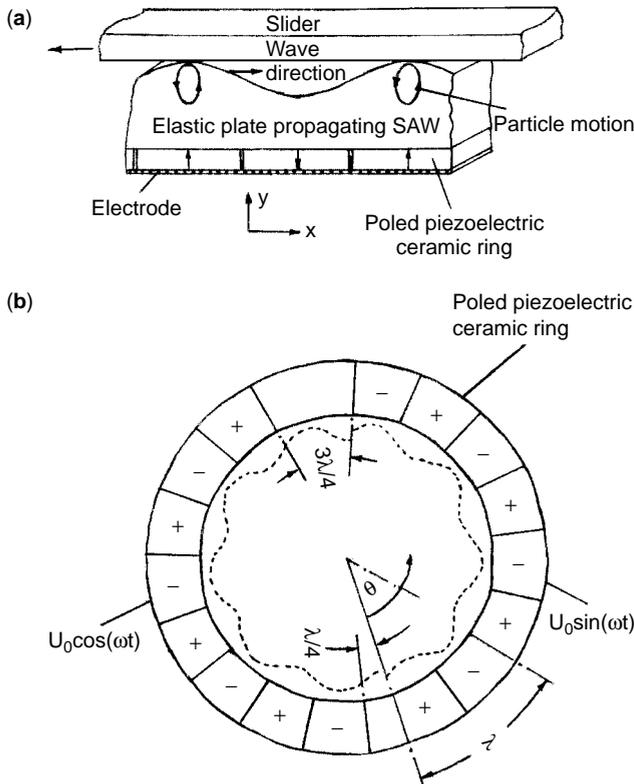


Figure 5. Operating principle of the ultrasonic rotary motor. (a) Side view, (b) plan view showing the poled segments and driving scheme.

is its thin design that makes it suitable for installation in analog instruments (21).

Future Trends in Piezoelectric Ceramics

Presently, PZT-based piezoelectric ceramics are leading materials for use in piezoelectric transducers, but recent concerns for the environment, safety, and health have prompted increasing research activity into Pb-free piezoelectric materials. To meet the extremely diverse applications, hybrid piezoelectric composites made by combining piezoelectric ceramics with other advanced materials are urgently required. Piezoelectric composites have some unique properties and functions such as improved dynamic response, high sensitivity to weak hydrostatic waves, damage resistance and control, which can be used to tailor or tune the overall performance of a smart structural system. Numerous efforts and exploratory approaches have been made to develop piezoelectric composites of piezoelectric ceramic and metal, piezoelectric ceramic and polymer, and piezoelectric ceramic and shape memory alloy (22–24), which represent a significant potential for advanced composites for smart systems. Numerical modeling and computer simulations of piezoelectric composites, in conjunction with some efforts in experimental characterization, will lead to the optimization of technical factors, such as structural design, geometry, processing parameters, and material selection, and in turn the improvement of the overall performance of piezoelectric composites. A major issue in piezoelectric composites is the creation of stresses, permanent strains, and cracks owing to a mismatch in thermal expansion and other physical properties of dissimilar

components in composites. To solve these problems, it is necessary to develop functionally graded composites that have gradients in composition, microstructure, and properties through one or more layers. Such an approach offers a number of advantages over the traditional methods of tailoring the compliance of composite materials, or structural elements, and opens up new horizons for novel applications.

FUNCTIONALLY GRADED MATERIALS

Background

Composite materials have been widely used for applications ranging from sporting and recreational accessories to advanced aerospace structural and engine components. Traditionally, the compositions and microstructures of composite materials are statistically homogenous, and have no significant spatial variations in properties. The development of space airplanes for the twenty-first century poses enormous technical problems, particularly for materials of superior heat resistance. The usual ceramic composites that combine a ceramic matrix and a dispersed phase are not expected to withstand the severe space environment or the high thermal stresses generated by the extreme temperature gradient experienced during reentry from space. Thus, materials that have superior stress relaxation, superior oxidation, thermal shock resistance, and other related characteristics are highly desirable. To meet these specific requirements, the concept of functionally graded materials (FGMs) has been proposed (25). By definition, FGMs are used to produce components that

feature engineered gradual transitions in microstructure and/or composition, whose presence is motivated by functional performance requirements that vary with location within the part. FGMs meet these requirements in a manner that optimizes the overall performance of the component. Thus, FGMs have grown to become one of the major current themes in structural materials research. They have also received considerable attention in a variety of nonstructural applications, where gradients in composition are deliberately introduced to optimize some physical properties and in areas where issues of mismatched material properties exist. The potential benefits derived from this new class of materials have led to increasing research activity in their design, processing, and applications.

Characterization

Functionally graded materials are composed of two or more phases, that are fabricated so that their compositions vary in some spatial direction and are characterized by nonlinear one-, two- or three-dimensional gradients that result in graded properties. They are distinguished from traditional composites by gradients of composition, phase distribution, porosity, grain size and texture, and particles or fiber reinforcement. From the viewpoint of their applications, FGMs can be classified as functionally graded coatings, functionally graded joints, and functionally graded materials, per se. Such a classification provides a good model for theoretical approaches and related numerical calculations for residual stress analyses (26). However, there is no significant difference in the structure and properties of the gradient volume among these FGM patterns. In contrast, based on their compositions, FGMs can be also classified as metal–ceramic, metal–metal, ceramic–ceramic, etc. The porosity, grain size and texture, or the gradient distribution (one-, two-, and three-dimensional) can also be used to classify FGMs. Note that the so-called back properties (e.g., modulus, thermal conductivity, electrical resistance, and specific heat) of FGMs (as for other materials) depend essentially on chemical composition rather than the structure, whereas the structurally sensitive properties such as strength, fracture toughness, and optical and magnetic characteristics depend on both crystal and microstructure. This observation should be considered when designing and analyzing gradient components because the structure of the final materials may affect both the desired properties and other incidental properties. Thus, the ultimate goal of FGM development is to fabricate components that have a predetermined concentration profile that best achieves the desired purpose for the material, and maintains other properties within limits that ensure acceptable performance.

Processing Methods

Several different physical and chemical methods are used to prepare FGMs (27,28). Each method has its own advantages and drawbacks, so it is difficult to predict a best-fit single method that will satisfy most of the requirements and technical limitations of the whole FGM spectrum. One of the potential answers may be optimal combination of

several methods. This combination will differ for each system, depending on the properties of the component materials.

The starting materials can be gases, liquids, or solids. If the starting materials are vapors, all CVD and PVD processes are available in principle. Compositional gradients can be easily obtained by continuously adjusting the ratios of the reactants in the mixture. The vapor method is one of the easiest ways to control the concentrations of phases being dispersed, but it is mainly suitable for obtaining compositional gradients in a thin film, or plate, across the material thickness.

Using liquids as starting materials, spraying is quite effective in achieving a graded structure. The major advantages of spraying are its flexibility and high deposition rates. Furthermore, plasma spraying can provide the possibility of coating a device that has complex shapes. Graded compositions can be obtained by adjusting the spray composition discretely or continuously. FGMs based on YSZ/Ni–Cr have been successfully prepared by this method (29).

Using solids, powder metallurgy or ceramic sintering are ideally suitable for fabricating FGMs because of the close microstructural control and versatility inherent in these techniques; they are widely used in industry because of the simple equipment and cost-efficient technology. Sizable and bulk functionally graded materials that have complex structures can be produced by this method, and transition phases in the thickness direction can be tailored from less than 1 mm to several centimeters, if desired. Several approaches have been proposed and tested to obtain a green body that has compositional gradients. The simplest is compositional distribution in a stepwise form, for example, many layers stacked in a block, which is then compacted and sintered to form FGMs (30). Another method uses centrifugal force, combined with gravity. The powder mixture is fed into the center of a centrifuge, from which it is projected toward the outer wall. The powder gradually forms a ring that has a through-thickness gradient (31). Self-propagating, high-temperature synthesis (SHS) is also a simple method that consists of simultaneous synthesis and forming (32, 33). The rapid propagation of a combustion reaction and low energy requirements limit elemental diffusion and allow one to conserve the graded composition designed in the green body. The SHS method is suitable for fabricating FGMs from components that have different high melting points and are chemically unreactive. The thin-sheet lamination method is a wet process, in which sheets that have different volume fractions are piled and pressed to form a graded green body and then sintered.

Other methods for producing functional or structural gradients such as sedimentation processes and magnetron sputtering have also been investigated. They have been used for gradient and multilayered thin coatings, but their application ranges are very limited. In general, the requirements and technical limitations of the whole FGM spectrum are quite diverse, so the methods for fabricating FGMs need to be versatile, but further work is needed to facilitate the transfer of production to industrial practice.

Modeling and Design

During the past 10 years, a variety of models have been established to describe the response of FGMs to thermal and mechanical loads, the chemical compatibility and degradation of interfacial layers, thermally activated diffusion, thermal stability, fabrication, and measurement of processing and/or service-induced residual stresses. Models developed for the spatial distribution of composition have been examined, followed by models developed for various aspects of FGM behavior and models for design and performance.

Models for Spatial Distribution of Composition. FGM models generally require an assumption for the spatial distribution of their constituent phases. For instance, consider an FGM that consists of two constituents denoted 1 and 2, respectively. Assume that the geometry is one-dimensional in the x direction, the direction of the functional gradient. A simple equation (34) describes the compositional gradient as

$$f_1(x) = \left(\frac{x_2 - x}{x_2 - x_1} \right)^n, \quad (9)$$

where $f_1(x)$ is the local volume fraction of phase 1 as a continuous function (the volume fraction of phase 2 is $1 - f_1(x)$ if the material is fully dense), x_1 and x_2 are the border regions of pure phase 1 and 2, respectively, and n is a variable functionally graded index, whose magnitude determines the curvature of $f_1(x)$. The curvature can be made concave upward, or downward, to a greater or lesser degree by selecting a proper functionally graded index n . Another approach to modeling the spatial distribution of composition is to select $f_1(x)$ varying discontinuously in a finite number of steps across the functionally graded direction. This would be appropriate to describe a multilayered FGM where the composition of each layer differs from one layer to the next. Based on knowledge of $f_1(x)$ and other information (such as the composition-dependent microstructure), one can determine the corresponding x dependence of effective values for evaluating physical properties such as thermal conductivity, Young's modulus, and the coefficient of thermal expansion. These, in turn, can be used to calculate the distributions of stress and temperature. The temperature distribution $T(x)$ under steady-state conditions can be determined by the equation (34),

$$\lambda(x) \frac{dT(x)}{dx} = \text{constant}, \quad (10)$$

where $\lambda(x)$ is the thermal conductivity. The solution to this equation is subject to appropriate boundary conditions. The determination of $f_1(x)$ is related to the design, whereas the calculation of $T(x)$ is related to the performance as well as fabrication. However, the calculation of $T(x)$ may indicate a need to revise the current $f_1(x)$, if the evaluated thermal stress has an excessively high value.

Systems Approach to FGM Modeling. A useful approach developed for the overall modeling of FGM processing, called the inverse design procedure, is based on the

properties of a homogenous composite material that has a certain composition ratio, materials and design databases, and known rules of mixture (35). Such an approach has been widely used in calculating the optimum compositional distribution for a material. The flow chart of the inverse design procedure is illustrated in Fig. 6. Here, the structure and the boundary conditions are specified initially. Then, several combinations of materials are assumed along with different assumptions for the spatially dependent mixture ratio. The distributions of stress and temperature can be calculated for these various combinations, and the calculations are repeated until optimum conditions are obtained. Based on these calculations, a FGM that excels in offering thermal relaxation can be prepared by various methods. Another system-based approach to the optimization process, used by Tanaka et al. (36), consists of the following sequence of steps: (1) select the initial compositions properly; (2) carry out a preliminary analysis of nonstationary heat conduction and thermal stress; (3) examine the design (failure) criteria at each step; (4) if the design criteria are violated, calculate a quantity known as the thermal stress sensitivity increment; (5) find the optimum composition profile that satisfies the design criteria; (6) repeat the analysis of nonstationary heat conduction, and thermal stress using the optimum composition profile determined; (7) if the design criteria are violated at another step, return to step (4). Note that these system-type approaches to FGM design are just ordered sequences of steps carried out to ensure that the resultant material performs adequately in its intended application.

Models Developed for Behavior of FGM. To use FGMs practically, performance tests should be established that include local evaluation of the microstructure and material properties to reveal the performance of the designed structure and property distributions and evaluation of the overall performance of FGM properties. To understand the performance of the FGM better, some models for describing the behavior of FGMs have been developed. A modified micromechanical model (37) is presented for the response of functionally graded metal–matrix composites subjected to thermal gradients. In this model, the actual microstructural details are explicitly coupled with the microstructure of the composite, which is particularly well suited for predicting the response to thermal gradients of thin-walled metal–matrix composites that have a finite number of large-diameter fibers in the thickness direction. A non-destructive method for detecting and evaluating the distribution of elastic parameters in the graded direction of an FGM by using ultrasonic waves has also been developed (38). The principle is based on the reflection of an impulse response. The acoustic impedance (defined as the product of density times the speed of sound) profile is first determined by ultrasonic waves, and then the profile of elastic modulus is evaluated using the equation (38)

$$V = \frac{E}{\rho(1 - \nu^2)} \quad (11)$$

where V is the speed of sound, ρ is the density, and ν Poisson's ratio. This method has been successfully used in

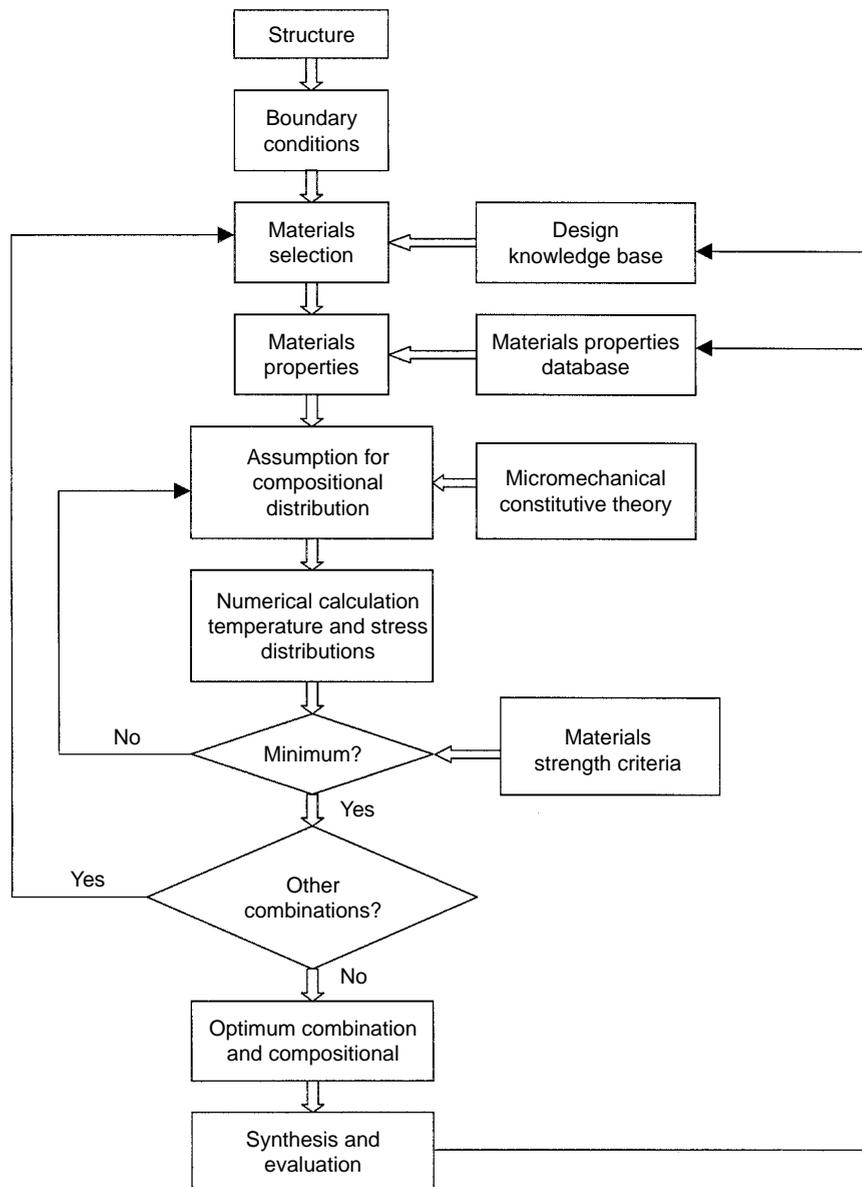


Figure 6. Flow chart of the inverse design of functionally gradient materials.

Ni-ZrO₂ FGMs to evaluate the accuracy of the design. A finite element model developed by Miller et al. (39) was intended to be a guide for both the design and fabrication of NiAl-Al₂O₃ FGMs. This model was applied to estimate residual stress as a function of the structure of the materials and the effects of thermal recycling. A coupled thermoelastic model is also being developed to study the thermal shock response of a metal/ceramic FGM to abrupt heating (40). Modeling studies of FGMs are clearly an essential ingredient to attaining the successful performance and have shed light on the future directions, but much work in modeling remains (41).

Applications of FGMs

The initial goal in developing FGMs was to obtain thermal barrier materials usable in space structures and fusion reactors and also in future space-plane systems. However, a

variety of nonstructural FGMs whose compositional gradients were introduced deliberately to optimize physical properties, such as electrical, magnetic and optical properties, have been developed recently (42–50). By controlling the compositional gradient in the radial direction, an optical fiber that has a graded refractive index can be prepared (48). Such graded index materials possess unique and useful optical properties that cannot be achieved by conventional optical materials, which have a constant refractive index. By this method, very low optical loss can be achieved in a wide wavelength range from the infrared to the ultraviolet spectrum (48). An important technological issue in nuclear power generation is the development of reactor core materials that have superlong lifetimes, high temperature strength, and high resistance to deformation and irradiation damage (51). The development of an FGM-type nuclear furnace is highly desired for this application. In optoelectronic devices, graded semiconducting layer

of InGaAs sandwiched between GaAs and InGaAs layers of fixed composition are used to increase the mean spacing of misfit dislocations and to control the density, and spatial distribution of threading dislocations (52). Furthermore, FGMs such as FGM semiconductors or superconductors that have gradients of quantum characteristics have been developed (50). Recently the application of FGMs in electronics is also increasing rapidly (50).

Piezoelectric Ceramic Materials. Recent trends in manufacturing electronic components are moving toward improved lightness, high density, and miniaturization. Therefore, electronic components that can be integrated with substrates or coating-materials and two- and/or three-dimensional composite components are strongly needed. The FGM technique is very suitable for fabricating such electronic components. By controlling the graded compositions between the component and substrate, the problems of spallation of components from the substrate can be effectively reduced and the stability of the electronic components can be greatly improved. PZT piezoelectric ceramics are widely used in electronic components, such as ultrasonic vibrators and ceramic filters (4,19). To improve the temperature stability of the piezoelectric and its dielectric properties, a third component that has a perovskite structure is usually incorporated into PZT to adjust the distributions of piezoelectric and dielectric properties and to increase the lifetime of the piezoelectric components (53–55). Piezoelectric bimorph benders are typical solid-state actuators that are composed of two piezoelectric ceramic strips bonded together by bonding agent (usually by epoxy resin) and fixed at one end to form a cantilever structure (4,10). The major problem of the device is that the bonding agent between the two oppositely poled piezoelectric ceramic strips may crack or peel off at low temperatures and may creep at a high temperature. Such problems may lead to deterioration of the electric-field induced displacement characteristics and reduced reliability and lifetime of the piezoelectric devices. To overcome these problems, a new type of piezoelectric actuator that has a sandwich structure has been recently developed, where the bonding agent is replaced by a FGM interlayer. The schematic structure of the FGM actuator is illustrated in Fig. 7a. The functional gradients of piezoelectric and dielectric properties vary in opposition to each other across the thickness of the FGM actuator, as shown in Fig. 7b. Thus, the FGM actuator acts like a monomorph-type piezoelectric device without a bonding agent and has improved durability and reliability in high-temperature operation. The compositions of the piezoelectric layer (that have high piezoelectric activity and low dielectric properties) and the dielectric layer (that has low piezoelectric activity and high dielectric properties) may be chosen by referring to the electrical property database of the $\text{Pb}(\text{Ni}_{1/3}\text{Nb}_{2/3})\text{O}_3\text{-PbZrO}_3\text{-PbTiO}_3$ (PNN-PZ-PT) ternary system. Because lattice parameters of the perovskite structure and the sintered characteristics of PNN are similar to those of PZT, they have better structural compatibility. By suitably selecting the compositions of the piezoelectric and dielectric layers, an FGM actuator can be prepared by powder metallurgy or ceramic sintering (56,57). The compositional and microstructural gradients

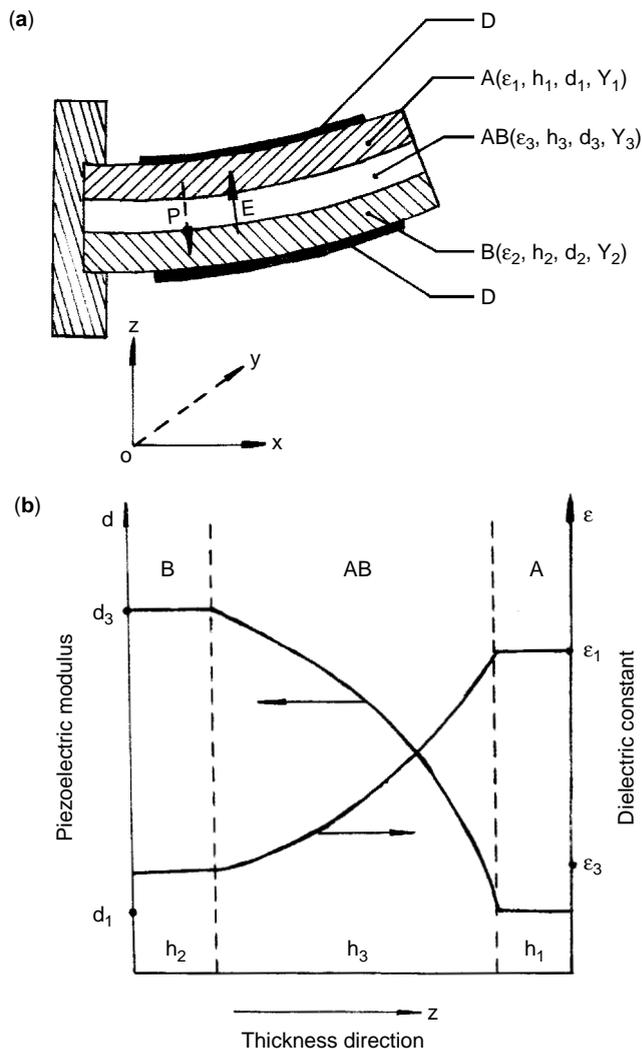


Figure 7. (a) Schematic structure of a functionally graded piezoelectric ceramic actuator. A: dielectric layer, B: piezoelectric layer, C: sandwich, D: electrode, E: applied electric field, and P: polarization; (b) variations of piezoelectric activity and dielectric activity across the thickness of the actuator.

across the section of the FGM actuator are shown in Figs. 8 and 9, respectively, and the related bending displacement of the FGM piezoelectric actuator is shown in Fig. 10.

Other types of monomorph piezoelectric ceramic actuators that have graded electrical resistivity across the thickness, as shown in Fig. 11, are fabricated by tape casting (58) or by uniaxially pressing two layers of different resistivities (59); graded resistivity is achieved by doping PZT ceramics with a gradual dopant concentration. Such gradual dopant concentration and resistivity gradients attained within a monolithic piezoelectric ceramic bar could result in a uniform stress gradient when a thickness poled monomorph is driven by an external electric field. This prevents a steep stress peak that is caused by a sharp poling gradient in a conventional bimorph. Then, significant increases in lifetime and reliability can be achieved. The application of an external electric field causes a monolithic ceramic bar to bend due to the differential strains induced

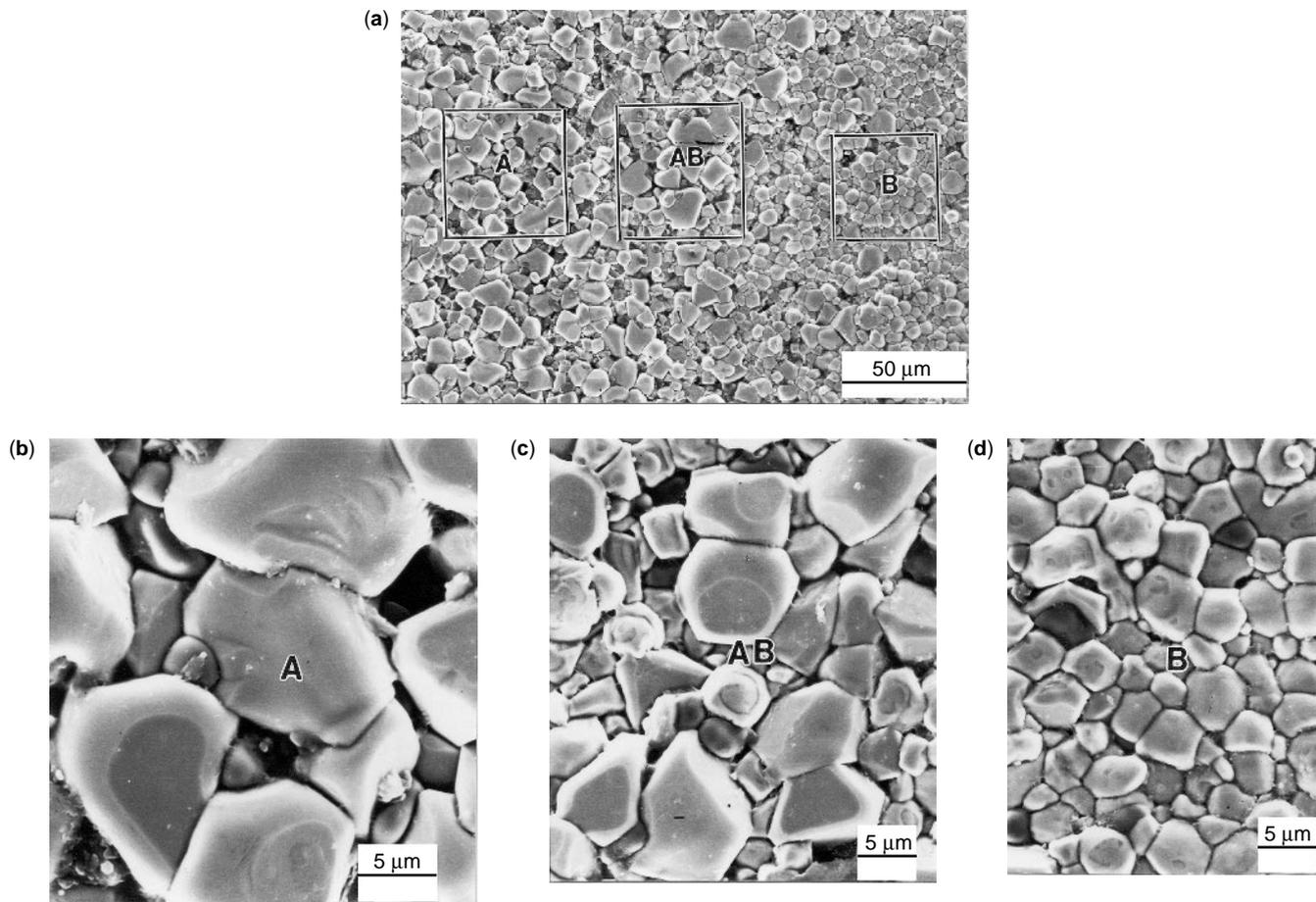


Figure 8. (a) General SEM observation on the morphologies of grains across the section of the FGM actuator; (b), (c), and (d) high magnification SEM images of regions A, AB, and B indicated in Fig. 8a. The grain sizes in the A, AB, and B regions are 10, 6, and 4 μm, respectively.

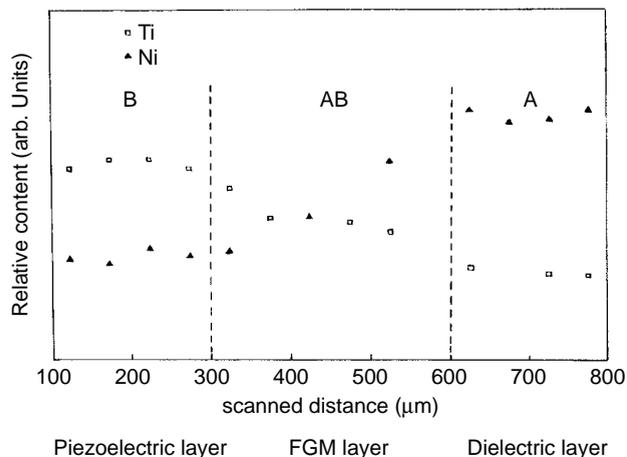


Figure 9. Typical compositional distributions of Ti and Ni elements across a section of the FGM actuator.

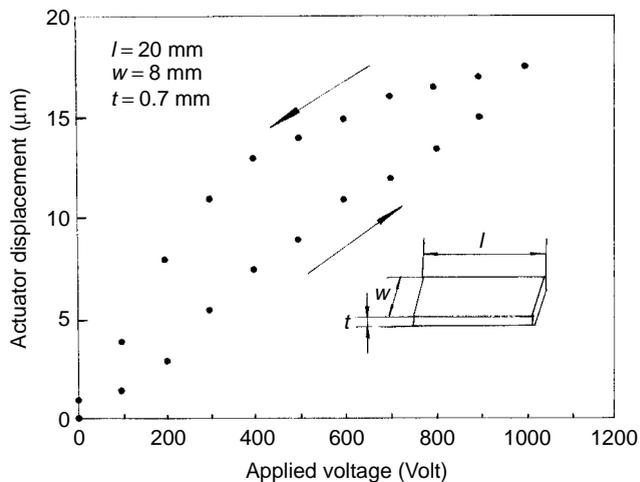


Figure 10. Measured microdisplacement versus applied voltage.

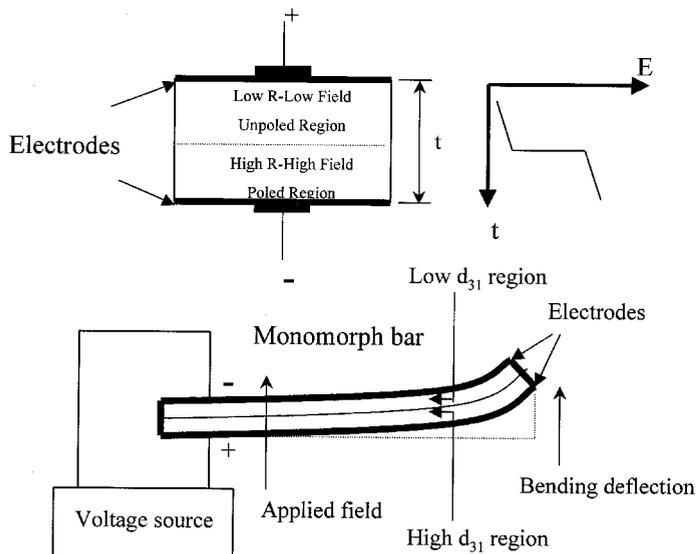


Figure 11. Schematic diagram of the field distribution and the structure of a monomorph that has graded electrical resistivity across its thickness.

by the piezoelectric effect. A monomorph that is initially unpoled PZT-5A on one side and PZT-5A modified by 2 mol% zinc borate on the other side shows a $28 \mu\text{m}/\text{cm}$ deflection in an external electric field of $5 \text{ kV}/\text{cm}$ (58). Chatterjee et al. also developed a multilayered rainbow type actuator that has graded piezoelectric properties across the thickness (60). Theoretical analyses of out-of-plane displacement and stress fields in such FGM piezoelectric actuator have been carried out by using a modified classical lamination theory (61). The results show that the linear profile has the best performance in higher out-of-plane displacement and moderate stress field and the concave displacement gave a larger out-of-plane displacement, but at the cost of higher stresses. It is possible to optimize the material property gradient profile within an FGM to provide out-of-plane displacements higher than those of standard piezoelectric bimorphs and still maintain moderate stress levels. The governing equations for a piezoelectric plate that has general symmetry and thickness-graded material properties are deduced from the three-dimensional equations of linear piezoelectricity by Mindlin's general procedure of series expansion (62). Recently, a more sophisticated model for calculating the bending behavior of FGM piezoelectric actuators that have arbitrary graded piezoelectric properties was proposed by T Hauke et al. (63). Based on their model, it seems possible to build bending actuators that have strongly reduced or vanishing internal mechanical stress induced by using a FGM. Some modeling results were verified experimentally by BaTiO_3 -based FGM piezoelectric actuators using different model structures.

Pyroelectric Ceramic Materials. In recent years, graded ferroelectrics have become of interest due to their unique properties (64–66) and the rapid development of ferroelectric-based heterostructures [e.g., ferroelectric semiconductor structures (67) and ferroelectric superconductor structures (68)]. For example, it has been demonstrated that a graded ferroelectric thin film can exhibit a large dc polarization offset. The sign of the offset is

determined by the direction of the composition gradient (64). Such an offset, reportedly, depends strongly on temperature and thus, can be used in pyroelectric detectors (66). The composition gradient can be tuned and tailored to produce properties that meet the designated requirements of practical applications, and many more efficient sensors, actuators, and energy converters are possible. Unlike in a simple structure of ferroelectric laminates, the polarization gradient in graded ferroelectric devices skews the potential wells to lower energy to favor populating the polarization states with lower energy and breaks the symmetry of the graded material, resulting in a self-poling effect. The “built-in” difference in free energy across a graded ferroelectric device is schematically shown in Fig.12, which can be likened to the “built-in” barrier potential of a p-n junction in semiconductor devices. The polarization gradient can be achieved in a variety of ways because the polarization of ferroelectric materials is a function of temperature, mechanical strain, and composition. Recently, several graded thin films, such as $(\text{Ba}_{1-x}\text{Sr}_x)\text{TiO}_3$ (BST) (66, 69, 70), PZT (71), $(\text{Pb,Ca})\text{TiO}_3$ (72–74), and $\text{Pb}_{1-3y/2}\text{La}_y(\text{Zr}_{0.4}\text{Ti}_{0.6})\text{O}_3$ (PLZT) (75), whose compositional gradients are normal to the

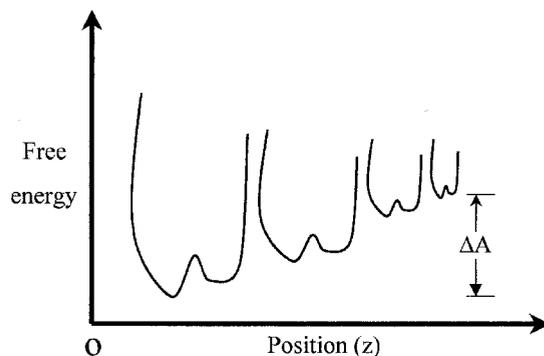


Figure 12. A graded potential well for electrons in a graded ferroelectric device skewed by its graded polarization below the lowest Curie temperature of the graded device.

growth surface, have been prepared by different methods. Such functionally graded devices exhibit several features, including broadening of the dielectric permittivity versus temperature plot, polarization offsets dependent on the direction of the compositional gradient, and electric field dependence of that offset, as shown in Fig. 13. A modified Slater model (76) was applied to explain these phenomena, in which the energy function for a ferroelectric material is characterized by

$$\xi = k_0(T - T_c)x^2 + Bx^4, \quad (12)$$

where x is the ion displacement relative to a central charge, T and T_c are the absolute and Curie temperatures, respectively, and k_0 and B are characteristic functions of temperature as well. Such a potential implies a

ferroelectric state when the temperature falls below the critical temperature T_c . When $T < T_c$, an energy extreme exists at $x = 0$ and also at

$$x = \pm \sqrt{\frac{k_0(T_c - T)}{2B}}. \quad (13)$$

The latter corresponds to the two spontaneous polarization states of ferroelectric dipoles. The self-poling phenomenon (“up” and “down” polarization hysteresis) is observed in graded ferroelectric thin films; however, its origin is still unclear, even though some legitimate interpretations have been proposed, including charge injection at the electrodes, nonuniformities in the thin film structures due to asymmetries in the growth process, leakage currents, and electrical breakdown of the films. Functionally graded ferroelectric thin films offers a significant approach to finding a method that combines the advantages of low-cost thin film processing with the high sensitivity of bulk ceramic and single-crystal materials. The unique properties observed in FGM ferroelectric materials may lead to new ferroelectric device applications; however, these unique properties have not been fully studied and are not well understood. The development of FGM ferroelectric thin films is still in its initial stage, so many problems remain unsolved and technical challenges lie ahead.

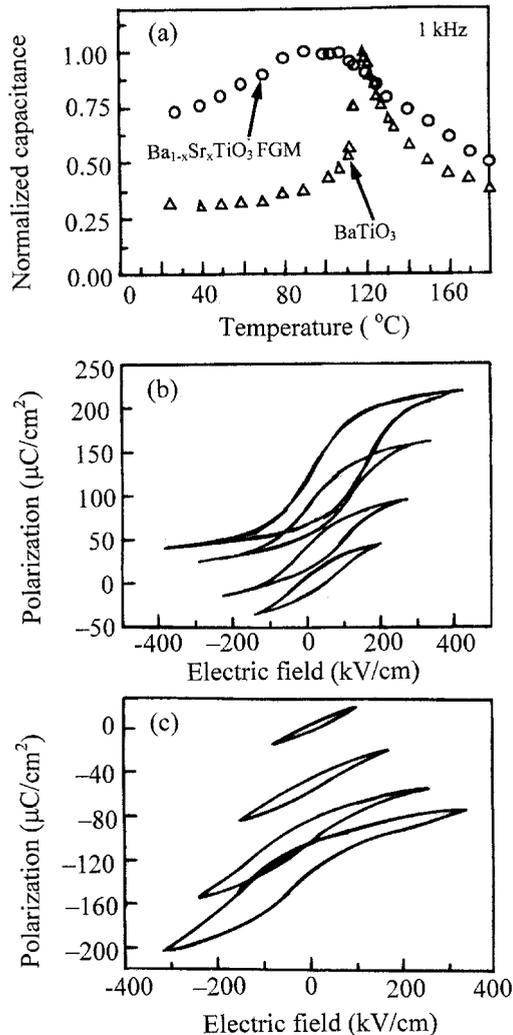


Figure 13. (a) Broad temperature dependence of the capacitance of a graded Ba_{1-x}Sr_xTiO₃ thin film at $x = 0, 0.47, 0.54, 0.70, 1.0$, compared with a single layer of BaTiO₃ thin film; (b) positive polarization offsets in upgraded (Pb,Ca)/TiO₃ thin films; (c) negative polarization in downgraded (Pb,Ca)/TiO₃ thin films. The magnitude of the offsets shown in Fig. 13b, c depends strongly on the driving electric field.

SUMMARY

In this article, we initially discussed current piezoelectric ceramic actuators and then focused on introducing piezoelectric ceramics, including the general characteristics, compositions, fabrication, and applications of piezoelectric ceramics, and finally reviewed the recent progress of FGMs and their applications in piezoelectric ceramic devices. FGMs, as a relatively new technology for solving the major problems and the development of the sharp interfaces at the joint of two dissimilar materials appear to have proven their worth in a variety of applications. Combinations of FGM technology with intelligent systems can be used to tailor, or tune, the overall performance of smart structural systems. It has been shown that the combination of theories, experiments, and computational simulations (such as neural network and fuzzy inference simulations) is an effective approach to developing FGMs. Because a completely or partially graded structure can be formed in any material and component, application of the FGM concept is virtually unlimited. The materials science developed in the twentieth century was based mainly on homogeneous material systems. It is likely that the twenty-first century will see the development of new materials science and technology for nonhomogeneous materials and systems that have optimized structures, like FGMs.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the helpful conversations with our colleagues at Nanjing University and Shanghai University (Jiading Campus). The authors thank Professor C.L.Choy of Hong Kong Polytechnic

University for his helpful comments on the manuscript. This work is supported by the National Natural Science Foundation of China, Nanjing University Talent Development Foundation, and a grant for a State Key Program for Basic Research of China.

BIBLIOGRAPHY

1. T. Kitamura, Y. Kodera, K. Miyahara, and H. Tamura, *Jpn. J. Appl. Phys.* **20**(4) (Suppl.): 97–100 (1981).
2. S. Takahashi, *Jpn. J. Appl. Phys.* **24**(2) (Suppl.): 41–45 (1985).
3. K. Uchino, M. Yoshizaki, H. Yamamura, N. Sakai, and H. Asakura, *Jpn. J. Appl. Phys.* **26**: 1046–1049 (1987).
4. K. Uchino, *Piezoelectric Actuator and Ultrasonic Motors*. Kluwer Academic, Boston, MA, 1996, pp.175–201.
5. A. Dogan, Q.C. Xu, K. Onitsuka, S. Yoshikawa, K. Uchino, and R.E. Newham, *Ferroelectrics* **156**(1–4): 1–6 (1994).
6. J.F. Fernandez, A. Dogan, Q.M. Zhang, J.F. Tressler, and R.E. Newham, *Sensors and Actuators A* **51**(1–3) 183–192 (1996).
7. G.H. Haertling, *Am. Ceram. Soc. Bull.* **73**(1): 93–96 (1994).
8. G.H. Haertling, *Soc. Photo Opt. Eng. Instr. Eng. (SPIE)* **3040**: 81–92 (1997).
9. P.S. Brody, *Ferroelectrics* **50**(1): 27–32 (1983).
10. K. Uchino, *Mater. Res. Soc. Bull.* **18**(4): 42–48 (1990).
11. L.E. Cross, S.J. Jang, R.E. Newham, S. Nomura, and K. Uchino, *Ferroelectrics* **23**(3): 187–192 (1980).
12. X.H. Zhu, Y. Tan, Q. Wang, Z.H. Cai, and Z.Y. Meng, *Jpn. J. Appl. Phys.* **32**(12A): 6623–6625 (1994).
13. L.E. Cross, *Ferroelectrics* **76**(3–4): 241–267 (1987).
14. J. Chen, Q.C. Xu, M. Blaszkiewicz, R. Meyer, Jr., and R.E. Newham, *J. Am. Ceram. Soc.* **75**(10): 2891–2892 (1992).
15. M.R. Alam, A. Kumar, N. Shu, H.L. Chan, and Q. You, *Appl. Surf. Sci.* **109/110**: 393–398 (1997).
16. W.G. Cady, *Piezoelectricity*. McGraw-Hill, NY, 1946, pp.1–20.
17. H. Jaffe, *J. Am. Ceram. Soc.* **41**(11): 494–498 (1958).
18. L. Benguigui, *Solid State Commun.* **11**: 825–828 (1972).
19. D. Berlincourt, *J. Acoust. Soc. Am.* **70**(6): 1586–1595 (1981).
20. D. Waller, T. Lqbal, and A. Safari, *J. Am. Ceram. Soc.* **72**(2): 322–324 (1989).
21. N. Lamberti, A. Iula, and M. Pappalardo, *IEEE Trans. Ultrasonics Ferroelectrics Frequency Control* **45**(1): 23–29 (1998).
22. Y. Sugawara, K. Onitsuka, S. Yoshikawa, Q. Xu, R.E. Newham, and K. Uchino, *J. Am. Ceram. Soc.* **75**(4): 996–998 (1992).
23. A. Bandyopadhyay, R.K. Panda, V.F. Janas, M.K. Agarwada, S.C. Danforth, and A. Safari, *J. Am. Ceram. Soc.* **80**(6): 1366–1372 (1997).
24. Z.G. Wei, R. Sandstrim, and S. Miyazaki, *J. Mater. Sci.* **33**(15): 3763–3783 (1998).
25. M. Niino, T. Hirai, and R. Watanabe, *Jpn. J. Soc. Comp. Mater.* **13**(6): 257–264 (1987).
26. Y. Itoh and H. Kashiwaya, *Jpn. J. Ceram. Soc.* **100**(4): 481–486 (1992).
27. M. Sasaki and T. Hirai, *JSME Int. J. Ser. I* **34**(2): 123–129 (1991).
28. A. Mortenson and S. Suresh, *Int. Mater. Rev.* **40**(6): 239–265 (1995).
29. S. Kitaguchi, *Jpn. Ceram. Soc.* **29**(3): 182–184 (1994).
30. R. Watanabe and A. Kawasaki, *Jpn. J. Soc. Powder Metall.* **39**(4): 279–286 (1992).
31. B. Ilchner and D. Delfosse, *Solid State Phenomena* **8/9**: 61–70 (1989).
32. J. Subrahmanyam and M. Vijayakumar, *J. Mater. Sci.* **27**(23): 6249–6273 (1992).
33. S.E. Niedzialek, G.C. Stangle, and Y. Kaieda, *J. Mater. Res.* **8**(8): 2026–2034 (1993).
34. A. Kawasaki and R. Watanabe, *Jpn. J. Inst. Met.* **51**(6): 525–529 (1987).
35. M. Niino and S. Maeda, *ISIJ Int.* **30**(9): 699–703 (1990).
36. K. Tanaka, Y. Tanaka, K. Enomoto, V.F. Poterasu, and Y. Sugano, *Computer Methods Appl. Mech. Eng.* **106**: 271–284 (1993).
37. J. Aboudi, S.M. Arnold, and M.J. Pindera, *Composite Eng.* **4**(1): 1–18 (1994).
38. Z.Z. Jin, in *Handbook of New Type Ceramics*, Z.Q. Qin, X.N. Qian, and P.F. He, eds., Jiangshu Science and Technology Press, Nanjing, 1996, pp. 591–595.
39. D.P. Miller, J.J. Lannutti, and R.D. Noebl, *J. Mater. Res.* **8**(8): 2004–2013 (1993).
40. Q.J. Zhang, L.M. Zhang, and R.Z. Yuan, *3rd Int. Symp. Struct. Functional Gradient Mater.* Lausanne, Switzerland, 1995, p.45.
41. A.J. Markworth, K.S. Ramesh, and W.P. Parks, Jr., *J. Mater. Sci.* **30**(9): 2183–2193 (1995).
42. K. Yamada, J. Sakamura, and K. Nakamura, *Mater. Sci. Forum* **308–311**: 527–532 (1999).
43. W.F. Shelley II, S. Wan, and K.J. Bowman, *Mater. Sci. Forum* **308–311**: 515–520 (1999).
44. S. Hoffmann, R. Liedtke, R. Slowak, U. Hasenkox, and R. Waser, *Mater. Sci. Forum* **308–311**: 544–548 (1999).
45. Y. Watanbe, Y. Nakamura, and Y. Fuki, *4th Porc. Symp. Functionally Graded Mater.* Tsukuba, Japan, 1996, pp. 713–718.
46. K. Niizuma and Y. Utsushikawa, *Mater. Sci. Forum* **308–311**: 567–572 (1999).
47. T. Endo, K. Uheda, and H. Takizawa, *4th Porc. Symp. Functionally Graded Mater.* Tsukuba, Japan, 1996, pp. 701–706.
48. X.R. Wang, H. Masumoto, Y. Someno, and T. Hirai, *Appl. Phys. Lett.* **72** (25): 3264–3266 (1998).
49. F. Nogata and H. Takahashi, *Composite Eng.* **5**(7): 743–751 (1995).
50. Y. Miyamoto, W.A. Kaysser, B.H. Rabin, A. Kawasaki, and R.G. Ford, *Functionally Graded Materials: Design, Processing and Applications*. Kluwer Academic, Boston, 1999, pp. 282–288.
51. M.M. Gasik and S. Ueda, *Mater. Sci. Forum* **308–311**: 603–607 (1999).
52. K.L. Kavanagh, J.C.P. Chang, J. Chen, J.M. Fernandez, and H.H. Wieder, *J. Vac. Sci. Technol. B* **10**(4): 1820–1823 (1992).
53. X.H. Zhu and Z.Y. Meng, *J. Mater. Sci.* **31**(8): 2171–2175 (1996).
54. X.H. Zhu, J. Xu, and Z.Y. Meng, *J. Mater. Sci.* **32**(16): 4275–4282 (1997).
55. J.H. Cho, I.K. Park, H.T. Chung, and H.G. Kim, *Jpn. J. Appl. Phys.* **36**(1A): 181–187 (1997).
56. X.H. Zhu, Q. Wang, and Z.Y. Meng, *J. Mater. Sci. Lett.* **14**(7): 516–518 (1995).

57. X.H. Zhu and Z.Y. Meng, *Sensors & Actuators A: Physical* **48**(3): 169–176 (1996).
58. C.C.M. Wu, M. Kahn, and W. Moy, *J. Am. Ceram. Soc.* **79**(3): 809–812 (1996).
59. S.U. Adikary, Z.Y. Meng, and D.R. Jin, *J. Mater. Sci. Technol.* **21**(6): 561–566 (200).
60. Controlled Composition and Crystallographic Changes in Forming Functionally Graded Piezoelectric Transducer, U.S. Pat. 5,900,274, 1999. D.K. Chatterjee, S.K. Ghosh, and E.P. Furlani.
61. A. Almajid, M. Taya, and S. Hudnut, *Int. J. Solids Struct.* **38**: 3377–3391 (2001).
62. P.C.Y. Lee and J.D. Yu, *IEEE Trans. Ultrasonics Ferroelectrics Frequency Control* **45**: 236–250 (1998).
63. T. Hauke, A. Kouvatov, R. Steinhausen, W. Seifert, H. Beige, H.T. Langhammer, and H.P. Abicht, *Ferroelectrics* **238**(1–4): 759–766 (2000).
64. N.W. Schubring, J.V. Mantese, A.L. Micheli, A.B. Catalan, and R.J. Lopez, *Phys. Rev. Lett.* **68**(11): 1778–1781 (1992).
65. J.V. Mantese, N.W. Schubring, A.L. Micheli, A.B. Catalan, M.S. Mohammed, R. Naik, and G.W. Auner, *Appl. Phys. Lett.* **71**: 2047–2049 (1997).
66. M.S. Mohammed, G.W. Auner, R. Naik, J.V. Mantese, N.W. Schubring, A.L. Micheli, and A.B. Catalan, *J. Appl. Phys.* **84**(6): 3322–3325 (1998).
67. Y.T. Kim and D.S. Shim, *Appl. Phys. Lett.* **71**: 3507–3509 (1997).
68. H. Lin, J.J. Wu, K. Xie, Y. Li, and A. Ignatiev, *Appl. Phys. Lett.* **65**(8): 953–955 (1994).
69. S. Hoffmann, R. Liedtke, R. Slowak, U. Hasenkox, and R. Waser, *Mater. Sci. Forum* **308–311**: 544–548 (1999).
70. X.H. Zhu, S.G. Lu, H.L.W. Chan, C.L. Choy and K.H. Wang, *Appl. Phys. A* in press.
71. M. Brazier, M. McElfresh, and S. Mansour, *Appl. Phys. Lett.* **72**(9): 1121–1123 (1998).
72. D.H. Bao, X. Yao, and L.Y. Zhang, *Appl. Phys. Lett.* **76**(19): 2779–2781 (2000).
73. D.H. Bao, X. Yao, and L.Y. Zhang, *Appl. Phys. Lett.* **77**(7): 1041–1043 (2000).
74. D.H. Bao, N. Mizutani, X. Yao, and L.Y. Zhang, *Appl. Phys. Lett.* **77**(8): 1203–1205 (2000).
75. I. Boerasu, L. Pintilie, and M. Kosec, *Appl. Phys. Lett.* **77**(14): 2231–2233 (2000).
76. J.V. Mantese, N.W. Schubring, A.L. Micheli, A.B. Catalan, M.S. Mohammed, R. Naik, and G.W. Auner, *Appl. Phys. Lett.* **71**(14): 2047–2049 (1997).

ADAPTIVE COMPOSITE SYSTEMS: MODELING AND APPLICATIONS

A. SULEMAN

Instituto Superior Técnico
Lisbon, Portugal

INTRODUCTION

In the past decade, technological developments in materials and computer sciences have evolved to the point where their synergistic combination have culminated in a new field of multidisciplinary research in adaptation. The

advances in material sciences have provided a comprehensive and theoretical framework for implementing multifunctionality in materials, and the development of high speed digital computers has permitted transforming that framework into methodologies for practical design and production. The concept is elementary: a highly integrated sensor system provides data on the structural environment to a processing and control system which in turn signals integrated actuators to modify the structural properties appropriately.

Multifunctional materials embedded in adaptive composite systems have presented exceptional promise in engineering design problems that require solutions in active vibration suppression, shape control, and noise attenuation. Piezoelectric materials, shape memory alloys, and magnetostrictive materials are the three most recognized types. These materials develop strains or displacements when exposed to electric, thermal, and magnetic fields, respectively.

ACTUATORS AND SENSORS

Actuation Technology

When a **shape memory** alloy is heated above a critical temperature, the material recovers its original predeformed shape. The most common commercially available shape memory alloy is Nitinol. This alloy is very ductile and can be deformed easily. In addition, it also has good strength and strain rate, it is corrosion resistant, and it is stable at high temperature. A limited number of efforts aimed at using shape memory alloys as actuators in composite structures have been made. Recent studies include the work by Boyd and Lagoudas (1) in developing a micromechanical model for shape memory composites, and Sullivan (2), who developed a model to predict shape memory composite behavior. Other significant theoretical studies on modeling systems that contain shape memory alloys include the works by Liang and Rogers (3), Feng and Li (4), and Graesser and Cozzarelli (5). Research on using shape memory alloys in active structural control include the works by Baz et al. (6), Ikegami et al. (7), and Maclean et al. (8).

Magnetostrictive materials change dimensions when placed in a magnetic field. Terfenol-D is the most popular commercially available magnetostrictive material. Recent research on magnetostrictive materials shows that they provide strokes significantly larger than their electromechanical counterparts; however, they tend to be difficult to implement in structural systems (9). Work on composites that incorporate magnetostrictive materials include the unimorphs by Honda et al. (10) and the microcomposite systems by Bi and Anjanappa (11).

Piezoelectric materials present two distinct characteristics: the “direct” piezoelectric effect occurs when a piezoelectric material subjected to a mechanical stress becomes electrically charged. Thus, these devices can be used to detect strain, movement, force, pressure, or vibration by developing appropriate electrical responses. The “converse” piezoelectric effect occurs when the piezoelectric material becomes strained when placed in an electric

field. The ability to induce strain can be used to generate a movement, force, pressure, or vibration by applying a suitable electric field. The most popular commercial piezoelectric materials are lead zirconate titanate (PZT) and polyvinylidene fluoride (PVDF). The potential of applying piezoelectric materials as distributed actuators in composite structures has resulted in several significant studies. Suleman and Venkayya (12) modeled a simple composite plate structure of piezoelectric layers using classical lamination theory; however, the first reported studies on adaptive composites include the works by Bailey and Hubbard (13), Crawley and de Luis (14), Leibowitz and Vinson (15), and Wang and Rogers (16).

The electrostrictive phenomenon is a nonlinear property that exists in all dielectric materials. When an electric field is applied across an electrostrictive material, the positive and negative ions are displaced, and a strain is induced in the material. The resulting strain is proportional to the square of the applied electric field and is independent of the applied electric field's polarity. Because the strain is proportional to the square of the electric field, the strain is always positive. This is analogous to the magnetostrictive behavior described earlier. The most popular electrostrictive material is lead magnesium niobate (PMN); however, this material is still not widely available commercially. These materials generally offer higher electrically induced strain with lower hysteresis than piezoelectric materials, due to nonlinearities; however, constitutive models for electrostrictors are not as mature as models for piezoelectrics. Hom and Shankar (17) formulated a fully coupled constitutive model for electrostrictive ceramic materials. Electrostrictive materials used as distributed actuator elements in adaptive composites have not been reported in the literature.

Magneto- and electrorheological fluids are multiphase materials that consist of a dispersion of polarizable particles in a carrier oil, and they exhibit the properties of a typical viscoelastic material. The use of electrorheological materials for vibration damping has been the subject of considerable research because these materials exhibit fast, reversible, and controllable changes in behavior. However, despite advances in sensing and controls, fundamental rheological research applied to vibration damping has lagged behind. Specifically, essential information on material-based structural reliability and controllability is still needed to implement such systems successfully.

Typically, the performance of an actuator is evaluated in terms of the following characteristics: displacement (the ability of the actuator to displace an object), force generation (the amount of force the actuator can produce), hysteresis (the degree of reproducibility in positioning operations), response time (how quickly an actuator can start the actuation process), bandwidth (the range of frequencies in which the actuator can operate effectively), temperature range of operation, repeatability and precision of the actuator, power required to drive the actuator, mass of actuator material required for a given displacement; and cost. Table 1 presents the general characteristics of commercially available actuators (18).

The piezoelectric PZT provides the potential for the greatest force handling capability. PZT also operates across

Table 1. Actuator Technology Assessment

	Nitinol	Terfenol-D	PZT
Energy	Heat	Magnetic field	Electric field
Hysteresis	High	Low	Low
Bandwidth	Low	Moderate	High
Accuracy	Poor	High	High
Response time	Low	Fast	Very fast
Power use	High	Moderate	Moderate
Maturity	New	New	Established

the highest bandwidth of the microactuators and has among the highest displacements. Electrostrictive PMN possesses the lowest hysteresis of any of the actuator materials. However, the temperature operating limits for PMN would require special insulation. Although PZT is preferred for most applications, future commercial applications may favor direct replacement of PZT with PMN because of its superior hysteresis efficiency.

The shape memory material, Nitinol, produces the greatest displacements but the weakest force of the actuators under consideration. This material is very ductile and consequently does not support as much force as other actuator materials. However, it does give it the advantage of being easily shaped into different actuator geometry. It is less desirable from the point of view of precision and economy of design due to its power consumption, accuracy, and hysteresis characteristics. It can also be observed that Terfenol-D and PZT compare very closely in these general characteristics. For deformation of thin structural elements, the most widely used multifunctional materials to date have been piezoelectric actuators. Piezoelectrics have higher bandwidths than are possible in shape memory alloys, they are more compact than magnetostrictive devices, and they are bidirectional by nature, unlike electrostrictive materials.

Sensing Technology

Optical fibers make excellent strain sensors because they are immune to electromagnetic interference. Optical fibers can be bonded to the surface of a structure or embedded directly into the structure. There are many types of optical fiber sensors. The more useful fiber-optic strain sensors uses the intrinsic properties of the optical fiber. In an intrinsic fiber measurement, one or more of the optical field parameters, which include frequency, wavelength, phase, mode index, polarization, index of refraction, and attenuation coefficient, are affected by the environment.

Piezoelectric sensors tend to operate best in dynamic situations because the induced charge imbalances created by straining the material dissipate with time. How quickly this occurs depends on the material's capacitance, resistivity, and output loading. Force transducers use piezoelectric elements to produce an electrical output that is proportional to the applied force. The force transducer is mounted in series with the force transmission path to expose the piezoelectric element directly to the forces that are to be measured. Because the piezoelectric is preloaded, the force sensor can measure both tensile and compressive

Table 2. Sensor Technology Assessment

	Fiber Optics	PZT
Sensitivity	Moderate	Moderate
Gauge length	Moderate	High
Bandwidth	High	Moderate
Resolution	High	Moderate
Temperature range	High	High

forces. High stiffness ensures high resonant frequency, and it will have a minimum effect on the structural integrity.

In the case of sensors, the technologies considered for adaptive composites must withstand the composite manufacturing process. Embedding issues make Nitinol a questionable choice. If the shape memory alloy had to be electrically insulated from the conductive composite, it would complicate the composite manufacturing process and increase the cost significantly.

The performance of any sensor can be evaluated in terms of sensitivity (amount of signal that a sensor produces for a given change in the variable), the length across which the measurement is made, bandwidth (the frequency range over which the sensor remains effective), response time (the speed at which the sensor can respond to a change in the variable), the temperature range across which the sensor can operate, repeatability and precision of the actuator, weight, and cost. Table 2 presents a relative assessment of the sensor types considered suitable for embedding in adaptive composite systems.

The temperature range is critical for an adaptive composite system because it is anticipated that sensors may be embedded and would therefore undergo the composite curing process. Embedding would be feasible for fiber optic sensors but not as desirable for PZT strain sensors. From the point of view of assembly and handling, strain gauges or even PZT are favored because of the availability of knowledge and experience with these techniques.

ADAPTIVE COMPOSITE MODELING

This section presents the recent advances and trends in finite-element modeling of adaptive composite systems, and emphasizes particularly the development of finite-element models for adaptive composites. Benjeddou (19) presented an excellent survey of finite-element developments in piezoelectric elements. Electromechanical coupling is the major feature added by a piezoelectric material to standard structural finite-element modeling.

For the last two decades, there has been increased research activity in finite-element modeling of adaptive composites. The primary interest has been in the analysis of piezoelectrically actuated composites, and early investigations were devoted to three-dimensional electromechanical elements. Among the reported studies, Tzou and Tseng (20) used variational methods to model finite-element piezoelectric solids. Ha et al. (21) developed an eight-node, three-dimensional, composite brick, finite-element for modeling the dynamic and static response of laminated composites that contain distributed piezoelectric ceramics

subjected to mechanical and electrical loading. The electrical potential is taken as a nodal degree of freedom that leads to an element that has four degrees of freedom per node. These models using three-dimensional finite elements can give accurate results by setting computationally expensive refined meshes that have acceptable aspect ratios. Classical plate theories have been proposed to analyze rectangular piezoelectric plates (22–25). Other plate formulations include the work reported by Chandrashekhara and Agarwal (26), who used a finite-element formulation based on first-order shear deformation theory to model the behavior of laminated composite plates that have integrated piezoelectric sensors and actuators. The model developed does not introduce voltage as an additional degree of freedom. Tzou and Ye (27) presented a laminated quadratic C^0 piezoelectric triangular shell finite-element using the layerwise constant shear angle theory that accounts for a constant approximation of the nonlinear cross-sectional warping applied to piezoelectric laminated systems. A model that contains an actuator element, an adhesive interfacial element, and an eight-node isoparametric plate element was developed by Lin et al. (28). An analytic solution is also derived and results are compared with the finite-element model. Chattopadhyay and Seeley (29) used a finite-element model based on a refined higher order theory to analyze piezoelectric materials surface bounded or embedded in composite laminates. The displacement field accounts for transverse shear stresses through the thickness and satisfies the boundary conditions at the free surfaces. Through numerical examples, they showed that the refined theory captures important higher order effects that are not modeled by the classical laminate theory. Recently, models using higher order theories for piezoelectric laminates can be found in Reddy and Mitchell (30) and Jonnalagadda et al. (31), among others.

Very few composite shell elements that have electromechanical properties have been reported in the literature. A four-noded shell element that extends the shallow shell shear deformation theory has been proposed, using an equivalent single-layer model for a three-layer shell (32). An eight-noded quadrilateral shell element (33) that has no electrical degrees of freedom using the 3-D degenerated shell theory has also been proposed, where the piezoelectric effect was treated as an initial strain problem. An axisymmetric three-node triangular shell element has also been developed to study mooney transducers (34). A 12-noded degenerated 3-D shell element that has a layerwise constant shear angle has been formulated (35). However, more research is required to understand and quantify the influence of the curvature on the piezoelectric actuators and sensors. Suleman and Venkayya (12) reported on an efficient finite-element formulation for vibration control of a laminated composite plate/shell that uses piezoelectric sensors and actuators. By modeling the plate and the sensor/actuator system with the four-noded, bilinear, Mindlin plate element, the problems associated with the solid element are eliminated, and modeling the plate and the sensor/actuator system with the four-noded, bilinear, Mindlin plate element considerably reduces the problem size.

Finite-Element Formulation

To derive the equations of motion for laminated composite structures (plate or shell) that have piezoelectrically coupled electromechanical properties, recall the generalized form of Hamilton's principle,

$$\delta \int_{t_1}^{t_2} (T - \Pi + W_e) dt = 0, \quad (1)$$

where T is the kinetic energy, P is the potential energy, and W_e is the work done by the electrical field. The kinetic and potential energies can be written in the form

$$\begin{aligned} T &= \int_V \frac{1}{2} \rho \dot{u}^T \dot{u} dV, \\ \Pi &= \int_V \frac{1}{2} \bar{S}^c{}^T \bar{T}^c dV, \end{aligned} \quad (2)$$

where \bar{S}^c and \bar{T}^c are the generalized electric strain and stress vectors. The work done by the electrical forces can be written as

$$W_e = \int_{V_p} \frac{1}{2} \bar{S}^e{}^T \bar{T}^e dV_p, \quad (3)$$

where \bar{S}^e is a vector of electrical field (potential/length), \bar{T}^e is a vector of electrical displacement (charge/area), and V_p is the volume pertaining to the piezoelectric material.

Constitutive Relationships

To apply piezoelectric materials in adaptive composite systems, the properties are defined relative to the poling direction through the thickness, and the material is approximately isotropic in the other two directions. In matrix form, the equations governing these material properties can be written as

$$\bar{T}^e = \mathbf{e}^T \bar{S}^c + \varepsilon \bar{S}^e, \quad (4)$$

$$\bar{T}^c = \mathbf{c}^T \bar{S}^c - e \bar{S}^e, \quad (5)$$

where \mathbf{e} is the dielectric permittivity matrix, ε is the dielectric matrix at constant mechanical strain, and \mathbf{c} is the matrix of elastic coefficients at constant electrical field strength.

Strain–Displacement Relationships

The strain–displacement relationships for an electromechanical composite element are based on the Mindlin first-order shear deformation theory. The matrix that relates the generalized strains to the displacements and rotations

can be written in the form

$$\bar{S}^c = \begin{bmatrix} \frac{\partial}{\partial x} & 0 & \frac{1}{R_{xx}} & 0 & 0 \\ 0 & \frac{\partial}{\partial y} & \frac{1}{R_{yy}} & 0 & 0 \\ \frac{\partial}{\partial y} & \frac{\partial}{\partial x} & \frac{2}{R_{xy}} & 0 & 0 \\ 0 & 0 & 0 & 0 & z \frac{\partial}{\partial x} \\ 0 & 0 & 0 & z \frac{\partial}{\partial y} & 0 \\ 0 & 0 & 0 & z \frac{\partial}{\partial x} & z \frac{\partial}{\partial x} \\ 0 & 0 & \frac{\partial}{\partial x} & 0 & 1 \\ 0 & 0 & \frac{\partial}{\partial y} & 1 & 0 \end{bmatrix} \begin{Bmatrix} u \\ v \\ w \\ \theta_x \\ \theta_y \end{Bmatrix}. \quad (6)$$

Electric Field Potential Relationships

For a laminated plate/shell that contains piezoelectric layers/patches polarized along the transverse axis, the in-plane electric fields vanish. The electric field potential relationships are $\bar{S}^e = -\nabla \bar{\phi}$, where $\bar{\phi}$ is the vector of potential degrees of freedom. It is assumed that these are constant throughout the plane of the piezoelectric patch/layer and vary linearly through the thickness. Thus $S_i^e = \frac{1}{t_{pi}} \phi_i$, where t_{pi} is the thickness of the i th piezoelectric layer/patch.

Substituting for the generalized stress and strain expressions in Eq. (1) gives the mass, elastic, and electrical stiffness matrices. For the entire structure, using the standard assembly technique for the finite-element method and applying the appropriate boundary conditions, we obtain the complete equations of motion for a piezoelectrically actuated adaptive composite structure:

$$\begin{bmatrix} M_{cc} & 0 \\ 0 & 0 \end{bmatrix} \begin{Bmatrix} \ddot{\bar{U}}_c \\ \ddot{\bar{U}}_e \end{Bmatrix} + \begin{bmatrix} \overbrace{K_{cc} & 0}^{\text{elastic}} \\ 0 & 0 \end{bmatrix} \begin{Bmatrix} \bar{U}_c \\ \bar{U}_e \end{Bmatrix} + \begin{bmatrix} 0 & \overbrace{K_{ce}}^{\text{electric}} \\ \overbrace{K_{ec}} & \overbrace{K_{ee}}^{\text{electric}} \end{bmatrix} \begin{Bmatrix} \bar{U}_c \\ \bar{U}_e \end{Bmatrix} = 0 \quad (7)$$

To avoid shear locking in thin structural elements, a reduced order integration technique is used to evaluate the shear stiffness term. For the actuation mechanism, the structure is given applied voltages on the layers/patches, and the static deformation is given by $\bar{U}^c = -K_{cc}^{-1} K_{ce} \bar{U}^e$. The sensing mechanism is governed by the equation $\bar{U}^e = -K_{ee}^{-1} K_{ec} \bar{U}^c$.

Finite-element techniques have also been developed for other active materials, such as electrostrictive (36,37), and magnetostrictive (38–40) systems. Boundary finite-element techniques have also been proposed for piezoelectric solids (41,42). However, most of the research activity

has been devoted to developing models for piezoelectric-based adaptive structural systems. It can be concluded that although relative maturity has been reached, there is a lack of curved two-dimensional and active constrained layer damping shell finite elements and some quadratic elements that save electric degrees-of-freedom representation.

APPLICATIONS

The application of adaptive composite systems in structural systems is vast and diverse. The literature indicates that applications of adaptive composite structures have been devoted mostly to dynamics and control analysis of adaptive beams and plates. Other applications include analysis of the thermal effect on active materials (43,44), noise suppression (45–47), composite delamination (48), precise position control (49), topology optimization (50), panel flutter suppression (51), and structural health monitoring and self-repair (52–55).

In this article, design problems in aircraft structures that require active solutions using adaptive composites to suppress vibration and control shape are presented. Adaptive composite solutions to vertical tail buffeting suppression, supersonic panel flutter envelope extension, and aeroacoustic noise suppression are discussed.

Vertical Tail Buffeting Suppression

The primary passive solutions to dynamic aeroelastic problems are increased stiffness and mass balance, and these were used as early as 1922. They still comprise the basic passive means of improving the response and stability of an aircraft. Increased stiffness has come to include composite tailoring which can greatly alter the stability characteristics of a given wing. Aeroelastic tailoring has made airplane configurations, such as a forward swept wing, feasible. Recently, aeroelastic tailoring has been used to enhance the performance of controllers. By the 1950s, the concept of active control to improve the aeroelastic performance of wings had emerged. One of the first aeroelastic control experiments began in 1972 in NASA Langley Research Center's Transonic Dynamics Tunnel (56,57). The model was a clipped delta wing that had a leading edge and a trailing edge actuator. Since that time, control experiments using flap actuators have studied a variety of control design techniques and objectives. The Active Flexible Wing (AFW) project used the flexibility of the wing in conjunction with active controls to provide greater maneuverability. Multifunctional control systems were also investigated in the AFW project. In addition to flutter suppression experiments, there have also been gust alleviation experiments.

Until now, the unifying factor of the previous work was that the actuators were all aerodynamic control surfaces. One of the reasons for these actuators is that they already exist on current aircraft. The development of aeroelastic control has evolved in the last couple of decades, a new actuation concept has emerged for structural control, and this is direct strain actuation. Direct strain actuators include materials such as piezoelectric ceramics and shape memory alloys. Piezoelectrics have several advantages

over hydraulic actuators because they have a higher frequency bandwidth of operation and because they act directly in the structure by straining it.

The DARPA/USAF Wright Laboratory program was to design, build, and test wind tunnel models to quantify performance improvements that could be achieved by incorporating smart materials such as PZTs for actuation and sensing systems in aircraft wings. The Smart Wing program performed wind tunnel tests at the NASA Langley Transonic Dynamic Tunnel to demonstrate the use of smart actuator systems in a realistically modeled aircraft operating environment. The wind tunnel test quantified aerodynamic improvements of two concepts: the use of embedded SMA wires in the trailing edge to provide a smooth variable contoured control surface and SMA torque tubes built into the wing structure that enabled the wing to be twisted or torqued.

Several active vibration suppression concepts have been investigated by a program shared between Daimler-Benz Aerospace Military Aircraft (DASA), Daimler-Benz Forschung (DBF), and Deutsche Forschungsanstalt für Luft und Raumfahrt (DLR); two concepts using aerodynamic control surfaces and two concepts using piezoelectric components. In the DASA concept, a thin surface of piezoactuators is set out to flatten the dynamic portion of the combined static and dynamic maximum bending moment loading case directly in the shell structure. The second piezo concept by DLR involves preloaded PZT block actuators at structural fixtures. Both piezoelectric strategies were aimed at straight open-loop performance related to concept weight penalty and input electric power.

The shape control of a flexible wing structure has a great potential for improving the aerodynamic lifting-surface performance. Significant reductions in the shock-induced drag can be achieved by small adaptive modifications to the wing cross-sectional profile. The results presented here are the experimental wind tunnel results obtained to suppress buffeting actively by using piezoelectric actuators to control the shape of the wing camber. Two methodologies are presented and compared. One uses the piezoelectric actuators in the traditional way, placing them at the root to generate bending moments that compensate for the mechanical vibration along the span of the wing. The other method consists of using the piezoelectrics to achieve airfoil camber shape control to control lift, so that the change of lift can be used to generate the same type of moments and hopefully result in using fewer piezoactuators and less energy.

The plate wing motion was controlled by six piezoceramic actuators bonded on the surfaces at the wing mount root portion and two shape control actuators near the wing tip. The piezo wafer sensors were located at the wing mount root, and the signal was sent to a digital signal processor through filters. The control signal was sent to power amplifiers. The amplified signal drove the piezoceramic actuators and attenuated vibration at the wing mount. This signal also drove the shape control actuators. A control law was designed based on a discrete system model.

Photographs of the test model system are presented in Fig. 1a. The dimensions were determined on the basis of the wind tunnel size, blowing air velocity, and the limitations

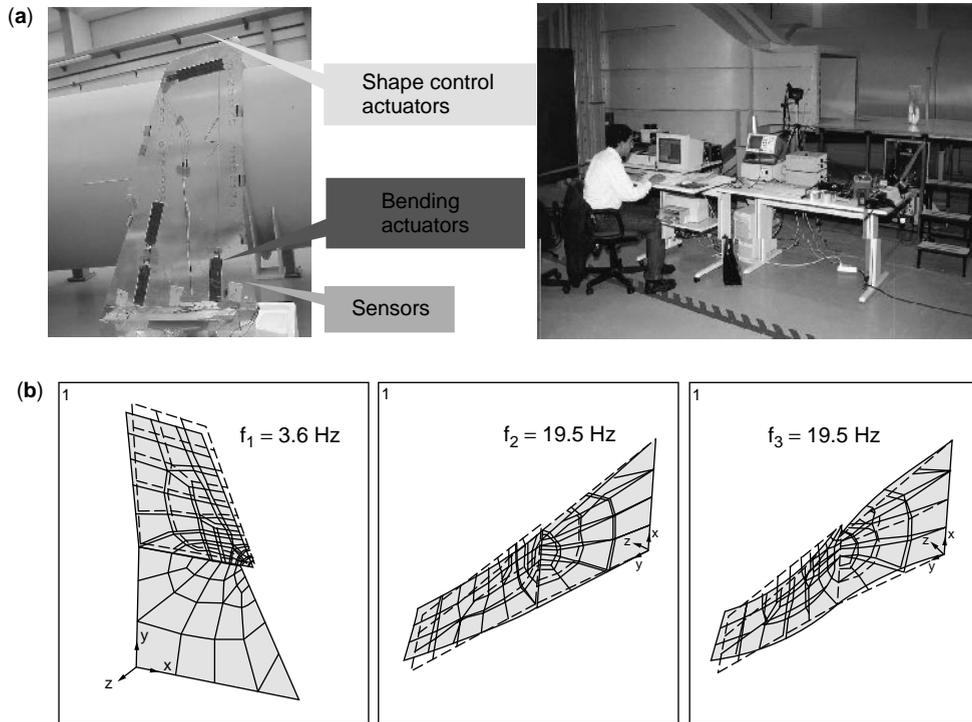


Figure 1. (a) The experimental test article that has surface-bonded piezoelectric actuators; (b) first three modal frequencies and shapes used in the controller design.

of the piezoceramic actuators. Structurally, it was aimed at having a fairly flexible wing that had low bending and torsion mode frequencies.

Six piezoceramic sensor patches (1.5' long; 1.00' wide; 0.0075' thick) are bonded to the top and bottom of the plate near the root. The actuators are divided in two groups, those to produce moments, two pairs near the leading edge and one near the trailing edge, all near the root to maximize the bending moment. The other group, which is responsible for the camber shape control, is located near the wing tip to change the wing camber and produce favorable changes in lift. A finite-element modal analysis was performed to generate natural frequencies and model shapes (Fig. 1b).

The experimental setup is based on a DSP state-of-the-art laboratory facility created for testing and validating the theoretical models and applying of active control design. A rigid plate was placed at the upstream end of the test section. The wake from this obstacle was used to generate buffeting flow. The position of this object was easily adjusted so that the resulting wake impinged on the model mounted downstream. Using the control law gain set to the desired value, the tunnel speed was increased to and then held constant at a preselected value. Buffet response measurements were made at a velocity of 5.5 m/s. The output signal from the PZT sensor was routed to a transfer function analyzer that was used to calculate the autocorrelation function of the response signal.

With respect to the gust alleviation tests, a special flap was mounted upstream of the model to deflect the flow so that a vertical wind component would appear (20% of the wind tunnel velocity). The wind tunnel speed was

maintained at 5 m/s, and the flap would change its incidence at 2.5 Hz, creating a vertical gust of 1 m/s that excited the bending on the wing. Maintaining the gust at this frequency and turning on the controller, the amplitude decreased considerably.

An appreciable buffeting reduction has been obtained, especially when using airfoil shape control, which combined with the root actuators to decrease the average buffeting amplitude from 32 to 47.5% (Table 3). The airfoil shape control also decreased the frequency of the vibration by 34%. The airfoil shape control has presented a feasible engineering design solution where the piezoelectric shape control actuators are used to create favorable changes in lift characteristics.

Supersonic Panel Flutter

Panel flutter is a self-excited, dynamic instability of thin plate or shell-like structural components of flight vehicles. It is a supersonic/hypersonic aeroelastic phenomenon that is often encountered in the operation of high-speed aircraft and missiles. The mode of failure for panel flutter is fatigue due to limit-cycle oscillations. Therefore, to increase the critical dynamic pressure or to suppress the limit-cycle oscillations is an important design consideration. Conventional design is to increase the panel stiffness, which in turn, may result in additional weight.

Many researchers have investigated the supersonic panel flutter suppression problem. Dowell (58) gave a comprehensive overview of panel flutter, and Olson (59) and Bismarck-Nasr (60) presented applications of the finite-element method to linear flutter problems for isotropic

Table 3. Active Wing Shape Control (Buffeting Suppression)

	Bending Control Only		Bending + Shape Control	
	Average Amplitude	Frequency	Average Amplitude	Frequency
Control law off	585 mm	1 Hz	620 mm	1 Hz
Control law on	395 mm	0.9 Hz	325 mm	0.66 Hz
Improvement	32%	10%	47.5%	34%

materials. Lee and Cho (61) and Liaw (62) investigated the use of composite panels in flutter problems due to stringent performance requirements in high-speed aircraft. Lee and Lee (63) performed supersonic flutter analysis of anisotropic panels taking into consideration the effects of panel geometry, boundary conditions, lamination scheme, flow directions, and thermal effects. Gray and Mei (64) conducted research on various theoretical considerations and analytical methods for investigating of nonlinear panel flutter. Zhou et al. (65) extended this work to include finite-element methods. Scott and Weishaar (66) proposed a linear panel flutter control using piezoelectric actuators and sensors. Paige et al. (67) extended this study to include anisotropic composites. Xue and Mei (68) studied the feasibility of applying shape memory alloys in linear panel suppression.

The flutter velocity of panels, or similarly the critical dynamic pressure, can be increased by making piezoelectric actuators induce in-plane tensile forces that alter the effective stiffness of the panel. The same voltage is applied to the top and bottom piezoelectric layers, resulting in uniform compression or tension in the plate.

Consider a case #1 where the patches cover only the center 6% of the plate area (Fig. 2). Here, the mass increases by 17% due to the addition of the piezo patches to the base structure. Obviously, the effective stiffness also increases. It was observed that the critical dynamic pressure increased from 267×10^3 to 323×10^3 N/m², an improvement of 27%. Note that this increase is solely due to the bonding of the piezo patches to the top and bottom surfaces of the aluminum panel. Subsequently, the piezo

patches were actuated with a voltage of 400 V and, in this instance, a further increase of 42% was attained, relative to the case where no voltage was applied. The effective stiffness was increased by attaching of the piezo patches in the first instance, and a further increase was obtained by actuating the piezos with an applied voltage.

Next, the performance of a patch that covered 25% of the plate area was assessed (configuration #2). Here, a substantial increase in mass was observed (69%). The addition of piezo patches with no voltage applied resulted in an increase of 92% in dynamic pressure. Further application of 400 V across each layer resulted in a smaller further improvement in dynamic pressure to 645×10^3 N/m², or 32% relative to the 0 V case. Thus, it is noted that an increase in the size of the piezo patches and/or actuating power does not necessarily result in better performance. In fact, the 25% patch configuration performed worse than the 6% patch case. Apparently, the larger piezo patch configuration resulted in a relatively much larger mass increase that offset the benefits of increased actuation capability.

Three more patch configurations were analyzed to probe this matter further. Let us call these configurations #3, #4, and #5. In configuration #3, five piezo patches are placed in a star-shaped form, resulting in a mass increase of 86%. The flutter dynamic pressure in the presence of an applied voltage of 400 V exhibits poor performance, a mere 5% increase in value relative to the no applied voltage case. In other words, a larger actuation capability, followed by a much larger mass increase, resulted in negligible improvement. For configuration #4, where four piezoelectric patches were arranged in a cross, the mass increased by 69%. The resultant critical dynamic pressure increased by 8% to 684×10^3 N/m² due to an applied voltage of 400 V. Finally, configuration #5, where the piezo patches were arranged along the perimeter of a square at the center of the plate, exhibited an increase of 20% in dynamic pressure and an increase of 52% in mass. Thus, in configuration #5, better performance was attained in the presence of a relatively smaller increase in mass, whereas in configuration #3, poor performance was exhibited with and mass increased.

Another aspect that draws attention is the fact that the 25% central patch configuration and the number #4 configuration provide a similar mass increase (69%). However, the 25% central patch configuration resulted in a 32% improvement in dynamic pressure, whereas the actuation in configuration #4 exhibited a poor 8% improvement. The performance parameters of interest are the added mass to stiffness ratio and the shape, size, and location of the piezoelectric patches on the adaptive composite structure.

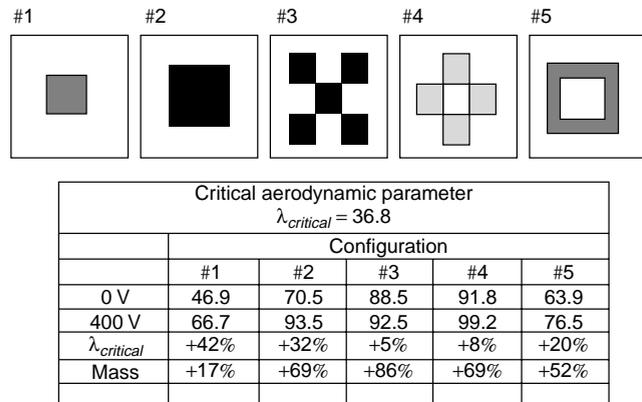


Figure 2. Critical dynamic pressure for five different piezoelectric actuator configurations.

Fuselage Noise Attenuation

During the last few years, there has been an increased understanding of the mechanisms involved in aircraft interior noise and the transmission paths from the sources to the cabin interior. Progress has been made in reducing interior noise levels, especially for propeller driven aircraft, when traditional passive noise control methods are supplemented by recent advances in active noise control. The technical advances have not always translated into large reductions in interior noise levels, but they have enabled existing sound levels to be maintained despite increases in the acoustic power generated by the sources. Because weight is at a high premium, optimization of airplane soundproofing is a design challenge.

Most aircraft use passive methods exclusively to control interior noise. The general concept of cabin soundproofing has remained constant for years, but design changes have occurred, particularly in more efficient treatments. The basic noise control treatment consists of one or more layers of porous material, an intervening sheet of heavy limp material, and a covering impervious trim (69). Stiffening the structure tends to incur a substantial weight penalty. Broadband passive damping is not useful because the response is primarily forced, rather than resonant. Narrowband damping using tuned vibration absorbers has yielded 10-dB reductions (70), but the problem is limited by the difficulty of keeping them tuned in a varying environment.

Active noise suppression methods include control of either the acoustic field or the structural vibration. There have already been flight tests using active control of the internal sound field that have provided significant reductions in noise levels (71). However, the many microphones and speakers required to obtain this performance create a moderate weight penalty and some difficulty in locating bulky speakers in the passenger compartment.

The first documented research on acoustic modeling was reported by Foxwell and Franklin (72) and later, Warburton (73) considered the effects of an acoustic field on the vibrations of a cylindrical shell. The use of finite-elements in acoustics was first suggested by Gladwell (74) who considered some simple one- and two-dimensional problems. Craggs (75) presented a model for studying sound transmission between enclosures that allowed for flanking as well as direct sound transmission.

The concept of controlling structural vibration to reduce transmitted noise is not new. Previous analytical and experimental work established that interior noise in aircraft is directly coupled to fuselage vibration (76–81). This observation implies that if the fuselage vibration can be reduced, then the associated interior noise levels should fall everywhere in the interior space, that is, global attenuation would be achieved. Piezoelectric actuators are well suited for this problem due to their high bandwidth. Jones and Fuller (82) used piezoelectric actuators bonded to an aircraft fuselage and obtained acceptable authority over internal noise using microphone feedback, indicating that the required actuator strength is feasible.

Jayachandran and Sun (83) studied the potential use of shallow, spherical-shell actuators, made of piezoelectric materials supported on a flexible foundation along

the edge. The actuators were modeled analytically and the effects of curvature, mount stiffness, mass, and other parameters on the natural frequencies, linear stroke, and volume velocity were studied. Shields et al. (84) investigated sound radiation from a plate into an acoustic cavity controlled by patches of active piezoelectric-damping composites. Yoon et al. (85) used the fact that anisotropy and the shape of a distributed piezopolymer actuator have coupling effects with vibration modes of structure to control the sound transmission actively through composite structures. Shen and Sun (86), on the other hand, compared the performance of acoustic microphones and strain induced by distributed piezoelectric actuators. A uniform cylindrical shell is used as a simplified model of a section of fuselage.

The referenced numerical investigations in noise attenuation concentrated on either flat plate or simple analytical models to represent adaptive composite structures. Other investigations more specifically implemented for aircraft components do not account for the effects of curvature of the fuselage to evaluate actuator performance properly (87–89). These models require refined meshes to capture the physics of the curved fuselage panel accurately. The model proposed in this investigation consists of a shell-type adaptive composite finite element that has electromechanical properties. The acoustic cavity model is based on a solid brick element.

The results presented in this section are based on finite-element computer simulations aimed at predicting the effectiveness of using piezoelectric actuators as passive control devices to attenuate the transmission of sound through an aircraft fuselage. A relatively simple model comprised of a finite, adaptive composite cylinder that has bonded piezoelectric patches was chosen to represent the aircraft fuselage. The properties of the cylinder material and its dimensions were selected to represent a typical aircraft fuselage and these values were obtained from Thomas et al. (90). The cylinder is 16 m long, and the propeller plane is at a distance of 3.5 meters along the axis of the cylinder. The diameter is 1.3 meters. The objective is to provide a quantitative measure of the actuating capabilities of the piezoelectric patches in internal noise attenuation applications (Fig. 3a).

Simulations were carried out at a frequency close to the blade passage frequency at 90 Hz. Forty acoustic and structural modes were used to model the dynamics of the coupled acoustic-structural system. Numerical results are obtained for the following geometric and material properties for the piezoactuators: $E_p = 63$ GPa, $\rho_p = 7650$ kg/m³, $\nu_p = 0.3$, $d_{31} = d_{32} = -180 \times 10^{-12}$ m/V, $L_p^z = 0.6$ m, $L_p^\theta = 0.6$ m, and $h_p = 1$ mm. It is assumed that the actuators are perfectly bonded to the aluminum fuselage shell structure. A constant voltage of 300 V was applied to each actuator patch.

The finite-element mesh for the structural-acoustic coupled system is shown in Fig. 3b. Two actuator configurations were selected to study the noise suppressing capabilities of the actuator geometry (Fig. 3c). It was observed that the response in the low-frequency range is very low because no modes are excited in this frequency range. Figure 3d presents the noise reduction attained

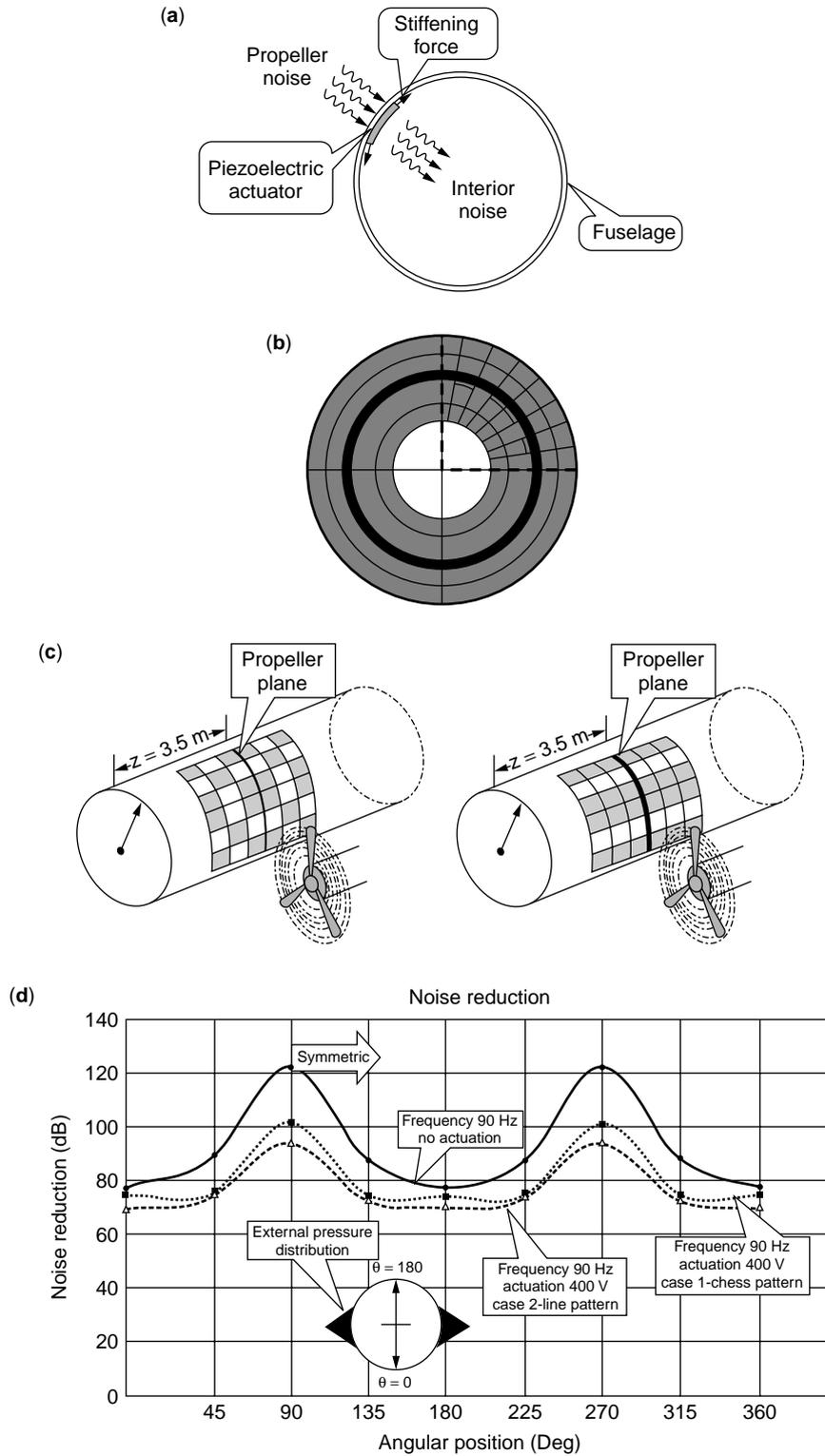


Figure 3. (a) Passive actuation mechanism for interior noise attenuation; (b) finite-element mesh for the aircraft fuselage and cavity with 16 elements along the fuselage length; (c) piezoelectric actuator configuration in chess and line layouts; and (d) noise reduction in the interior of the fuselage for both actuation configurations.

in the interior of the fuselage as determined from the numerical simulations at 90 Hz. It can be observed that it is possible to attain noise reduction levels of the order of 10–15 dB for the frequency simulated in this study. The results presented are compared to those with no actuation in the presence of the piezoelectric actuators. The effect of adding the actuators in the dynamics of the coupled system was taken into account. Noise reduction levels of the order of 17% were attained for the chess-type actuator pattern, and for the line-type configuration, the noise reduction levels obtained were of the order of 22%. This demonstrates that the actuator layout is an important design parameter in applying piezoelectric devices for noise suppression, as was observed in the panel flutter suppression problem.

The results presented here are simply intended to demonstrate the feasibility of applying piezoelectric actuators passively to suppress noise in the interior of closed acoustic cavities such as the interior of an aircraft fuselage. The design concept is based on the effect of stiffening the structure by actuating of the piezoactuators.

CONCLUDING REMARKS

An overview of the state of the art in adaptive composite systems design and applications has been presented. This area of research has been commonly known as the field of smart structures. Two key technological developments

have combined to establish the potential feasibility of smart/adaptive structures. The first is the development of multifunctional materials (materials science) and their use in devices such as distributed actuators and sensors (mechanical engineering). The second development in the field of computer science and electrical engineering, is the advent of ever faster and compact computers, new control algorithms, and signal processing technologies.

The adaptive structures design concepts presented here are feasible engineering solutions to problems that require active vibration suppression and shape control in aircraft structures. However, for the adaptive composite systems to see widespread commercial applications, lighter multifunctional materials, new manufacturing techniques, and cost reduction will have to be integrated into the design process. There is also a need to link the engineered solutions to the market needs. Materials supply is not a viable commercialization position because synthesis and assembly will ultimately address the design, shaping, and functional issues of adaptive materials and structures. Only vertical integration of the manufacturing process will ensure that materials science research will be directly connected to design, implementation, and production of adaptive composite systems.

The multidisciplinary nature of the field of adaptive composite systems requires a complex and sophisticated effort for the transition from research to full implementation and commercialization. Figure 4 depicts technologies and other aspects of research that must be addressed to enable adaptation of structural composite structures and systems.

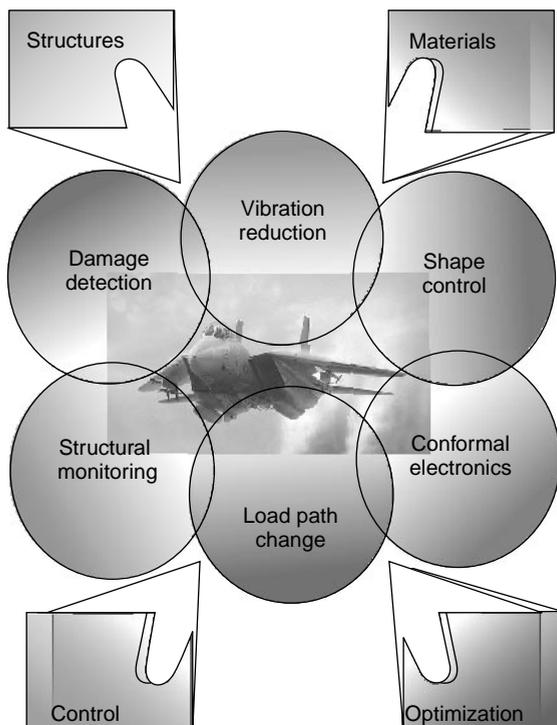


Figure 4. Technologies that need to be addressed to enable adaptation in aircraft structures.

BIBLIOGRAPHY

1. J.G. Boyd and D.C. Lagoudas, Thermomechanical response of shape memory composites. *J. Intell. Mater., Syst. Struct.* **5**, 333–346 (1993).
2. B.J. Sullivan, Analysis of properties of fiber composites with shape memory alloy constituents. *Proc. 2nd Intell. Conf. Intell. Mater.*, pp. 1194–1209 (1994).
3. C. Liang and C.A. Rogers, One-dimensional thermomechanical constitutive relations for shape memory materials. *J. Intell. Mater., Syst. Struct.* **1**, 207–234 (1990).
4. Z.C. Feng and D.Z. Li, Dynamics of a mechanical system with a shape memory alloy bar. *J. Intell. Mater., Syst. Struct.* **7**, 399–410 (1996).
5. E.J. Graesser and F.A. Cozzarelli, A proposed three-dimensional constitutive model for shape memory alloys. *J. Intell. Mater., Syst. Struct.* **5**, 78–89 (1994).
6. A. Baz, K. Iman, and J. McCoy, Active vibration control of flexible beams using shape memory alloys. *J. Sound Vibr.* **140**, 437–456 (1990).
7. R. Ikegami, D.G. Wilson, J.R. Anderson, and G. Julien, Active vibration control using nitinol and piezoelectric ceramics. *J. Intell. Mater., Syst. Struct.* **1**, 189–206 (1990).
8. B.J. Maclean, G.J. Patterson, and M.S. Misra, Modeling of a shape memory integrated actuator for vibration control of large space structures. *J. Intell. Mater., Syst. Struct.* **2**, 72–94 (1990).

9. A.E. Clark, Giant magnetostriction materials from cryogenic temperatures at 250°C. *Proc. SPIE* **1543**, 374–381 (1992).
10. T. Honda, K.I. Arai, and M. Yamaguchi, Fabrication of actuators using magnetostrictive thin films. *Proc. IEEE, Micro Electro-Mech. Syst.*, Oiso Jpn., 1994, pp. 51–56.
11. J. Bi and M.A. Anjanappa, Active vibration damping using magnetostrictive miniactuators. *Proc. SPIE* **2190**, 171–181 (1994).
12. A. Suleman and V.B. Venkayya, A simple finite element formulation for a laminated composite plate with piezoelectric layers. *J. Intell. Mater., Syst. Struct.* **6**, 776–782 (1995).
13. T. Bailey and J.E. Hubbard, Distributed piezoelectric polymer active vibration control of cantilever beams. *AIAA J.* **25**, 606–610 (1985).
14. E.F. Crawley and J. de Luis, Use of piezoelectric actuators as elements of intelligent structures. *AIAA J.* **25**, 1373–1385 (1987).
15. M. Leibowitz and J.R. Vinson, *Intelligent Composites*, CCM Report 91–54. Center for Composite Materials, College of Engineering, University of Delaware, Newark, 1991.
16. B.T. Wang and C.A. Rogers, Laminated plate theory for spatially distributed induced strain actuators. *J. Compos. Mater.* **25**, 433–452 (1991).
17. C.L. Hom and N. Shankar, A fully coupled constitutive model for electrostrictive ceramic materials. *J. Intell. Mater., Syst. Struct.* **5**, 795–801 (1994).
18. E. Prasad, *The Development of Piezoelectric and Electrostrictive Sensors and Actuators for Incorporation into Smart Structures*, CSA Stear-9 Report, Sensor Technology Ltd., 1997.
19. A. Benjeddou, Advances in finite element modeling of smart structural elements. *NATO-ASI Adv. Compos. Struct.*, Portugal, 1998, pp. 171–192.
20. H.S. Tzou and C.I. Tseng, Distributed piezoelectric sensor/actuator design for dynamic measurement/control of distributed parameter systems: A piezoelectric finite element approach. *J. Sound Vibra.* **138**, 17–34 (1990).
21. S.K. Ha, C. Keilers, and F. Chang, Finite element analysis of composite structures containing distributed piezoceramic sensors and actuators. *AIAA J.* **30**, 772–780 (1992).
22. C.K. Lee and F.C. Moon, Laminated piezopolymer plates for torsion and bending sensors and actuators. *J. Acoust. Soc. Am.* **85**(6), 2432–2439 (1989).
23. E.F. Crawley and K.B. Lazarus, Induced strain actuation of isotropic and anisotropic plates. *AIAA J.* **29**(6), 944–951 (1991).
24. B.T. Wang and C.A. Rogers, Laminated plate theory for spatially distributed induced strain actuators. *J. Compos. Mater.* **25**, 433–452 (1991).
25. K.Y. Lam, X.Q. Peng, and J.N. Reddy, A finite element model for piezoelectric composite laminates. *Smart Mater. Struct.* **6**, 583–591 (1997).
26. K. Chandrashekhara and A.N. Agarwal, Active control of laminated composite plates using piezoelectric devices: A finite element approach. *J. Intell. Mater., Syst. Struct.* **4**, 496–508 (1993).
27. H.S. Tzou and R. Ye, Analysis of piezoelectric structures with laminated piezoelectric triangle shell elements. *AIAA J.* **34**, 110–115 (1996).
28. C.C. Lin, C. Hsu, and H. Huang, Finite element analysis on deflection control of plates with piezoelectric actuators. *Compos. Struct.* **35**, 423–433 (1996).
29. A. Chattopadhyay and C.E. Seeley, A higher order theory for modeling composite laminates with induced strain actuators. *Composites* **28B**, 243–252 (1997).
30. J.N. Reddy and J.A. Mitchell, Refined nonlinear theories of laminated composite structures with piezoelectric laminae. *J. Indian Acad. Sci.* **20**, 721–747 (1995).
31. K.D. Jonnalagadda, T.R. Tauchert, and G.E. Blandford, Higher-order displacement formulation for a piezo-thermoelastic laminate. *Mech. Electromagn. Mater. Struct.*, ASME, AMD-161/MD-42, pp. 145–156 (1993).
32. R. Lammering, The application of a finite shell element for composites containing piezoelectric polymers in vibration control. *Comput. Struct.* **41**, 1101–1109 (1991).
33. S.R. Thirupathi and N.G. Naganathan, A composite shell finite element for the analysis of smart structures. *Proc. SPIE Smart Struct. Mater.* **1916**, 424–438 (1993).
34. V.V. Varadan, Y.H. Lim, and V.K. Varadan, Closed loop finite element modeling of active/passive damping in structural vibration control. *Smart Mater. Struct.* **5**, 685–694 (1996).
35. H.S. Tzou and R. Ye, Pyroelectric and thermal strain effects of piezoelectric devices. *Adapt. Struct. Comp. Mater.*, ASME AD-45/MD-54, pp. 125–132 (1994).
36. C.L. Hom and N. Shankar, A finite element method for electrostrictive ceramics. *Int. J. Solids Struct.* **33**, 1757–1779 (1996).
37. L. Hom and N. Shankar, A numerical analysis of relaxor ferroelectric multi-layered actuators and 2-2 composite arrays. *Smart Mater. Struct.* **4**, 305–317 (1995).
38. K.S. Kannan and A. Dasgupta, Finite element scheme for modeling the magnetoelastic response of magnetostrictive smart structures. *Proc. 2nd Int. Conf. Intell. Mater., Technomic, 1994*, pp. 719–730.
39. K.S. Kannan and A. Dasgupta, Finite element modeling of multifunctional composites with embedded magnetostrictive devices. *Adapt. Struct. Compos. Mater. Appl.*, ASME AD-45/MD-54, pp. 21–28 (1997).
40. K.S. Kannan and A. Dasgupta, A nonlinear Galerkin finite element theory for modeling magnetostrictive smart structures. *Smart Mater. Struct.* **6**, 341–350 (1997).
41. L.R. Hill and T.N. Farris, Three dimensional piezoelectric boundary elements. *Proc. SPIE Smart Struct. Mater.* **3039**, 406–417 (1997).
42. R.K. Rajapakse, Boundary element method for piezoelectric solids. *Proc. SPIE Smart Struct. Mater.* **3039**, 418–427 (1997).
43. K. Chandrashekhara and R. Tenetti, Thermally induced vibration suppression of laminated plates with piezoelectric sensors and actuators. *Smart Mater. Struct.* **4**, 281–290 (1995).
44. S.W. Zhou and C.A. Rogers, Heat generation, temperature, and thermal stress of structurally integrated piezo-actuators. *J. Intell. Mater., Syst. Struct.* **6**(3), 372–399 (1995).
45. A. Baz, Active noise control of piston-cavity systems. *Adapt. Struct. Mater. Syst.*, ASME AD-45, pp. 121–131 (1997).
46. J. Kim, V.V. Varadan, and V.K. Varadan, Finite element optimization methods for the active control of radiated sound from a plate structure. *Smart Mater. Struct.* **4**, 318–326 (1995).
47. J. Kim, V.V. Varadan, and B. Ko, Finite element modeling of active cabin noise control problems. *Proc. SPIE, Smart, Struct. Mater.*, **3039**, 305–314 (1997).
48. A.S. Islam and K.C. Craig, Damage detection in composite structures using piezoelectric materials. *Smart Mater. Struct.*, **3**, 318–328 (1994).

49. L.D. Jones, D.V. Newton, and E. Garcia, Adaptive devices for precision control. *Proc. SPIE, Smart Struct. Mater.* **1917**, 648–659 (1993).
50. F. Pourki and H. Sosa, Kinematics and computation of workspace for adaptive geometry trusses. *Proc. SPIE, Smart Struct. Mater.* **1917**, 11052–1062 (1993).
51. J.Y. Shen and L. Sharpe, Jr., A finite element model of the aeroelastic analysis of hypersonic panels. Part III. Flutter suppression. *Proc. SPIE, Smart Struct. Mater.* **3039**, 315–323 (1997).
52. D. Jung, A. Hegeman, N.R. Sottos, S.R. White, and P.H. Gembelle, Self-healing composites using embedded microspheres. *ASME Int. Mech. Eng. Congr. Expo.*, Dallas, TX, 1997.
53. S.R. White, Magnetostrictive tagging of composites for health monitoring. *Proc. Int. Compos. Expo.*, Cincinnati, OH, 1999, pp. 22-E1-E6.
54. S.R. White, Health monitoring and repair of smart composite materials. *Proc. 1st Korea-U.S. Workshop Compos. Mater.*, Seoul National University, Seoul, Korea, 1998, pp. 1–11.
55. S.R. White and J.B. Berman, Thermomechanical response of SMA composite beams with embedded nitinol wires in an epoxy matrix. *J. Intell. Mater. Syst. Struct.* **9**(5), 391–400 (1998).
56. J. Heeg., Flutter suppression via piezoelectric actuation. *NASA Tech. Memo. NASA TM-104120* (1991).
57. J. Heeg, J.M. Miller, and R.V. Doggett, Attenuation of empennage buffet response through active control of damping using piezoelectric material. *NASA Tech. Memo. NASA TM-107736* (1993).
58. E.H. Dowell, Panel flutter: A review of the aeroelastic stability of plates and shells. *AIAA J.* **8**, 385–399 (1970).
59. M.D. Olson, Some flutter solutions using finite elements. *AIAA J.* **8**, 747–752 (1970).
60. M.N. Bismarck-Nasr, Finite element analysis of aeroelasticity of plates and shells. *Appl. Mech. Rev.* **45**(12), 461–482 (1992).
61. I. Lee and M.H. Cho, Supersonic flutter analysis of clamped symmetric composite panels using shear deformable finite elements. *AIAA J.* **29**(5), 782–783 (1991).
62. D.G. Liaw, Supersonic flutter of laminated thin plates with thermal effects. *J. Aircr.* **30**, 105–111 (1993).
63. D.M. Lee and I. Lee, Vibration analysis of anisotropic plates with eccentric stiffener. *Comput. Struct.* **57**, 99–105 (1995).
64. C.E. Gray and C. Mei, Large amplitude finite element flutter analysis of composite panels in hypersonic flow. *AIAA J.* **31**, 1090–1099 (1993).
65. R.C. Zhou, Z. Lai, D.Y. Xue, J.K. Huang, and C. Mei, Suppression of nonlinear panel flutter at elevated temperature with piezoelectric actuators. *AIAA J.* **33**, 1098–1105 (1995).
66. R.C. Scott and T.A. Weisshaar, Controlling panel flutter using adaptive Materials. *J. Aircr.* **31**, 213–222 (1994).
67. D.A. Paige, R.C. Scott, and T.A. Weisshaar, Active control of composite panel flutter using piezoelectric materials. *Proc. SPIE Smart Struct. Mater.* **1917**, 84–97 (1993).
68. D.Y. Xue and C. Mei, A study of the application of shape memory alloys in panel flutter control. *Proc. 5th Int. Conf. Recent Adv. Struct. Dyn.*, 1994, pp. 412–422.
69. J.F. Wilby, Aircraft interior noise. *J. Sound Vibr.*, **190**, 545–564 (1996).
70. D.G. McMartin, Collocated structural control for reduction of aircraft cabin noise. *J. Sound Vibr.*, **190**, 105–119 (1996).
71. S.J. Elliott, P.A. Nelson, I.M. Stothers, and C.C. Boucher, In-flight experiments on the active control of propeller-induced cabin noise. *J. Sound Vibr.*, **140**, 219–238 (1990).
72. J.H. Foxwell and R.E. Franklin, The vibrations of a thin walled stiffened cylinder in an acoustic field. *Aeronaut. Q.*, **1**, 47–48 (1959).
73. G.B. Warburton, Vibration of a cylindrical shell in an acoustic medium. *J. Mech. Eng. Sci.* **3**, 69–75 (1961).
74. G.M.L. Gladwell, A variational formulation of damped acousto-structural vibration problems. *J. Sound Vibr.*, **4**, 172–186 (1966).
75. A. Craggs, The transient response of a coupled plate-acoustic system using plate and acoustic elements. *J. Sound Vibr.*, **15**, 509–528 (1971).
76. C.R. Fuller, S.D. Snyder, C.H. Hansen, and R.J. Silcox, Active control of interior in model aircraft fuselages using piezoceramic actuators. *J. Aircr.* **28**, 208–215 (1991).
77. J.V. Warner and R.J. Bernard, Digital control of local sound fields in an aircraft passenger compartment. *AIAA J.* **28**, 284–289 (1990).
78. A. Grewal, F. Nitzsche, D.G. Zimcik, and B. Leigh, Active control of aircraft cabin noise using collocated structural actuators and sensors. *AIAA Pap.* **96-1274**, (1996).
79. R.J. Silcox, C.R. Fuller, and L.H. Lester, Propeller modeling effects on interior noise in cylindrical cavities with application to active control. *AIAA-Pap.* **89-1123**, (1989).
80. H.C. Lester and S. Lefebvre, Piezoelectric actuator models for active sound and vibration control of cylinders. *Recent Adv. Act. Control Sound Vibr.*, VA Tech., 1991, pp. 3–26.
81. D. Snyder and C.H. Hansen, The effect of modal coupling characteristics on one mechanism of active noise control. *Recent Adv. Act. Control Sound Vibr.*, VA Tech., 1991, pp. 708–727.
82. J.D. Jones and C.R. Fuller, Active control of structurally coupled sound fields in elastic cylinders. *Int. J. Anal. Exp. Modal Anal.*, **5**, 123–140 (1989).
83. V. Jayachandran and J.Q. Sun, Modeling shallow-spherical-shell piezoceramic actuators as acoustic boundary control elements. *Smart Mater. Struct.* **7**, 72–84 (1998).
84. W. Shields, J. Ro, and A. Baz, Control of sound radiation from a plate into an acoustic cavity using active piezoelectric-damping composites. *Smart Mater. Struct.* **7**, 1–11 (1999).
85. K.W. Yoon, S.J. Kim, and W.D. Kim, Application of distributed PVDF actuators to the control of sound transmission through composite plate. *Proc. SPIE Soc. Opt. Eng.* **3041**, 846–855 (1997).
86. W. Shen and J.Q. Sun., Study of shell interior noise control. *Proc. SPIE Int. Soc. Opt. Eng.* **3041**, 812–818 (1997).
87. R.L. Clark and C.R. Fuller, Optimal placement of piezoelectric actuators and polyvinylidene fluoride error sensors in active structural acoustic control approaches. *J. Acoust. Soc. Am.* **92**, 1521–1533 (1992).
88. R.L. Clark and C.R. Fuller, Control of sound radiation with adaptive structures. *J. Intell. Mater., Syst. Struct.* **2**, 431–452 (1991).
89. E. Szechenyi, Approximate methods for the determination of natural frequencies of stiffened and curved plates. *J. Sound Vibr.* **14**, 401–418 (1971).
90. D.R. Thomas, P.A. Nelson, and S.J. Elliott, Active control of the transmission of sound through a thin cylindrical shell. Part I. The minimization of vibrational energy. *J. Sound Vibr.* **167**, 91–111 (1993).

ADAPTIVE SYSTEMS, ROTARY WING APPLICATIONS

ADITI CHATTOPADHYAY
 JOHN RAJADAS
 Arizona State University
 Tempe, AZ

INTRODUCTION

Research and development of smart materials and smart structural systems has been actively pursued during the last decade. Smart materials are defined as materials that can adapt to external stimuli, such as load or environment. The control capabilities of the material can be enhanced by material composition, processing, defect or microstructure, and conditioning to adapt to various levels of stimuli. When these special types of materials are integrated into the structure, both the sensing and the actuating mechanisms become integral parts of the structure through induced strain actuation. The technological applications of this class of materials span a wide range. Today, a large variety of materials and techniques exist (piezoceramics, magnetostrictives, electrostrictives, electrorheological fluids, shape-memory alloys and fiber optics) that serve as sensors and actuators in smart structures. A material that is subjected to a mechanical load and accumulates an electric charge is said to have piezoelectric properties. Conversely, when the piezoelectric material is subjected to an applied electrical field, it induces strain. Thus, piezoelectric materials can be used as active vibrational dampers for structures where they act as both sensors and actuators. When placed in discrete locations on a structure, these materials can sense movement and can control that motion via localized strains. The strain is either compressive or tensile depending on the polarity of the applied electric field. Although these materials generate very low strain, they cover a wide range of actuating frequency and therefore have more practical applications. Although piezoelectric materials are relatively linear and bipolar, they exhibit hysteresis. Magnetostriction is the change in any dimension of a magnetic material caused by applying an external field. These materials are monopolar and nonlinear and exhibit hysteresis. At room temperature and higher, rare-earth transition-metal compounds such as Terfenol-D offer considerable promise as materials for magnetomechanical actuators. These materials generate low strains and moderate forces across a broad frequency range. However, these actuators are often bulky. Electrostrictive materials are similar in characteristics to piezoelectrics and have slightly better strain capability but high sensitivity to temperature. These materials are also monopolar and nonlinear but exhibit negligible hysteresis. Electrorheological (ER) fluids consist of a suspension of fine semiconducting particles in a dielectric liquid that solidifies in a sufficiently large electric field. More precisely, the particles align themselves and thereby form columns that connect the two electrodes. The viscosity of the fluid clearly undergoes dramatic changes as this process takes place, and this effect has been used in a variety

of studies to suppress vibrations by increasing damping. The very short response time of these fluids represents a very important feature from the standpoint of vibrational control. Fiber optics are increasingly becoming popular because like piezoceramics, they can be easily integrated into composite structures. Shape-memory alloys (SMA) operate via a martensitic phase transformation and its subsequent reversal. The shape-memory effect is driven by cooling and heating cycles. It is characterized by recoverable strains as high as 10% and can generate a significant restoring force. The most common commercially available SMA is Nitinol™.

An intelligent or smart structure is generally considered an integrated system of structure, sensors, actuators, and a control module. Currently, there is a great deal of interest in applying smart structure technology to various physical systems. Active vibration, noise, aeroelastic stability, damping, shape control, health monitoring, process monitoring, and damage and delamination detection are a few examples. A significant amount of the early development was devoted to application to space-related systems. Recently, studies have also been conducted to investigate using this concept to improve fixed wing performance. During the last decade, significant research efforts have also been directed toward applying these concepts to rotary wing aircraft which is the topic of this article. A review of the applications of smart materials and structures in aeronautical systems is presented in (1). The potential for improving performance, handling qualities in stall, and increasing fatigue life is discussed. Actuators that use smart materials require relatively slow-acting shape and shape-distribution changes in this respect. A similar review is made of applications for improving aeroelastic divergence, flutter instabilities, and tail buffeting on fixed-wing aircraft and reducing vibrations, improving external acoustics, and providing flight controls for rotary wing aircraft—all of these require a high-frequency response.

Vibration and noise are two long-standing problems that have limited the expansion of military and commercial applications of rotorcraft. Helicopters are particularly susceptible to high vibratory loads compared to fixed wing aircraft due to the high unsteady load and complex aerodynamic environment in which the rotor must operate and the existence of the complex structural, aerodynamic, and dynamic coupling. The structural flexibilities in the rotor blade, lead to vibration and control problems during operation (especially at high operating speeds) even though lightweight and stiff advanced materials such as composites and ceramics may be regularly used to build them. Alleviating rotor noise and vibration presents a major challenge to the aviation engineering community and is essential for increased survivability, reliability, and maintainability of rotorcraft. The flexibility of the rotor system in rotary wing aircraft makes it possible to impose changes in shape, mechanical properties and stress-strain fields upon it. The use of smart structures will allow tailoring the imposed changes according to the conditions sensed by the rotor. Most importantly, by using distributed sensors and actuators, control can be achieved across a much broader bandwidth than by using a swash plate control system which is limited to n/rev for an n -bladed rotor.

Recently, a significant amount of research has been conducted to control vibration using both active and passive methods. Passive techniques, such as nonstructural tuning masses and additional stiffening members, can help reduce vibrational amplitudes. However, such techniques are subject to significant weight penalties. Active vibrational suppression techniques offer definite advantages compared to conventional methods because of reduced weight penalties. An early study of different techniques for controlling rotary wing vibration by using smart structures was done (2).

Friedmann and Millott (3) present a concise review of the state of the art for vibrational reduction in rotorcraft by using active controls. The principal approaches presented for vibrational reduction in helicopters are higher harmonic control (HHC) and individual blade control (IBC). In HHC, time-dependent pitch control is introduced into the nonrotating frame through the swash plate. In the IBC approach, each blade is individually controlled in the rotating frame. Although the HHC method has been successful in reducing hub vibration, a similar reduction in airframe vibrational levels is not always possible because conventional HHC methods employ a single frequency of the form $n\Omega$. Other factors, including mechanical complexity and weight penalty, also limit its application. In addition, if the hub moves as a result of a dynamic rotor-airframe coupling, its filtering characteristics for frequencies other than n/rev are severely degraded, this makes the HHC scheme ineffective. In the IBC approach, control laws are synthesized independently and are aimed at minimizing the modal contribution of each blade to the overall rotor vibration. Therefore, vibrations due to manufacturing variability among the blades can be addressed effectively. The procedure is not restricted by the $(n + 1)\Omega$ phenomenon specific to the swash plate and hence can be more effective in addressing airframe vibrations due to rotor-airframe coupling.

A major improvement in helicopter performance can be achieved by implementing active rotor control technology. Smart material actuator technology for operation "on the blade" is now becoming available and promises to overcome the size, weight, and complexity of hydraulic and electric on-rotor actuation which should alleviate the problem. Blade actuation by smart materials offers the chance for an electrical system integrated into the blade itself. A study of different blade actuating systems shows that, in principle, there is a possibility of achieving this goal. A significant amount of research has been conducted by using induced strain actuation for improved IBC. These include using distributed induced strain actuation for active/passive control of structural response for improved vibratory response, an actively controlled flap located on the blade for possible noise and vibration reduction, actuator driven servoflaps, an active twist concept for controlling vibration and noise, and novel active/passive damping treatments for increased modal damping. Unfortunately, the development of high force/high deflection actuators, the most critical issue in all of these applications, has not been addressed adequately and still remains a topic of future research. Finally, in all applications involving smart materials, accurate modeling of the material response and the integrated smart

structural response is important because control authority can be significantly mispredicted if inaccurate theories are used. In the following sections, the current state of the art in improving rotor vibratory response and aeromechanical characteristics is reviewed. Advantages and disadvantages of these techniques are discussed. Numerical results from feasibility studies are summarized, and accomplishments already achieved and the goals to be pursued in the near future are discussed.

ACTIVE/PASSIVE CONTROL OF STRUCTURAL RESPONSE

Both the first elastic flatwise bending (second for hingeless rotors) and the first elastic torsion modes of a single blade require special attention in vibrational control. Theoretical investigations supported by wind tunnel and flight tests confirmed that these modes are responsible for the larger amplitude loads at $3/\text{rev}$ in four-blade hingeless rotors. IBC, based on a collocated actuator-sensor arrangement along the blade and tailored to act specifically on the bending and the torsional modes, is expected to reduce further the overall dynamic response of helicopters. Induced strain actuation using both active and passive control has been investigated by several researchers. Nitzsche and Breitbach (4,5) investigated the feasibility of employing adaptive material to build sensors and actuators for attenuating the higher harmonic loads of the helicopter rotor based on the individual blade control (IBC) concept. The results indicate that the adaptive material has enough power in a number of applications to accomplish the task without saturating the applied electrical field. Individual blade control of bearingless composite helicopter rotors aimed at vibrational attenuation by using adaptive structures is presented by Nitzsche (6). A periodic model-following regulator was developed in the frequency domain to approximate the aeroelastic response of a single blade in forward flight to its hover characteristics. Two control strategies are investigated: a robust one using distributed control and a cost-efficient one using lumped control. Selective linearization of the periodic system at determined frequencies is demonstrated in both cases. Active and passive control of the aeroelastic response of helicopter rotors was achieved by using smart materials to tailor the blade root flexibility (7). An individual blade controller designed to attenuate the aeroelastic response of helicopter rotors in forward flight by tailoring the blade root attachment conditions was developed. A feasibility analysis indicates that the open-loop controller that incorporates both passive and active designs is energy efficient and may be realized by available adaptive structures. A thin-walled cantilever beam is used to model a helicopter blade (8,9). The integrated distributed actuator design methodology was based on the converse piezoelectric effect and aimed at actively control of the vertical and lateral eigen vibrational characteristics of the beam.

A helicopter rotor system that employs an adaptive rotor blade was developed (10). The unique feature of this system is the replacement of the traditional swash-plate-based collective control by adaptive rotor blades that generate lift by using active camber control based on a

shape-memory alloy. A control scheme that allows a smooth variation of camber changes was developed. However, appropriate heating and cooling schemes necessary in active SMA are not discussed.

An adaptive nonlinear neural network control algorithm that can be used for smart structure actuators and sensors to control the shape and suppress the vibrations of flexible beams was developed (11). The algorithm couples an explicitly model-based adaptive component that employs time-varying estimates of the beam's material properties and an adaptive neural network, learning in real time, which is used to estimate the additional actuator moments needed to offset the effects of periodic external disturbances. Experimental results for a thin beam that has bonded piezoceramic sensors and actuators demonstrate the algorithm's ability to track desired bending profiles and reject vibrations caused by external disturbances and to maintain this performance despite changes in the material properties of the structure or in the properties of the external disturbance. The development of one-dimensional pure bending, coupled bending and extension, and combined bending, extension, and torsion models of isotropic beams that use induced-strain actuation is discussed (12). An adhesive layer of finite thickness between the crystal and beam is included to incorporate shear lag effects. Experimental tests evaluated the accuracy and limitations of the models. The bending and coupled bending and extension models show acceptable correlation with static test results, whereas the combined extension, bending, torsional model indicates the need for model refinement.

Chattopadhyay et al. (13) investigated the reduction of vibratory loads at a rotor hub using smart materials and closed-loop control. A comprehensive composite theory was used to model the smart box beam. The theory, which is based on a refined displacement field, is a three-dimensional model that approximates the elasticity solution, so that the beam's cross-sectional properties are not reduced to one-dimensional beam parameters. The finite-state, induced-flow model is used for predicting the dynamic loads. Significant reductions in the amplitudes of dynamic hub loads are shown by using closed-loop control. Detailed parametric studies assessed the influence of number of actuators and their locations in vibratory hub load reduction.

All of this work suggests that distributed induced strain actuation for IBC is a promising concept. However, the lack of appropriate rotating blade actuating systems is a major barrier that needs to be overcome before the impact of these concepts can be truly realized in the area of vibration and noise control. The limited output capability of the materials and the dynamic operating environment must be fully addressed and resolved to take advantage of this technology.

PASSIVE/ACTIVE CONTROL OF DAMPING

Both airframes and rotors are prone to vibratory motions that can become severe when conditions of resonance, limit cycle, chaos, or aeromechanical instability are approached. Current advanced rotor designs tend toward hingeless and bearingless rotors to reduce life-cycle costs, improve hub

designs, and achieve superior handling qualities. Due to stress and weight considerations, advanced rotors are designed as soft in-plane rotors. These rotor systems are susceptible to aeromechanical instabilities such as air and ground resonance due to the interaction of a poorly damped regressing lag mode and the body mode. The introduction of damping into the system may alleviate aeroelastic flutter, reduce gust loading, increase fatigue life expectancy of rotor blades, improve maneuverability and handling, and diminish interior noise levels. In conventional articulated rotors, these instabilities are usually mitigated by using hydraulic dampers. But hingeless and bearingless helicopter rotor designs have elastomeric lag dampers. Elastomeric dampers have received a significant amount of attention for damping in helicopter design due to the variety of advantages they exhibit over conventional dampers. A new technology called Fluidlastic[®] which combines fluids and bonded elastomeric elements to provide unique dynamic performance in solving vibrational control and damping problems in helicopters was introduced (14). The fluids add a viscous component to the energy dissipation mechanism in the dampers. These devices accommodate all relative motion by flexing the elastomer, which eliminates dynamic seals and achieves a long, predictable service life. Incorporation of the fluid facilitates the design of devices that have a much broader range of dynamic characteristics than can be achieved by elastomers alone. Substantial energy dissipation, as required in lead-lag dampers, can be provided. Large inertial forces that have little damping, as required in tuned vibrational absorbers and pylon isolators, can also be achieved. However, elastomeric dampers are sensitive to temperature, exhibit significant loss of damping at extreme temperatures, and it has been known that they cause limit cycle oscillations in rotor blades. They are expensive and are susceptible to fatigue. These dampers also exhibit complex nonlinear behaviors that introduce additional modeling issues. As a result, a variety of alternatives to auxiliary lag dampers are currently under consideration. The elimination of lag dampers would further simplify the hub and reduce weight, aerodynamic drag, and maintenance costs. However, the design of a damperless, aeromechanically stable configuration is a true challenge. Several concepts have shown some promise, but no generally accepted solution has evolved for eliminating lag dampers (15,16).

Recently, new concepts for enhancing structural damping characteristics were introduced in the study of adaptive structures. Such active damping techniques, based on combinations of viscoelastic, magnetic, and/or piezoelectric materials, magnetorheological (MR) fluids, shunted electric circuits, and active nonlinear control strategies, have emerged as candidates for improving structural performance and reliability. A numerical study of electrorheological (ER) dampers (17) used two models, the Newtonian and the Bingham plastic models, to characterize ER fluid behavior. Damping performance of two damper configurations, the moving electrode and the fixed electrode, were studied and the effects of electrode gap sizes, the field strength, and the ER fluid model were quantified. The study provides a basis for designing ER-fluid based dampers. Magnetorheological fluid dampers are

attractive candidates for augmenting lag mode damping in helicopter rotors, where additional damping is required to avert instabilities during specific flight conditions. MR fluids change properties dramatically as a magnetic field is applied. This active damping component presents an advantage over passive elastomeric and fluid–elastomeric dampers. The feasibility of using MR fluid-based dampers for lag damping augmentation in helicopters was explored by Marathe et al. (18). An analytical study to examine the influence of MR lag dampers on rotor aeroelasticity was carried out by integrating a MR damper model and a rotor aeromechanical model. Two different control schemes are presented, the on-off scheme and the feedback linearization scheme. The two control schemes are compared for lag transient response in ground resonance and their ability to reduce damper load in forward flight. The study examines damper sizing, as related to the magnetic field saturation limits and lag perturbation amplitudes, to produce the desired damping. It is shown that a MR damper whose size is comparable to an elastomeric damper can provide sufficient damping for ground resonance stabilization and can significantly reduce periodic damper loads by judiciously choosing an operating scheme. An extensive comparative study of fluid–elastomeric and MR dampers was done (19). The MR dampers were tested when the magnetic field was turned off (off condition) and when it was turned on (on condition). The dampers were tested individually and in pairs, under different preloads, and under conditions of single and dual frequency excitation. The fluid–elastomeric and MR (off) damper behavior was linear, whereas the MR (on) behavior was nonlinear and the stiffness and damping varied as the displacement amplitude changed.

The concept of active constrained layer (ACL) damping treatment in reducing vibrational and augmenting stability has also been investigated. An ACL system consists of a viscoelastic damping layer sandwiched between an active piezoelectric layer and the host structure. The piezoelectric layers have sensing and control capabilities that actively tune the shear of the viscoelastic layer according to the structural response. Using such active/passive control capabilities, the energy dissipation mechanism of the viscoelastic layer is enhanced, and the damping characteristics of the host structure can be improved. Thus, the ACL configurations capitalize synergistically on both passive and active techniques. The merits of ACL for suppressing vibration of flat plates are discussed in (20–22). The primary concern in ACL configuration is the fact that the viscoelastic layer reduces the actuating ability of the piezoelectric layer. An enhanced active constrained Layer (EACL) concept was introduced by Liao and Wang (23,24) to improve the active action transmissibility of ACL by introducing edge elements. The edge elements are modeled as equivalent springs mounted at the boundaries of the piezoelectric layer. The viscoelastic layer model becomes the current active constrained layer system model as the stiffness of the edge elements approaches zero. Some important characteristics of enhanced active constrained layer damping (EACL) treatments for vibrational controls are addressed in (23). Analysis indicates that the edge elements can significantly improve the active action transmissibility of the current active constrained layer

damping treatment. The authors suggest that although the edge elements slightly reduce viscoelastic material passive damping, the EACL still has significant damping from the viscoelastic material (24). However, debonding of the edge elements, which are made of aluminum and are bonded to the piezoelectric and viscoelastic layers still needs attention before these treatments can be effective in practical applications.

Limited efforts have been reported in using these novel damping treatments to improve helicopter aeromechanical stability. Badre-Alam et al. (25) developed a simplified beam model that uses an enhanced active constrained layer (EACL) damping treatment to represent the flexbeams of soft in-plane bearingless main rotors. The goal was to demonstrate the feasibility of using such damping treatment to improve lag mode damping and aeromechanical stability. A derivative controller based on the flexbeam tip transverse velocity was used in this study. Design optimization was studied to understand the influence of various design parameters, such as viscoelastic layer thickness, PZT actuator thickness, and edge element stiffness, on PZT actuator electrical field levels, induced axial stress levels, and available lag damping. All of this work used a constrained layer treatment that covered the entire beam surface. It is well known that segmentation of the active constraining layer can effectively increase passive damping in low-frequency vibrational modes by increasing the number of high shear regions. The use of segmented constrained layer (SCL) damping for improved rotor aeromechanical stability was recently addressed by Chattopadhyay et al. (26–28). The rotor blade load-carrying member was modeled by using a composite box beam of arbitrary wall thickness. The SCLs were bonded to the upper and lower surfaces of the box beam to provide passive damping. A finite element model based on a hybrid displacement theory was used to capture the transverse shear effects accurately in the composite primary structure and in the viscoelastic and piezoelectric layers within the SCL. Detailed numerical studies assessed the influence of the number of actuators and their locations on aeromechanical stability. Ground and air resonance analysis models were implemented in the rotor blade built around the composite box beam by using segmented SCLs. Results indicate that surface-bonded SCLs significantly increase rotor lead-lag regressive modal damping in the coupled rotor-body system (26). A hybrid optimization procedure was also used (27) to address the complex design problem of properly placing actuators along the blade spar as well as to determine the optimal ply stacking sequence of the host structure for improved modal damping. Active control, using SCLs was addressed in (28). For rotary wing applications, the controller must be designed to deal with the time-variant characteristics of the dynamic model due to rotor rotation. A transformation matrix was introduced to transform the time variant problem to a time invariant problem. A linear quadratic Gaussian (LQG) controller was designed for the transformed system on the basis of the available measurement output. Numerical studies showed that the control system was effective in improving helicopter aeromechanical stability across a wide range of operating speeds (28).

TRAILING EDGE FLAPS

Smart structure activated trailing edge flaps can actively alter the aerodynamic loads on rotor blades. Using a suitable feedback control law, such actuators could be used to counter the vibrations induced by periodic aerodynamic loading on the blades, eliminate the bandwidth constraints, and reduce the weight penalties of previous actuating methods. The flaps are actuated by piezostack, bimorph, or magnetostrictive actuators. It is, however, still unclear if there is a single actuating mechanism that addresses both noise and vibrational reduction and still has enough control authority to act as an extra control effector in its own right. The uncertainties about the actuating mechanism, the precise amount of flap deflection available, and the accuracy of current constitutive models of the actuators lead to significant difficulties in analyzing the potential of the concept for helicopter applications. Today, the available materials are not quite ready for real "smart" applications. The preferred materials, such as piezoceramics, have tensile strength that is too low and very low active strains.

The feasibility of using piezoelectric materials as integrated actuators for a trailing edge helicopter blade flap was first demonstrated by Spangler and Hall (29). A model of the dynamic behavior of this actuator was developed for a fixed airfoil section using an aeroelastic Rayleigh–Ritz procedure. A set of scaling laws for the dynamic similarity between scaled test articles and representative full-scale blades was developed and experimental tests were done to verify actuator feasibility and effectiveness. The results include the amplitude response of the flap angle to the applied electric field as well as the lift and moment on the airfoil due to flap deflection. A feasible blade actuating system for individual blade control is now a hinged flap at the outer third of a rotor blade (30). This flap can be controlled by a smart (piezoelectric) actuator. Initial analysis shows that such a system will work as desired. Vibrational reduction in a four-bladed helicopter rotor using an actively controlled flap located on the blade was studied by Millott and Friedmann (31). This study examined the influence of using a more sophisticated fully flexible blade model on the potential of active flap control for reducing vibrations. A deterministic feedback controller was implemented to reduce the 4/rev hub loads. Trend studies show that the torsional stiffness of the blade and the spanwise location of the active flap are important parameters. A time-domain simulation of the helicopter response to control was carried out to investigate the validity of the quasi-static frequency domain approach to formulating control strategies.

Samak and Chopra (32) developed electromechanical actuators based on the concept of mechanical amplification using piezo and electrostrictive stacks as drivers to achieve high force and high displacement actuation. Two different actuators were developed. The first, using a piezostack, actuates a "Flaperon," which consists of a small movable surface to trip the boundary layer that is located on the top surface of a wing model using a NACA0012 airfoil. The second actuator, using an electrostrictive stack as a driver, was designed to move a leading edge droop flap hinged

at the 25% chord of a wing model using a VR-12 airfoil. Experiments to evaluate the actuators were conducted. The dynamic response of the actuators across a 33 Hz frequency range showed that 7 lb of actuator force was obtainable in both cases; the flaperon actuator produced 15 mils of dynamic displacement at 15 Hz, and the droop flap actuator produced about 6 mils of displacement at 16 Hz. A helicopter trailing edge flap using piezostack actuators was also designed and tested by Spencer and Chopra (33). A flap actuator was constructed from two piezoelectric stacks arranged in series, and mechanical amplification was provided through a hinged arm. A test model using NACA0012 airfoil was constructed, and the flap performance of the model was tested in no-load conditions. An analytical model was formulated to calculate the predicted flap deflection using quasi-steady-state aerodynamics. The model showed that the high force piezoelectric stacks maintained flap deflection across a wide range of free-stream velocities.

A parametric design of a plain trailing edge flap system was studied for a typical helicopter (34). The results show that plain flaps are effective in implementing multi-cyclic control. Significant reductions in fixed system 4/rev hub loads are predicted. The effects of variations in flap length, spanwise location, chord, and aerodynamic effectiveness could be largely offset by compensating adjustments to the flap commands. A comprehensive aeroelastic analysis was done to investigate plain trailing edge flaps for vibrational control (35). Results of the analysis are compared with experimental wind tunnel data and that of another comprehensive analysis. Correlation between predicted and measured frequencies and forward flight trim control predictions are good. Significant discrepancies are observed in 3/rev in-plane bending. Using a trailing edge flap moving $\pm 4^\circ$ at 5/rev, the analysis overpredicts the 5/rev flatwise bending moments due to the flap motion, whereas the blade torsional moment is predicted fairly well. Varying the phase angle of the flap motion had a significant effect on the blade 4/rev flatwise and in-plane bending loads.

A new actuating mechanism for a smart rotor that has an active trailing edge flap uses a piezo-activated composite bending–torsion coupled beam as an actuator (36). Spanwise variation in the beam layup and piezoceramic element phasing maximize the twist response and minimize the bending response. The surface-bonded piezoceramic elements are excited to induce spanwise and chordwise bending of the beam, which results in an induced twist via structural coupling. The composite beam has an outboard integral flap, and the flap deflection corresponds to the tip twist of the beam. Two one-eighth scale model rotor blades that used the actuator beam were hover tested at various rotor speeds and collective settings to evaluate the performance of the flap drive system. The actuator beam was excited at various frequencies and rms voltages, and 4/rev peak-to-peak flap deflections of 3–4° were achieved at an operating speed of 900 rpm. Two active rotor blade concepts were developed using a piezo-induced bending–torsion coupled composite actuator beam (37). The first is an active moving blade, and the second is a rotor blade that has an active linear twist. A spanwise variation in the beam layup and piezoceramic element phasing maximizes the twist

response and minimizes the bending response. Two operational one-eighth Froude scale model rotor blades that have a 90% main blade section and 5 and 10% span active tips were used for hover testing. In the controllable twist configuration, a nonrotating twist amplitude of 0.8° is achieved at $100V_{\text{rms}}$ and 75Hz. In hover, at 875 rpm, this reduces to 0.5° at 5/rev. The actuating power per blade is estimated at 1.2% of the hover power required at 8° collective. In the moving blade tip configuration, a nonrotating deflection amplitude of 1.7° is achieved at $100V_{\text{rms}}$. Nonrotating dynamic tests show resonant amplification for all frequencies up to 5/rev. The deflection at $120V_{\text{rms}}$ increases from a low frequency 2.25° to 3.5° at 4/rev, and for $90V_{\text{rms}}$, the low-frequency deflection amplitude of 1.5° doubles to 3.0° at 5/rev.

Koratkar and Chopra (38) reported on the analysis and testing of a Froude scaled rotor model that uses piezoceramic-bimorph actuated trailing-edge flaps. Piezoceramic benders mounted into the spar of the blade actuate the flap, and a rod-cusp mechanism is used to amplify the actuator tip displacement. An analytical model was developed by coupling an unsteady-state aerodynamic model for the flap hinge moment and a structural finite element beam model for the piezoceramic bender. Two operational one-eighth (Froude) scale rotor models that had piezo-bimorph actuated trailing-edge flaps were tested. Flap deflections of $\pm 6^\circ$ (4/rev excitation) were achieved in hover at the Froude scaled operating speed of 900 rpm. Flap authority of 10% for the rotor thrust was achieved at 4° collective for a 4/rev excitation. The analytical model was used to design a Mach scaled rotor model that had piezo-bimorph actuated trailing-edge flaps. The analysis indicates that two side by side, tapered, eight-layered bimorphs excited by a bias voltage (2:1 amplification) can generate flap deflections of $\pm 6^\circ$ at the Mach scaled operating speed of 2100 rpm. An active rotor blade that can be flown in two different configurations (with a moving blade tip or as a controllable twist blade) was designed and tested (39). The inboard 90% of the blade that uses a bending-torsion coupled actuator located in the spar is common to both blade types. The actuator (both blades) is a piezo-induced bending-torsion coupled composite beam. Two operational one-eighth scale model rotor blades were used for hover testing. In the controllable twist configuration, a nonrotating twist amplitude of 0.8° is achieved at $100V$ and 75Hz. In hover, at 875 rpm, this reduces to 0.5° at 5/rev. The actuating power per blade is estimated at 1.2% of the hover power required at 8° collective. In the moving blade tip configuration, a nonrotating deflection amplitude of 1.7° is achieved at $100V$. Nonrotating dynamic tests show resonant amplification for all frequencies up to 5/rev. The deflection at $120V_{\text{rms}}$ increases from a low frequency 2.25° to 3.5° at 4/rev, and for $90V_{\text{rms}}$, the low-frequency deflection amplitude of 1.5° doubles to 3.0° at 5/rev. The rotor blades that had nonactivated moving tips were successfully flown in hover at 900 rpm and collective settings from $4-8^\circ$.

Advances in developing a Froude scaled helicopter rotor model featured a trailing-edge flap driven by piezoceramic bimorph actuators for active vibration suppression (40). Dynamic performance of the actuator and the actuator-flap

assembly were examined. A beam model of the piezo bimorph, including rotor moment effects was formulated to illustrate better the physical mechanisms that affect the system in a rotating environment. An analytic model and validation tests of Froude and Mach scaled rotors featured piezoelectric bender actuated trailing-edge flaps for active vibrational suppression (41). A finite element structural formulation in conjunction with time domain unsteady-state aerodynamics was used to develop the rotor blade and bender-flap coupled response in hover. A two-bladed Froude scaled rotor that used piezobender actuation was tested to investigate the feasibility of piezobender actuation and validate the analytic model. Flap deflections of ± 4 to $\pm 8^\circ$ for 1 to 5/rev bender excitation were achieved at the Froude scaled operating speed of 900 rpm. The trailing-edge flap activation resulted in a 10% variation in the rotor thrust levels at 6° collective pitch. The analytic model shows good correlation with experimental flap deflections and oscillatory hub loads for different rotor speeds and collective settings. Two Mach scaled rotor blades using piezo-bender actuation were also designed and tested. An eight-layered, tapered bender was used, and the bender performance was improved by selectively applying large electric fields in the direction of polarization to individual piezoceramic elements.

Lee and Chopra (42) developed an actuator for a trailing-edge flap on a full-scale helicopter rotor blade using a high performance piezoelectric stack device. The actuator was designed and constructed using two identical piezostacks that were selected from commercially available actuators. An analytical model was formulated by using quasi-steady-state aerodynamics to calculate the aerodynamic requirement of the flap actuator. A new amplification device that used a double-lever mechanism was also designed. An amplification factor of 19.4 was obtained experimentally under nonrotating conditions. The flap actuator can achieve the required flap deflections for active vibrational control. A systematic approach was used to design an active trailing-edge flap actuator for helicopter vibrational suppression (43). The prototype actuator was developed by using two piezostacks that had a double-lever amplification mechanism. An extended piezostack testing methodology, including both high preload and dc/ac excitation, was used to evaluate the performance of the piezostacks. The measured actuation stroke at 600 g of centrifugal loading included more than 90% of nonrotating actuating capability. Koratkar and Chopra (44,45) reported on the development of a Mach scaled rotor model that uses piezoelectric bender actuated trailing-edge flaps for individual blade control of helicopter vibration. An analytic model was developed for the coupled actuator-flap-rotor dynamic response in hover and was validated by using experimental data obtained from a Froude scaled rotor that had piezo bender actuated trailing-edge flaps. Future work in this area will involve designing, fabricating, and testing two Mach scaled rotor blades using piezo bender actuation.

A full scale demonstration system to provide active control of noise and vibrations as well as inflight blade tracking for the MD-900 helicopter was conceptually sized and designed (46). Active control is achieved via a trailing-edge flap and trim tab, both driven by on-blade smart material

actuators. Fulton and Ormiston (47) studied a trailing-edge control surface (elevons) used to generate local aerodynamic lift and pitching moments. A two-blade, hingeless rotor that used 10% chord on-blade elevons driven by piezoceramic bimorph actuators was tested in forward flight for advance ratios from 0.1 to 0.3 at low to moderate thrust coefficients. The fundamental dynamic response characteristics and the effectiveness of elevon control in reducing blade vibratory loads were studied. Deflections of $+1-5^\circ$ at a nominal rotor speed of 760 rpm were achieved, and nonlinear actuator response characteristics were evaluated. Aeroelastic and structural dynamic responses were evaluated across a wide range of rotor speeds using sine sweep excitation of the elevon up to 100 Hz. Preliminary experimental results include actuator effectiveness, the effects of low Reynolds number on elevon pitch moments, elevon control reversal, and the variation of flap bending mode responses to rotor speed and elevon excitation.

A new, robust individual blade control (IBC) control methodology for vibrational suppression using a piezo-actuated trailing-edge flap was explored (48). The controller uses a neural network, learning in real time, to cancel the effects of periodic aerodynamic loads on the blades adaptively, greatly attenuating the resulting vibrations. Complete proof of the stability and convergence of the proposed neurocontrol strategy is provided, and numerical simulation results for a one-eighth Froude scale blade model demonstrate that the controller can virtually eliminate blade vibration from a wide variety of unknown, periodic, disturbance sources. An innovative approach to vibration reduction in rotorcraft that consists of modeling the smart actuating mechanism by using a simple low-order linear model that matches test data (with an associated variation or uncertainty) was introduced by Sahasrabudhe et al. (49). The model is used along with a helicopter flight dynamic model to optimize flap sizing and placement for minimum fixed frame vibration. The effectiveness of the flap in reducing interaxis coupling and as a redundant control for primary actuator failure was analyzed. Straub and Charles (50) studied a bearingless rotor that had trailing-edge flaps for active control of rotor aerodynamics and dynamics. The flaps are controlled by piezoelectric actuators installed in the blade. Two aeroelastic codes are used to explore the aerodynamics and dynamics of this rotor. Results show that a simple but efficient code may be used to predict rotor response and conduct concept development and active control law development. An advanced code must be used to obtain more accurate loads and to study the coupled blade/flap dynamics and stability.

A numerical study to evaluate the aeroelastic steady-state response of a hingeless rotor blade that used trailing-edge flap controls was conducted by Yillikci (51). A new aerodynamic environment due to flap control is formulated on the basis of Theodorsen's unsteady-state oscillating airfoil aerodynamics representation, including unsteady-state trailing-edge flap motions. Aeroelastic responses of elastic rotor blades that have flap and pitch controls were compared, and the responses of different blade configurations that had variable chord and flap geometries were analyzed. Control flap actuation using the magnetostrictive material Terfenol-D was studied by designing a minimum

weight actuator, subject to a set of actuation and stress constraints (52). The resulting device can reduce vibration by more than 90% under cruise conditions.

C-blocks are midrange piezoelectric actuators that show promise for use in dynamic applications, such as noise and vibrational control. They can be combined in series or parallel to generate tailorable performance and exploit the advantages of bender and stack architectures. An analytical model of an individual C-block actuator was developed, and the possible configurations for achieving improved force and deflection characteristics were investigated (53). The use of C-block actuators in trailing-edge flap control was addressed (54). It was found that several C-blocks, attached in series and in parallel configurations were necessary to achieve the required amount of lift authority. Although important issues such as power requirements were not investigated, this study shows that it is possible to attach the C-block actuators directly to the trailing-edge flap and thereby avoid the need for excessive hinge moments caused by other types of actuating mechanisms. A more recent effort in using C-block actuators is discussed in (55). Planar structural actuation that uses anisotropic, active materials is addressed by Bent et al. (56). The mechanisms for creating anisotropic actuators and the impact of anisotropy at the individual lamina level and at the laminated structural level are discussed. Models for laminated structures were developed using an augmented classical laminated theory and incorporating induced stress terms to accommodate anisotropic actuator materials. Four anisotropic actuators that have different material anisotropies were compared. A laminate incorporating piezoelectric fiber composite actuators was manufactured and tested.

SERVOFLAP

Although a vast amount of research based on the IBC concept has been reported, direct use of piezoelectric devices is still infeasible due to excessive power requirements and lack of sufficient induced strain for adequate control. For this reason, research efforts have been directed toward servoflap designs to suppress aeroelastic instabilities. Though greater efficiency has been reported, servoflap concepts increase the complexity of rotor designs and involve expensive aeroelastic computations. Giurgiutiu et al. (57) investigated the engineering feasibility of induced strain actuators for active vibrational control. Rotor blade vibrational reduction based on higher harmonic control-individual blade control (HHC-IBC) principles is a possible area for applying the induced strain actuator (ISA). Recent theoretical and experimental work on achieving HHC-IBC through conventional and ISA means has been reviewed. A benchmark specification for a tentative HHC-IBC device based on the aerodynamic servoflap principle operated through an ISA was developed. Preliminary studies based on force, stroke, energy, and output power requirements show that available ISA stacks coupled with an optimally designed displacement amplifier might meet the benchmark specifications. Straub (58) investigated the feasibility of using smart materials for primary and

active control on the AH-64 helicopter. The results suggest that embedded actuator concepts, that is, pitch, twist, and camber control, were not practical. Servoflap control using hinged control surfaces driven by discrete actuators emerged as the most suitable candidate for smart material actuation. Preliminary data show that rotor control using smart materials might be feasible if a combination of smart materials is used and the rotor design is driven toward low control loads and motions.

The basic concepts for analyzing, designing, and experimenting on induced-strain actuators for rotor blade aeroelastic vibrational control are presented in (59). A trailing-edge servoflap actuated by a hydraulically amplified, large-displacement, induced-strain actuator is considered. The principle of high-power induced-strain actuation and the energy and energy density of several commercially available induced-strain actuators and results from static and dynamic tests on a full scale model are presented. Prechtl and Hall (60) describe a servoflap that uses a piezoelectric bender to deflect a trailing-edge flap for a helicopter rotor blade. The design uses a new flexure mechanism to connect the piezoelectric bender to the control surface. The efficiency of the bender was improved by tapering its thickness over its length. The authority of the actuator was also increased by implementing a nonlinear circuit to control the applied electric field; this allowed a greater range of actuator voltages.

An analytically developed smart material actuator that employed ETREMA TERFENOL-D[®] to demonstrate its effectiveness for helicopter rotor servoflap control was introduced by Ghorayeb et al. (61). The design enables control of the rotor blade flap by an actuator embedded in the blade itself. A series of loading conditions characterized by an additive process of steady-state, cyclic, and active control functions were considered for sustained flight. Optimization of the overall system gave rise to a system gain of 3.7 for sustained motion. Magnetostrictive actuators used in conjunction with an extension-torsion coupled composite tube were studied for actuating a rotor blade trailing-edge flap to control helicopter vibration actively (62). Thin-walled beam analysis based on the Vlasov theory was used to predict the axial force-induced twist and extension in the composite tube. Tests of the magnetostrictive actuator/composite tube systems showed good correlation between measured and predicted twist values.

A flap-operated, individual blade control (IBC) system for reducing vibration in the Army UH-60A helicopter is described (63). An alternative actuating technology using magnetostrictive materials was adapted. The advantages of the design include all-electric operation, simplicity, reliability, low mass, low voltage, and insensitivity to centripetal acceleration. The actuator requirements are derived for 17% chord width flaps integrated into the outer half-span of the trailing edge. Four flaps per blade for improved control effectiveness using opposing actuation, particularly for the first flatwise bending mode which has a node in the blade region under control, were used. An optimized magnetostrictive actuator was developed using Terfenol-D[®], material. The impact of magnetostrictive flap actuators on the development of practical IBC systems for

reducing vibration is discussed. Straub and Merkley (64) presented the results of a study to define conceptually an on-blade smart material actuator for primary and active control of a servoflap rotor. For a previously developed hybrid actuator concept, the design of the cyclic and active (high speed) control actuator and the feasibility of the collective (low speed) actuator and stroke multiplier were investigated. Sizing of actuator components based on AH-64 servoflap requirements shows that collective control using shape-memory alloys is well within the capability of the material. Cyclic and active control using magnetostrictive material leads to a reduced maneuver envelope due to weight and volume constraints.

Teves et al. (65) introduced an active control technology in the rotating system that uses pitch link actuators and results in the first flying four-bladed helicopter whose blades are individually controlled. To reduce the number of free control parameters, the blade control technology was tested in the harmonic mode. Recently, the design of large stroke, electromechanical actuators to power a trailing-edge servoflap system for feedback control of helicopter rotor vibration, acoustics, and aerodynamic performance was addressed by Prechtl and Hall (66). A new high-efficiency discrete actuator, the X-frame actuator, is described.

ACTIVE TWIST

A smart rotor that actively controls blade twist by using embedded piezoceramic elements as sensors and actuators to minimize rotor vibrations was developed (67). A one-eighth Froude-scale, bearingless helicopter rotor model that uses banks of torsional actuators capable of manipulating blade twist at frequencies from 5 to 100 Hz was tested for vibrational suppression capability. Experimental results show that tip twist amplitudes of the order of 0.5° are attainable in forward flight by the current actuator configurations. Test results also show that partial reduction of vibration is possible. Open-loop phase shift control of blade twist in the first four rotor harmonics produced measured changes in rotor thrust up to 9% of the steady-state values that resulted in 3 and 8% reductions in rotor pitching and rolling moments, respectively. Results from hover tests of a one-eighth dynamically scaled (Froude scale) helicopter model that had embedded piezoceramic elements (vibrational suppression) were reported (68). The twist performance of several rotor blade configurations as investigated using accelerometers embedded in the blade tip. A dynamically scaled (Froude scale) helicopter rotor blade was developed that has embedded piezoceramic elements as sensors and actuators to control blade vibrations (69). Each blade of a bearingless rotor model has banks of piezoelectric actuators embedded in the top and bottom surfaces at $\pm 45^\circ$ angles with respect to the beam axis. A twist distribution along the blade span is achieved by in-phase excitation of the top and bottom actuators at equal potentials, and a bending distribution is achieved by out-of-phase excitation. A uniform strain beam theory was formulated to predict analytically the static bending and torsional response of composite rectangular beams that have piezoelectric actuators embedded. Parameters such as bond thickness,

actuator skew angle, and actuator spacing were investigated by experiments and then validated by theory. The static bending and torsional response of the rotor blades was experimentally measured and correlated with theory. The viability of a vibrational reduction system based on piezoactuation of blade twist was investigated.

The structural and aerodynamic characteristics of a new class of active flight control surface were developed (70). The control surface uses a symmetrical, subsonic aerodynamic shell that is supported at the quarter-chord by a main spar and actively pitched by an adaptive torqueplate. Details of the structural mechanics of the torqueplate and several actuator elements, an interdigitated electrode (IDE) and constrained directionally attached piezoelectric (CDAP) elements are presented. An experimental torqueplate specimen constructed from PSI-5A-S2 piezoceramic shows high torsional deflections and stiffness as well as excellent correlation with theory. An active moving blade tip for a smart rotor that is torsionally actuated via a piezo-induced bending-torsion coupled composite beam is presented (91). A novel spanwise variation in the beam layout and piezoceramic element phasing is used to maximize the twist response and minimize the bending response. The bending results in an induced twist via structural coupling, and this twist deflects the moving tip. A variety of tests were conducted using a proof-of-concept actuator beam.

An interdigitated electrode piezoelectric in fiber composite (IDEPEC) method uses smart materials to achieve main spar twist (72). Active materials are embedded in the composite plies of rotor blades actuated by applied electrical fields. Active twist for a one-sixteenth scale CH-47 rotor blade is demonstrated, and a preliminary design for a one-sixth scale blade is presented. The benefits of actively twisting the main rotor spar are demonstrated, and the influence that active twist may have on rotor design constraints and traditional ground rules is illustrated. Overall system performance, including economic merit, weight, and power consumption for this method of actuation is evaluated, and the feasibility of applying this process to rotorcraft is discussed. An active blade designed for controlling rotor vibrations and noise was introduced (73). Active fiber composites were integrated within the composite rotor blade spar to induce shear stresses, which results in a distributed twisting moment along the blade. The design of an active blade model based on a one-sixth Mach scale Chinook CH-47D is reviewed. The requirements for the active fiber composites subjected to 160 kt., 3 g maneuver loads and the experimentally determined actuator capabilities are reviewed. The testing of a half-span active blade test article is described. Twist actuation performance is compared with model predictions. An analytical effort to examine the effectiveness of embedded piezoelectric active fiber composite laminae for alleviating adverse vibratory loads on helicopter rotor blades in high-speed, high-thrust forward flight conditions is detailed (74). Structural and piezoelectric actuating properties for a conceptual full-scale active fiber composite rotor blade were developed using a classical laminated plate theory approach. The out-of-plane bending and torsional dynamic responses of the active fiber composite blade, both with and without piezoelectric twist control, are calculated by using an aeroelastic

analysis code. The dynamic stall-limited maximum rotor thrust versus forward flight speed trends for the active fiber composite blade are compared with those of a baseline, conventional rotor blade structure. The maximum stall-limited rotor thrust using active blade twisting is approximately 5–10% greater than that achievable by using a conventional passive blade structure. A corresponding 10–15% increase in dynamic stall-limited forward flight speed is also shown when the active fiber composite blades are used. Torsional load buildup due to dynamic stall effects is delayed by 4% in rotor thrust and 10% in forward flight speed.

The vibratory load reduction at the rotor hub by using smart materials and closed-loop control was investigated by Chattopadhyay et al. (75). The principal load-carrying member in the blade is represented by a composite box beam, of arbitrary thickness, that has surface-bonded, self-sensing piezoelectric actuators. A comprehensive shear deformation theory was used to model the smart box beam. An integrated procedure for rotor vibratory load analysis was developed by coupling an unsteady aerodynamic model with the rotor blade dynamic model based on the smart composite box beam theory. The dynamic deformations of the blade in all three directions, flap, lead-lag, and torsion, were included in the analysis. The results showed significant reductions in the amplitudes of rotor dynamic loads by using closed-loop control. Aeroelastic modeling procedures used in the design of a piezoelectric controllable twist helicopter rotor are described (76). Two aeroelastic analytical methods were developed for active twist rotor studies and were used in designing the wind tunnel model blade. The first procedure uses a simple flap-torsion dynamic representation of the active twist blade and is intended for rapid and efficient control law and design optimization studies. The second technique employs a commercially available comprehensive rotor analysis package and is used for more detailed analytical studies. Analytical predictions of hovering light-twist actuating frequency responses are presented. Forward flight fixed system $n\Omega$ vibration suppression capabilities of the model active twist rotor system are also presented. Frequency responses predicted by using both analytical procedures agree qualitatively for all design cases considered; the best correlation results when uniform blade properties are assumed. The development of the active twist rotor for hub vibration and noise reduction studies is described (77). The rotor blade is integrally twisted by direct strain actuation by distributing embedded piezoelectric fiber composites along the span of the blade. The development of the analytical framework for this type of active blade is presented. The requirements for the prototype blade and the final design results are also presented. Active rotor blade tips offer an alternative approach for main rotor active vibrational control. The tips are actively pitched via a piezodriven bending-torsion coupled actuator beam that runs down the length of the blade. A near Mach scale rotor model design (tip Mach = 0.47) that has adaptive blade tips was reported (78).

A new type of flightworthy solid-state adaptive rotor system uses directionally attached piezoelectric (DAP) torqueplates to control Hiller servopaddles (79). The servopaddles change the rotor disk tilt and thereby induce

changes in forces and moments for flight control. The servopaddles were constructed from PZT-5H piezoceramic actuator sheets bonded symmetrically at 45° . An aluminum substrate and a high temperature cure was used for pre-compression. The classical laminate theory (CLT) along with strip theory aerodynamics and inertial relations was used in analytical modeling. Because rotor moments are proportional to servopaddle deflections at a fixed rotational speed, it was possible to cancel them out by balancing an aeroelastic coupling between the center of mass, the aerodynamic center, and the elastic axis. Bench testing of the rotor system showed good agreement between theory and experiment. The design and testing of a one-twelfth scale solid-state adaptive rotor for helicopter flight control via active pitch manipulation is described (80). The pitch angle is adjusted by using piezoceramic directionally attached piezoelectric (DAP) torqueplates mounted between the rotor shaft and the blade root. Analytical models based on classical laminated plate theory for steady torqueplate deflection are presented and verified using the solid-state adaptive rotor test article (Froude scaled).

MODELING

The presence of a sensor and actuator introduces discontinuities in both the geometric and material properties of a structure. Accurate and validated modeling techniques are also necessary to account for the presence of multiple and interacting failure modes. The efficient implementation of smart materials requires developing accurate analytical models that incorporate their unique properties and the discrete nature of their positions in the structure. Important issues that need to be addressed include: (1) a transducer model for smart materials that addresses the coupling effects or interactions among all possible fields, electrical, mechanical, and thermal; (2) a constitutive model that addresses the application of large induced strain when driven by a high electrical field, that is, a hysteresis model; (3) in a composite primary structure, a laminated smart structure model that efficiently predicts transverse shear effects accurately; and (4) modeling the presence of defects to predict the life of the structure accurately. Although, a significant amount of research effort has been directed to using smart materials for rotary wing control, all of the issues of basic modeling have not yet been fully resolved.

In the process of modeling rotor blades that use induced strain actuators, composite box beams can be used to model the principal load-carrying member in the blade. These simplified reduced-order models can be designed to meet the stiffness, mass, and twist distributions of a reference blade. The use of these models offers physical insight and reduces the computational time requirement, compared to the detailed finite element model. Therefore, they can be used efficiently within an optimization framework. Composite box beam analysis has been investigated by a number of researchers. Chandra et al. (81) developed a formulation for a composite thin-walled box beam that has distributed actuators. The results of the static analysis were correlated with experimental data.

Chattopadhyay and Seeley (82) developed a rotating composite box beam model that has embedded actuators and sensors, and results were obtained by developing a formal optimization technique. Chen and Chopra (83,84) used a one-dimensional, uniform strain model in their investigation of using distributed piezo actuators to twist the blade actively. Later, they refined the one-dimensional uniform strain theory to include closely spaced actuator interactive effects, based on a block-force assumption, and developed a Bernoulli–Euler plate theory to predict the static response of smart composite beams (85).

In all of this work involving box beams that have embedded actuators and sensors, the analysis was based on the classical laminate theory (CLT), thereby neglecting out-of-plane stresses and strains. The need for modeling transverse shear deformation in composites has been well documented. Transverse shear effects increase as laminate thickness and material anisotropy increase. In modeling composite laminates that have distributed sensors and actuators, it is important to have a general framework for evaluating the transverse shear effects accurately. Full three-dimensional analyses can accurately predict displacements and stresses, even at a local level, but are too computationally expensive to be used for practical analysis of aerospace structures. First-order shear deformation theories that use shear correction factors have been used by some researchers to analyze composites that use embedded adaptive materials (86). The dynamical modeling of rotating blades carrying a tip mass and surface-bonded piezoelectric actuators is addressed in (87). The blade is modeled as a thin-walled beam, including anisotropy and secondary warping. Transverse shear is included using shear correction factors. Layerwise theory was used in to ascertain the level of model complexity necessary to represent piezoelectric actuation of beam structures accurately (88). A refined hybrid plate theory that combines the layerwise theory and an equivalent single-layer theory with linear piezoelectricity was developed by Mitchell and Reddy to model smart composite laminates (89).

Several researchers have shown that the higher order theory can accurately describe transverse shear deformation in composites without being computationally prohibitive. Chattopadhyay et al. developed a higher order theory for modeling smart composite structures (90–92). In modeling plates of arbitrary thickness and surface-bonded self-sensing piezoelectric actuators, it was shown that the first natural frequency is highly overestimated by the classical theory, especially for thicker plates, whereas the first-order and the higher order theories agree well. Larger differences between the theories were observed in the second bending mode. Even the first-order theory, it was shown, overpredicts these frequencies (90). These results indicate that it is necessary to include an accurate description of the transverse shear stresses, which are important in composites due to the large ratio between Young's moduli and shear moduli. Because a helicopter rotor blade is very flexible, shear deformation plays an important role. Therefore, an accurate description of the displacement field is necessary. The higher order theory was extended by Chattopadhyay et al. (92) to model a composite box beam that has surface-bonded piezoelectric actuators and

sensors. The new theory, based on a refined higher order displacement field of an eccentric plate, is a three-dimensional model, which approximates the elasticity solution so that the box beam's cross-sectional properties are not reduced to one-dimensional beam parameters. Both in-plane and out-of-plane warping are included automatically in the formulation.

Coupling of thermal, piezoelectric, and mechanical fields plays an important role in analyzing smart structures. However, these issues are ignored in most applications. In most of the work, a one-way coupling that considers only the effect of a known field on another field is used. For example, it is assumed that the piezoelectric field can be calculated directly from the applied voltage, which is then introduced in the unknown displacement field as an induced strain. However, piezoelectric actuation changes the strain field during active control of the structure, and the new strain field, in return, also affects piezoelectric distributions. This is referred to as "two-way" interaction/coupling. Two-way coupling between piezoelectric and mechanical fields was included in the hybrid plate theory developed by Mitchell and Reddy (89). The piezothermoelastic behavior of composite plates was addressed by Tauchert (93) using CLT and by Lee and Saravanos using layerwise theory (94). In this work, a known thermal field was used to study the effect on mechanical and piezoelectric fields. The interactions between thermal and mechanical fields and between thermal and piezoelectric fields were ignored. Recently, a coupled thermal–piezoelectric–mechanical (t–p–m) model was developed by Chattopadhyay et al. to address the two-way coupling of smart composites (95,96). As shown in their work, coupling affects plate deformation and can lead to mispredicting the control authority of smart composite plates.

In all of this work, the material is usually assumed perfect and is not debonded or delaminated. Seeley and Chattopadhyay addressed debonded in piezoelectric actuators, which causes significant reductions in control authority (97). Recently Chattopadhyay et al. developed a more general framework for modeling adaptive composite beams and plates that have delaminations in the primary structure by using a refined third-order displacement field (91).

The applications of piezoelectric materials in smart structures are based mostly on a linear piezoelectric model. This implies both low electrical fields and low mechanical strains. However, to obtain greater actuating authority and increased induced strain for practical applications, an electric field is required, whose magnitude exceeds the limit of linear piezoelectric constitutive relationships. Piezoelectric actuators exhibit only mild nonlinear response at low voltage levels, but it is well known that the response can be profoundly nonlinear at high field strength (98). The behavior of piezoelectric actuators under high applied voltage was studied by Tiersten using a theory linear in displacement gradients but cubic in electric field (99). Chattopadhyay et al. (92) used the experimental results of Crawley and Lazarus (98) to develop an analytical model of the nonlinear effect. In this model, the coupling coefficients also depend on the actual strain in the actuator. The benefit of this model is that the final governing equations remain linear.

Piezoceramics exhibit constitutive nonlinearity involving hysteresis due to the variation in polarization if the magnitude of the applied electric field is higher than that of the coercive electric field. In addition, hysteretic behavior is also related to mechanical loading. Different hysteretic loops are observed as mechanical stresses vary. Several phenomenological theories have been proposed during the last two decades to match experimental results to assumptions motivated by certain piezoelectric characteristics (100,101). These theories are not completely satisfactory due to weak physical modeling or difficulties in adapting the models to accommodate varying operating conditions. Issues such as the crystal structure and interactions between the crystallites, the unique crystallite orientation distributions due to polarization switching, and stress and strain states are neglected in these models. A phenomenological hysteretic model based on a simplified piezoelectric microstructure was recently developed by Zhou and Chattopadhyay (102). The internal constraint of 180° and 90° polarization switching was considered. The energy losses due to orientation switching and the inclusions in piezoelectric materials were modeled. The three-dimensional stress state was addressed by using a new form of Gibbs free energy. The research yielded an explicit form of the constitutive relationships governing hysteresis.

The applications of SMA controllers result from their large constrained recovery despite their slow response time and high energy requirement, compared to piezoelectric materials. Birman (103) published a comprehensive review of work in the areas of SMA constitutive modeling and applications. A good overview of shape-memory alloy characteristics and applications was recently presented by Otsuka and Wayman (104). Lagoudas and Bo (105) studied the use of more than one active material by modeling piezoelectric and SMA layers in a composite laminate. Hebda and White (106) investigated SMA transformation behavior within an elastomeric matrix and concentrated on the two-way shape memory effect (TWSME). Epps and Chandra (107) embedded SMA wires in a composite beam to tuning the beam actively. A thermomechanical model under multiple nonproportional loading cycles was reported by Lagoudas et al. (108).

Simple classical Euler–Bernoulli beam theories or classical laminate theories have been used in most applications of SMA fibers. The SMA fibers were heated by applying electrical current using some control device, and it was assumed that the heat would not affect the composite. This important effect was addressed by Turner et al. (109) who observed significant changes in the properties, and natural frequencies of laminates by including the thermal effect. However, the model was macromechanical, so that it depends only on measuring fundamental engineering properties. In embedded SMA actuators, the large load transfer over a thin shear layer exacerbates the problem of debonding and, even using improved composite design and fabrication, debonding is expected at some point during the useful life of an active composite. Therefore, modeling must account for incipient damage to enable designing damage-tolerant active composites. Therefore, although significant research has been reported in modeling that uses SMA fibers, a critical gap must be bridged between

micromechanics and structural lamination theories. The strain energy absorption of SMA and the performance of SMA hybrid composites for augmenting damping in rotary wing structures have not yet been fully explored. Several issues require further investigation to make these concepts useful in rotorcraft application.

Several modeling approaches that describe the behavior of elastomeric materials used in helicopter rotor lag dampers are examined and evaluated in (110). Two of the models are then used in a simulation of a helicopter rotor startup, and the simulation results are compared to each other as well as to flight test data. The simple models (complex modulus models) that predict damper energy dissipation well are inadequate for predicting forced response. Elastomeric and Fluidlastic(R) damper activities at Boeing Helicopters and issues that must be considered in designing and analyzing such systems are discussed in (111). Bench tests were performed on various damper configurations to understand their dynamic characteristics.

FUTURE DIRECTIONS

Smart materials and composite structures have demonstrated a huge potential for enhancing the performance, reliability, and agility of rotary wing aircraft. Significant advances have been made in piezoelectric, electrostrictive, and magnetostrictive materials; shape-memory alloys; and active fiber composites. However, practically implementing of these proposed schemes still requires substantial research in a number of areas. The following is a brief description of some suggested research directions.

Multifunctional Systems

Multifunctionality will be a cornerstone of future military platforms and structures, and it is envisioned that these systems will have sensors and control systems of the order of tens of thousands. The excellent strength, stiffness, and low weight of heterogeneous materials and their extensive tailorability make them the material of choice for many present and future Army applications. Integration with arrays of distributed actuators, sensors, and microelectromechanical systems (MEMS) can lead to the development of new generations of efficient multifunctional systems. These information technology devices that are fabricated from active materials are self-contained data processors and embedded components of control devices. They can potentially be used for a multitude of rotary wing applications, including vibrational and noise control, detection and active mitigation of structural damage, and stability augmentation. The control and optimal distribution of these types of devices still remains a major concern. Data fusion techniques are also needed to optimize the number of embedded devices, their distribution, and material behavior within the system, so that crucial mission information is efficiently processed and maintained. Furthermore, the reliability of these embedded devices, not as stand-alone devices, but as integrated subsystems within systems that may comprise combinations of heterogeneous materials, such as resin and metal matrix composites,

high-strength laminated metals, toughened ceramics, and functionally graded materials must be addressed. New actuating materials and control systems that have the desired strain and frequency responses are needed so that these devices, which can range in size from the micro- to the mesoscale, can be easily integrated within larger scale heterogeneous host structures for tailored applications. Finally, research in developing a multidisciplinary analytical computational and experimental framework is necessary for designing and optimizing such systems.

Health Monitoring

Structural damage of heterogeneous systems can occur during production or field operations. Cyclic loading, excessive vibrations, and low velocity and/or high velocity impacts can cause structural damage in a rotary wing environment. Inability to detect damage in heterogeneous structures that may comprise combinations of composites, ceramics, and metals is a factor that limits their use in practice. Therefore, the development of integrated health monitoring and inspection capabilities is a vital research issue. The application of active materials to the development of novel sensing techniques and the ability to interpret sensor signals effectively and accurately in nearly real time are fundamental for improving the reliability of physical systems. Sensors based on active material can revolutionize health monitoring, damage detection, and non-destructive evaluation. However, this potential can be ensured only by deploying ultrareliable sensing techniques. Advances in the development of microsystems have made it possible for microsystems to be the enabling technology in developing viable health monitoring strategies for complex composite structures. Miniaturized sensory devices could be incorporated into heterogeneous structures to signal the presence, location, and extent of local and global failure modes, such as fiber breakage, fiber pull-out, delamination, and large matrix structural cracking. This requires interdisciplinary approaches that integrate research activities focused on sensors, actuators, signal processing and interpretation, structural modeling, system integration, electronics, and computational techniques.

Given the complexity of studying damage phenomena in solids, it may be necessary to incorporate a hierarchy of analytical networks and achieve structures that have large sensory array capabilities to ensure structural integrity and reliability based on continuous and dependable structural health monitoring, status inspection, and damage detection with minimal human intervention/involvement. Because practical structures are often subjected to impact loads and high strain rates that cause large nonlinear deformation and failure, the physical modeling of heterogeneous structures that contain integrated microdevices must be addressed. Investigation of physical models for nonconservative and nonlinear structural and sensor/actuator response and robust hierarchical control system models for fault-tolerant design of distributed embedded devices will be important. Appropriate numerical and analytical techniques and optimization procedures also need to be developed. The precision and reliability of the health monitoring system strongly rely on the accuracy

and analysis that relate sensor measurements to physical changes due to fracture initiation and progression. It would be desirable to integrate sensors and actuators based on active material, the structural components, the power supply, and the localized control circuit into a single unit.

Enhanced Structural Damping

Successfully incorporating advanced, active damping schemes and/or treatments into rotary wing aircraft would lead to a more comfortable ride, quieter interiors for greater crew and passenger comfort, safer flight due to enhanced rotor stability, more economical operation because of extended fatigue life of airframe and rotor blades, and improved maneuverability. Therefore, developing a more effective and economical structural damping approach that can adjust its mechanical properties to appropriate specifications will be beneficial in designing future rotorcraft systems. Because currently used damping models are fragmented and ad hoc, they must be generalized to become useful in design and analysis. Although recent research has demonstrated the feasibility of tailoring damping methodologies on a small scale, their synthesis on a practical scale is still not proven. Innovative approaches to modeling damping, compatible with current finite element codes, and the introduction of damping mechanisms into a structure are needed for vibrational suppression, noise reduction, increased system performance, extended service life, and lower maintenance costs. Thus, fundamental research is required to devise actuating schemes, develop, and validate experimentally new modeling methods and controller designs for damping mechanisms. The control issues in this initiative are on the frontier of distributed nonlinear control. These issues range from structural modeling of control elements, through coupling and feedback to the elements of the system, to the study of the overall dynamics of the control system. Improving the performance of active vibrational suppression systems by using nonlinear approaches and subsequently exploring the new design space holds the promise of higher energy dissipation capabilities, more robustness, and adaptability to changing demands and diminished requirements from resources and infrastructure. However, a number of hurdles in system concept definition, nonlinear systems modeling, and design need to be addressed.

Improved Actuators

Much progress has been made recently in using active materials in actuator design for damping and/or induced strain. However, successful application of such actuators will depend upon their cost and ease of system integration and retrofitting. Therefore, innovative concepts to address issues in the development of design tools, methodology, modeling, simulation, prototyping, and real-time software/hardware implementation are necessary for a number of designs of actuators based on active materials. Low-cost, high-bandwidth, high-authority (force/displacement characteristics), innovative actuators for both military and commercial applications need to be developed, and the associated driving electronics and control systems must also be addressed.

Aerodynamic Control

The application of smart materials, intelligent systems and control logic concepts to aerodynamic problems has been very limited, primarily because of the lack of knowledge and understanding of the way these techniques might best be used in aerodynamic flow control. There is a real need for better understanding of the fluid physics and fluid mechanisms that will become available from using advanced materials and intelligence technologies. Although major improvements in aerodynamic performance and control from using unsteady flow control techniques seem feasible, their use will depend upon developing techniques to create the flow interactions that govern the manner in which these new flow control ideas are implemented. Much better understanding must be established concerning the benefits and difficulties of using the various advanced actuators that are needed to create physical interaction with flow. Research activity in this area might combine concepts derived from the areas of adaptive structures, MEMS devices, control theory, modeling and numerical analysis techniques, and other relevant technologies to evaluate the various possible actuators for such applications.

BIBLIOGRAPHY

1. R.G. Loewy, *Smart Materials Struct.* **6**(5): 11–42 (1997).
2. F. Nitzsche, *Aeronaut. J.* **103**(1027): 429–434 (1999).
3. P.P. Friedmann and T.A. Millott, *J. Guidance Control Dynamics*, **18**(4): 664–673 (1995).
4. F. Nitzsche and E. Breitbach, *J. Aircraft* **31**(5): 1178–1188 (1994).
5. F. Nitzsche and E. Breitbach, *Proc. 33rd AIAA/ASME/ASCE/AHS/ASC Struct. Dynamics Mater. Conf.*, Dallas, TX, April 1992.
6. F. Nitzsche, *AIAA-94-1766*, 1994, pp. 298–308.
7. F. Nitzsche, *Proc. SPIE* **2443**: 20–27 (1995).
8. L. Librescu, O. Song, and C.A. Rogers, *Int. J. Eng. Sci.* **31**(5): 775–792 (1993).
9. O. Song, L. Librescu, and C.A. Rogers, *J. Intelligent Mater. Syst. Struct.*, **5**(1): 42–48 (1994).
10. R.L. Roglin and S.V. Hanagud, *Smart Mater. Struct.*, **5**(1): 76–88 (1996).
11. M.G. Spencer, R.M. Sanner, and I. Chopra, *J. Intelligent Mater. Syst. Struct.* **9**(3): 160–170 (1998).
12. C. Walz and I. Chopra, *Smart Mater. Struct.* **5**(1): 98–113 (1996).
13. A. Chattopadhyay, Q. Liu, and H. Gu, *AIAA J.* **38**(4): (2000).
14. D.P. McGuire, *Proc. Am. Helicopter Soc. 50th Annu. Forum*, Washington DC, 1994, pp. 295–303.
15. R.A. Ormiston, *Proc. 22nd Euro. Rotorcraft Forum*, Brighton, England, September 17–19, 1996.
16. F. Gandhi, *Proc. R. Aeronaut. Soc. Innovation Rotorcraft Technol. Conf.*, London, UK, June 24–25, 1997.
17. G.M. Kamath and N.M. Wereley, *Proc. SPIE* **2443**: 120–133 (1995).
18. S. Marathe, F. Gandhi, and K.W. Wang, *J. Intelligent Mater. Syst. Struct.* **9**(4): 272–282 (1998).
19. G.M. Kamath, N.M. Wereley, and M.R. Jolly, *J. Am. Helicopter Soc.* **44**(3): 234–248 (1999).

20. A. Baz and J. Ro, *Smart Mater. Struct.* **5**(3): 272–280 (1996).
21. A. Baz and J. Ro, *Shock Vib.* **2**(1): 33–42, (1995).
22. E.M. Austin, D.J. Inman, and S.C. Huang, *Proc. SPIE Conf. Smart Struct. Mater. 1996*, CA, Feb. 1996.
23. W.H. Liao and K.W. Wang, *Trans. ASME J. Vib. Acoust.* **120**(4): 894–900 (1998).
24. W.H. Liao and K.W. Wang, *Trans. ASME J. Vib. Acoust.* **120**(4): 886–893 (1998).
25. A. Badre-Alam, K.W. Wang, and F. Gandhi, *Smart Mater. Struct.* **8**(2): 182–196 (1999).
26. Q. Liu, A. Chattopadhyay, H. Gu, and X. Zhou, *Smart Mater. Struct.* **9**: 523–532 (2000).
27. Q. Liu and A. Chattopadhyay, *Intelligent Mater. Syst. Struct.* **11**(6): 492–500 (2000).
28. A. Chattopadhyay, J-S. Kim, and Q. Liu, *J. Vib. Control*, Special Issue on Active Constrained Layer Damping (in press).
29. R.L. Spangler and S.R. Hall, *AIAA-90-1076*, 1990, pp. 1589–1599.
30. H. Strehlow and H. Rapp, *Proc. 19th Eur. Rotorcraft Forum*, 1993.
31. T.A. Millott and P.P. Friedmann, *AIAA-94-1306*, 1994, pp. 8–22.
32. D.K. Samak and I. Chopra, *Smart Mater. Struct.* **5**(1): 58067 (1996).
33. B.T. Spencer and I. Chopra, *Proc. SPIE* **2717**: 120–130 (1996).
34. J. Milgram and I. Chopra, *J. Am. Helicopter Soc.* **43**(2): 110–119 (1998).
35. J. Milgram, I. Chopra, and F. Straub, *J. Am. Helicopter Soc.* **43**(4): 319–332 (1998).
36. A.P.F. Bernhard and I. Chopra, *Proc. Am. Helicopter Soc. Natl. Tech. Specialists Meet.*, Williamsburg, VA, 1995, pp. 107–137.
37. A.P.F. Bernhard and I. Chopra, *Proc. Am. Helicopter Soc. 53rd Annu. Forum*, Virginia Beach, VA, 1997, pp. 249–273.
38. N.A. Koratkar and I. Chopra, *J. Intelligent Mater. Syst. Struct.* **8**(7): 555–570 (1997).
39. A.P.F. Bernhard and I. Chopra, *J. Am. Helicopter Soc.* **44**(1): 3–15 (1999).
40. O. Ben-Zeev and I. Chopra, *Smart Mater. Struct.* **5**(1): 11–25 (1996).
41. N.A. Koratkar and I. Chopra, *Proc. SPIE* **3329**: 266–290 (1998).
42. R. Lee and I. Chopra, *Proc. SPIE* **3329**: 321–332 (1998).
43. T. Lee and I. Chopra, *AIAA-99-1502*, 1999, pp. 2403–2413.
44. N.A. Koratkar and I. Chopra, *AIAA-99-1501*, 1999, pp. 2380–2402.
45. N.A. Koratkar and I. Chopra, *Proc. Am. Helicopter Soc. 55th Annu. Forum*, Montreal, Can., 1999, pp. 558–578.
46. F.K. Straub and R.J. King, *Proc. SPIE* **2717**: 66–77 (1996).
47. M.V. Fulton and R.A. Ormiston, *Proc. Am. Helicopter Soc. 53rd Annu. Forum*, Virginia Beach, VA, 1997.
48. M.G. Spencer, R.M. Sanner, and I. Chopra, *AIAA-98-2099*, 1998, pp. 3326–3336.
49. V. Sahasrabudhe, P.M. Thompson, and B.L. Aponso, and P.C. Chen, *Proc. Am. Helicopter Soc. 54th Annu. Forum*, Washington DC, 1998, pp. 1057–1069.
50. F.K. Straub and B.D. Charles, *Proc. Am. Helicopter Soc. 55th Annu. Forum*, Montreal, Can., 1999, pp. 523–532.
51. Y.K. Yillikci, *Proc. 18th Eur. Rotorcraft Forum*, 1992.
52. T.A. Millot and P.P. Friedmann, *J. Guidance Control Dynamics* 900–913.
53. C.E. Seeley, A. Chattopadhyay, and D. Brei, *AIAA J.* **34**(1): 123–128 (1996).
54. C.E. Seeley, A. Chattopadhyay, and L. Mitchell, *Smart Mater. Struct.* **6**: 134–144 (1997).
55. D. Brei, J. Clement, and R. Barrett, *Proc. 40th AIAA/ASME/ASCE/AHS Struct. Dynamics Mater. Conf., Adaptive Struct. Forum*, April 1999, Paper No. AIAA-99-1503.
56. A.A. Bent, N.W. Hagood, and J.P. Rodgers, *J. Intelligent Mater. Syst. Struct.* **6**(3): 338–349 (1995).
57. V. Giurgiutiu, Z. Chaudhry, and C.A. Rogers, *J. Intelligent Mater. Syst. Struct.* **6**(5): 583–597 (1995).
58. F.K. Straub, *Smart Mater. Struct.* **5**(1): 1–10 (1996).
59. V. Giurgiutiu C.A. Rogers, and R. Rusovici, *J. Intelligent Mater. Syst. Struct.* **7**(2): 192–202 (1996).
60. E.F. Prechtl and R. Steven, S.R. Hall, *Smart Mater. Struct.* **5**(1): 26–34 (1996).
61. S.R. Ghorayeb, T.T. Hansen, and F.K. Straub, *Proc. SPIE* **2443**: 28–36 (1995).
62. C.M. Bothwell, R. Chandra, and I. Chopra, *AIAA J.* **33**(4): 723–729 (1995).
63. D. Bushko, R. Fenn, M. Gerver, J. Berry, F. Frank Philips, and D.J. Merkley, *Proc. SPIE* **2717**: 80–90 (1996).
64. F.K. Straub, and D.J. Merkley, *Smart Mater. Struct.* **6**(3): 223–234 (1997).
65. D. Teves, V. Klöppel, P. Richter, *Proc. 22nd Eur. Rotorcraft Forum*, 1996.
66. E.F. Prechtl and S.R. Hall, *Smart Mater. Struct.* **8**(1): 13–30 (1999).
67. P.C. Chen and I. Chopra, *J. Intelligent Mater. Syst. Struct.* **8**(5): 414–425 (1997).
68. P.C. Chen, and I. Chopra, *AIAA J.* **35**(1): 6–16 (1997).
69. P.C. Chen, and I. Chopra, *Smart Mater. Struct.* **5**(1): 35–48 (1996).
70. R. Barrett, *Smart Mater. Struct.* **4**(2): 65–74 (1995).
71. A.P.F. Bernhard and I. Chopra, *Proc. SPIE* **2717**: 63–79 (1996).
72. R.C. Derham and N.W. Hagood, *Proc. Am. Helicopter Soc. 52nd Annu. Forum*, Washington DC, June 1996, pp. 1242–1252.
73. J.P. Rodgers and N.W. Hagood, *Proc SPIE* **3329**: 291–308 (1998).
74. W.K. Wilkie, K.C. Parkt, and W.K. Belvin, *AIAA-98-2002*, 1998, pp. 2458–2472.
75. A. Chattopadhyay, Q. Liu, and H. Gu, *AIAA J.* **38**(7): 1125–1131 (2000).
76. W.K. Wilkie, M.L. Wilbur, P.H. Mirick, C.E.S. Cesnik, and S. Shin, *Proc. Am. Helicopter Soc. 55th Annu. Forum*, Montreal, Can., May 1999, pp. 545–557.
77. C.E.S. Cesnik, S. Shin, W.K. Wilkie, and M.L. Wilbur, *Proc. Am. Helicopter Soc. 55th Annu. Forum*, Montreal, Can., May 1999, pp. 533–544.
78. A.P.F. Bernhard, and I. Chopra, *Proc. Am. Helicopter Soc. 55th Annu. Forum*, Montreal, Can., May 1999, pp. 579–598.
79. R. Barrett, M. Schijesman, and P. Frye, *Smart Mater. Struct.* **7**(3): 422–431 (1998).
80. A.R. Barrett and J. Stutts, *Smart Mater. Struct.* **6**(4): 491–497 (1997).
81. R. Chandra, A.D. Stemple, and I. Chopra, *J. Aircraft* **27**(7): 619–626 (1990).

82. A. Chattopadhyay and C.E. Seeley, *Smart Mater. Struct.* **4**: 170–178 (1995).
83. P.C. Chen and I. Chopra, *Proc. SPIE North Am. Conf. Smart Struct. Mater.*, Albuquerque, NM, 1993.
84. P.C. Chen and I. Chopra, *Proc. 36th AIAA/ASME/ASCE/AHS Struct. Dynamics Mater. Conf., Adaptive Struct. Forum*, New Orleans, LA, April 13–14, 1995.
85. P.C. Chen and I. Chopra, *Proc. Am. Helicopter Soc. Natl. Tech. Specialists' Meet. Rotorcraft Struct.*, Williamsburg, VA, Oct. 30–Nov. 2, 1995.
86. K. Chandrashekhara and A.N. Agarwal, *J. Intelligent Mater. Syst. Struct.* **4**: 496–508 (1993).
87. O. Song and L. Librescu, *Acta Mechanica* **134**: (3–4) 169–197 (1999).
88. D.H. Robbins and J.N. Reddy, *Composite Struct.* **41**(2): 265–279 (1991).
89. J.A. Mitchell and J.N. Reddy, *Int. J. Solids Struct.* **32**(16): 2345–2367 (1995).
90. A. Chattopadhyay and C.E. Seeley, *Composites Part B*, **28B**: 243–252 (1997).
91. D. Dragomir-Daescu, A. Chattopadhyay, and H. Gu, *AIAA J.* **137**(2): (1999).
92. A. Chattopadhyay, H. Gu, and Q. Liu, *Composites Part B* **30**: 603–612 (1999).
93. T.R. Tauchert, *J. Thermal Stresses* **15**: 25–37 (1992).
94. H. Lee and D.A. Saravanos, *Proc. AIAA/ASME/ASCE/AHS/ASC 37th Struct. Struct. Dynamics Mater. Conf.*, Salt Lake City, UT, 1996, pp. 1781–1788.
95. H. Gu, A. Chattopadhyay, J. Li, and X. Zhou, *Int. J. Solids Struct.* **37**: 6479–6497 (2000).
96. X. Zhou, A. Chattopadhyay, and H. Gu, *AIAA J.* **38**(10): 1939–1948 (2000).
97. C.E. Seeley and A. Chattopadhyay, *Int. J. Solid Struct.* **36**: 1823–1843 (1999).
98. E.F. Crawley and K.B. Lazarus, *Proc. 30th AIAA/ASME/ASCE/AHS/ASC Struct. Struct. Dynamics Mater. Conf.*, Mobile, AL, 1989, pp. 2000–2010.
99. H.F. Tiersten, *J. Appl. Phys.* **74**: 3389–3393 (1993).
100. J. Sirohi and I. Chopra, *SPIE Smart Struct. Integrated Syst.* **3329**: 626–646 (1998).
101. M. Omura, H. Adachi, and Y. Ishibashi, *Jpn. J. Appl. Phys.* **30**(9B): 2384–2387 (1991).
102. X. Zhou and A. Chattopadhyay, *J. Appl. Mech.* **68**: 270–277 (2001).
103. V. Birman, *Appl. Mech. Rev.* **50**(11): 629–645 (1997).
104. K. Otsuka and C.M. Wayman, (eds.), *Shape Memory Materials*. Cambridge University Press, Cambridge, UK, 1998.
105. D.C. Lagoudas and Z. Bo, *Smart Mater. Struct.* **3**: 309–17 (1994).
106. D.A. Hebda and S.R. White, *Proc. ASME AMD 206/MD* **58**: 111–119 (1995).
107. J. Epps and R. Chandra, *Smart Struct. Mater.* **2443**: 76–88 (1995).
108. Z. Bo, D.C. Lagoudas, and D. Miller, *J. Eng. Mater. Technol.* **121**: 75–85 (1999).
109. T.L. Turner, *7th Int. Conf. Recent Adv. Struct. Dynamics*, The Institute of Sound and Vibration Research, University of Southampton, England, July 24–27, 2000.
110. D.L. Kunz, *AIAA J.* **35**(2): 349–354 (1997).
111. B. Panda, E. Mychalowycz, and F.J. Tarzanin, *Smart Mater. Struct.* **5**(5): 509–516 (1996).

AIRCRAFT CONTROL, APPLICATIONS OF SMART STRUCTURES

JOHANNES SCHWEIGER

European Aeronautic Defense
and Space Company
Military Aircraft Business Unit
Muenchen, Germany

INTRODUCTION

Probably the most famous photograph in aviation, (Fig. 1) depicts the first successful manned powered flight on 17 December 1903 by the Wright brothers. It also shows the first successful airplane that had an active structure. The Wright brothers' design must definitely be called smart in many aspects and design features. They were among the first pioneers in aviation who had realized that uncoupled control about all three vehicle axes was required. They had done systematic experimental aerodynamic research to achieve the maximum possible aerodynamic performance. And they had learned how to design and manufacture lightweight structures in their bicycle shop. Their solution for adequate roll control of the airplane, moreover, was more than one century ahead of the state of the art in aviation technology. As we are approaching the centennial celebrations of this remarkable event, no single airplane exists yet that uses a smart structural concept to control the flight of the vehicle.

Rather than fighting the low torsional stiffness of their braced biplane wing design, they used this characteristic positively. By the sideways motions of the pilot, who lay on a sliding cradle, the wires attached to the cradle twisted the wing tips in opposite directions, thus producing the desired aerodynamic loads to roll the airplane. Figure 2 from Orville Wright's book (1) demonstrated this principle, which is also a very good example of the importance of integrated or multidisciplinary design concepts, especially in aeronautics. Unfortunately, this knowledge was lost and forgotten over the years, mainly because of more expert knowledge in single disciplines and more formal and

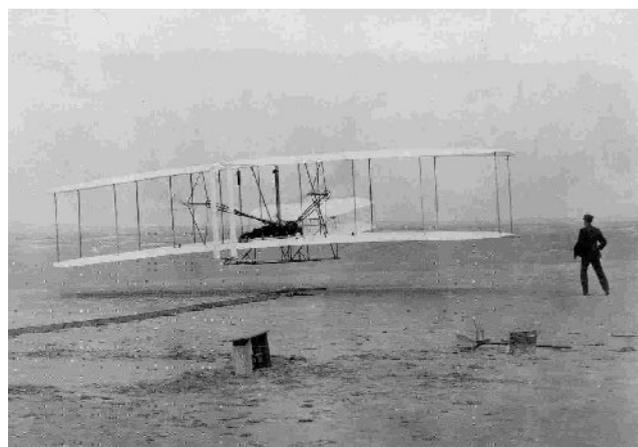


Figure 1. First flight of Wright Flyer I on 17 Dec. 1903 (available as a postcard).

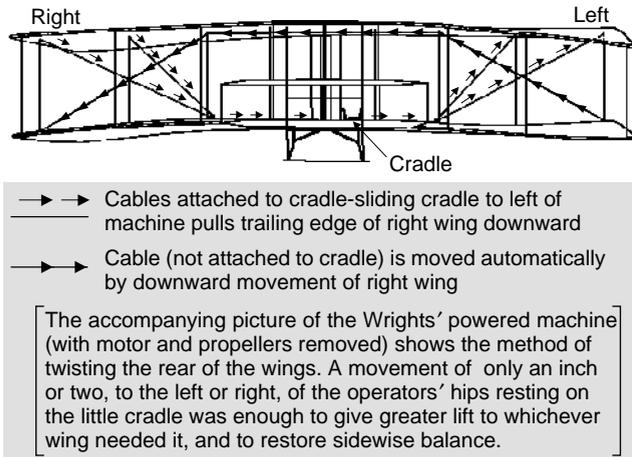


Figure 2. Active structural concept of the Wright Flyer I for roll control [adopted from (1)].

bureaucratic processes in designing new airplanes. Only in recent years, some prophets in aerospace are trying to spread the news about this old idea again and develop some new ones. Weisshaar (2), for example, in 1986, cited the success of the Wright Flyer as a good example of the need for integrated design methods. The Wright Flyer also demonstrates that smart aircraft structures do not necessarily rely only on advanced active materials.

Even earlier than that, active structural concepts were studied. Alois Wolfmüller (1864–1948), the producer of the world's first motorcycle, bought the No. 2 model of the production glider "Normal-Segelapparat" (normal soaring apparatus) from Otto Lilienthal in 1894 (3). Both aviation pioneers were communicating about improvements in performance and maneuverability by controlling the air loads through flexible wing twist. Wolfmüller tried to improve performance by introducing a flexible hinge in the wings to modulate aerodynamic control forces by flexible deformations.

Today, aircraft control is achieved by control surfaces attached to the main aerodynamic surfaces. These devices—aileron, elevator, and rudder—create the required forces and moments to control the motion of the aircraft about all three axes in space. Depending on the size and speed of the aircraft, these surfaces are actuated manually or by hydraulic systems. If the Wright brothers had used separate ailerons to roll the airplane, the additional structural weight might have been too much for the available power from the engine.

The idea of active or smart structures to control air vehicles is as old as the earliest known attempts to fly in heavier-than-air machines. Early attempts by humankind to fly were usually based on efforts to understand and copy the flight of birds. Besides the difficulty in controlling an unstable flying vehicle, which requires to day's high computing power or the complex neural network of animals to control their muscular systems, it is even more difficult to sense and actuate the dynamic motions of continuously deforming, flexible aerodynamic surfaces.

In most of these efforts, the pilot was supposed to actuate birdlike aerodynamic surfaces to produce the required

lift force, create the forward thrust, stabilize the vehicle, and apply the necessary control inputs in time. For a multitude of reasons, all of these early efforts were doomed to failure from the beginning:

- very limited knowledge of the laws of aerodynamics and flight mechanics;
- insufficient load capacity of the available materials, only primitive manufacturing techniques, and only a rudimentary understanding of structural mechanics;
- not enough sustained power output from the human body to produce the required thrust or lack of engines that had sufficient power density;
- efforts in trying to copy structural design principles from nature did not take into account the scaling laws of physics; the resulting designs were too heavy, too fragile, or too flexible; and
- and, in most cases, also the very complex efforts required to stabilize and control a flying vehicle without natural stability.

Only in recent years, after almost one century since the first successful powered flight, was it possible to design, build, and fly vehicles powered by the human body for the sole purpose of winning trophies.

For these reasons, the first successes in aviation were possible only by using design concepts for almost "rigid" surfaces and natural stability of the vehicles. Nevertheless, a major contribution to the success of the Wright brothers was their "Smart Structures" flight control system for the roll axis. They were among the first pioneers in aviation who had realized that uncoupled control about all three vehicle axes was required. They had done systematic experimental aerodynamic research to achieve maximum possible aerodynamic performance. They had a combustion engine that had sufficient energy density available "just in time," and they had learned how to design and manufacture lightweight structures in their bicycle shop.

SMART STRUCTURES FOR FLIGHT IN NATURE

Although complete plants do not fly, they have to withstand aerodynamic loads by using proper aeroelastic reactions, and sometimes their existence relies on their aerodynamic performance. The seeds of some plants are optimized in shape for long distance flight and mass distribution by millions of generations in a genetic optimization process. The interest in micro air vehicles in recent years has increased aerodynamic research efforts in this area (4). Although not for free flight, the leaves of trees and their stems are built to withstand strong winds. The joints between leaves and branches must have the right amount of flexibility in bending and torsion to reduce aerodynamic loads and at the same time avoid excessive unsteady loads from flutter.

More interesting for aircraft applications of smart structures is the flight of animals. As mentioned before, early aviators tried to learn from birds. However, the required knowledge about the physics of flight was missing in these early attempts. The complex interactions of steady

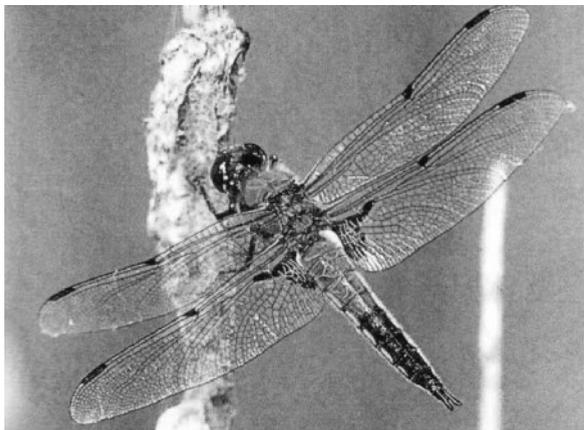


Figure 3. “Structural” design concept for dragonflies.

and unsteady aerodynamic forces and the active motions and passive deformations of wings and feathers are still not completely understood today. We are only now beginning to understand the functions and importance of the individual components for the efficiency of animal flight. Pendleton (5) gives some good examples from the prehistoric flying saurian to modern birds that have feathers.

But even more astonishing are the achievements in the art of flying for another species. Insects show by far the widest variety and most advanced structural concepts of active and passive control. An ordinary fly can land and take off from the tip of your nose or from the ceiling of a room. Dragonflies like that shown in Fig. 3 use a combination of passive mass balance to prevent flutter at the wing tips and an advanced structural design that has stiff chitin “spar” elements to support the membrane skins and flexible hinges of resilium to adjust the shape for all flight conditions.

The variety and large number flying members in the family of insects has been attributed to their ability to fly, which offers advantages of reaching and conquering new territories more easily.

In the context of formal optimization methods in aeronautics, genetic algorithms became fashionable in recent years. The question in the context of technical products is, can we really wait as long as in nature to get better products? Or should we continue to rely on gradient based methods, which can be seen as targeted “artificial genetic manipulations” similar to the “biological engineering” approaches of today?

GENERAL REMARKS ON ASPECTS OF AIRCRAFT DESIGN

One reason that we have not seen more progress in the application of smart structures in aeronautics may be the lack of understanding of the interactions between the different classical disciplines in aircraft design and between these disciplines and the specialists from the smart materials area for each others’ needs and capabilities.

To assess the possibilities of smart structure applications for aircraft control, it is advisable to look at some aspects of aircraft design first. The structural engineer is

usually concerned with the strength of a design for critical conditions and may be surprised in some cases, how relatively small deformations of the structure can have severe impacts on the aerodynamic performance or on maneuverability. The aerodynamic and flight control engineers want the structure “as rigid as possible,” but at the same time want different shapes for different flight conditions. These requirements may attract the smart structure specialists to make the design deformable and at the same time meet tight requirements for the external shape in changing flight conditions. This requires a look at the stiffness of the structure from the viewpoint of strength, as well as at the additional internal loads created by active deformations. And finally, aeroelastic aspects like flutter stability and effectiveness of deformed aerodynamic surfaces have to be considered.

An aircraft design is always a compromise for the aerodynamic engineer between different flight conditions: a small flat wing for cruising with low drag at high speed and a large, cambered wing for take-off and landing at low speed. This can be met partially by extendable control surfaces attached to a fixed surface by complex kinematic systems. On the other hand, the complexity of these systems increases the structural weight. Therefore, it looks attractive to replace these mechanisms and integrate their functions into one actively deformable structure. To consider such options, it is necessary to look at basic structural design requirements for aircraft wings. They have to carry loads from 2.5 to 9.0 times the total weight of the airplane. And they must be strong enough in torsion to keep the aerodynamic shape and transmit the loads from control surfaces. To achieve this at an acceptable structural weight, sophisticated lightweight design concepts were developed during the first half century of manned aviation. Besides improvements in materials and manufacturing, lightweight design by shape—within the prescribed aerodynamic shape is the main principle.

Simplified, this principle leads to the placement of material on the external shape of the airfoil and to closed cross sections of maximum area. This provides maximum strength for a fixed amount of material. At the same time, this structure also has maximum stiffness. This fact must be kept in mind when active deformable structural concepts are considered for airframe components.

TRADITIONAL ACTIVE OR ADAPTIVE AIRCRAFT CONTROL CONCEPTS

One kind of active aircraft control concept was developed and demonstrated in the early 1980s: artificial stabilization of an airplane’s flight path by fast motions of the control surfaces, commanded by a digital flight control system. Today, this system helps to reduce trim drag and increase the agility of all modern fighter aircraft by avoiding negative contributions from a stabilizer surface to the total lift force to establish natural static stability.

Most airplanes have wings adaptive by additional deployable surfaces like slats at the leading edge or Fowler flaps at the trailing edge to provide additional lift for take-off and landing or to provide clean flow conditions

for extreme maneuvers. The resulting effects are usually a combination of increased wing surface, increased camber, and accelerated air flow. These concepts are well described in most aerodynamic text books (6), where more sophisticated and exotic concepts like telescope wings can also be found. Variable wing sweep is rather common for combat aircraft, among them the American F-111. The Mission Adaptive Wing (MAW) demonstrator version of this aircraft used a complex mechanical system to adjust the wing camber for changing flight conditions (5). However, this system proved too complex and heavy for applications on production aircraft.

Enhancement of active flutter stability by control surface deflections to create unsteady aerodynamic damping forces was also demonstrated in the 1970s (7). There are two main reasons that we do not see them on today's airplanes:

- flutter within the flight envelope in failure of the system is considered too critical;
- the performance of these systems was not very good because of the loss of static aeroelastic effectiveness of the control surfaces as airspeed increased; it is typically close to zero for ailerons mounted near the wing tip at the trailing edge.

The last point also applies to the effectiveness of maneuver load alleviation systems by symmetrical deflected outboard aileron surfaces.

THE RANGE OF ACTIVE STRUCTURES AND MATERIALS APPLICATIONS IN AERONAUTICS

Besides aircraft control, including load alleviation, aerodynamic performance improvement by shape, and static aeroelastic applications for maneuverability and static stability enhancement, which will be the main points of the following sections, the following main additional applications in aeronautics are being addressed by theoretical and experimental research:

Health monitoring. Because piezoelectric materials can serve as actuators and sensors as well, these materials can be used to monitor changing static or dynamic internal loading conditions that result from airplane maneuvers or from failures in the structure. Although these functions have been widely discussed in the literature for two decades, the author knows of no application to a production airplane to date.

Vibration control. This application can be subdivided into two categories:

- to treat local or global vibrations of structural components. This was also the first flight test demonstration of piezo materials applied to a skin panel affected by engine noise of the Rockwell B-1B bomber (8).
- to reduce vibrational levels on equipment that is mounted internally or externally on the aircraft structure. This can be done to support the integrity of the equipment or to improve

the performance of equipment acting as sensor systems. Because rather small deformations and forces have to be handled here, the most promising and near term applications of smart materials on airplanes in this area were already predicted in the survey paper by Crowe and Sater (9).

Active flutter suppression. A lot of research has been dedicated to this task. The NASA wind tunnel test program "Piezoelectric Aeroelastic Response Tailoring Investigation"(PARTI) (10) demonstrated this technology. Applications to real aircraft are rather unlikely because of the previously mentioned safety aspects. Panel flutter as a special case for the instability of individual skin panels at supersonic speed is also addressed in many publications. The required effort for installing and controlling active devices is inappropriate compared to a simple structural reinforcement of critical panels.

Acoustic noise reduction. Several research projects are dealing with cabin noise reduction, especially for turboprop airplanes.

AIRCRAFT STRUCTURES

Definition of a Structure

To design active structural systems for airplane control, it is necessary to understand the functions of the structure, the requirements that define its properties, and the reasons that existing aircraft structures are built in a specific way.

The main functions of the airframe are

- to bear and carry the loads acting on the vehicle;
- to provide space and support for engines, equipment, payload, and fuel; and
- to satisfy the tight rigidity and smoothness required by aerodynamics for the external shape.

In other words, the external shape of the vehicle is defined by aerodynamic performance requirements, and the structure has to meet these requirements within this shape at minimum weight. This requires carefully combining the shape and load-carrying purpose of the individual structural elements with the best available material for this purpose. Early designs used fabric to collect the aerodynamic pressure loads on a wing surface and transfer them to the fuselage through a wooden framework supported by thin wires. As soon as it was discovered that the aerodynamic drag of these wires is 100 times higher than that of a carefully designed airfoil of the same thickness, design efforts concentrated on cantilevered wings where the loads could be completely transferred internally. Aluminum skins paved the way for modern monocoque airframes, where the main loads are carried by a monolithic shell structure which needs mainly only the internal structure to maintain the shape. The shape of this shell provides highest resistance against bending and torsion loads at minimum weight.

Rigidity of Wing Structures

Considering active deformations of a wing structure, it is useful to look at the differences in the possible passive deformations under different loading conditions. The lift force is the largest component of the aerodynamic forces. It corresponds to multiples of the total weight of the airplane, 2.5 times for transport aircraft and currently up to 9 times for fighters. At the same time, the external shape of a wing has the smallest dimension in its height and the largest in a spanwise direction. This means that the structure needs the largest cross sections of its skins on the upper and lower surfaces. Therefore, to deform a wing in bending actively would be difficult but not impossible. However, the bending deformation of an unswept wing has no impact on the aerodynamic characteristics or loading conditions. Only swept wings are sensitive in this respect. Swept forward, bending increases the local aerodynamic angle of attack in the streamwise direction, as indicated in Fig. 4. This increases the bending moment and causes structural divergence at a flight speed, called the divergence speed, which depends on the bending stiffness and geometric properties of the wing. The same effect reduces the bending moment on a swept-back wing under load, which acts as a passive load alleviation system. But to control these deformations actively by internal forces would mean stretching and compressing the skins in the spanwise direction—a rather difficult task.

The aerodynamic drag forces that act in the streamwise direction are smaller than the lift forces by a factor of 10. At the same time, the shape of the airfoil creates a high static moment of resistance in this direction. For these reasons, the loads in this direction need no special attention in the structural design. An active deformation would be both very difficult and meaningless.

Torsional loads on a wing can be very high, depending on the chordwise center of pressure locations and on additional forces from deflected control surfaces. A center of pressure in front of the fictive elastic axis through the wing cross sections causes torsional divergence at a certain flight speed, and a center of pressure too far behind the elastic axis twists the wing against the desired angle

of attack or control surface deflection. Therefore, the wing torsional deformation is very sensitive to the loads acting on the wing, to the spanwise lift distribution which is important in the aerodynamic drag, and to the effectiveness of the control surfaces. As mentioned before, a closed torque box that has a maximum cross section is desirable for the structural designer. But the possibility of adjusting torsional flexibility would also allow several options for active control of aerodynamic performance, load distribution, and control effectiveness. This active control by internal forces could theoretically be achieved by active materials in the skins or by an internal torque device fixed at the wing root and attached to the wing tip. Such a device, based on a shape memory alloy, has already been demonstrated on a wind-tunnel model (4). For practical applications, the parameters that define the torsional stiffness of a structure that has a closed cross section, should be kept in mind. Torsional stiffness is proportional to the square of the complete cross-sectional area and linearly proportional to the average thickness of the skin. This demonstrates how difficult it would be to modify the stiffness by changes in the skins or by an internal torque tube that has a smaller cross section.

The most often mentioned application of active structures for aircraft application is camber control and the integration of control surface functions into the main surface by camber control. This would mean a high chordwise bending deformation of the wing box. As mentioned before for spanwise bending, the skins would have to be stretched and compressed considerably, but in this case based on a smaller reference length and a smaller moment arm. For this reason, we do not see chordwise bending deformations on conventional wings under load. Aeroelastic tailoring, addressed later, by adjusting the carbon fiber plies in thickness and direction to meet desired deformation characteristics, was also addressing camber control in the 1970s and 1980s as one specific option. Because of the previously mentioned constraints, there has been no application of this passive aeroelastic control feature on a realistic wing design.

Structures and Mechanisms

To a certain extent, the main functions of structures and mechanisms are opposite: a structure must provide rigidity, and a mechanism must provide large defined motions between parts. If active structures are considered, both functions must be integrated into the structure, the performance of this system should be better, and the total weight should be lower compared to a conventional design. This shows the difficulty of developing active structures for aircraft control.

Therefore, the intention of making the structure more flexible to allow deformations is a contradiction. Hinges are required to allow deformations without producing internal forces. If this function is desired within the structure, the structure has to become more flexible in distinct small regions. But this attempt will create high internal loads in regions that have small cross sections, and the desired deformations for aircraft control functions will produce a high number of load cycles.

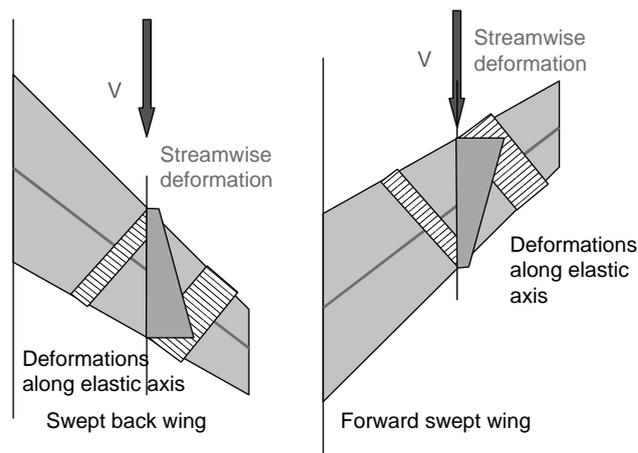


Figure 4. Bending deformation of swept wings and impacts on the aerodynamic angle-of-attack.

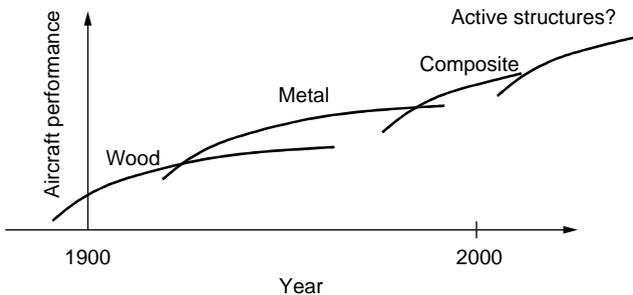


Figure 5. A/C performance improvements from materials.

Passive Materials for Aircraft Structures

Lightweight aircraft structures are obtained by optimal shape and the best suitable materials for the load levels and type of loading. Figure 5 shows the achievements produced in aircraft design by new materials. Unfortunately for active structural concepts, today's high performance composite materials are stiffer than previous aluminum structures. Therefore, it is a mistake for some active aircraft structures researchers to talk about "highly flexible" composite structures for their designs. On the other hand, today's skepticism about future applications of smart materials may be as wrong as the statement in one of the earliest textbooks on aircraft structures, where the author states that metal will never be used on aircraft structures, because its density is too high compared to wood (11). Obviously, the author was only considering iron at that time.

The figure also indicates the typical performance trends for new technologies. When they are introduced, they are inferior to the best available state-of-the-art technology at the time. The book, 'The Innovator's Dilemma,' by Clayton M. Christensen (12) describes this trend for several new, disruptive technologies. First applications are typical on low-cost, low-performance products.

Typical Load Requirements for Aircraft Structures

A typical fighter aircraft has to be designed to carry a load nine times its own weight. Applied to a car that has an empty weight of 1 metric ton plus a $\frac{1}{2}$ ton payload, this would mean an external load of 14.5 tons! The wings for a transport aircraft have to carry 2.5 times the total weight. This shows that airplane structures have to be strong, which means that they are also rather stiff. The upper and lower skins of the torque box have a typical thickness of the order of 10 to 20 millimeters, compared to the body of a car that is less than 1 millimeter thick.

SMART MATERIALS FOR ACTIVE STRUCTURES

Smart materials for active structures applications are mainly interesting because of their high energy density. On the other hand, their strain or stroke capacity is rather limited, compared with other materials for aircraft structures and with other actuators. And they are rather heavy. Probably the most complete survey paper on this topic entitled "Smart aircraft structures" by Crowe and Sater (9) was presented in 1997 at the AGARD Symposium on

Future Aerospace Technology in the Services of the Alliance. It classifies the different concepts and gives an overview on recent and ongoing research activities. It also predicts future applications in real systems that have different purposes and for different classes of airplanes.

So far, piezoelectric materials and shape memory alloys (SMA) are addressed mainly for potential applications. Whereas piezos are usually applied as patches in multiple layers, distributed over the surface, or as concentrated stack actuators, SMAs have been investigated mainly as wires or torque tubes for active deformation of aerodynamic surfaces. To date, the results achieved are not very promising for aircraft control by active structures. For this purpose, which means large static deformations in a rather short time, piezo materials respond fast enough, but their stroke is very limited. SMAs, on the other hand, could provide larger deformations and higher forces but are not fast enough for flight control inputs, and their thermal energy supply issues are rather complex within the airframe and aircraft environment.

Because of these limitations, the Defense Advanced Research Projects Agency (DARPA) launched an ambitious research program in 1999, called "Compact Hybrid Actuators" (CHAP), to multiply the stroke and force output of current actuators by a factor of 10.

THE ROLE OF AEROELASTICITY

The Reputation of Aeroelasticity

Some years after the Wright brothers' success using their active wing, designers began to fear the flexibility of the structure. The famous MIT Lester B. Gardner Lecture "History of Aeroelasticity" by Raymond L. Bisplinghoff (13) quotes many of the early incidents involving aeroelastic phenomena and the famous comment from Theodore von Kármán "Some fear flutter because they do not understand it, and some fear it because they do." Also quoted from a review paper on aeroelastic tailoring by T. A. Weisshaar (2): "As a result, aeroelasticity helped the phrase "stiffness penalty" to enter into the design engineer's language. Aeroelasticity became, in a manner of speaking, a four-letter word....it deserves substantial credit for the widespread belief that the only good structure is a rigid structure." The role of aeroelasticity in aviation is depicted in Fig. 6. It shows the impact on aircraft performance over the years,

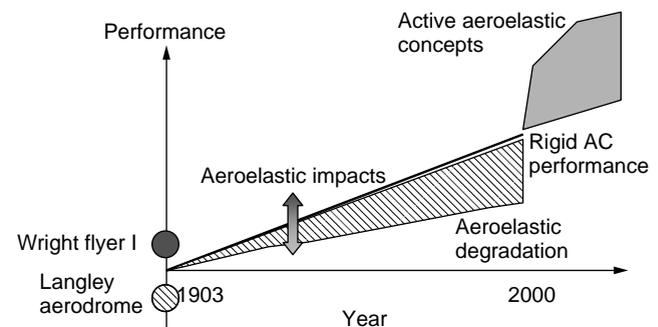


Figure 6. The impact of aeroelasticity on aircraft performance.

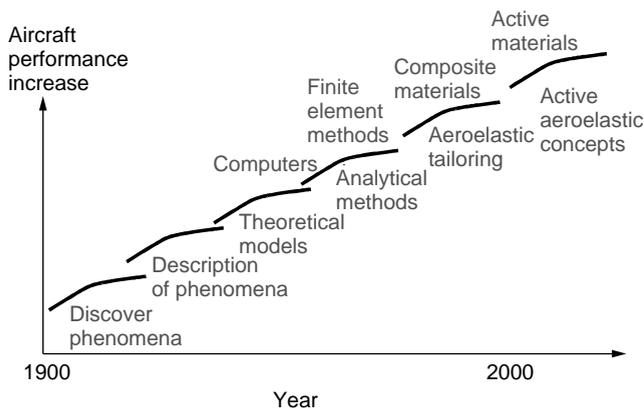


Figure 7. Relationship between aircraft performance, advances in aeroelasticity, and external stimuli.

caused mainly by increasing speed. But the upper dot in 1903 also indicates that aeroelasticity can act positively, if properly used and understood today and on faster airplanes. Smart structural concepts will help to reverse this negative trend of aeroelastic impact on aircraft performance.

Similar to Fig. 5, the progress in aeronautics can also be connected to the progress in aeroelasticity and related external stimuli and events, as shown in Fig. 7.

Aeroelastic Effects

Because of the difficulty in describing aeroelastic effects by using proper theoretical models that involve a good description of the structure, its flexibility, and structural dynamic characteristics, as well as its steady and unsteady aerodynamic properties, solutions were limited in the early years of aviation to selected cases that had only a few degrees of freedom. More general solutions required the storage space and short computing time of modern computers. The aeroelastic triangle, (Fig. 8), cited the first time by Collar (14) in 1946, describes the involved types of forces in the different aeroelastic phenomena. Looking at these forces and interactions, it becomes obvious that smart structures for aeronautical applications will have a close relationship to aeroelasticity in most cases.

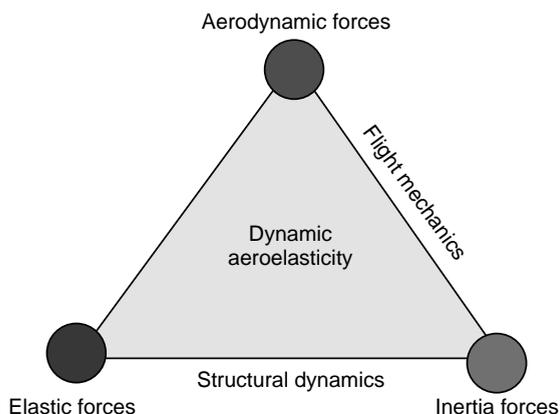


Figure 8. Aeroelastic triangle.

Static Aeroelasticity. No inertial forces are involved, by definition, in static aeroelastic effects. This is true for aileron reversal, an effect, where the rolling moment due to a control surface deflection changes sign at a certain flight speed due to opposite deformation of the fixed surface in front of the control surface. This effect has to be avoided within the flight envelope of the aircraft to avoid disturbing the pilot when he moves the stick to roll the airplane.

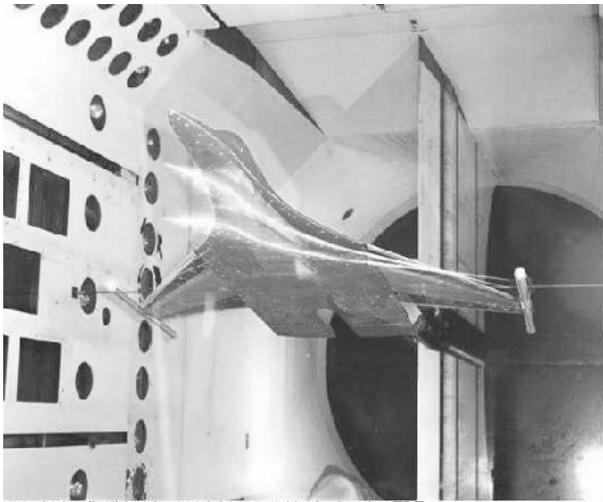
If the Wright Brothers had used conventional ailerons on their first airplanes, they might have experienced aileron reversal because of the low torsional stiffness of their wings, even at very low speeds. On the other hand, the Wright brothers' main competitor, Samuel P. Langley, was very likely less fortunate in using his Aerodrome designs because of insufficient aeroelastic stability (13) after scaling up the successful smaller unmanned vehicle to larger dimensions.

It is not sufficient to avoid aileron reversal in fighter airplanes. Even under the worst flight conditions, a high roll rate must be achieved to provide high agility. This is usually done by reinforcing the wing structure because the basic static design of a fighter wing yields rolling moment effectiveness slightly above or even below zero under the worst flight conditions. The basic design of the American F-18 had to be revised after delivery of the first batch of production aircraft. An additional weight of 200 lb per wing side was added to the Israel Lavi lightweight fighter to provide sufficient roll power. In addition to the loss of roll power, the adverse deformation of the control surface requires larger control surface deflections, which result in higher hinge moments and require stronger actuators. The difficulty of predicting the most effective distribution of additional stiffness for improved roll effectiveness, especially in conjunction with the introduction of modern composite materials that have highly anisotropic stiffness properties in airframe design, inspired the development of formal mathematical structural optimization methods (15).

Aileron reversal usually has the most severe static aeroelastic impact on aerodynamic forces and moments. But all other aerodynamic performance or control characteristics of an airplane are affected as well by static aeroelastic deformations and aerodynamic load redistributions, to a more or less severe degree. Weisshaar (16), for example, mentions the excessive trim drag due to aeroelastic wing deformations on the delta wing of a supersonic transport aircraft.

Roll control improvement by active concepts was and still is the most often studied application of active concepts for static aeroelastic phenomena in aircraft. Although active structures or materials are not involved, the Active Aeroelastic Wing project (17), or Active Flexible Wing project, as it was called before, is currently on the way to flight test trials in 2001 on a modified F-18. This concept originates in several theoretical studies and wind tunnel demonstrations in the 1980s. A summary of these activities was presented in a special edition of the *Journal of Aircraft* in 1995 (18). Figure 9 depicts the wind tunnel model installed in the Transonic Dynamics Tunnel at the NASA-Langley Research Center.

Losses of static aeroelastic effectiveness in lateral stability and rudder yawing moment are well-known design drivers for vertical tails. Surprisingly, almost nobody has



Active flexible wing model mounted in the langley TDT
 NASA langley research center 3/1/1991 Image # EL-1996-00022

Figure 9. Active aeroelastic wing model mounted in the NASA Transonic Dynamics Wind Tunnel (from the Internet).

looked so far into smart structural concepts to obtain better designs. Sensburg (19) suggested a smart passive solution, called the diverging tail, achieved by aeroelastic tailoring of the composite skins and modifying the fin root attachment to a single point aft position to achieve higher yawing moments compared to a rigid structure.

Aeroelastic divergence was the most severe instability for early monowing airplanes. If the wing main spar is located too far behind the local aerodynamic center of pressure (at 25% chord), a lack of torsional stiffness causes the wing structure to diverge and break at a certain speed. As Anthony Fokker describes in his book (20), sufficient strength of the design had already been demonstrated by proof load and flight tests for his D-VIII, (Fig. 10), when regulations called for a reinforced rear spar that has strength proportional to the front spar. This redistribution of stiffness caused torsional divergence under flight loads. This example also demonstrates the potential effects

and impacts of applying smart structures to an airplane structure.

The introduction of high-strength composite materials that had the possibility of creating bending-torsion coupling effects from anisotropic material properties caused a renaissance of the forward swept wing in the late seventies (2), this had been ruled out before for higher sweep angles because of the bending-torsion divergence, as depicted in Fig. 4.

Static aeroelasticity also includes all effects on aerodynamic load distributions, the effectiveness of active load alleviation systems by control surfaces, and flexibility effects on aerodynamic performance. In this case, the variable inertial loads from the payload or fuel on structural deformations have to be considered simultaneously.

Dynamic Aeroelasticity. Flutter is the best known dynamic aeroelastic stability problem. It belongs to the category of self-excited oscillatory systems. In this case, any small external disturbance from a control surface command or atmospheric turbulence that excites the eigenmodes of the structure, creates additional unsteady aerodynamic forces at the same time. Depending on the mass and stiffness distribution and on the phase angles between the vibrational modes involved, aerodynamic forces dampen the oscillations or enforce them in the case of flutter.

Active control for enhancing flutter stability by aerodynamic control surfaces was fashionable in the late 1970s (21). In this case, the effectiveness of the system depends on the static aeroelastic effectiveness of the activated control surfaces. Mainly because of safety aspects, but also because of limited effectiveness, none of these systems has entered service so far. Active control by active structural devices was a popular research topic in recent years (10), but, for the same reasons, it is doubtful that we will ever see applications.

Panel flutter is a special case, where only individual skin elements of the structure (panels) are affected. This usually happens at low supersonic speeds, and only



Front spar Rear spar

Figure 10. Fokker D-VIII monoplaner, where aeroelastic divergence caused several fatal accidents after reinforcement of the rear spar (modified by author, photo from the Internet).

structural elements that have low static load levels, like fairings, can usually be affected. Active control by smart materials is possible, but there are no considerable impacts on aircraft effectiveness.

Buffeting is forced vibration where turbulent flow generated by one aerodynamic surface excites this surface itself or another surface located in the turbulent flow region. Here also, aerodynamic control surfaces located on the affected part can be used to counteract the vibrations. Compared to flutter, the aerodynamic effectiveness of these surfaces is additionally reduced because of the turbulent flow conditions. Active structural systems are more effective in this case. For this reason and because the required active deformations are small, the first large-scale active structural application in aircraft dealt with the buffeting problem of fighter aircraft vertical tails under extreme maneuver conditions. After several theoretical (22) and small-scale experimental studies (23), full-scale ground tests were performed in a joint Australian–Canadian–USA research program (24) on an F-18 and in a German program for a simplified fin structure of the Eurofighter (25). In both cases, piezoelectric material was used.

Aeroelastic Tailoring and Structural Optimization

Weisshaar (26) was one of the first researchers who tried to give aeroelasticity a better reputation when modern fiber-reinforced composite materials that had highly anisotropic directional stiffness were considered for primary aircraft structures. They provided the possibility of tailoring the materials' directional stiffness within the composite lay-up to meet desired deformation characteristics for improved aeroelastic performance. Together with formal mathematical optimization methods for the structural design, this approach allowed minimizing the impact of aeroelasticity.

Any improvement of a technical system is often referred to as an optimization. In structural design, this expression is mainly used today for formal analytical and numerical methods. Some years after the introduction of finite element methods (FEM) for analyzing aircraft structures, the first attempts were made to use these tools in an automated design process. Although the structural weight is usually used as the objective function for optimization, the major advantage of these tools is the fulfillment of aeroelastic constraints, not the weight saving. Other than static strength requirements, which can be met by adjusting the dimensions of the individual finite elements, the sensitivities of the elements to aeroelastic constraints cannot be expressed so easily. The option of tailoring the composite material's properties by individual ply orientations and different layer thicknesses for the individual orientations required and inspired the development of numerical methods (27).

OVERVIEW OF SMART STRUCTURAL CONCEPTS FOR AIRCRAFT CONTROL

Classification of Concepts

Active structural concepts for aircraft control can be subdivided into these categories:

- the purpose of the active system,
- the types of devices to activate the structure.

In the first case, the intended concepts are aiming at improving

- control effectiveness,
- aerodynamic drag reduction by adaptive shape,
- load alleviation by adaptive deformation, and
- stabilizer effectiveness for trim and static stability.

As mentioned earlier, the majority of the concepts aim to improve roll control power because it usually has the highest sensitivity to structural deformations.

A classification by actuation devices can be given by

- activation of a passive structure by conventional or novel aerodynamic control surfaces,
- active structural elements, and
- actuators or connecting elements that have adaptive stiffness between structural components.

An additional classification can be made by

- concepts, where aeroelastic effects are intentionally used, and
- concepts without special aeroelastic considerations.

As far as aeroelasticity is addressed by concepts, the intended improvements aim at the high-speed part of the flight envelope, where aeroelastic effects become more important. When aeroelastic effects are exploited in a positive sense, this also means that active aeroelastic effects can usually be used beneficially only at higher speeds. An exception is shown later.

Fictitious Control Surface Concepts

To evaluate the potential benefits of smart structural concepts, as well as the required energy to activate them, it is useful to start with a “virtual concept,” assuming that the intended structural deformation is created by any device. Khot, Eastep, and Kolonay (28) call this the “fictitious control surface” concept. They investigated the aeroelastic loss of roll control power for a conventional trailing edge control surface and then tried to retwist the wing by supplying the same amount of strain energy that was created by the aileron deflection. The main purpose of this effort was analytical evaluation of the energy required to maintain a constant roll rate as dynamic pressure increases. The result, however, an increase in energy versus dynamic pressure at the same gradient as the reduction of effectiveness, may be misleading. The achievable rolling moment from a deformation depends on the position, where the deformation is initiated by an internal force or by a control surface deflection. Similarly to a rigid wing, where a trailing edge surface is much more effective than a leading edge surface, there are more or less effective regions on a flexible wing, where a deflection

of a control surface or a deformation of the structure by internal forces results in different rolling moments and requires different efforts to create the deflection or deformation.

Variable Shear Stiffness Spar Concept

Similarly to the fictitious control surface, a study by Griffin and Hopkins (29) used a “fictitious” variable stiffness spar concept to modulate the rolling moment effectiveness of a generic F-16 wing model. They assumed a small outboard trailing edge control surface on an analytical F-16 wing model for roll control, which would operate in a conventional mode at low dynamic pressures, and the negative “postreversal” effectiveness could be enhanced by turning the spar web shear stiffness off at high dynamic pressures. This concept was explained simply by “link elements” attached to the upper and lower spar caps by bolts and removable pins. The basic principle of this concept was also experimentally verified by an aeroelastic wind tunnel model for an unswept, rectangular wing that had removable spars (30). Unfortunately, no reference was found to show that more technical smart structural solutions were ever investigated for this concept.

Innovative Control Effector Program

In the Innovative Control Effector (ICE) program from NASA-Langley (31,32), the positions and required amount of small, “fictitious control surfaces” were determined by a genetic optimization process for an advanced “blended wing-body” configuration. These “control effectors” are elements of the surface grid in the analytical aerodynamic model that create the “virtual” shape change. See (31) for an excellent overview of all research activities within the NASA “Morphing Program.”

Active Flow Control Actuators

Synthetic jet actuators were also developed and tested as a part of the NASA Morphing Program (31). This device is based on a piezoelectrically driven diaphragm, which sucks and blows air through a small orifice. It was originally developed for cavity noise control. The power output needs to be multiplied to use it for aircraft control, where much higher forces are required.

Innovative Aerodynamic Control Surface Concepts

Although there are no active structural components involved, these concepts can also be considered “smart structures.” In this case, the active deformation of the structure is actuated by aerodynamic control surfaces. The January/February 1995 edition of the *Journal of Aircraft* (33) was a special issue, dedicated to the U.S. “Active Flexible Wing Program,” which started in 1985 and later turned into the “Active Aeroelastic Wing Program.” This basic idea was improving roll effectiveness for a fighter aircraft wing by combining two leading edge and two trailing edge control surfaces, which could also be operated beyond reversal speed. This concept was demonstrated on an aeroelastic wind tunnel model by tests that started in 1986. After

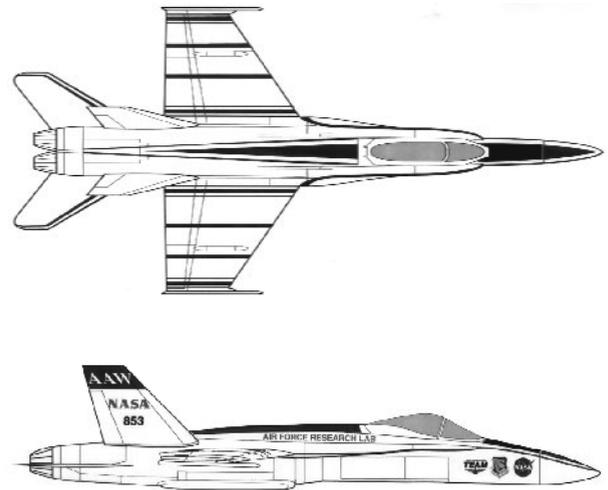


Figure 11. Active aeroelastic wing demonstrator aircraft [from (5)].

theoretical studies on F-16 and F-18 wings, reported by Pendleton (5,17,34), the F-18, depicted in Fig. 11, was selected as the candidate for flight test demonstrations, that are expected to start in 2001. For this purpose, the wing structure was returned into the original stiffness version, which had shown aileron reversal in early flight tests.

Flick and Love (35) studied wing geometry sensitivities for potential improvements from active aeroelastic concepts based on a combination of leading and trailing edge surfaces. The results shown in Fig. 12 (5) indicate only very small advantages for low aspect ratio wings. The theoretical studies of a generic wing model of the Eurofighter wing by the first author, however, also resulted in large improvements for this configuration, as can be seen in Fig. 6. Active aeroelastic concept research by TsAGI in Russia already demonstrated impressive improvements in flight tests. In addition to using leading edge control surfaces to improve roll performance, a small control surface was also mounted at the tip launcher. Figure 13 from (36) shows the achievable improvement compared to the trailing edge aileron only. Note the size of the special surface compared with a conventional aileron.

For high aspect ratio transport aircraft wings, especially in combination with a winglet, similar devices could be used for roll control and also for adaptive induced drag reduction, or load alleviation, as indicated in Fig. 14 for a concept, called active wing tip control (AWTC) by Schweiger and Sensburg (37). In this case, the winglet root provides sufficient space and structural rigidity to integrate the control device and its actuation system. In flutter stability, the forward positions of the masses increase the flutter stability, which is reduced by the aft position of the winglet.

Of course, the static aeroelastic effectiveness of a control surface is also important for dynamic applications like flutter suppression or load alleviation for buffeting of vertical tails. This fact is very often forgotten in favor of optimizing the control laws.

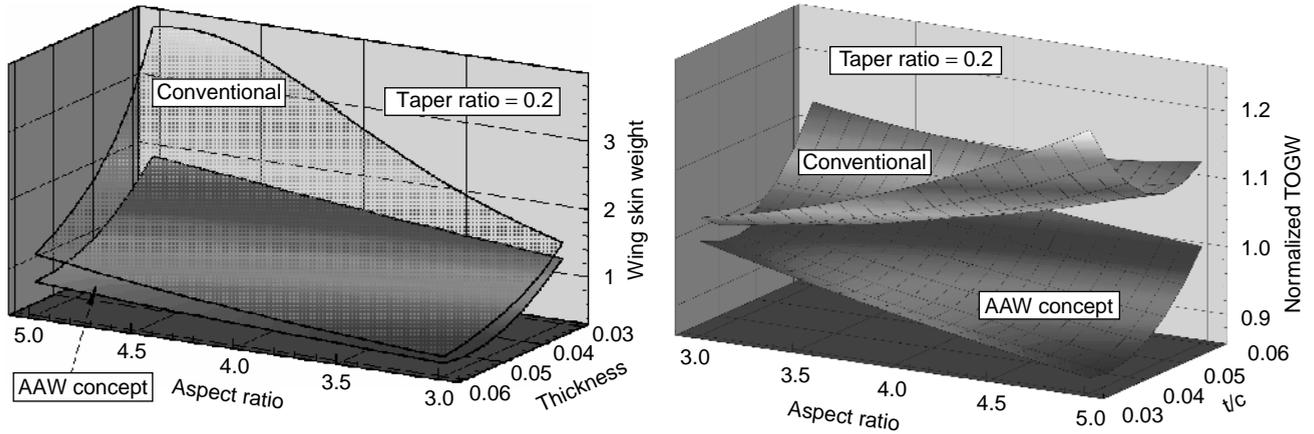


Figure 12. AAW technology advantages and wing geometry sensitivities for lightweight fighter wings [results from (35) figures from (5)]

Active Structures and Materials Concepts

Dynamic applications for flutter suppression (10) or buffeting load alleviation (38,39) by piezoelectric material were demonstrated on wind tunnel models and in full-scale ground tests. The involved mass and complexity, mainly for the electric amplifiers, precludes practical applications at the moment. For dynamic applications, however, a semi-active solution using shunted piezos (40) that have very little energy demand is an interesting option.

The use of piezoelectric materials for static deformations is limited by the small strain capacity, as well as by the stiffness of the basic structure. Because of these facts, some researchers realized rather early that it is not advisable to integrate the active material directly into the load-carrying skins. To achieve large deflections, it is necessary to amplify the active material's stroke and to uncouple the to-be-deformed (soft) part of the passive structure from the (rigid) main load-carrying part. Because this usually causes a severe "strength penalty" for the main

structures of conventional airplanes, practical applications are limited to unusual configurations like small UAVs or missiles (41). As an example, Barrett (42) developed such a device, where the external shell of a missile fin is twisted by a PZT bender element.

Compared to piezoelectric materials, which respond very fast, shape memory alloys (SMA) are rather slow, but they can produce high forces. This precludes applications for the speed of flight control motions or higher and allows only adaptation to very slow processes like the pre-described trajectory of a transport airplane and the parallel reduction of fuel mass.

Two typical applications of SMAs were investigated in the DARPA/AFRL/NASA Smart Wing Program (5,31), a SMA torque tube to twist the wing of a 16% scale wind tunnel model of a generic fighter aircraft and SMA wires to actuate the hingeless trailing edge control surfaces. The ratio between the torque tube cross section and the wing torque box cross section should be kept in mind. To replace conventional control surfaces, efforts to create deformation of a realistic structure still need to be addressed. And,

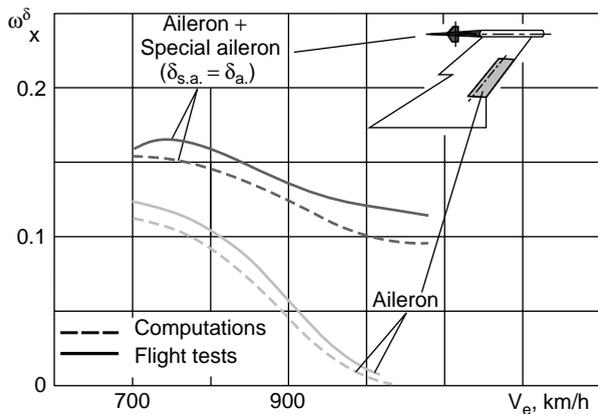


Figure 13. AAW technology in Russia: Flight test results and comparison with analysis for a special wing tip aileron [results adapted from (36)].

Active Wing Tip Control Device

- for:
- Drag reduction
- Increased roll control
- Load alleviation

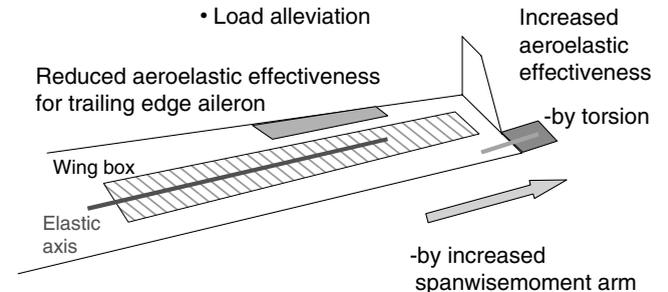


Figure 14. Advantages of an active wing tip control device on a transport aircraft wing.

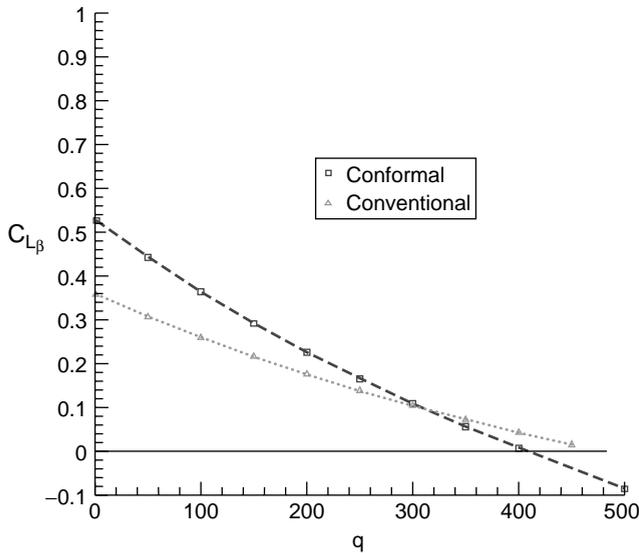


Figure 15. Comparison of rolling moment effectiveness for conventional and conformal trailing edge control surface [from (5)].

what is even more important than the limitations of actuation speed, aeroelastic aspects should be kept in mind from the beginning to evaluate and optimize the effectiveness of such concepts. As depicted in Fig. 15 (5), the effectiveness of the conformal trailing edge control surface is better than the conventional control surface a low speed but gets worse as dynamic pressure increases. As mentioned in this reference, such concepts are not developed to replace existing systems but to demonstrate the capabilities of active materials. If this is the case, realistic applications still need to be discovered.

Smart materials applications on small RPVs are currently investigated at the Smart Materials Lab of the Portuguese Air Force together with the Instituto Superior Técnico in Lisbon (43).

Other Innovative Structural Concepts

Because of the limited stroke of active materials and the inherent stiffness of a minimum weight aircraft structure, some researchers try to amplify the stroke by sophisticated

kinematic systems and enable larger deformations more easily by “artificially” reducing the structural stiffness. So far, all of these concepts show the following disadvantages:

- high complexity for the actuation system,
- higher energy demand compared with the actuation of conventional control surfaces,
- additional internal loads in the structure from the forced deformation,
- additional structural weight from the reduced strength,
- reduced static aeroelastic effectiveness because of additional flexibility in the rear wing area, and
- reduced aeroelastic stability (flutter) from the reduced stiffness.

As one example, such systems are described by Monner, et al. (44). That paper summarizes active structural research by the German aerospace research establishment DLR on an Airbus type transport aircraft wing.

An old idea, the pneumatic airplane, as depicted in Fig. 16, may be useful, if applied to small UAVs (for storage), or, on larger airplanes to selected structural elements, like spar webs, to adjust the shape by variable pneumatic stiffness to control the aeroelastic load redistribution.

Adaptive All-Movable Aerodynamic Surfaces

Adaptive rotational attachment or actuation stiffness for all-movable aerodynamic surfaces can be seen as a special class of active aeroelastic structural concepts. If properly designed, this concept will also provide superior effectiveness compared to a rigid structure at low speeds. Other active aeroelastic concepts show their advantages only as speed increases, in the same way as negative aeroelastic effects increase.

As an example, a fixed root vertical tail can be made more effective, if the structure is tailored so that the elastic axis is located behind the aerodynamic center of pressure. This wash-in effect, for example, increases the lateral stability compared to a conventional design on a swept-back vertical tail, as depicted in Fig. 17. The so-called diverging tail (19) has improved effectiveness but also experiences higher bending moments.



Figure 16. Goodyear Inflatoplane (1950s) (from the Internet).

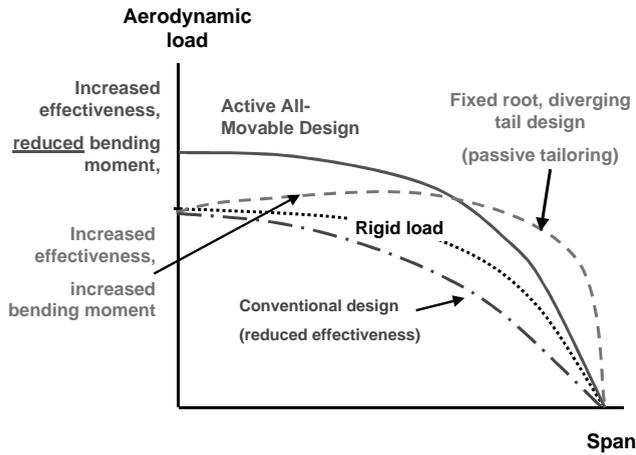


Figure 17. Aerodynamic load distribution for different design approaches on a vertical tail.

Instead of tailoring the structure, which essentially always creates a (minimized) weight increase, the tail can be designed as a reduced size all-movable surface. The location of the spigot axis is used to tailor the wash-in effect, and the attachment stiffness is adjusted to the desired effectiveness. This also allows obtaining the required effectiveness at low speeds using a smaller tail. As described in (45), the proper shape of the surface in conjunction with the spigot axis location also enhances flutter stability.

Figure 18 depicts the effectiveness of different spigot axis locations at different Mach numbers (and dynamic pressure) using variable stiffness.

The crucial element of the all-movable surface using adaptive attachment stiffness is the attachment/actuation

component. This can, for example, be a mechanical spring that has variable stiffness and a conventional hydraulic actuator or, as a more advanced system, a hybrid actuator using smart material elements, such as magnetorheological fluids. The objective of the current DARPA program “Compact Hybrid Actuators” is to develop such components at high energy density and 10 times the stroke of current systems.

Of course, such systems can also be used for horizontal stabilizers or outboard wing sections. Compared to a horizontal tailplane, where the fuselage flexibility causes losses of aeroelastic effectiveness, a forward surface can exploit additional benefits from the fuselage flexibility.

QUALITY OF THE DEFORMATIONS

The amount of internal energy required for the desired deformations depends strongly on the static aeroelastic effectiveness involved. As depicted in Fig. 19, the aerodynamic loads can either deform the structure in the wrong direction and require additional efforts to compensate for the deformations caused by external loads, or the internal, initial deformation is used so that the desired deformations are only triggered and the major amount of energy required is supplied by the air at no cost. In the first case, the required deformation generated by internal forces already creates a high level of internal strain in the structure, resulting in reinforcement and extra structural weight. In the second case, the required internal actuation forces and the strain levels are much smaller. For a favorable solution, the design process must reduce the total load level of the structure in the “design case,” thus reducing the total weight required for the structure and actuation system

Actuator position and stiffness variation for aeroelastic effectiveness vs. dynamic pressure

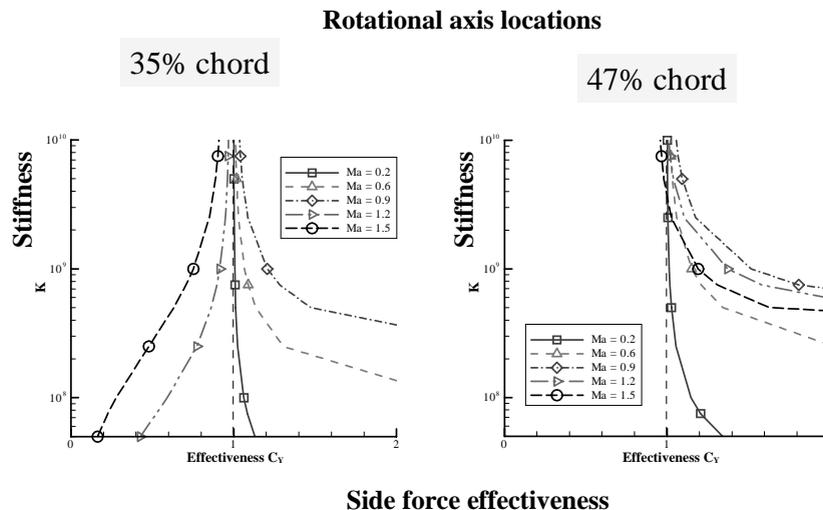


Figure 18. Achievable aeroelastic effectiveness using variable attachment stiffness for different locations of the spigot axis.

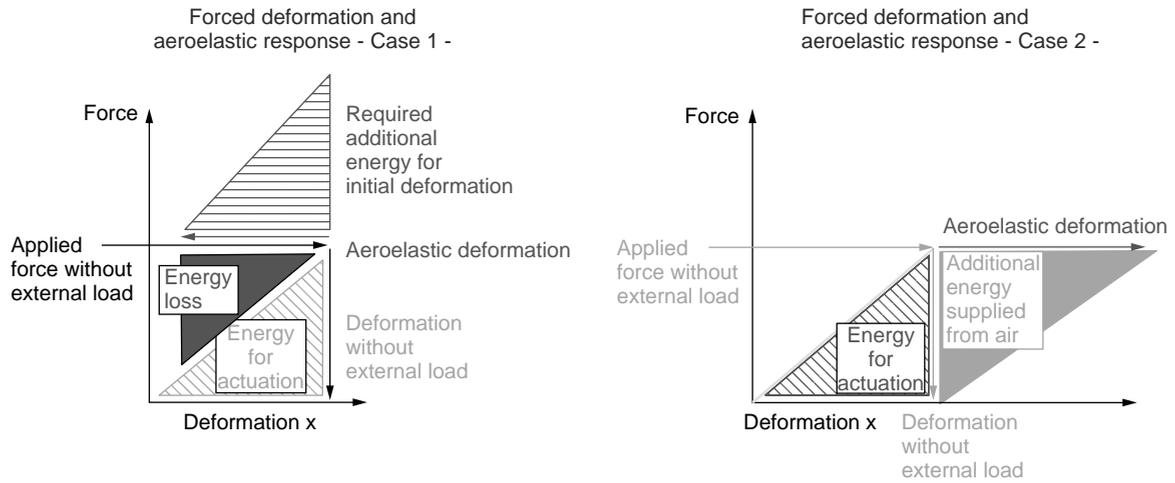


Figure 19. Forced structural deformation and aeroelastic response of different design approaches.

compared to a conventional design, as indicated in Fig. 20, and achieving better performance.

The effort required and the results achievable for a specific type of deformation depend strongly on the typical properties of the wing structure, which in most cases can be described as a beam. First of all, lift forces create bending deformations in the direction of the lift force. Because drag forces are much smaller (1/10) and because of the shape of the airfoil, in-plane bending deformations can be neglected. Depending on the chordwise location of the resulting lift force relative to the beam (torque box) shear center location, it is possible to twist the wing. This can, for example, be used to reduce the bending deformation. Because of the high stiffness of a modern wing in a chordwise section and because of the resulting aerodynamic pressure distribution that usually acts in one direction, a chordwise bending deformation (camber) is very difficult to achieve by internal or external forces. This is also true for a reduced thickness rear section of a wing, for example, to replace a deflected control surface.

ACHIEVABLE AMOUNT OF DEFORMATION AND EFFECTIVENESS OF DIFFERENT ACTIVE AEROELASTIC CONCEPTS

Classical active aeroelastic concepts rely on the adaptive use of aerodynamic control surfaces and their aeroelastic effectiveness under various flight conditions. In conventional designs, the aeroelastic effect is more pronounced as airspeed increases, as demonstrated in Fig. 21 for potential losses and gains.

Conventional active aeroelastic concepts exploit the increasing effectiveness in the upper half of this figure, as well as the recovering effectiveness of a conventional aileron beyond the reversal speed. A combined operation of leading and trailing edge surface results in an achievable roll rate, as indicated in Fig. 22.

To exploit aeroelastic effects more beneficially, increasing the aeroelastic sensitivity of the design in a wider range of the flight envelope is required. This can be achieved,

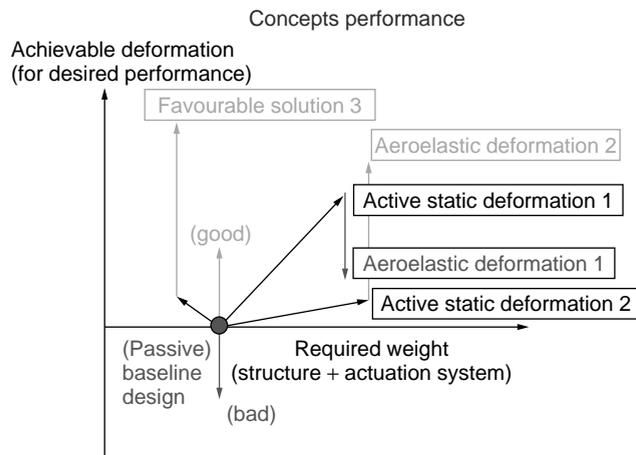


Figure 20. Performance of active structural concepts in weight and performance.

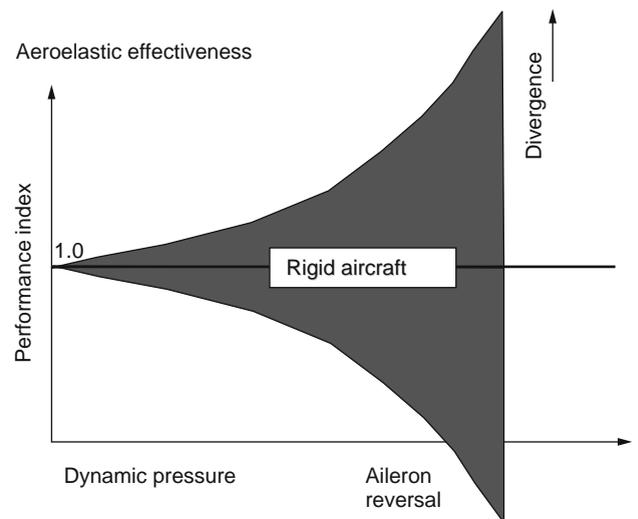


Figure 21. Typical range of aeroelastic effectiveness.

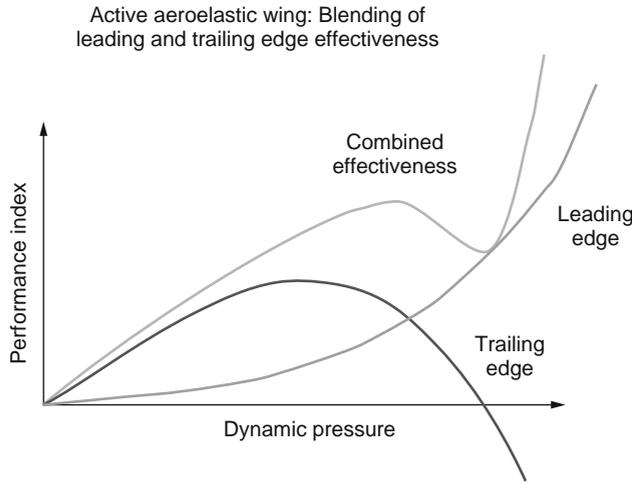


Figure 22. Achievable roll performance by combining leading and trailing edges.

for example, by an all-movable aerodynamic surface that has adaptive rotational attachment stiffness. This also provides high effectiveness at low speeds, and excessive loads from diverging components or flutter instabilities at high speeds can be avoided.

The usable aeroelastic effectiveness for conventional concepts is rather limited between take-off and cruise speed. Aileron reversal usually occurs between the cruise speed and limit speed, and too high an effectiveness of leading edge surfaces must be avoided at the limit speed. On the other hand, adaptive all-movable concepts can provide high effectiveness at all speeds and avoid excessive loads at the high end of the speed envelope, as indicated in Figs. 23 and 24. This means, for example, that a stabilizer surface can be built smaller than would be required by “rigid” aerodynamic low-speed performance.

NEED FOR ANALYZING AND OPTIMIZING THE DESIGN OF ACTIVE STRUCTURAL CONCEPTS

Of course, active materials and structural components, together with the stimulating forces, need a correct

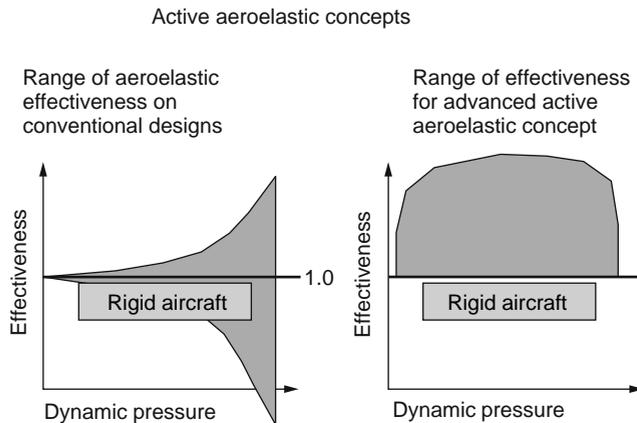


Figure 23. Aeroelastic effectiveness of conventional and adaptive-all-movable active aeroelastic concepts.

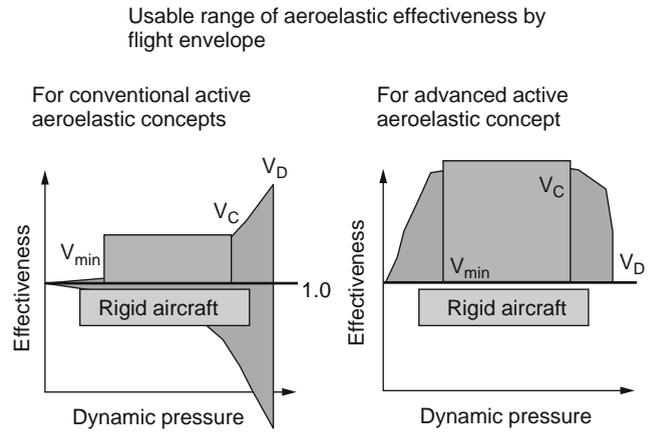


Figure 24. Usable range of aeroelastic effectiveness for conventional and advanced active aeroelastic concepts.

description in theoretical structural or multidisciplinary analysis and optimization (MDO) models and methods. Once this is provided, the actively deforming structure needs another approach for static aeroelastic analysis. The deflections of selected control surfaces of an aircraft that has conventional control surfaces can be prescribed for aeroelastic analysis. For an actively deformed structure, initial deformations without external loads first need to be determined, for example, by static analysis.

As described before, the deformations achievable in conjunction with the distribution of external aerodynamic loads are essential for the effectiveness of active structural concepts for aircraft control. This requires efficient tools and methods for simultaneous, multidisciplinary analytical design. The best design involves optimizing

- external shape,
- arranging the passive structure (topology),
- sizing the passive structure,
- placing and sizing the active elements, and
- a control concept for the active components.

The aims of this approach are the optimum result for the objective function (minimum weight, aerodynamic performance), fulfillment of all constraints like strength, and also optimization of additional objectives, such as minimum energy. As depicted in Fig. 25 for the optimization of a passive structure that has different constraints for the required rolling moment effectiveness, the energy required to actuate the control surface can be considerably reduced, even if the required (low) roll rate is already met.

MDO does not mean combining single discipline analytic tools by formal computing processes. It means first a good understanding of what is going on. This is essential for a conventional design. Only from this understanding can the creative design of an active concept start.

It is also very important to choose the proper analytic methods for individual disciplines. Usually, not the highest level of accuracy is suitable for the simulation of important effects for other disciplines. This also refers to refining the analytic models, where local details are usually not

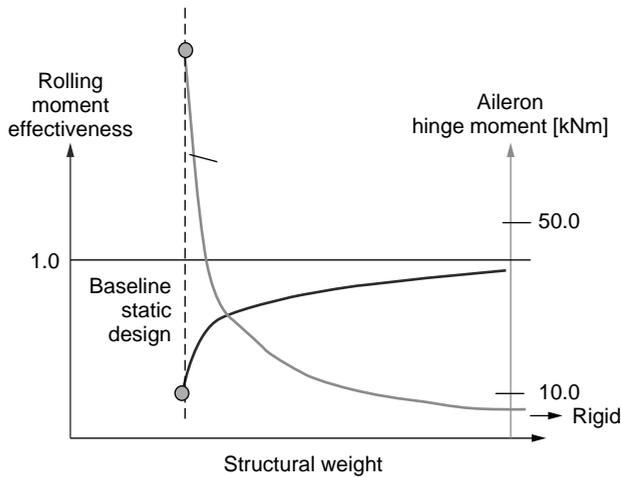


Figure 25. Optimization of the rolling moment and hinge moment of a trailing edge aileron of a low aspect ratio fighter wing.

interesting for interactions. It is more important to keep the models as versatile as possible for changes in the design concepts and to allow the simulation as many variants as possible. This also means an efficient process for generating models, including the knowledge of the user for this process. Fully automated model generators can create terrible results, if the user cannot interpret or understand the modeling process.

Any improvement in a technical system is often referred to as an optimization. In structural design today, this expression is mainly used for formal analytic and numerical methods. Some years after the introduction of finite element methods (FEM) for analyzing aircraft structures, the first attempts were made to use these tools in an automated design process. Although the structural weight is usually used as the objective function for optimization, the major advantage of these tools is the fulfillment of aeroelastic constraints, not the weight saving. Other than static strength requirements, which can be met by adjusting the dimensions of individual finite elements, the sensitivities of the elements to aeroelastic constraints cannot be expressed so easily.

In the world of aerodynamics, the design of the required twist and camber distribution for a desired lift at minimum drag is also an optimization task. Assuming that minimum drag is achieved by an elliptical lift distribution along the wingspan, this task can be solved by a closed formal solution and potential flow theory. More sophisticated numerical methods are required for the 2-D airfoil design or for Euler and Navier–Stokes CFD methods, which are now maturing for practical use in aircraft design.

Formal optimization methods have been used for conceptual aircraft design for many years. Here, quantities such as direct operating costs (DOC) can be expressed by rather simple equations, and the structural weight can be derived from empirical data. Formal methods such as optimum control theory are also available for designing the flight control system.

So, one might think that these individual optimization tasks could easily be combined into one global aircraft optimization process. The reasons that this task is not so

simple is the different natures of the design variables of individual disciplines and their cross sensitivities with other disciplines. The expression multidisciplinary optimization (MDO) summarizes all activities in this area, which have intensified in recent years. It must be admitted that today most existing tools and methods in this area are still single discipline optimization tasks that have multidisciplinary constraints.

To design and analyze active aeroelastic aircraft concepts, especially when they are based on active materials or other active structural members, new quantities are required to describe their interaction with the structure, the flight control system, and the resulting aeroelastic effects.

SUMMARY, CONCLUSIONS, AND PREDICTIONS

In the same way as it was wrong in the past to demand that an aircraft design to be as rigid as possible, it's wrong now to demand a design that is as flexible as possible.

It is sometimes said that smart structural concepts can completely replace conventional control surfaces. But this looks very unrealistic, at least at the moment. The major difficulties for successful application are the limited deformation capacity of active materials, as well as their strain allowables, which are usually below those of the passive structure. However, this can be resolved by proper design of the interface between the passive and active structures. But the essential difficulties are the stiffness and strain limitations of the passive structure itself. It cannot be expected that the material of the passive structure just needs to be replaced by more flexible materials without an excessive weight penalty. It is also not correct to believe that an active aeroelastic concept becomes more effective, if the flexibility of the structure is increased. Aeroelastic effectiveness depends on proper aeroelastic design, which needs certain rigidity of a structure to produce the desired loads. A very flexible structure would also not be desirable from the standpoints of aerodynamic shape, stability of the flight control system, and transmission of static loads.

Because large control surface deflections are required at low speeds, where aeroelastic effects on a fixed surface are small, it is more realistic to use conventional control surfaces for this part of the flight envelope and use active aeroelastic deformations only at higher speeds. This would still save weight on the control surfaces and their actuation system due to the reduced loads and actuation power requirements.

To produce usable deformations of the structure also at low speeds, all-movable aerodynamic surfaces that have a variable attachment stiffness are an interesting option. This concept relies on development efforts for active devices that have a wide range of adjustable stiffness.

The reasons that we have not seen more progress to date in successfully demonstrating smart structural concepts in aeronautics may be that

- specialists in aircraft design do not know enough about the achievements in the area of smart materials and structures, and
- smart materials and actuation system specialists, who try to find and demonstrate applications in

aeronautics, do not know or care enough about real-world conditions for airplane structures.

What we need is more awareness on both sides, as well as stronger efforts to learn from each other and work together.

Although there are strong doubts about useful applications of smart structures for aircraft control, it should always be remembered how often leading experts have been wrong in the past in their predictions, in many cases even on their own inventions. Norman R. Augustine quotes some of them in his famous book "Augustine's Laws" (46):

- "The [flying] machines will eventually be fast; they will be used in sport but they should not be thought of as commercial carriers." – Octave Chanute, aviation pioneer, 1910.
- "The energy produced by the breaking down of the atom is a very poor kind of thing. Anyone who expects a source of power from the transformation of these atoms is talking moonshine. – Ernest Rutherford, physicist, ca. 1910.
- "Fooling around with alternative currents is just a waste of time. Nobody will use it, ever. It's too dangerous ... it could kill a man as quick as a bolt of lightning. Direct current is safe." – Thomas Edison, inventor, ca. 1880.

Also quoted by Augustine (46), the eminent scientist Niels Bohr remarked: "Prediction is very difficult, especially about the future."

At the moment it looks more realistic that new hybrid, concentrated active devices, positioned between a passive but properly aeroelastically tailored main aerodynamic surface and the corresponding control surfaces are showing the like. Hopefully this article will inspire useful applications of smart structures and prevent some unnecessary research.

BIBLIOGRAPHY

1. O. Wright, *How We Invented the Airplane*. Dover, Mineola, NY, 1988.
2. T.A. Weisshaar, Aeroelastic tailoring—creative use of unusual materials. AIAA-87-0976-CP
3. W. Schwipps, *Schwerer als Luft—die Frühzeit der Flugtechnik in Deutschland*. Bernhard & Graefe Verlag, Koblenz, Germany, 1984.
4. M.I. Woods, J.F. Henderson, and G.D. Lock, *Aeronaut. J.* (2001).
5. E. Pendleton, B. Sanders, P. Flick, and O. Sensburg, *Int. Forum Aeroelasticity Struct. Dynamics*, Madrid, Spain, 2001.
6. J.J. Bertin and M.L. Smith, *Aerodynamics for Engineers*. Prentice-Hall, Englewood Cliffs, NJ, 1979.
7. O. Sensburg, H. Hönlinger, T.E. Noll, L.J. Huttsell, *J. Aircraft* **19** (5): 1981.
8. C.R. Larson, E. Falanges, S.K. Dobbs, *SPIEs 5th Annu. Int. Symp. Smart Struct. Mater.*, San Diego, CA, 1998.
9. C.R. Crowe and J.M. Sater, *AGARD Symp. Future Aerosp. Technol. Serv. Alliance*, Paris, 1997, CP-600, Vol. I.
10. A.R. McGowan, J. Heeg, and R.C. Lake, *Proc. 37th AIAA/ASME/ASCE/AHS/ASC Struct. Struct. Dynamics Mater. Conf.*, Salt Lake City, UT, 1996.
11. O.L. Skopik, *Wie berechnet und konstruiert man selbst ein Flugzeug*. Vienna, Austria, 1917.
12. C.M. Christensen, *The Innovator's Dilemma*. Harvard Business School Press, Boston, 1997.
13. H.I. Flomenhoft, *The Revolution in Structural Dynamics*. Dynaflo Press, Palm Beach Gardens, FL, 1997.
14. A.R. Collar, *J. R. Aeronaut. Soc.* 613–636, August 1946.
15. J. Schweiger, J. Simpson, F. Weiss, E. Coetzee, and Ch. Boller, *SPIEs 6th Annu. Int. Symp. Smart Structures Mater.*, Newport Beach, CA, 1999.
16. T.A. Weisshaar, RTO Meet. *Proc. 36 (AC/323(AVT)TP/17) from the RTO AVT Specialists' Meet. Struct. Aspects Flexible Aircraft Control*, Ottawa, Canada, October 1999.
17. E.W. Pendleton, D. Bessette, P.B. Field, G.D. Miller, K.E. Griffin, and *J. Aircraft* **37**(4), (2000).
18. *J. Aircraft* **32**(1), Special Section (1995).
19. O. Sensburg, G.; Schneider, V.; Tischler, and V. Venkayya, *Specialists' Meet. Struct. Aspects Flexible Aircraft Control. RTA Meet. Design Issues*, Ottawa, Canada; October 1999.
20. A.H.G. Fokker and B. Gould, *Flying Dutchman. The Life of Anthony Fokker*. Henry Holt, NY, year unknown.
21. T. Noll, H. Hönlinger, O. Sensburg, and K. Schmidt, Active Flutter Suppression Design and Test, a Joint U:S-F.R.G. Program. ICAS- CP 80-5.5. Munich, Germany, 1980.
22. K.B. Lazarus, E. Saarmaa, and G.S. Agnes, Active Smart Material System for Buffet Load Alleviation. Bellingham, WA, 1995, *SPE* Vol. 22447/179.
23. R.M. Hauch, J.H. Jacobs, K. Ravindra, and C. Dima, Reduction of Vertical Tail Buffet Response Using Active Control. AIAA-CP-95-1080, Washington, 1995.
24. R.W. Moses, *40th AIAA-SDM Con.*, St. Louis, 1999.
25. J. Simpson and J. Schweiger, *SPIE 5th Annu. Int. Symp. Smart Struct. and Mater.*, San Diego, CA, 1998.
26. T.A. Weisshaar, *J. Aircraft* **17**(6), (1980).
27. J. Schweiger, J. Krammer, and H. Hörnlein, *6th AIAA/NASA/ISSMO Symp. Multidisciplinary Anal. Optimization*, Seattle, WA, 1996, CP-4169.
28. N. Khot, F. Eastep, and R. Kolonay, A Method for Enhancement of the Rolling Maneuver of a Flexible Wing. AIAA-CP-96-1361.
29. K. Griffin and M. Hopkins, *36th AIAA Struct. Struct. Dynamics Mater. Conf.*, New Orleans, 1995.
30. C.L. Giese, G.W. Reich, M.A. Hopkins, and K.E. Griffin, An Investigation of the Aeroelastic Tailoring for Smart Structures Concept. AIAA-96-1575-CP, 1996.
31. A.R. McGowan, L.G. Horta, J.S. Harrison, and D.L. Raney, *RTO Meet. Proc. 36 (AC/323(AVT)TP/17) RTO AVT Specialists' Meet. Struct. Aspects Flexible Aircraft Control*, Ottawa, Canada, October 1999.
32. S.L. Padula, J.L. Rogers, and D.L. Raney, Multidisciplinary Techniques and Novel Aircraft Control Systems. *8th AIAA/NASA/ISSMO Symp. Multidisciplinary Anal. Optimization*, Long Beach, CA, 2000, AIAA-CP-2000-4848.
33. *J. Aircraft* **32**(1), Special Section (1995).
34. E. Pendleton, K.E. Griffin, M.W. Kehoe, and B. Perry, *Conf. Proc. AIAA-96-1574-CP*.
35. P.M. Flick and M.H. Love, *Specialists' Meet. Struct. Aspects Flexible Aircraft Control, RTA Meet. Design Issues*, Ottawa, Canada, October 1999.

36. S.I. Kuzmina, G.A. Amiryants, F.Z. Ishmuratov, V.A. Mosunov, and V.V. Chedrik, *Int. Forum Aeroelasticity Struct. Dynamics*, Madrid, Spain, 2001.
37. J. Schweiger and O. Sensburg, *Int. Forum Aeroelasticity Struct. Dynamics*, Madrid, Spain, 2001.
38. R.W. Moses, CP, *40th AIAA-SDM Conf.*, St. Louis, 1999.
39. J. Simpson and J. Schweiger, Industrial Approach to Piezoelectric Damping of Large Fighter Aircraft Components. *SPIE 6th Annu. Int. Symp. Smart Struct. Mater.*, San Diego, CA, 1998.
40. A.R. McGowan, *SPIE 5th Annu. Int. Symp. Smart Struct. Mater.*, Newport Beach, CA, 1999.
41. T.A. Weisshaar, *41st Annu. Isr. Conf. Aerosp. Sci.*, Tel Aviv and Haifa, February 2001.
42. R. Barrett, in *Smart Materials and Structures* No. 5, IOP, UK, 1996.
43. P. Costa, P.A. Moniz, and A. Suleman, *42nd AIAA SDM Conf.*, Seattle, WA, 2001, AIAA-2001-1361.
44. H.P. Monner, E. Breitbach, Th. Bein, and H. Hanselka, *Aeronaut. J.* **104**(1032), (2000).
45. J. Schweiger, F. Weiss, T. Kullrich, *8th AIAA / USAF / NASA / ISSMO Symp. Multidisciplinary Anal. Optimization*, Long Beach, CA, September 2000.
46. N.R. Augustine, *Augustine's Laws*. AIAA, Reston, VA, 1997.

ARCHITECTURE

D. MICHELLE ADDINGTON
 DANIEL S. SCHODEK
 Harvard University
 Cambridge, MA

INTRODUCTION

An inextricable link has existed historically between a building's characteristics—form, appearance, and function—and the characteristics of the different materials that were available and suitable for construction. As exemplified by historical building traditions in stone and wood, early architects sought to understand intuitively the intrinsic physical behavior of commonly available materials to exploit their properties in designing and constructing buildings. Conversely, later innovations in the type and availability of materials strongly impacted the development of new architectural forms as architects began to respond to changing societal demands and new building functions emerged. This trend is illustrated by the development of steel in the nineteenth century and the related emergence of long-span and high-rise building forms. Today, architects are beginning to look forward to using the developments in smart materials to bring new solutions to long-standing problems and also to exploit the potential of smart materials in developing new building functions, forms, and responses. The wide variety of smart materials available has great potential for use within the field, but, in this area, their applications remain only marginally explored.

MATERIAL CONSIDERATIONS IN ARCHITECTURE

Unlike materials used for specific applications or products such as in refractory linings or engine blocks that are fundamentally chosen on the basis of performance criteria and cost, the choice of materials for architectural use has always been based on very different types of criteria. Performance and cost obviously play a role, but the final selection is often based on appearance and aesthetics, ease of constructability in terms of labor skill, local or regional availability, as well as the material used in nearby existing buildings. The multimodal nature of the selection process coupled with the wide-ranging array of building types, uses, and locales has resulted in a material palette that encompasses all of the major material classes.

TRADITIONAL MATERIAL CLASSIFICATIONS IN ARCHITECTURE

The Construction Specification Institute (or CSI) devised a classification system in 1948 that is used throughout the architectural design and building construction industries. The classification system is bipartite: the first half is devoted to the broad classes of materials typically used in buildings, including paint, laminate, and concrete, and the second half categorizes standard building components such as doors, windows, and insulation. The emphasis in both major groupings is on application, not on fundamental behavior or properties. For example, in Division 6 the characteristics of wood are discussed in relationship to their relevance to the intended application: the grade of wood suitable for load-bearing roof structures or the type of wood suitable for finish flooring.

The CSI index serves as a template for communication among architects, contractors, fabricators, and suppliers. After the preliminary design of a building is completed and approved, architects prepare construction documents (known as CDs) that will serve as the “instructions” for constructing the building. Accompanying each set of CDs are the “Construction Specifications”: a textual document that defines each building element documented in the CDs and specifies the material or component. The Construction Specifications serve as a binding contract that construction professionals and contractors must follow. Trade associations and manufacturers of building products routinely write their material and product specifications in CSI format to streamline the specification process for architects, and many architectural firms maintain an internal set of Construction Specifications that is used as the baseline for all of their projects.

TRADITIONAL TECHNOLOGY CLASSIFICATION IN ARCHITECTURE

The CSI index also categorizes the technologies used in architectural design and construction. Unlike the standard technology classifications used in engineering sciences that categorize according to process and product, the CSI specifications categorize by system. As in the CSI

material classes, the focus of the technology classes is also on application. The technologies are divided into two major groups: the first is devoted to building operational systems such as HVAC, lighting, and plumbing systems, and the second is devoted to building construction systems such as structural, drainage, and vertical circulation systems. The specifications for the building operational systems are almost entirely supplied by manufacturers.

PROPOSED CLASSIFICATION SYSTEM FOR SMART MATERIALS

The introduction of smart materials into architecture poses a challenge to the normative classification system. A smart material may be considered as a replacement for a conventional material in many components and applications, but most smart materials have inherent “active” behaviors, and, as such, are also potentially applicable as technologies. For example, electrochromic glass can be simultaneously a glazing material, a window, a curtain wall system, a lighting control system, or an automated shading system. The product would then fall into many separate categories, rendering it particularly difficult for the architect to take into consideration the multimodal character and performance of the material. Furthermore, many smart materials are introducing unprecedented technologies into the field of design, and are also making more commonplace many technologies, such as sensors, which previously had only limited application in highly specialized functions. Table 1 describes a proposed organization in which smart materials establish a sequential relationship between materials and technologies. The proposed organization also maintains the fundamental focus on application of the traditional classification system.

TAXONOMY OF SMART MATERIALS

Four fundamental characteristics are particularly relevant in distinguishing a smart material from the traditional materials used in architecture: (1) capability of property change (2) capability for energy exchange, (3) discrete size/location, and (4) reversibility. These characteristics can potentially be exploited either to optimize a material property to match transient input conditions better or to optimize certain behaviors to maintain steady-state conditions in the environment.

Smart Material Characteristics

Property Change. The class of smart materials that has the greatest volume of potential applications in architecture is the property-changing class. These materials undergo a change in a property or properties—chemical, thermal, mechanical, magnetic, optical, or electrical—in response to a change in the conditions of the material’s environment. The conditions of the environment may be ambient or may be produced via a direct energy input. Included in this class are all color-changing materials, such as thermochromics, electrochromics, and photochromics, in which the intrinsic surface property of the molecular spectral absorptivity of visible electromagnetic radiation is modified by an environmental change (incident solar radiation, surface temperature) or an energy input to the material (current, voltage).

Energy Exchange. The next class of materials predicted to have a large penetration into architecture is the energy-exchanging class. These materials, which can also be called “first law” materials, change an input energy into another form to produce an output energy in accordance with the first law of thermodynamics. Although the energy

Table 1. Proposed Classification System for Smart Materials and Systems

Category	Fundamental Material Characteristics	Fundamental System Behaviors
Traditional materials: Natural materials (stone, wood) Fabricated materials (steel, aluminum, concrete)	Materials have given properties and are “acted upon”	Materials have no or limited intrinsic active response capability but can have good performance properties
High performance materials: Polymers, composites	Material properties are designed for specific purposes	
Smart materials: Property-changing and energy-exchanging materials	Properties are designed to respond intelligently to varying external conditions or stimuli	Smart materials have active responses to external stimuli and can serve as sensors and actuators
Intelligent components: Smart assemblies, polyvalent walls	Behaviors are designed to respond intelligently to varying external conditions or stimuli in discrete locations	Complex behaviors can be designed to respond intelligently and directly to multimodal demands
Intelligent environments	Environments have designed interactive behaviors and intelligent response—materials and systems “act upon” the environment	Intelligent environments consist of complex assemblies that often combine traditional materials with smart materials and components whose interactive characteristics are enabled via a computational domain

converting efficiency of smart materials such as photovoltaics and thermoelectrics is typically much less than those of conventional energy conversion technologies, the potential utility of the energy is much greater. For example, the direct relationship between input energy and output energy renders many of the energy-exchanging smart materials, including piezoelectrics, pyroelectrics and photovoltaics, excellent environmental sensors. The form of the output energy can further add direct actuating capabilities such as those currently demonstrated by electrostrictives, chemoluminescents and conducting polymers.

Reversibility/Directionality. Some of the materials in the two previous classes also exhibit the characteristic of either reversibility or bidirectionality. Many of the electricity converting materials can reverse their input and output energy forms. For example, some piezoelectric materials can produce a current from an applied strain or can deform from an applied current. Materials that have a bidirectional property change or energy-exchange behavior can often allow further exploitation of their transient change rather than only of the input and output energies and/or properties. The energy absorption characteristics of phase changing materials can be used either to stabilize an environment or to release energy to the environment, depending on the direction in which the phase change is taking place. The bidirectional nature of shape-memory alloys can be exploited to produce multiple or switchable outputs, allowing the material to replace components composed of many parts.

Size/Location. Regardless of the class of smart material, one of the most fundamental characteristics that differentiates smart materials from traditional materials is the discrete size and direct action of the material. The elimination or reduction in secondary transduction networks, additional components, and, in some cases, even packaging and power connections allows minimizing the size of the active part of the material. A component or element composed of a smart material can be much smaller than a similar construction using traditional materials and also will require less infrastructural support. The resulting component can then be deployed in the most efficacious location. The smaller size coupled with the directness of the property change or energy exchange renders these materials particularly effective as sensors: they are less likely to interfere with the environment that they are measuring, and they are less likely to require calibration.

Relevant Properties and Behaviors

Architectural materials are generally deployed in very large quantities, and building systems tend to be highly integrated into the building to maintain homogeneous interior conditions. Materials and systems must also withstand very large ranges of transient exterior conditions. The combination of these two general requirements tends to result in buildings of high thermal and mechanical inertia. Therefore, even though the typical building uses

several different materials for many functions, there are only a few areas in which the characteristics of smart materials can be useful. The transient environmental conditions experienced by most buildings often results in oversizing systems to accommodate the full range of the exterior environmental swing. The swings may be instantaneous, as in the case of wind, diurnal, or seasonal. These conditions include those that affect both heat transfer and daylight transmission through the building envelope (also known as the building façade or exterior skin) as well as those that create dynamic loading on the building's structural support system. For the building envelope, the property-changing class of smart materials has the most potential application, whereas the energy-exchanging class is already finding application in building structural systems.

Buildings consume two-thirds of the electrical energy generated in the United States, and the majority of that electrical energy is used to support the building's ambient environmental systems, primarily lighting and HVAC (heating, ventilating, and air conditioning) systems. The intent of these systems is to effect a desired state in the interior. That state may be defined by a specified illuminance level or by an optimum temperature and relative humidity. Because conditions are generally maintained at a steady state, the primary need is for more efficacious control. Energy-exchanging materials have potential application as discrete sources, particularly for lighting delivery systems, and also as secondary energy supply sources. The most significant applications of smart materials in buildings, however, has been and will continue to be as sensors and actuators for the control systems of these ambient environmental systems.

Smart Material Mapping

The material properties and/or characteristics that are most relevant to architectural requirements are mapped in Table 2 against examples of smart material applications.

CATEGORIES OF APPLICATIONS

One of the major difficulties in incorporating smart materials into architectural design is the recognition that very few materials and systems are under single environmental influences. For example, the use of a smart material to control conductive heat transfer through the building envelope may adversely impact daylight transmission. Furthermore, because most systems in a building are highly integrated, it is difficult to optimize performance without impacting the other systems or disrupting control system balancing. As an example, many ambient lighting systems include plenum returns through the luminaires (lighting fixtures) that make it particularly difficult to decouple HVAC from lighting systems. The following discussion establishes four major categories of applications for smart materials and takes into account the material/behavior mapping described in Table 2 but also considers the complex systems that are affected. The four categories—glazing materials, lighting systems, energy systems, and

Table 2. Mapping of Smart Materials to Architectural Needs

Architectural Need	Relevant Material Characteristic	Smart Material Application
Control of solar radiation transmitting through the building envelope	Spectral absorptivity/transmission of envelope material	Electrochromics Photochromics Liquid crystal displays Suspended particle panels
	Relative position of envelope material	Louver control systems <ul style="list-style-type: none"> • exterior radiation sensors (photovoltaics) • interior daylight sensors (photoelectrics) • controls (shape-memory alloys)
Control of conductive heat transfer through the building envelope	Thermal conductivity of envelope material	Thermotropics Phase change materials
Control of interior heat generation	Heat capacity of interior material	Phase change materials
	Relative location of heat source	Fiber-optic systems Thermoelectrics Photoluminescents Light-emitting diodes
Secondary energy supply systems	Lumen/watt energy conversion ratio	Photovoltaics
Optimization of lighting systems	Conversion of ambient energy to electrical energy	
	Daylight sensing Illuminance measurements Occupancy sensing Relative location of source	Photovoltaics Photoelectrics Fiber optics Electroluminescents
Optimization of HVAC systems	Temperature sensing Humidity sensing Occupancy sensing CO ₂ and chemical detection Relative location of source and/or sink	Pyroelectrics Hygrometers Photoelectrics Biosensors Thermoelectrics Phase change materials
	Euler buckling Inertial damping	Piezoelectric Magnetorheological Electrorheological Shape-memory alloys Fiber optics
Control of structural vibration	Strain sensing	

monitoring/control systems—are also intended to be consistent with the more normative and identifiable classification systems of architecture.

Glazing Materials

Whether serving as windows or as glass curtain walls, glazing materials are extensively used on the building envelope. Originally incorporated and developed during the twentieth century for aesthetic reasons, the current use of glazing materials also considers the delivery of daylight into the building's interior. The majority of developments in high-performance glazing materials have focused on thermal characteristics—spectral selectivity to reduce radiant transmission to the interior or low emissivity to reduce radiant loss to the exterior. Glazing introduces the problematic condition in which, depending on the exterior environmental conditions, performance criteria that have been optimized for one set of conditions may be undesirable in a matter of hours or even moments later. The ideal glazing material would be switchable—managing the radiant transmission between exterior to interior to transmit solar radiation when the envelope is conducting heat out

(typical winter daytime condition) and reflect solar radiation when the envelope is conducting heat into the building (typical summer daytime condition). Photochromics, thermochromics and thermotropics have been proposed as switchable glazing materials, although only thermotropics are currently being developed commercially for this application. The basic operation of these materials is that either high incident solar radiation (photochromic) or high exterior temperature (thermochromic or thermotropic) produces a property change in the material that increases its opacity, thereby reducing radiant transmission to the interior. When incident solar radiation lessens or when the exterior temperature drops, the material reverts to a more transparent quality, allowing more solar radiation to transmit to the interior.

There are numerous circumstances, however, for which this type of switching is neither desirable nor useful. Direct solar radiation into the building can create overheated zones in particular locations, even in the dead of winter. Winter sun altitude is also much lower, thereby significantly increasing the potential for glare if solar radiation is not controlled. During the summer, reducing the radiant transmission may increase the need for

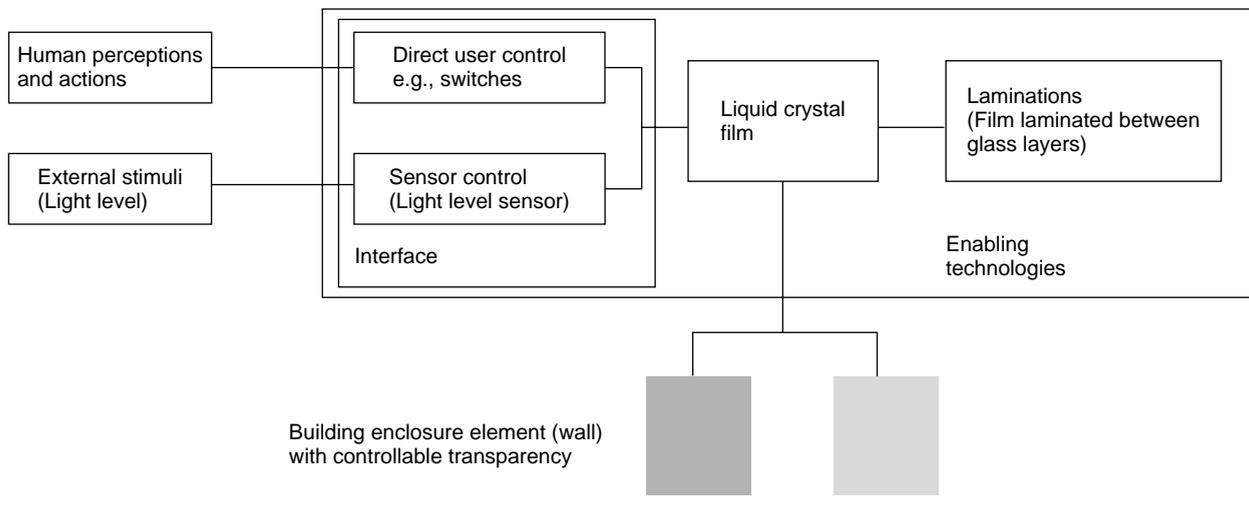


Figure 1. Typical current use of a smart material in architecture. Only a single behavior is controlled.

interior lighting systems and, because all electrically generated light has a lower lumen/watt ratio than daylight, might exacerbate the building's internal heat gains. As a result, the majority of efforts to develop smart glazing have focused on the electrically activated chromogenics—electrochromics, liquid crystal panels, and suspended particle panels (see Fig. 1). By using an electrical input to control transparency, these materials can be more easily incorporated into the control schemes for energy management systems and/or lighting control systems. The optimum balance among lighting needs, heating/cooling requirements, and occupant comfort can be determined, and the transparency can be adjusted to meet these demands in highly transient conditions.

Lighting Systems

Most high efficiency lighting systems—fluorescent, HID (high intensity discharge)—are relatively unsuitable for low-level lighting or task lighting. Furthermore, the typical ambient lighting system requires enormous infrastructure for support: electronic control systems, ballasts, integrated cooling, light diffusers/distributors (often part of the luminaire or lighting fixture). The efficiency and economics of these systems drop as the overall lighting requirements become smaller or more discrete. Ambient systems are also difficult to dim and to focus, so that very low-efficiency incandescent/halogen systems are still widely used for task or discrete lighting requirements. The low efficiency of the typical lighting system results in producing a substantial amount of heat and can be responsible for as much as 30% of a commercial building's cooling load. The development of fiber-optic lighting systems allows decoupling the delivered light from the primary energy conversion processes for generating light. This has the dual advantage of allowing light delivery to any location in a building, which is much more efficacious than using ambient lighting systems to deliver light, as well as removing the heat source from the occupied space. Current applications for fiber-optic systems

include many museums and retail display areas, where the removal of the heat source can profoundly improve the environmental conditions of the objects under display and the discrete nature of the light allows better highlighting and focusing.

Ambient lighting systems are generally designed to provide a standard illuminance level throughout a space at a specified height (usually three feet above the floor). The human eye, however, responds to the relative luminance contrast between surfaces in the field of vision. A lighting level of 100 footcandles may be too low for reading if the surrounding surfaces provide little contrast and may be too high if the surfaces provide high contrast. The division of light into smaller and more discrete sources allows optimizing contrast within the field of vision. Furthermore, the design of lighting for managing contrast enables using lower levels of lighting. Sources produced by the various luminescents—chemo, photo, electro—are starting to find application in architectural interiors, particularly as emergency lighting systems, because they have low and in some cases no input power requirements. LED (light-emitting diode) systems are also being developed as low energy lighting delivery systems. The latest developments in polymer LED technology have produced lighting fixtures that have precise color control. They provide excellent color rendition and also allow for color variation—features that are difficult to achieve in standard lighting systems.

Energy Systems

The majority of buildings in the United States are connected to a utility grid and as such have little need for primary energy conversion on-site. There are numerous circumstances, however, where secondary energy conversion can be quite useful, including back-up power generation, peak demand control, and discrete power for remote needs. For these situations, photovoltaic energy systems are increasingly becoming popular because they can be

readily deployed on roofs or integrated directly into the building envelope to take advantage of the incident solar radiation. Two other developments in smart materials hold greater promise for managing energy needs within a building. The large interior heat loads of most buildings coupled with a diurnal exterior temperature swing has encouraged investigation into thermal mass systems for maximum exploitation of a building's thermal inertia. Although theoretically sound, thermal mass systems have three major problems: (1) very slow response time, (2) the inability to switch off the phenomenon when it is not desirable, and (3) the large embodied energy required to provide the necessary mass of material. Phase change materials offer the advantages of thermal mass and very few of its disadvantages. The materials can be tuned to particular temperatures and can have very rapid responses. Much less mass is required, and therefore, the materials can be packaged and distributed throughout the building much more efficiently and strategically. By layering phase change materials and other smart materials, such as electrochromics or thermotropics, there may be a potential to add switching capability that allows activating or deactivating of the inertial behavior of the materials.

The removal of heat generated in a building is becoming an increasing concern as point loads from lighting, computers, and other electrical equipment escalate. Ambient HVAC systems do not distinguish between human-generated and equipment-generated cooling needs. The ability to manage and remove the heat generated by a point load without affecting the ambient environmental system could improve the operation of the ambient system and significantly reduce the energy requirements. Thermoelectrics are currently being explored for their potential to manage point loads discretely. Already serving as heat sinks in the majority of microprocessor cooling packages, thermoelectrics could be incorporated into integrated cooling for many other types of point sources. Although the devices are not practical for cooling air directly because of their low coefficient of performance (COP), they are ideal for managing the conjugate heat transfer that is characteristic of most nonhuman heat sources encountered in a building.

Monitoring and Control Systems

The increasing push to reduce the energy used by building HVAC systems has led to tighter buildings to reduce infiltration and to larger resets for the control equipment. This combination of an impermeable building envelope and more variable interior conditions has led to an increase in occupant complaints and indoor air quality problems. Many of the strategies intended to reduce energy can impact human health adversely, and much discussion of the appropriate compromise between the two requirements continues. One solution that holds promise is DCV, or "demand controlled ventilation." DCV adjusts interior ventilation depending on the presence of occupants; it reduces ventilation when no occupants are in a room or zone and increases ventilation as more occupants enter. Because the

human need for fresh air is linked to activity, simple occupancy sensors are not enough. The level of carbon dioxide in a room has been proposed as a good surrogate for the amount of fresh air needed in a space, but many concerns have arisen in regard to other chemical contamination, such as finish material outgassing, that is not connected to occupancy. Chemical sensing for building monitoring has previously been too expensive to incorporate and too slow to be useful. New developments in smart sensors for environmental monitoring, particularly biosensors, hold great promise for optimizing the controls of ambient HVAC systems.

The need to control various kinds of motions and, in particular, vibrations in a structure appears in many forms. At the level of the whole building structure, excitations resulting from seismic or wind forces can result in damage to both primary structural systems and nonstructural elements. User discomfort can also result. Many pieces of delicate equipment in buildings also need to be protected from external vibrations by using similar strategies. Alternatively, many pieces of equipment used in buildings can produce unwanted vibrations that can propagate through buildings. In response to these needs, methods of mitigating structural damage have been proposed that seek to control overall structural responses via controllable smart damping mechanisms used throughout a structure. Several smart base isolation systems for mitigating structural damage in buildings exposed to seismic excitations have also been proposed. These dampers are based on various electro- or magnetorheological fluids or piezoelectric phenomena. Piezoelectric sensors and actuators, for example, have been tested for use in vibrational control of steel frame structures for semiconductor manufacturing facilities.

Active control can be used to modify the behavior of specific structural elements by stiffening or strengthening them. Structures can adaptively modify their stiffness properties, so that they are either stiff or flexible as needed. In one project, microstrain sensors coupled with piezoceramic actuators were used to control linear buckling, thereby increasing the buckling load of the column severalfold.

Several new technologies provide capabilities for damage detection in structures. Various kinds of optical-fiber sensors have been developed for monitoring damage in materials as diverse as concrete and fiber-reinforced plastic composite laminate structures. Optical fibers are usually embedded in the material. Strain levels can be measured via wavelength shifts and other techniques. Crack development in structures made of concrete, for example, has been monitored via optical-fiber sensors, and special distributed systems have been developed for use in the structural health monitoring of high-performance yachts. Distributed fiber-optic systems have also been proposed for leak detection in site applications involving infrastructure systems. Other site-related structural applications include using optical-fiber sensors for ground strain measurement in seismically active areas. Other applications where smart materials serve as sensors include the use of embedded temperature sensors in carbon-fiber structures.

FUTURE DESIGN APPROACHES IN ARCHITECTURE

The previous sections have outlined and discussed smart materials in conjunction with needs currently defined in architecture. In some cases, smart materials have been proposed as replacements for conventional materials, and in other cases, smart materials have been proposed for improving the functionality of standard building systems. All of these developments can be positioned into the third category titled “Smart materials” in Table 1. The impact of incorporating these materials into standard architectural practice will be significant, particularly in regard to energy use and building performance, but far more interesting potentials derive from reconsidering smart materials as fundamental conceptual elements in design rather than only as improvements to existing elements.

As architectural design has always traditionally involved integrated systems and materials—the building envelope construction depends on the building’s structural system, the building’s HVAC system depends on the envelope construction—then the greatest potential may come from using smart materials to *dis-integrate* certain components, behaviors, or environments within the building. A smart component would be one that functions intelligently without infrastructural support and also will not disrupt the performance of surrounding systems. An example of a smart component might be a luminaire that can sense relative luminances within a visual field and self-adjust its focus, dimming capacity, and position.

Several different smart materials would be involved in the development of this component, including sensors and actuators, electroluminescents or LEDS, and perhaps even shape-memory alloys. A smart assembly would operate at the next level of functionality beyond the smart component. There are many “high-tech” assemblies currently used in architecture. These assemblies integrate several types of components and technologies to achieve multiple functions. For example, many of the most advanced envelope systems incorporate mechanical shading systems, thermal and ventilation control systems, and multiple layers of glass into a highly integrated assembly intended to preserve view without incurring energy penalties. A smart assembly would be designed to manifest the same behaviors, but do so in the most strategic manner (see Fig. 2). Shading could be accomplished at the micron or molecular level by using smart materials, and thermal control could take place discretely and transiently by selective placement of phase change materials and thermoelectrics. The smart assembly would maximize functionality and minimize the number of components.

Many development activities have been focused on proposals for “smart rooms” (see Fig. 3). Most of these proposals accept the building as a traditional structure and seek to insert certain technologies into a room to add increased functionality. Ubiquitous computing, teleconferencing, smart boards, voice and gesture recognition systems, and wireless communication systems are among the many smart technologies being developed for incorporation

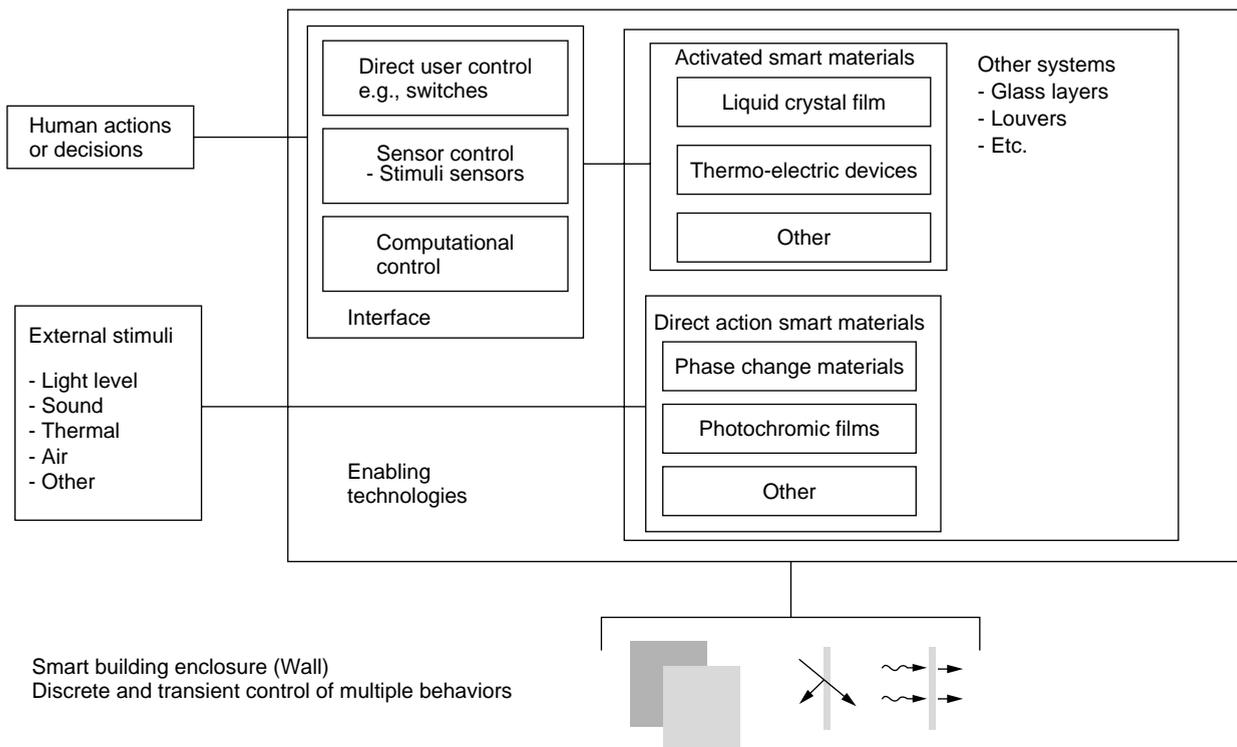


Figure 2. Control of multiple behaviors via smart building assemblies.

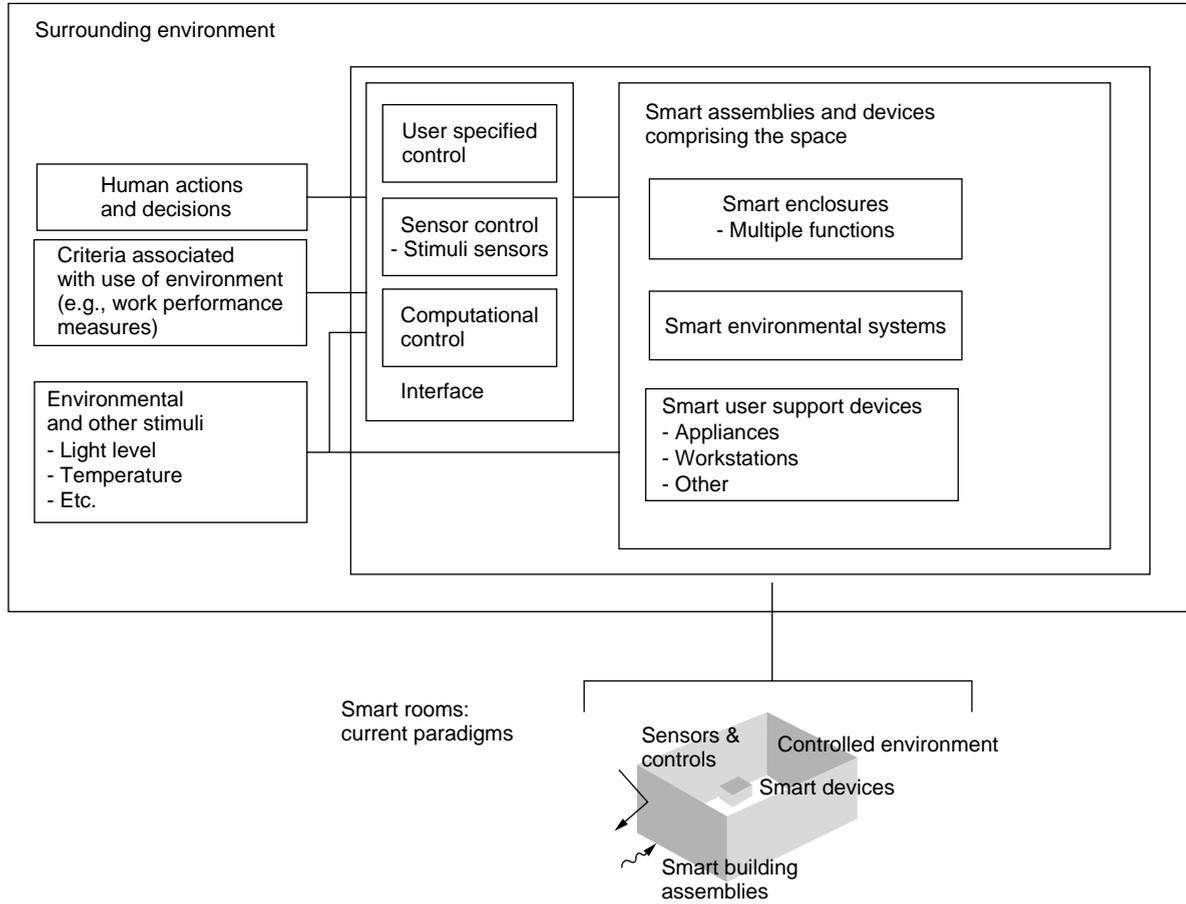


Figure 3. Smart rooms: In the current paradigm of a smart room, new smart devices are added to increase functionalities. The controlling interface is visibly and operationally present.

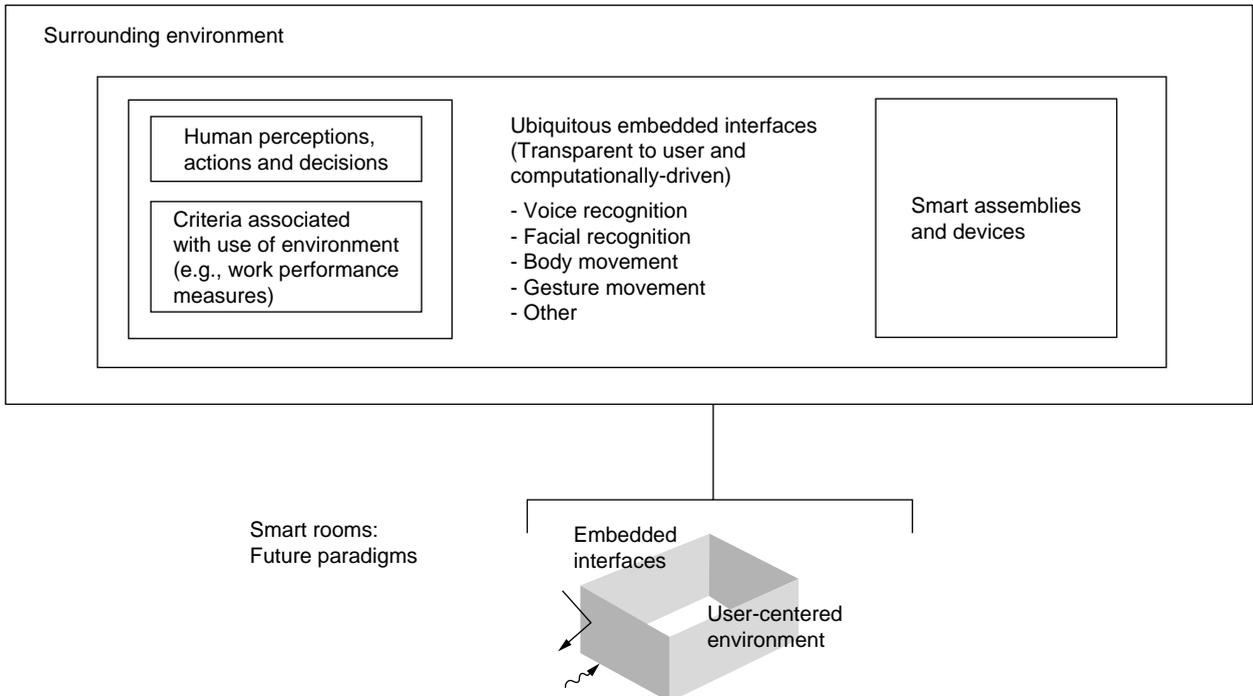


Figure 4. Smart rooms—future paradigms: The interface will disappear to the user.

into buildings. A more interesting and provocative question might be, “What would a room or building of the future be like if we could exploit smart materials and technologies to redesign the environment?” Smarter structures and construction materials might allow significant reductions in the size of the static building components—buildings could become thinner, lighter, and more flexible. The energy intensive ambient systems in buildings could be reduced or even eliminated if we allowed full interactivity between the occupant and the environmental behavior (see Fig. 4). Ambient lighting systems could be replaced by discrete sources that respond to the viewer. HVAC systems could be minimized if only the zone around an occupant were conditioned. Fundamentally, actions could be discrete and direct—the minimum necessary at the point and time for maximum effect.

BIBLIOGRAPHY

1. D.M. Addington, *Boundary Layer Control of Heat Transfer in Buildings*, Harvard University Dissertation, Cambridge, MA, 1997.
2. D.M. Addington, Discrete Control of Interior Environments in Buildings, *Proc. ASME Fluids Eng. Div.*, 1998.
3. E. Allen, *Fundamentals of Building Construction: Materials and Methods*. J Wiley, NY, 1999.
4. R.E. Christenson and B.F. Spencer, Coupled Building Control Using Smart Damping Strategies, *SPIE 7th Int. Symp. Smart Struct. Mater.*, 2000.
5. J. Hecht, *City of Light*. Oxford University Press, NY, 1999.
6. N.K. Khartchenko, *Advanced Energy Systems*. Taylor & Francis, Washington, DC, 1998.
7. R.B. Peterson, Micro Thermal Engines: Is there Any Room at the Bottom, *Proc. ASME Heat Transfer Div.*, 1999.
8. K. Satori, Y. Ikeda, Y. Kurosawa, A. Hongo, and N. Takeda, Development of Small-Diameter Optical Fiber Sensors for Damage Detection in Composite Laminates, *SPIE 7th Int. Symp. Smart Struct. Mater.* 2000.
9. R. Shekarriz and C.J. McCall, State-of-the-Art in Micro- and Meso-Scale Heat Exchangers. *Proc. ASME Adv. Energy Syst. Div.*, 1999.
10. M.D. Symans, G.J. Madden, and N. Wongprasertt, Analytical and Numerical Study of a Smart Sliding Base Isolation System for Seismic Protection of Buildings, *SPIE 7th Int. Symp. Smart Struct. Mater.*, 2000.

BATTERY APPLICATIONS

ARUMUGAM MANTHIRAM
The University of Texas at Austin
Austin, TX

INTRODUCTION

Batteries are the major power sources for portable electronic devices and toys. They are also used in automobiles for starting, lighting, and ignition (SLI batteries). At present, the worldwide battery market exceeds \$30 billion per year. Rapid technological advances and miniaturization in electronics have created an ever-increasing demand for compact, lightweight batteries. For example, popular portable electronic devices such as cellular phones, laptop computers, and camcorders require batteries of high energy density. Additionally, a need for more efficient use of available energy resources as well as air-quality control have created enormous interest in electric vehicles. For example, the major automobile manufacturers around the globe are engaged in developing advanced batteries for electric vehicles in response to increased environmental regulations and legislative mandates. The advanced and high energy density batteries have become possible due to the discovery and development of smart materials and processes. This article, after providing a brief introduction to the basic electrochemical concepts and the principles involved in batteries, presents the materials and electrochemical aspects of high energy density (lithium-ion) batteries.

ELECTROCHEMICAL CONCEPTS

A battery is an electrochemical cell that converts the chemical energy of a reaction directly into electrical energy. This section covers briefly the fundamental principles of electrochemical cells. For more detailed information, readers are referred to several excellent texts available in the literature (1–4).

Electrochemical Cells

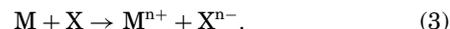
Figure 1 shows a schematic of an electrochemical cell that consists of three components: an anode or negative electrode, a cathode or positive electrode, and an electrolyte or ionic conductor. During the electrochemical reaction, the anode M is oxidized and it gives up electrons to the external circuit:



and the cathode X accepts the electrons from the external circuit and is reduced:



The electrolyte, on the other hand, acts as a medium for charge transfer between the anode and cathode as ions inside the cell. The overall cell reaction is given by adding the two half-cell reactions (1) and (2):



The amount of electricity that passes through an electrochemical cell is related by the Faraday law to the masses of reactants involved and products formed. If a current of I amperes flows in the circuit for a time of t seconds, then the amount of charge Q transferred across any interface in the cell is equal to It coulombs. Now, in accordance with the Faraday law, the number of moles N_m of the reactants M or X [see Eqs. (1) and (2)] consumed by the passage of It coulombs is given by

$$N_m = \frac{It}{nN_{Ae}}, \quad (4)$$

where n , N_A , and e are, respectively, the number of electrons given up or accepted by each M or X, Avogadro's number, and the charge on an electron. The product N_{Ae} is called the Faraday constant F , which is equal to $96,487 \text{ C mol}^{-1}$, and Eq. (4) can be reduced to

$$N_m = \frac{It}{nF}. \quad (5)$$

Consequently, the theoretical capacity Q of the electrode is given by

$$Q = It = nFN_m. \quad (6)$$

One gram-equivalent weight of an electrode, for example, theoretically has a capacity of $96,487 \text{ C}$ or 26.8 Ah . Gram-equivalent weight is defined as the atomic or molecular weight in grams divided by the number of electrons n involved in the reaction.

Thermodynamics of Electrochemical Cells

The driving force for an electrochemical cell to deliver electrical energy to an external circuit is the decrease in the standard free energy ΔG° of the cell reaction [Eq. (3)]. The free energy ΔG° is related to the standard cell potential E° by

$$\Delta G^\circ = -nFE^\circ, \quad (7)$$

where n and F are, respectively, the number of electrons involved in the reaction and the Faraday constant. The cell potential E° is the difference between the electrode potentials of the cathode and anode. The values of E° for various electrochemical couples are given in terms of standard reduction/oxidation potentials in textbooks and handbooks

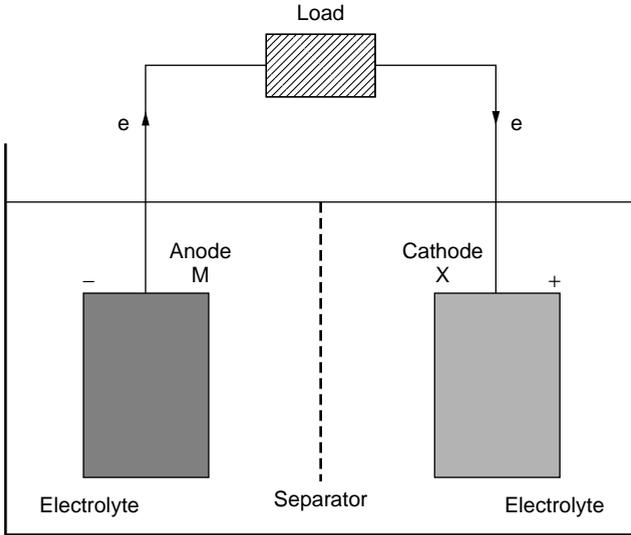


Figure 1. Schematic of an electrochemical cell.

(5,6). A positive value of E^o means that the cell reaction occurs spontaneously. The standard potential E^o is the equilibrium potential when all of the cell components are in their standard states. For example, the solution species have unit molar activities, the gases have pressures of 1 atmosphere, and the solid phases are in their most stable form in their standard states. For conditions other than the standard state, the cell potential E is given by the Nernst equation,

$$E = E^o - \frac{RT}{nF} \ln \frac{a_{M^{n+}} a_{X^{n-}}}{a_M a_X}, \quad (8)$$

where R is the gas constant, T is the absolute temperature, and $a_{M^{n+}}$, $a_{X^{n-}}$, a_M , and a_X are the activities of the products and reactants involved in cell reaction (3). At room temperature $T = 298$ K, the Nernst equation can be simplified to

$$E = E^o - \frac{0.0591}{n} \log \frac{a_{M^{n+}} a_{X^{n-}}}{a_M a_X}. \quad (9)$$

The cell potential also depends on the temperature and pressure. The dependences are related to the thermodynamic quantities by

$$\left(\frac{\partial E}{\partial T} \right)_P = \frac{\Delta S}{nF} \quad (10)$$

and

$$\left(\frac{\partial E}{\partial P} \right)_T = -\frac{\Delta V}{nF}, \quad (11)$$

where ΔS is the entropy change and ΔV is the volume change. Thus, the measurement of the cell potential can be used to determine thermodynamic quantities such as ΔG , ΔS , enthalpy change ΔH , and equilibrium constants.

Polarization Losses in Electrochemical Cells

The amount of electrical energy that an electrochemical cell can deliver is related to the free energy change of the cell reaction [Eq. (7)]. However, when a current I is passed through the cell, part of the energy is lost as waste heat due to polarization losses in the cell. The polarization loss can be classified into three types: activation polarization, concentration polarization, and ohmic polarization. Activation polarization is related to the kinetics of electrode reactions. Concentration polarization is related to the concentration differences of the reactants and products at the electrode surfaces and in the bulk as a result of mass transfer. Ohmic polarization, usually referred to as internal IR drop, is related to the internal impedance of the cell, which is a sum of the ionic resistance of the electrolyte and the electronic resistance of the electrodes.

The different polarization losses are indicated schematically in Fig. 2 as a function of operating current (2). The operating (measured) cell voltage E_{op} is given by

$$E_{op} = E_{oc} - \eta, \quad (12)$$

where E_{oc} is the open-circuit voltage and η is the overvoltage from polarization. The overvoltage η is a measure of the deviation of the cell voltage E_{op} from the equilibrium open-circuit voltage E_{oc} . The overvoltage η from the three different polarizations is given by

$$\eta = \eta_a + \eta_c + IR_i, \quad (13)$$

where η_a is the activation polarization at the anode and cathode, η_c is the concentration polarization at the anode and cathode, I is the load or operating current, and R_i is the internal resistance of the cell. The degree of polarization increases and the measured cell voltage E_{op} decreases as current increases. Therefore, the cell will operate close to the open-circuit voltage E_{oc} and deliver most of the expected energy only at very low operating currents. Obviously, the

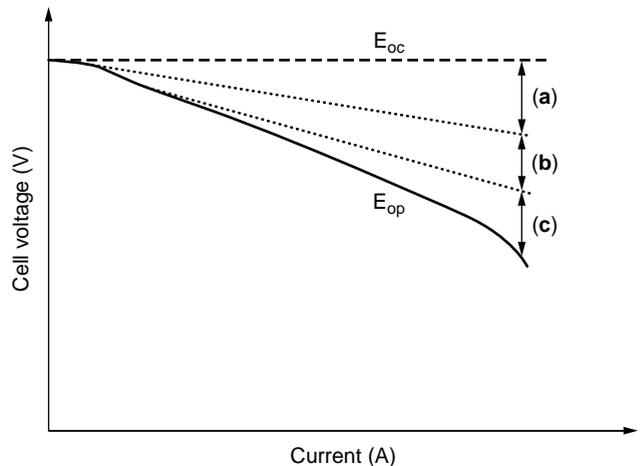


Figure 2. Variation of cell voltage with operating current illustrating polarization losses: (a) ohmic polarization, (b) activation polarization, and (c) concentration polarization.

intrinsic properties of the electrodes and electrolytes as well as the engineering design of the cell will influence the polarization losses and hence the performance and efficiency of electrochemical cells.

BATTERIES

Performance Parameters

As mentioned in the previous section, a battery is an electrochemical device that converts stored chemical energy directly into electrical energy. The performance characteristics of a battery are assessed in terms of several parameters discussed later (1–3). The cell voltage E_{op} is the difference between the electrode potentials of the cathode E_c and anode E_a :

$$E_{op} = E_c - E_a. \quad (14)$$

Although the theoretical capacity Q of a cell or half-cell is given by Eq. (6), it is often convenient to calculate the specific capacity Q_{sp} for purposes of comparison. The specific capacity Q_{sp} is obtained by dividing the capacity Q of the cell or half-cell by the mass m or volume V of the cell or half-cell and is usually expressed in terms of Ah/kg or Ah/L:

$$Q_{sp} = \frac{Q}{m}, \quad (15)$$

or

$$Q_{sp} = \frac{Q}{V}. \quad (16)$$

The available energy ε of a cell is given by the product of the cell capacity Q and the average operating voltage E_{op} and again is usually given in terms of either gravimetric energy density (specific energy ε_{sp}) in Wh/kg or volumetric energy density in Wh/L:

$$\varepsilon_{sp} = Q_{sp} E_{op}. \quad (17)$$

The power P delivered by the cell is given by the product of the current I flowing and the associated cell voltage E_{op} and is generally given in terms of gravimetric power density (specific power P_{sp}) in W/kg or volumetric power density in W/L:

$$P_{sp} = I_{sp} E_{op} \quad (18)$$

where I_{sp} is the current density (current per unit weight or volume).

The discharge characteristic of a battery is another important parameter, which is given in terms of a plot of cell voltage versus capacity. The discharge profile and the final capacity obtainable depend on the current density I_{sp} used. Figure 3 compares the discharge profiles for various current densities. A useful way of defining the influence of current density on discharge curves is in terms of C

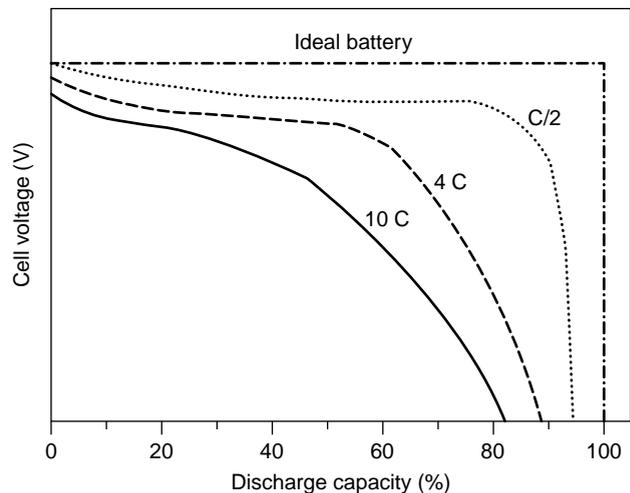


Figure 3. Discharge profiles at various C rates that illustrate the influence of current density.

rates:

$$C = \frac{I_d}{Q_n}, \quad (19)$$

where I_d and Q_n are, respectively, discharge current and nominal capacity. For example, a C rate of τ implies that the nominal capacity of the cell is delivered in $1/\tau$ hours under the specified current density. In an ideal battery, the discharge voltage drops sharply to zero when the chemical reaction reaches completion and the stored energy is fully consumed (Fig. 3). The discharge curves deviate from the ideal curve as the discharge rate (or current density) increases due to the polarization losses discussed in the previous section.

Coulometric and energy efficiencies and cycle life of secondary (rechargeable) batteries are some additional important parameters. Coulometric efficiency q_c is defined as

$$q_c = \frac{Q_d}{Q_c}, \quad (20)$$

where Q_d and Q_c are, respectively, the discharge and charge capacities. The reciprocal of the coulometric efficiency is the charging factor f :

$$f = \frac{1}{q_c}. \quad (21)$$

Energy efficiency q_E is defined as

$$q_E = q_c \frac{E_d}{E_c}, \quad (22)$$

where E_d and E_c are, respectively, the average discharge and charge voltages. A $q_C < 1$ implies the occurrence of unwanted side reactions that produce heat during the charging process. Intrinsic cell materials characteristics, cell engineering, and cell operating conditions such as current density and temperature can all influence q_C . A $q_E < q_C$ implies a deviation of the discharge and charge curves

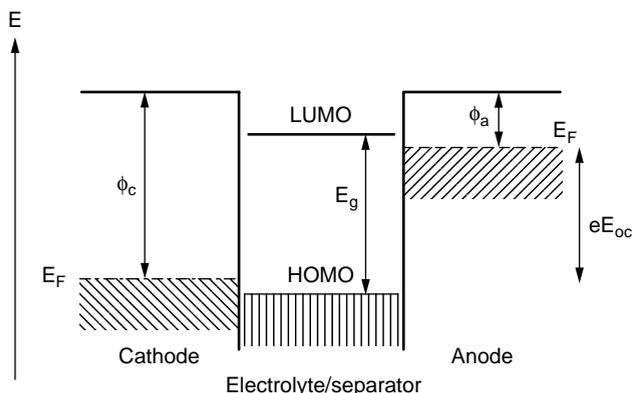


Figure 4. Schematic energy diagram of a cell that has an open circuit.

from the open-circuit voltage profile. Again, polarization losses arising from materials characteristics, cell engineering, and operating conditions can influence q_E .

The cycle life of a battery is the number of times it can be charged and discharged repeatedly before the cell capacity falls below a limiting value. Generally, the limiting value is set around 70 to 80% of the nominal capacity. The cycle life depends on the reversible characteristics (structural and chemical stability) of the electrode materials, cell engineering, and operating conditions such as temperature, current density, and depth of discharge.

Design Considerations

The equilibrium cell voltage E_{oc} and the capacity Q of a battery are determined by the intrinsic properties of the electrode materials. The cell voltage can be maximized by choosing anode materials that have a smaller work function ϕ_a and cathode materials that have a larger work function ϕ_c . In other words, the anode should be a good reducing agent that has a large negative reduction potential, and the cathode should be a good oxidizing agent that has a large positive reduction potential. A schematic energy diagram of an open circuit is shown in Fig. 4. The open-circuit voltage E_{oc} of the cell is given by

$$E_{oc} = \frac{\phi_c - \phi_a}{e} \quad (23)$$

or by the difference between the electrode potentials of the cathode and anode [Eq. (14)]. Thermodynamic stability considerations also require that the Fermi energies E_F of the cathode and anode lie within the band gap E_g of the electrolyte, as shown in Fig. 4, so that no unwanted reduction or oxidation of the electrolyte occurs. This implies a limitation of

$$\phi_c - \phi_a < E_g. \quad (24)$$

Alkali and alkaline-earth metals that have a smaller ϕ_a or a larger negative reduction potential are attractive anodes, and higher valent transition-metal compounds that have a larger ϕ_c or larger positive reduction potentials are attractive cathodes to maximize the cell voltage. The cell capacity, on the other hand, is determined by the atomic or molecular weight of the elements or compounds used as electrodes and the degree of reaction (number of electrons involved) per mole of the electrode material [Eq. (6)]. Lightweight elements such as hydrogen, lithium, or oxygen and low molecular weight compounds are preferred as electrodes to maximize cell capacity.

In addition to high cell voltage and capacity, several other criteria are important in designing a battery to achieve high efficiency and minimal energy loss. The electrolyte should have good ionic conductivity, but should be an electronic insulator to avoid internal short-circuiting. High ionic conductivity in the electrolyte is essential to minimize the IR drop or ohmic polarization. Using a given electrolyte, the IR drop due to electrolyte resistance can be reduced, and the rate capability can be improved by a higher electrode interfacial area and thin separators. The electrode should have a high electronic conductivity and diffusion rate for the ions to minimize the IR drop. The electronic conductivity of the electrodes can be improved by adding electrically conducting additives such as carbon. The electrode reaction rates at the cathode and anode should be high enough to minimize activation polarization. This is commonly achieved by using a porous electrode design, which can reduce the local current density by providing high surface area. Adequate flow or passage of electrolytes is essential to facilitate mass transfer and minimize concentration polarization. Electrode porosity and pore size, optimum separator thickness and structure, and concentration of the reactants in the electrolytes are important factors in minimizing concentration polarization. In addition to these points, the electrolyte should have

Table 1. Major Primary Battery Systems

Battery	Anode	Cathode	Cell Reaction	Cell Voltage (V)	Capacity (Ah/kg) ^a
Leclanche	Zn	MnO ₂	Zn + 2MnO ₂ → ZnO · Mn ₂ O ₃	1.6	224
Magnesium	Mg	MnO ₂	Mg + 2MnO ₂ + H ₂ O → Mn ₂ O ₃ + Mg(OH) ₂	2.8	271
Alkaline MnO ₂	Zn	MnO ₂	Zn + 2MnO ₂ → ZnO + Mn ₂ O ₃	1.5	224
Mercury	Zn	HgO	Zn + HgO → ZnO + Hg	1.34	190
Zinc-air	Zn	O ₂	Zn + 0.5O ₂ → ZnO	1.65	658
Li-SO ₂	Li	SO ₂	2Li + 2SO ₂ → Li ₂ S ₂ O ₄	3.1	379
Li-MnO ₂	Li	MnO ₂	Li + MnO ₂ → LiMnO ₂	3.1	286

^a Based only on active cathode and anode materials.

Table 2. Major Secondary Battery Systems

Battery	Anode	Cathode	Cell Reaction	Cell Voltage (V)	Capacity (Ah/kg) ^a
Lead–acid	Pb	PbO ₂	Pb + PbO ₂ + 2H ₂ SO ₄ → 2PbSO ₄ + 2H ₂ O	2.1	120
Nickel–cadmium	Cd	NiOOH	Cd + 2NiOOH + 2H ₂ O → 2Ni(OH) ₂ + Cd(OH) ₂	1.35	181
Nickel–hydrogen	H ₂	NiOOH	H ₂ + 2NiOOH → 2Ni(OH) ₂	1.5	289
Nickel–metal hydride	MH	NiOOH	MH + NiOOH → M + Ni(OH) ₂	1.35	206
Lithium-ion	Li	Li _{0.5} CoO ₂	0.5Li + Li _{0.5} CoO ₂ → LiCoO ₂	3.7	137

^a Based only on active cathode and anode materials.

good chemical stability and should not undergo any direct reaction with the electrodes. In rechargeable batteries, chemical reversibility on the electrodes is crucial to maintaining good capacity retention. Raw materials and fabrication costs, cell safety, and environmental factors are additional considerations.

Types of Batteries

Batteries can be classified into two types: primary (non-rechargeable) and secondary (rechargeable) batteries. Electrode materials undergo irreversible chemical reactions in primary batteries, but they exhibit reversible chemical reactions in secondary batteries. Some major primary and secondary battery systems are given in Tables 1 and 2 (2). The tables give the cell reactions, voltage, and capacity for each system. Most of the primary and secondary systems are based on aqueous electrolytes; the lithium-based primary systems in Table 1 and the lithium-ion system in Table 2 are based on nonaqueous electrolytes. The aqueous systems are limited in cell voltage (≤ 2.1 V) due to a smaller separation E_g between the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) of water (Fig. 4) and the consequent vulnerability of water to reduction/oxidation reactions at higher cell voltages. The use of nonaqueous electrolytes that have a larger E_g , on the other hand, permits higher cell voltages in lithium-based systems.

SMART BATTERIES

The discovery of smart materials and the development of new processes have revolutionized the electronics industry over the years. The continued reduction in the sizes and weights of popular portable electronic devices such as cellular phones and laptop computers has driven the parallel, development of smart batteries to power them. In this regard, lithium-ion batteries have become appealing because they offer higher energy density (volumetric and gravimetric) compared to other rechargeable systems (Fig. 5) such as lead-acid, nickel–cadmium, and nickel–metal hydride batteries (7). Lithium-ion batteries are smaller and lighter compared to other systems. Lithium-ion batteries have become a commercial reality since Sony Corporation introduced them in 1990 as a result of the discovery of new materials over the years. The history, principles, current status, and future challenges of lithium-ion technology are briefly discussed in the following sections.

For more detailed information, the readers are referred to several references in the literature (1–3, 8–13).

Lithium-Ion Batteries

Maximizing the energy density of a battery requires using electrode materials that offer both high capacity and high cell voltage E_{oc} . High cell voltage can be achieved by using an anode material that has a smaller work function ϕ_a and a cathode material that has a larger work function ϕ_c , as shown in [Eq. (23)]. Lithium metal—the lightest solid in the periodic table—has a high specific capacity and a high standard oxidation potential (small ϕ_a) and is an attractive anode for achieving high energy density. Because of this objective, batteries that consist of metallic lithium as an anode and a lithium insertion compound $Li_xM_yX_z$ (M = transition metal and X = nonmetal) as a cathode became appealing during the 1970s; a lithium insertion compound is a host matrix into/from which the guest species Li^+ can be reversibly inserted/extracted. This concept of a secondary lithium battery was initially demonstrated using a layered metal sulfide TiS_2 as the cathode and a nonaqueous electrolyte consisting of a lithium salt such as $LiClO_4$ dissolved in an organic solvent such as propylene carbonate. In this cell, the Li^+ ions produced at the anode by oxidation of the metallic lithium during discharge migrate through the electrolyte and are inserted into the van der Waals gap between the sulfide layers of TiS_2 , and the electrons flow through the external circuit from the anode to the cathode to give Li_xTiS_2 . During the charging process, the Li^+ ions are extracted from Li_xTiS_2 and the electrons

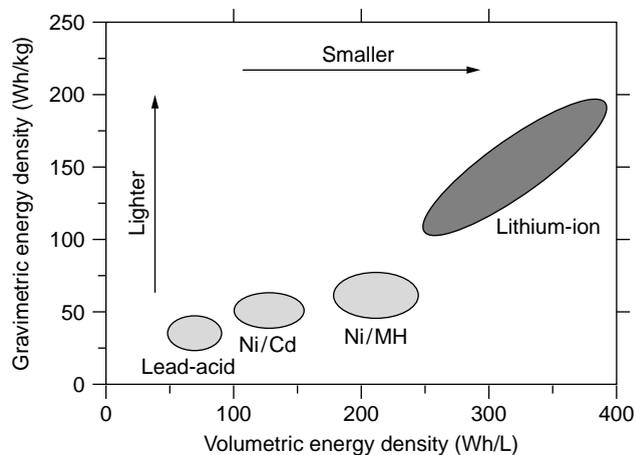


Figure 5. Comparison of the gravimetric and volumetric energy densities of various rechargeable battery systems.

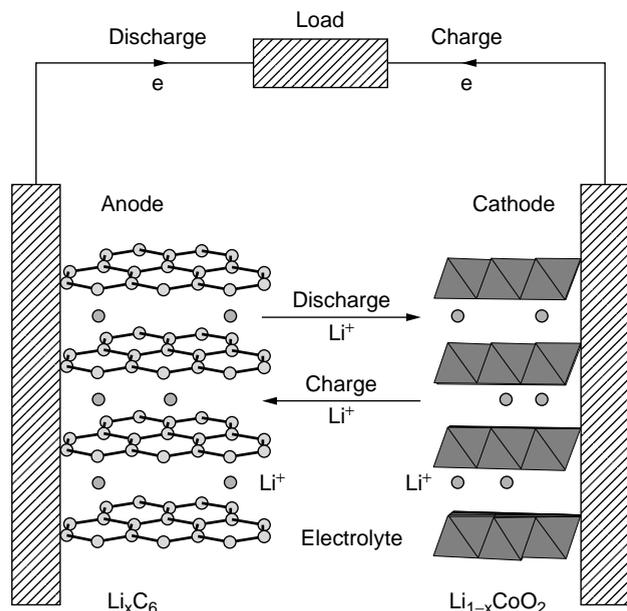


Figure 6. Schematic of the charging/discharging process in a lithium-ion cell.

flow from the cathode to the anode through the external circuit.

In this cell, however, the chemical reactivity of metallic lithium with the nonaqueous electrolyte results in forming a passivating film on the anode. Although the passivating film prevents further corrosion, it leads to nonuniform plating of lithium during charging, which results in total cell failure due to dendritic short-circuiting and also in serious safety problems due to local overheating. These difficulties of the metallic lithium anode forced the use of lithium insertion compounds as both anodes and cathodes. These cells are called lithium-ion cells or rocking-chair cells because the lithium ion shuttles or rocks between the cathode and anode hosts during the charging/discharging process (Fig. 6). This strategy, however, requires careful selection of cathode and anode pairs to maintain high cell voltage (>3 V) and to minimize the added weight of the insertion compound anode.

Although the concept of secondary lithium batteries was initially demonstrated by using a sulfide cathode, it was recognized during the 1980s that it is difficult to achieve high cell voltage using sulfide cathodes because an overlap of the higher valent M^{n+} :d energies and the top of the S:3p energy and the formation of S_2^{2-} ions lead to an inaccessibility of higher oxidation states for M^{n+} in a sulfide $Li_xM_yS_z$; the stabilization of the higher oxidation state is essential to maximize the work function ϕ_c and thereby the cell voltage E_{oc} [Eq. (23)]. On the other hand, the location of O:2p energy much below the S:3p energy and a larger increase of the M^{n+} :d energies in an oxide compared to those in a sulfide, due to a larger Madelung energy, make the higher valent states accessible in oxides. Accordingly, transition-metal oxide hosts were pursued as cathodes during the 1980s (14–16).

Figure 7 compares the electrochemical potential ranges of some lithium insertion compounds versus metallic

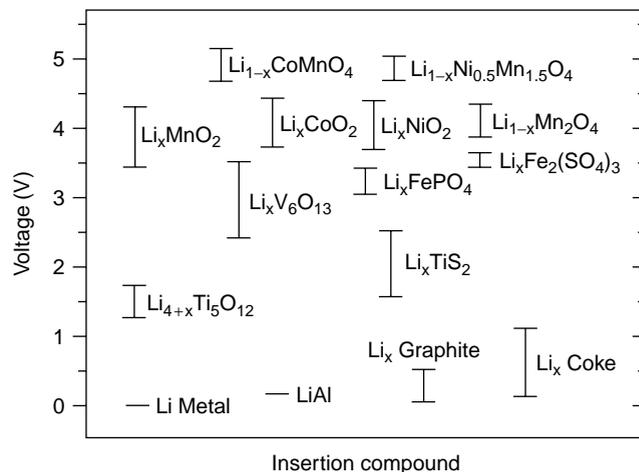


Figure 7. Electrochemical potential ranges of some lithium insertion compounds with reference to metallic lithium.

lithium. Among them, $LiCoO_2$, $LiNiO_2$, and $LiMn_2O_4$ oxides that have a higher electrode potential of 4 V versus metallic lithium have become attractive cathodes for lithium-ion cells. Graphite and coke that have lower electrode potentials <1 V versus metallic lithium and are lightweight have become attractive anodes. In a lithium-ion cell made from, for example, a $LiCoO_2$ cathode and a carbon anode (Fig. 6), the lithium ions migrate from the $LiCoO_2$ cathode to the Li_xC_6 anode through the electrolyte, and the electrons flow through the external circuit from the cathode to the anode during the charging process. Exactly the reverse reaction occurs during the discharging process.

A lithium insertion compound should have several features to be a successful electrode (cathode or anode) in lithium-ion cells:

- The cathode should have a high lithium chemical potential ($\mu_{Li(c)}$), and the anode should have a low lithium chemical potential ($\mu_{Li(a)}$) to maximize the cell voltage:

$$E_{oc} = \frac{\mu_{Li(c)} - \mu_{Li(a)}}{F}. \quad (25)$$

The voltage is determined by the energies involved in both electron transfer and Li^+ transfer. The energy involved in electron transfer is related to the work functions of the cathode (ϕ_c) and anode (ϕ_a) as shown in Eq. (23), whereas that involved in Li^+ transfer is determined by the crystal structure and the coordination geometry of the site into/from which Li^+ ions are inserted/extracted (17). If we consider only electron transfer, then E_{oc} can be given by Eq. (23). This implies that the M^{n+} ion in the insertion compound $Li_xM_yO_z$ should have a high oxidation state to be used as a cathode and a low oxidation state to be used as an anode.

- The insertion compound $\text{Li}_x\text{M}_y\text{O}_z$ should allow insertion/extraction of a large amount of lithium x to maximize the cell capacity. This depends on the number of available lithium sites and the accessibility of multiple valences for M in the insertion host.
- The lithium insertion/extraction process should be reversible and have no or minimal changes in the host structure across the entire range x of lithium insertion/extraction to provide a good cycle life.
- The insertion compound $\text{Li}_x\text{M}_y\text{O}_z$ should have good electronic conductivity σ_e and Li^+ -ion conductivity σ_{Li} to minimize polarization losses during the discharging/charging process and thereby to support a high current and power densities.
- The insertion compound $\text{Li}_x\text{M}_y\text{O}_z$ should be chemically stable and should not react with the electrolyte across the entire range x of lithium insertion/extraction.
- The Fermi energies of the cathode and anode in the entire range x of lithium insertion/extraction should lie within the band gap of the electrolyte, as shown in Fig. 4, to prevent any unwanted oxidation or reduction of the electrolyte.
- The insertion compound $\text{Li}_x\text{M}_y\text{O}_z$ should be inexpensive, environmentally benign, and lightweight.

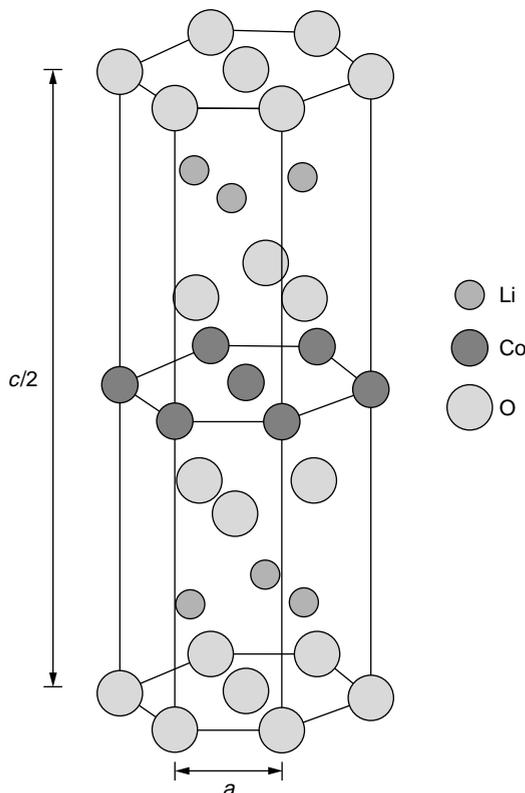


Figure 8. Crystal structure of layered LiCoO_2 .

Layered Cobalt Oxide Cathodes

LiCoO_2 has a layer structure in which the Li^+ and Co^{3+} ions occupy the alternate (111) planes of a rock salt structure, as shown in Fig. 8, to give a layer sequence of $-\text{O}-\text{Li}-\text{O}-\text{Co}-\text{O}-$ along the c axis. This structure has an oxygen stacking sequence of ABCABC along the c axis, and the Li^+ and Co^{3+} ions occupy the octahedral interstitial sites of the cubic close-packed oxygen array. Accordingly, it is designated as an O3 layer structure. The structure provides reversible extraction/insertion of lithium ions from/into the lithium planes. Two-dimensional motion of the Li^+ ions between the strongly bonded CoO_2 layers provides fast lithium-ion diffusion (high σ_{Li}), and the edge-shared CoO_6 octahedral arrangement that has a direct $\text{Co}-\text{Co}$ interaction provides good electronic conductivity σ_e necessary for a high rate. A large work function ϕ_c for the highly oxidized $\text{Co}^{3+/4+}$ couple provides a high cell voltage of around 4 V, and the discharge voltage does not change significantly as the degree of lithium extraction/insertion x in $\text{Li}_{1-x}\text{CoO}_2$ changes (Fig. 9). These features have made LiCoO_2 an attractive cathode, and most of the commercial lithium-ion cells are currently made from LiCoO_2 .

However, only 50 % of the theoretical capacity of LiCoO_2 that corresponds to a reversible extraction of 0.5 lithium per Co (practical capacity of 140 Ah/kg) can be practically used. The limitation in practical capacity has been attributed in the literature (18) to an ordering of Li^+ ions and consequent structural distortions around $x = 0.5$ in $\text{Li}_{1-x}\text{CoO}_2$. However, it has been shown more recently that the limited capacity could be due to the tendency of $\text{Li}_{1-x}\text{CoO}_2$ to lose oxygen (or react with the electrolyte) at

a deep charge when $(1-x) < 0.5$ (19). Figure 10 shows the variation of the oxidation state of cobalt and the oxygen content as the lithium content $(1-x)$ varies. The data in Figure 10 were obtained by chemically extracting lithium from LiCoO_2 using the oxidizing agent NO_2PF_6 in a non-aqueous (acetonitrile) medium and determining the oxidation state of cobalt by a redox (iodometric) titration. Constancy of the cobalt oxidation state and an oxygen content significantly less than 2 at low lithium contents demonstrate the chemical instability of $\text{Li}_{1-x}\text{CoO}_2$ cathodes at a deep charge when $(1-x) < 0.5$. The tendency of $\text{Li}_{1-x}\text{CoO}_2$ to lose oxygen at a deep charge is consistent with the recent X-ray absorption spectroscopic (20) and electron energy loss spectroscopic (21) data. The spectroscopic data indicate that the holes (removal of electrons) are introduced

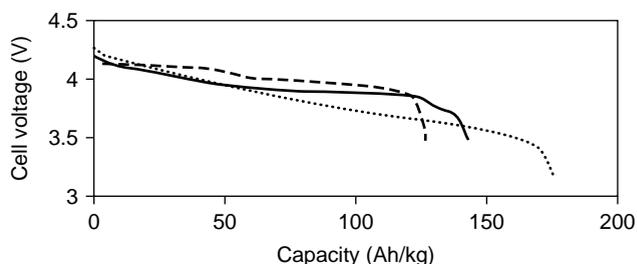


Figure 9. Typical discharge curves of layered LiCoO_2 (—), layered $\text{LiNi}_{0.85}\text{Co}_{0.15}\text{O}_2$ (···), and spinel LiMn_2O_4 (---).

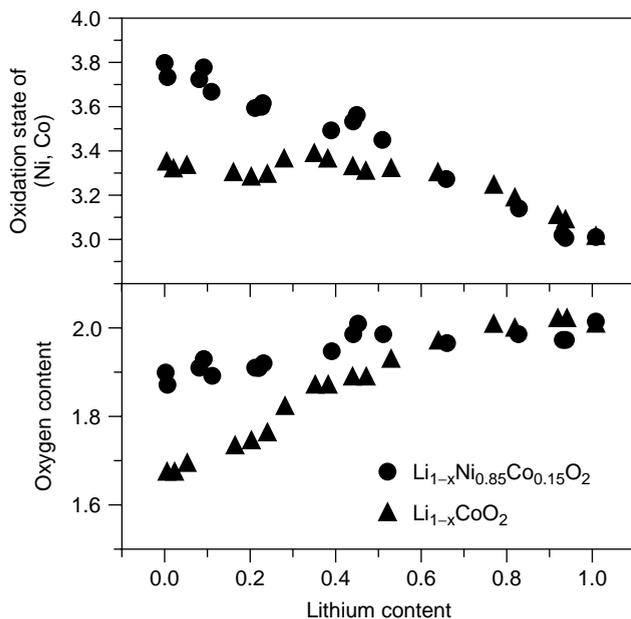


Figure 10. Variations of the oxidation state of the transition-metal ions and oxygen content as lithium content varies in $\text{Li}_{1-x}\text{CoO}_{2-\delta}$ and $\text{Li}_{1-x}\text{Ni}_{0.85}\text{Co}_{0.15}\text{O}_{2-\delta}$.

into the O:2p band rather than the Co:3d band during the electrochemical extraction of lithium. Introduction of a significant amount of holes into the O:2p band will lead to evolution of oxygen from the lattice. However, note that neutral oxygen in the presence of electrolytes in lithium-ion cells may not be evolved under conditions of overcharge when $(1-x) < 0.5$. Instead, the cathode may react with the electrolyte due to the highly oxidized nature of the deeply charged $\text{Li}_{1-x}\text{CoO}_2$ cathode.

Figure 11 shows the X-ray diffraction patterns of the $\text{Li}_{1-x}\text{CoO}_2$ samples that were obtained by chemically extracting lithium from $\text{Li}_{1-x}\text{CoO}_2$. The samples maintain the initial O3 layer structure (CdCl₂ structure) for $0.35 \leq (1-x) \leq 1$. For lithium contents $(1-x) < 0.35$, a second phase begins to form as indicated by the appearance of a shoulder on the right-hand side of the (003) reflection centered around $2\theta = 20^\circ$. The intensity of the new reflection increases as the lithium content decreases further, and the end member $\text{CoO}_{2-\delta}$ consists of reflections corresponding only to the new phase. The X-ray diffraction pattern of the new phase could be indexed on the basis of a two-phase mixture consisting of a major P3 phase and a minor O1 (CdI₂ structure) phase. The P3 and O1 phases have oxygen stacking sequences of ABBCCA and ABABAB, respectively. The Li⁺ ions occupy prismatic (trigonal prism) and octahedral sites, respectively, in the P3 and O1 structures. The formation of the P3 and O1 phases from the initial O3 structure is due to sliding of the oxide ions, as shown in Fig. 12. The driving force for the sliding appears to be structural instability caused by the formation of oxygen vacancies at low lithium contents (Fig. 10). The observed transformation of the O3 phase at low lithium content is consistent with that found in electrochemically

prepared samples (22). The tendency to lose oxygen and the associated structural transitions limit the practical capacity of LiCoO_2 cathodes.

Layered Nickel Oxide Cathodes

LiNiO_2 has an O3 layer structure (Fig. 8) like LiCoO_2 , and the $\text{Ni}^{3+/4+}$ couple that has a larger ϕ_c provides a high cell voltage of around 4 V. However, LiNiO_2 suffers from a few drawbacks: (1) difficulty in synthesizing LiNiO_2 as a perfectly ordered phase without mixing Li⁺ and Ni³⁺ ions in the lithium plane (23,24), (2) Jahn–Teller distortion (tetragonal structural distortion) associated with a low spin $\text{Ni}^{3+}:d^7$ ion (25), (3) irreversible phase transitions during the charge/discharge process, and (4) safety concerns in the charged state. As a result, LiNiO_2 is not a promising material for commercial cells. However, some of these difficulties have been overcome by partially substituting cobalt for nickel. For example, the composition $\text{LiNi}_{0.85}\text{Co}_{0.15}\text{O}_2$, has been shown to exhibit attractive electrochemical properties (26). It has a reversible capacity of around 180 Ah/kg (Fig. 9) and excellent cyclability. This capacity is 30% higher than that of LiCoO_2 , and it corresponds to 65% of the theoretical capacity. The substitution of cobalt for nickel has been found to suppress the cation disorder and Jahn–Teller distortion, as indicated by X-ray absorption fine structure studies (25). The higher capacity of $\text{LiNi}_{0.85}\text{Co}_{0.15}\text{O}_2$ has made it an attractive alternate for LiCoO_2 .

However, the reason for the higher capacity of $\text{LiNi}_{0.85}\text{Co}_{0.15}\text{O}_2$ compared to the analogous LiCoO_2 cathodes was not clear in the literature. The structural stability of the $\text{LiNi}_{0.85}\text{Co}_{0.15}\text{O}_2$ cathodes during long-term cycling, particularly under mild heat, also remained to be assessed. Recent experiments on $\text{Li}_{1-x}\text{Ni}_{0.85}\text{Co}_{0.15}\text{O}_2$ samples obtained by chemically extracting lithium from $\text{LiNi}_{0.85}\text{Co}_{0.15}\text{O}_2$ show that the higher capacity of $\text{LiNi}_{0.85}\text{Co}_{0.15}\text{O}_2$ compared to that of LiCoO_2 is due to its resistance to losing oxygen at low lithium contents. Figure 10 compares the variations of the average oxidation state of the transition-metal ions and the oxygen contents as lithium content varies in $\text{Li}_{1-x}\text{Ni}_{0.85}\text{Co}_{0.15}\text{O}_2$ and $\text{Li}_{1-x}\text{CoO}_2$. The data show that the former system exhibits better stability without losing much oxygen at a deep charge. Figure 13 shows the X-ray diffraction patterns of the $\text{Li}_{1-x}\text{Ni}_{0.85}\text{Co}_{0.15}\text{O}_2$ samples that were obtained by chemically extracting lithium from $\text{LiNi}_{0.85}\text{Co}_{0.15}\text{O}_2$. In this case, the initial O3 structure is maintained for a wider lithium content $0.23 \leq (1-x) \leq 1$, and the new phase is formed at a lower lithium content $(1-x) < 0.23$. More importantly, the X-ray diffraction pattern of the end member $\text{NiO}_{2-\delta}$ could also be indexed on the basis of an O3 structure but had smaller lattice parameters compared to the initial O3 phase. The observation of an O3 structure for the chemically prepared $\text{NiO}_{2-\delta}$ agrees with that found for the electrochemically prepared sample (27,28). The absence of a significant amount of oxygen vacancies appears to prevent the sliding of oxide ion layers and the structural transformation. The absence of oxygen loss and the maintenance of the initial O3 structure to a much

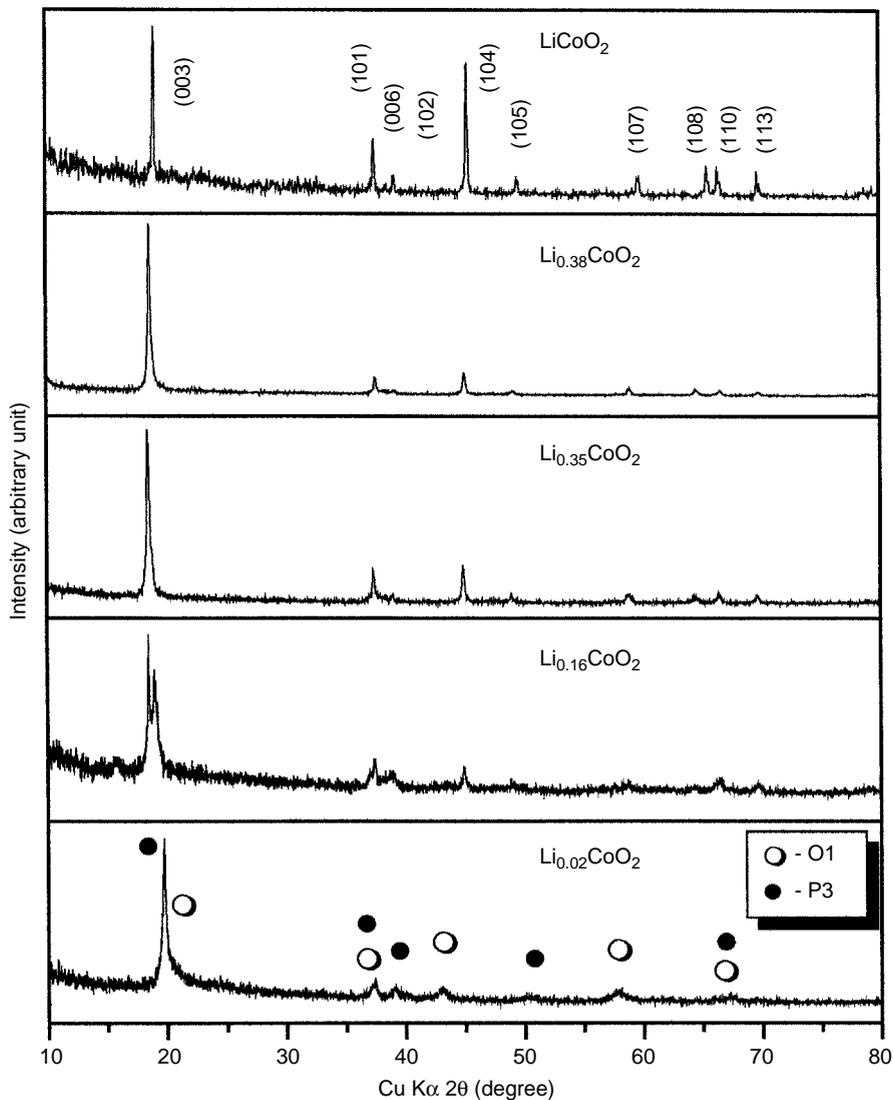


Figure 11. X-ray diffraction patterns of $\text{Li}_{1-x}\text{CoO}_{2-\delta}$ that were synthesized by chemical delithiation.

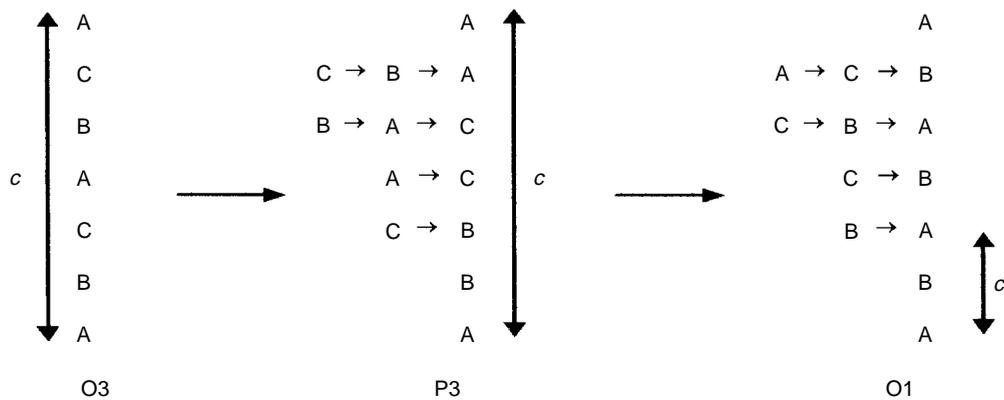


Figure 12. Schematic of the transformation of $\text{O3} \rightarrow \text{P3} \rightarrow \text{O1}$ structures by sliding of oxide layers.

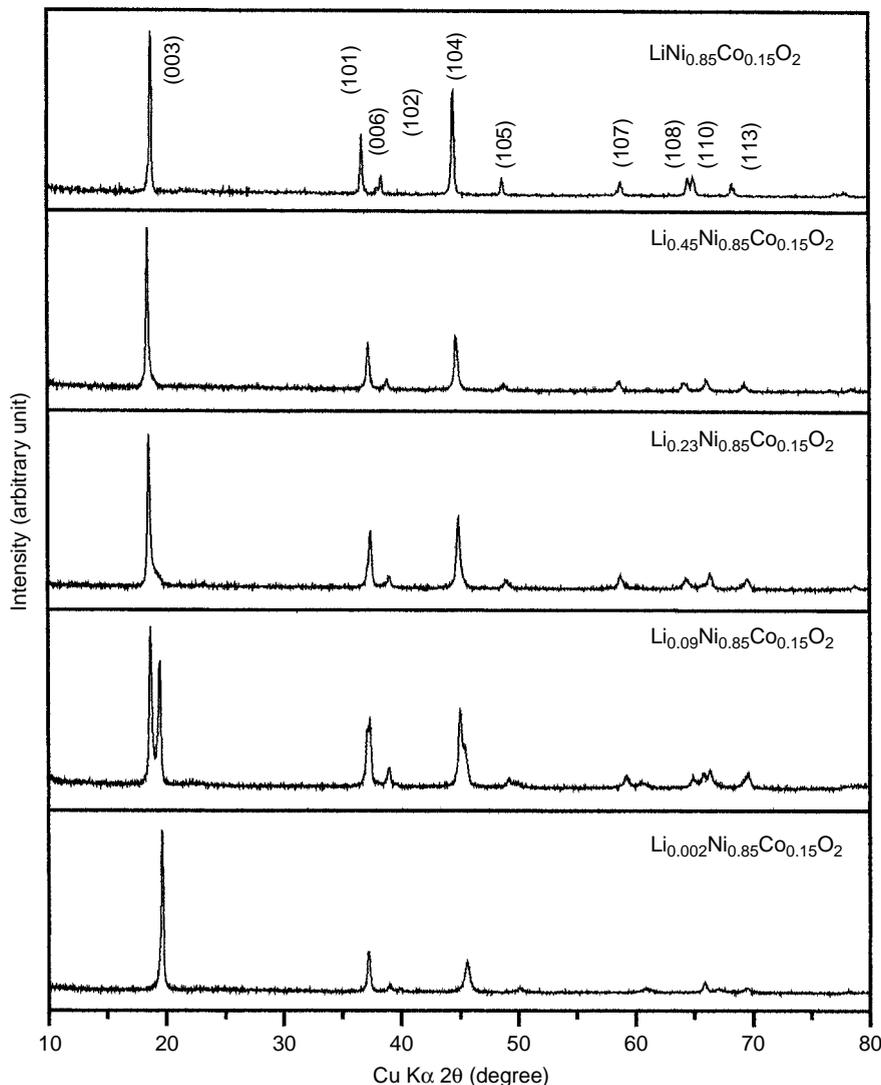


Figure 13. X-ray diffraction patterns of $\text{Li}_{1-x}\text{Ni}_{0.85}\text{Co}_{0.15}\text{O}_{2-\delta}$ that were synthesized by chemical delithiation.

lower lithium content ($1 - x$) compared to that in $\text{Li}_{1-x}\text{CoO}_2$ permit a higher capacity in the $\text{Li}_{1-x}\text{Ni}_{0.85}\text{Co}_{0.15}\text{O}_2$ system.

The differences in oxygen loss behavior between the $\text{Li}_{1-x}\text{CoO}_2$ and the $\text{Li}_{1-x}\text{Ni}_{0.85}\text{Co}_{0.15}\text{O}_2$ systems can be understood by considering qualitative energy diagrams for $\text{Li}_{1-x}\text{CoO}_2$ and $\text{Li}_{1-x}\text{NiO}_2$ (Fig. 14). In LiCoO_2 that has a $\text{Co}^{3+}:3d^6$ configuration, the t_{2g} band is completely filled, and the e_g band is empty. As lithium is extracted from LiCoO_2 , the Co^{3+} ions are oxidized to Co^{4+} , which is accompanied by removal of electrons from the t_{2g} band. Because the t_{2g} band overlaps the top of the O:2p band, deeper lithium extraction where $(1 - x) < 0.5$ results in a removal of electrons from the O:2p band as well. The removal of a significant amount of electron density from the O:2p band will result in oxidation of the O^{2-} ions and an ultimate loss of oxygen from the lattice. In contrast, the LiNiO_2 system that has a $\text{Ni}^{3+}:3d^7$ configuration involves the removal of electrons only from the e_g band. For

$\text{LiNi}_{0.85}\text{Co}_{0.15}\text{O}_2$, the electrons will be removed from the e_g band for $(1 - x) > 0.15$. Because the e_g band lies well above the O:2p band, this system does not lose oxygen down to a lower lithium content. The band diagrams in Fig. 14 are consistent with the recent spectroscopic evidence for the introduction of holes into the O:2p band rather than into the Co:3d band in LiCoO_2 (20,21) and into the Ni:3d band in $\text{Li}_{1-x}\text{NiO}_2$ and $\text{Li}_{1-x}\text{Ni}_{0.85}\text{Co}_{0.15}\text{O}_2$ (29,30).

To assess the structural stability of $\text{Li}_{1-x}\text{Ni}_{0.85}\text{Co}_{0.15}\text{O}_2$ cathodes during long-term cycling, chemically prepared $\text{Li}_{1-x}\text{Ni}_{0.85}\text{Co}_{0.15}\text{O}_2$ samples were subjected to mild heat and examined by X-ray diffraction (31). The data show a decrease in the c/a ratio of the unit cell parameters of, for example, $\text{Li}_{0.35}\text{Ni}_{0.85}\text{Co}_{0.15}\text{O}_2$ when heated at $T > 50^\circ\text{C}$ due to migration of the Ni^{3+} ions from the nickel plane to the lithium plane. Interestingly, the cobalt oxide $\text{Li}_{0.35}\text{CoO}_2$ that has a similar degree of lithium extraction (charging) shows little or no decrease in the c/a ratio when heated under similar conditions. Thus, the $\text{Li}_{1-x}\text{Ni}_{0.85}\text{Co}_{0.15}\text{O}_2$

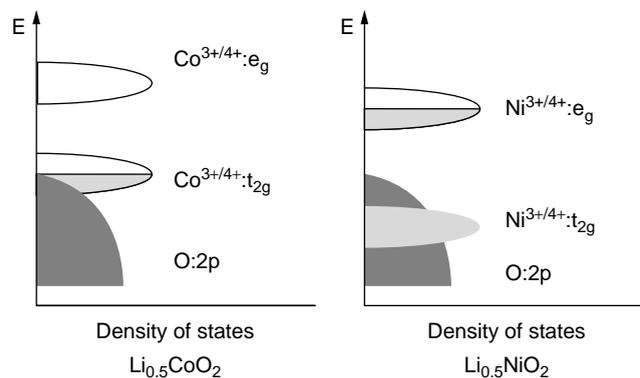


Figure 14. Comparison of the qualitative energy diagrams of $\text{Li}_{1-x}\text{CoO}_2$ and $\text{Li}_{1-x}\text{NiO}_2$.

cathodes experience structural instability under mild heat, whereas the $\text{Li}_{1-x}\text{CoO}_2$ cathodes do not under similar conditions. Although the $\text{LiNi}_{0.85}\text{Co}_{0.15}\text{O}_2$ cathode has higher capacity (180 Ah/kg) than the LiCoO_2 cathode (140 Ah/kg) and is more resistant to losing oxygen from the lattice compared to LiCoO_2 , the structural instability experienced due to cation migration may become an issue under cycling at higher temperatures ($T > 50^\circ\text{C}$). The differences in the structural stability between the two systems can be explained by considering the mechanism of cation migration. The migration of transition-metal ions from octahedral sites in the transition-metal plane to the octahedral sites in the lithium plane needs to occur via the neighboring empty tetrahedral sites, as shown in Fig. 15. While the low spin $\text{Co}^{3+}:3d^6$ ion that has strong octahedral site stabilization energy (32) is unable to migrate to the neighboring tetrahedral site, but the low spin $\text{Ni}^{3+}:3d^7$ ion that has moderate octahedral site stabilization energy is able to move to the tetrahedral site under mild heat.

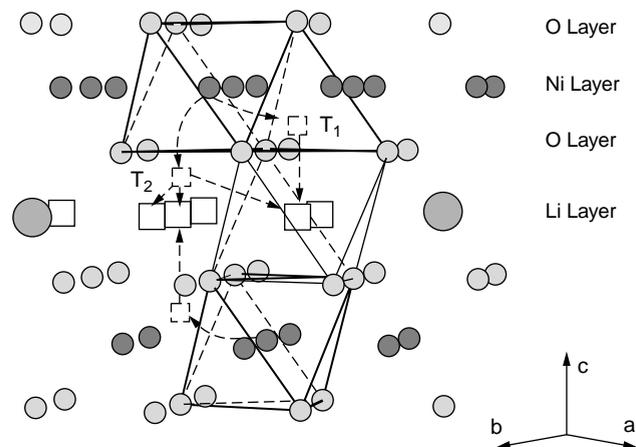


Figure 15. Schematic representation of the diffusion processes of nickel ions in $\text{Li}_{1-x}\text{Ni}_{0.85}\text{Co}_{0.15}\text{O}_2$. Dotted and solid squares refer to tetrahedral site and lithium-ion vacancy, respectively. T_1 and T_2 refer to tetrahedral sites at $(0, 0, 0.125)$ and $(0, 0, 0.375)$, respectively.

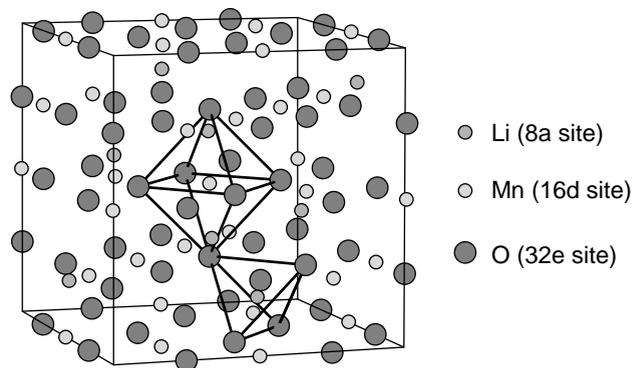


Figure 16. Crystal structure of LiMn_2O_4 spinel.

Spinel Manganese Oxide Cathodes

Although LiCoO_2 and $\text{LiNi}_{0.85}\text{Co}_{0.15}\text{O}_2$ are attractive candidates, both Co and Ni are expensive and relatively toxic. These considerations have created much interest in manganese oxides because Mn is inexpensive and environmentally benign (33–35). In this regard, LiMn_2O_4 that crystallizes in a three-dimensional cubic spinel structure (Fig. 16) has become appealing (16). In the LiMn_2O_4 spinel, the Li^+ and the $\text{Mn}^{3+/4+}$ ions occupy, respectively, the 8a tetrahedral and 16d octahedral sites of the cubic close-packed oxygen array. A strong edge-shared octahedral $[\text{Mn}_2]\text{O}_4$ array permits reversible extraction of the Li^+ ions from the tetrahedral sites without collapsing the three-dimensional spinel framework. An additional lithium-ion can also be inserted into the empty 16c octahedral sites of the spinel framework to give the lithiated spinel $\text{Li}_2[\text{Mn}_2]\text{O}_4$. However, electrostatic repulsion between the Li^+ ions in the 8a tetrahedral and 16c octahedral sites, which share common faces, causes a displacement of the tetrahedral Li^+ ions into the neighboring empty 16c sites to give an ordered rock salt structure that has a cation distribution of $(\text{Li}_2)_{16c}[\text{Mn}_2]_{16d}\text{O}_4$. Thus, theoretically, two lithium ions per LiMn_2O_4 formula unit could be reversibly inserted/extracted. Although the edge-shared MnO_6 octahedral arrangement that has direct Mn–Mn interaction provides good electrical (small polaron) conductivity σ_e , the interconnected interstitial sites in the three-dimensional spinel framework provide good lithium-ion conductivity σ_{Li} .

The lithium extraction/insertion from/into the 8a tetrahedral and 16c octahedral sites of the $\text{Li}[\text{Mn}_2]\text{O}_4$ spinel occurs in two distinct steps (16). The former occurs at around 4 V (Fig. 9) maintaining the cubic spinel symmetry; in contrast, the latter occurs at around 3 V by a two-phase mechanism involving the cubic spinel $\text{Li}[\text{Mn}_2]\text{O}_4$ and the tetragonal lithiated spinel $\text{Li}_2[\text{Mn}_2]\text{O}_4$. Although both involve the $\text{Mn}^{3+/4+}$ couple, the 1 V difference between the two processes reflects the differences in the site energies (17) as differentiated by Eq. (23) and (25). A deep energy well for the 8a tetrahedral Li^+ ions and a high activation energy required for the Li^+ ions to move from one 8a tetrahedral site to another via an energetically unfavorable neighboring 16c site lead to a higher voltage of 4 V (33). The cubic to tetragonal transition from $\text{Li}[\text{Mn}_2]\text{O}_4$ to $\text{Li}_2[\text{Mn}_2]\text{O}_4$ is due

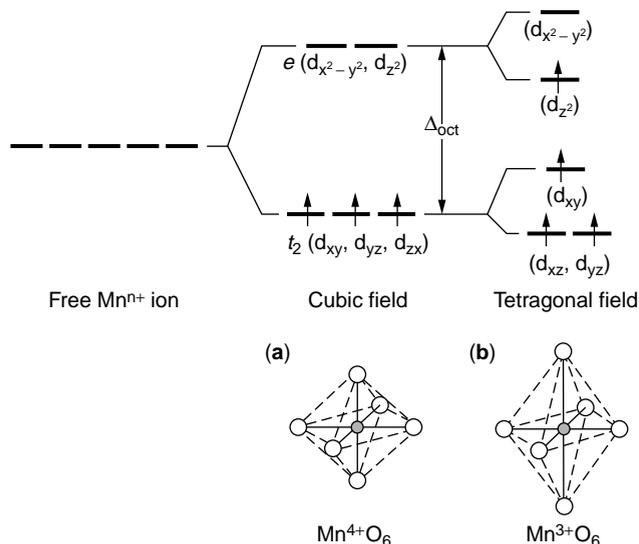


Figure 17. Illustration of the Jahn–Teller distortion in manganese oxides: (a) $\text{Mn}^{4+}:\text{3d}^3$ that has cubic symmetry (no Jahn–Teller distortion) and (b) $\text{Mn}^{3+}:\text{3d}^4$ that has tetragonal symmetry (Jahn–Teller distortion).

to the Jahn–Teller distortion of the single electron in the e_g orbitals of a high spin $\text{Mn}^{3+}:\text{3d}^4$ ion (Fig. 17). A cooperative distortion of the MnO_6 octahedra that have long Mn–O bonds along the c axis and short Mn–O bonds along the a and b axes results in macroscopic tetragonal symmetry for $\text{Li}_2[\text{Mn}_2]\text{O}_4$.

Although, in principle, two lithium ions per LiMn_2O_4 formula unit could be reversibly extracted/inserted from/into the $\text{Li}[\text{Mn}_2]\text{O}_4$ spinel framework, the cubic to tetragonal transition is accompanied by a 16% increase in the c/a ratio of the unit cell parameters and a 6.5% increase in unit cell volume. This change is too severe for the electrodes to maintain structural integrity during the discharge/charge cycle, and so LiMn_2O_4 exhibits drastic capacity fade in the 3 V region. As a result, LiMn_2O_4 has limited practical capacity of around 120 Ah/kg (Fig. 9) that corresponds to an extraction/insertion of 0.4 lithium per Mn in the 4 V region. Furthermore, even though it has limited capacity, LiMn_2O_4 tends to exhibit capacity fade in the 4 V region as well, particularly at elevated temperatures (50°C). The capacity fade in the 4 V region has been attributed to a dissolution of manganese into the electrolyte originating from a disproportionation of Mn^{3+} into Mn^{4+} and Mn^{2+} (36) and the formation of tetragonal $\text{Li}_2[\text{Mn}_2]\text{O}_4$ on the surface of the particles under conditions of nonequilibrium cycling (37).

The difficulties of lattice distortions in the LiMn_2O_4 spinel have motivated strategies to suppress Jahn–Teller distortion. One way to suppress Jahn–Teller distortion is to increase the average oxidation state of manganese because $\text{Mn}^{4+}:\text{3d}^3$ does not undergo Jahn–Teller distortion. The oxidation state of manganese can be increased either by aliovalent cationic substitutions or by increasing the oxygen content in LiMn_2O_4 . Using this strategy, Thackeray et al. (33,38,39) pioneered the Li–Mn–O phase diagram. For example, substituting Li for Mn in $\text{Li}_{1+x}\text{Mn}_2\text{O}_4$ increases the

oxidation state, and the end member $\text{Li}_4\text{Mn}_5\text{O}_{12}$ ($x = 0.33$) has an oxidation state of 4+ for Mn. Similarly, the oxidation state increases as oxygen content increases in $\text{LiMn}_2\text{O}_{4+\delta}$, and the end member $\text{Li}_2\text{Mn}_4\text{O}_9$ ($\delta = 0.5$) will have an oxidation state of 4+ for Mn. However, these defective spinels are difficult to synthesize by conventional high-temperature procedures, and more recently solution-based syntheses have been pursued to obtain them (40,41). It is also difficult to extract lithium from $\text{Li}_4\text{Mn}_5\text{O}_{12}$ and $\text{Li}_2\text{Mn}_4\text{O}_9$ because Mn^{4+} is difficult to oxidize further and therefore they are not suitable for lithium-ion cells that have carbon anodes. Both $\text{Li}_4\text{Mn}_5\text{O}_{12}$ and $\text{Li}_2\text{Mn}_4\text{O}_9$ exhibit most of their capacity in the 3 V region that corresponds to insertion of additional lithium into the 16c sites. Nevertheless, the Jahn–Teller distortion has been shown to be delayed until late in the discharge process in both systems (33,38,39); the cubic symmetry without Jahn–Teller distortion has been shown to be preserved to $x = 2.5$ in $\text{Li}_{4+x}\text{Mn}_5\text{O}_{12}$ and $x = 1.7$ in $\text{Li}_{2+x}\text{Mn}_4\text{O}_9$.

Other Oxide Cathodes

The difficulties of the LiMn_2O_4 spinel also motivated the investigation of several nonspinel manganese oxides, particularly by employing low-temperature synthesis (33–35). LiMnO_2 obtained by conventional synthesis does not crystallize in the $O3$ structure of LiCoO_2 ; it adopts an orthorhombic rock salt structure in which the oxygen array is distorted from the ideal cubic close packing (42). However, LiMnO_2 isostructural with layered LiCoO_2 can be obtained by ion exchange of NaMnO_2 (43) or by partial substitution of Mn by Cr or Al (44). Unfortunately, both the orthorhombic LiMnO_2 and the layered LiMnO_2 ($O3$ structure) that have close-packed oxygen arrays tend to transform to spinel-like phases during electrochemical cycling. In this regard, $\text{Na}_{0.5}\text{MnO}_2$ —designated as $\text{Na}_{0.44}\text{MnO}_2$ in the literature—adopts a non-close-packed structure and has drawn some attention because it does not transform to spinel-like phases (45,46). However, only a small amount of lithium could be extracted from the ion-exchanged sample $\text{Na}_{0.5-x}\text{Li}_x\text{MnO}_2$ although additional lithium could be inserted into $\text{Na}_{0.5-x}\text{Li}_x\text{MnO}_2$. Therefore, it is not attractive for lithium-ion cells that use carbon anodes. Nevertheless, it has been shown that it is a promising candidate for lithium polymer batteries that employ metallic lithium anodes (45). Additionally, it has been shown that amorphous manganese oxides $\text{Li}_x\text{Na}_y\text{MnO}_z\text{I}_7$ synthesized in nonaqueous media exhibit high capacity (300 Ah/kg) and good cyclability (47,48). However, the capacity occurs across the wide voltage range of 4.3 to 1.5 V that has a continuously sloping discharge profile, which is not desirable for commercial cells. Not much lithium could be extracted from the initial material $\text{Li}_x\text{Na}_y\text{MnO}_z\text{I}_7$, and therefore, these amorphous oxides also are not attractive for lithium-ion cells fabricated using carbon anodes.

To improve the cyclability of the LiMn_2O_4 spinel, partial substitution of Mn by several other transition metals $M = \text{Cr}, \text{Co}, \text{Ni}$ and Cu in $\text{LiMn}_{2-y}\text{M}_y\text{O}_4$ has been pursued (49–52). These substitutions, however, result in the development of two plateaus that correspond to the removal of lithium from the 8a tetrahedral sites: one around 4 V that

corresponds to the oxidation of Mn^{3+} to Mn^{4+} and the other around 5 V that corresponds to the oxidation of the other transition-metal ions. The capacity in the 4-V region decreases, whereas that in the 5-V region increases as y in $\text{LiMn}_{2-y}\text{M}_y\text{O}_4$ increases. Although an increase in the cell voltage is attractive from the view point of energy density, the $\text{LiMn}_{2-y}\text{M}_y\text{O}_4$ oxides are prone to suffer from oxygen loss and safety concerns in the 5-V region.

LiVO_2 adopts the *O3* structure of LiCoO_2 . Although lithium can be readily extracted from LiVO_2 , the vanadium ions migrate to the lithium planes for $(1-x) < 0.67$ in $\text{Li}_{1-x}\text{VO}_2$ (53). Similarly, the LiV_2O_4 spinel is plagued by the migration of vanadium ions during the charge/discharge process (54). Interestingly, a number of other vanadium oxides such as $\text{VO}_2(\text{B})$ —a metastable form of VO_2 obtained by solution-based synthesis (55)—and V_6O_{13} (56) exhibit high capacity and good cyclability. However, these oxides that have no lithium are not attractive for lithium-ion cells made using carbon anodes. LiCrO_2 also crystallizes in the *O3* structure of LiCoO_2 , but it is difficult to extract lithium from this material.

Iron oxides offer significant advantages in both cost and toxicity compared to other oxides. Although LiFeO_2 obtained by conventional procedures adopts a different structure, layered LiFeO_2 that has the *O3* structure of LiCoO_2 can be obtained by an ion-exchange reaction of NaFeO_2 . Unfortunately, the layered LiFeO_2 does not exhibit good electrochemical properties because the high spin $\text{Fe}^{3+}:\text{3d}^5$ ion tends to move around between the octahedral and tetrahedral sites. This problem could, however, be overcome by designing complex iron oxides that consist of poly ions such as $(\text{SO}_4)^{2-}$ and $(\text{PO}_4)^{3-}$. For example, it was shown in the 1980s that $\text{Fe}_2(\text{SO}_4)_3$ that has a framework structure that consists of FeO_6 octahedra which share all six corners with $(\text{SO}_4)^{2-}$ tetrahedra has a capacity of around 130 Ah/kg and a flat discharge voltage of 3.6 V (57). However, the poor electronic conductivity of the Fe-O-X-O-Fe ($X = \text{S}$ or P) linkages leads to poor rate capability. Nevertheless, following this initial concept of using poly ions, LiFePO_4 that crystallizes in an olivine structure has recently been shown to be a promising material; it exhibits a flat discharge voltage of around 3.4 V (58). However, LiFePO_4 appears to suffer from limited rate capability due to inferior electronic (σ_e) and lithium-ion (σ_{Li}) conductivity compared to other oxide cathodes such as LiCoO_2 and LiMn_2O_4 .

Carbon Anodes

Carbon has become the material of choice for anodes in lithium-ion cells (59,60) due to its light weight and low electrochemical potential that lies close to that of metallic lithium (Fig. 7). Although the intercalation of alkali metals into graphitic carbon was known for some time, the recognition of a practical carbon anode began in 1989 (61). A subsequent announcement by Sony Corporation in 1990 commercializing lithium-ion cells that have carbon anodes and LiCoO_2 cathodes intensified the interest in carbon. Carbon materials can be broadly classified into two categories: soft carbon (graphitic carbon) and hard carbon (glassy carbon). The former has a better ordering of graphene layers compared to the latter, and they differ significantly in

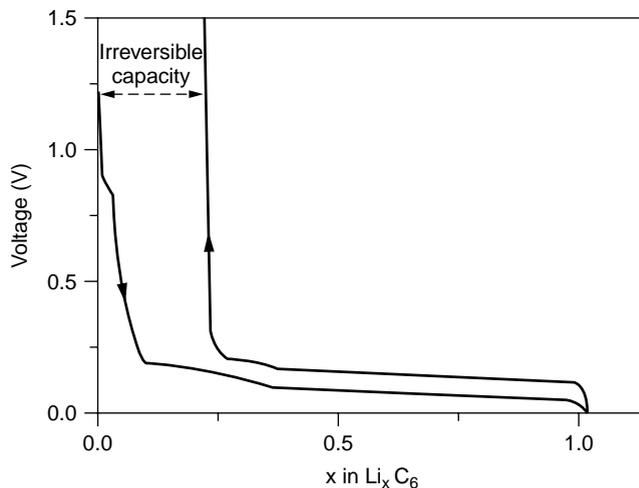


Figure 18. Typical first discharge/charge curves of graphite. The difference between the discharge and charge curves is irreversible capacity loss.

their physical and chemical properties. The hard carbons are typically obtained by thermal decomposition of, for example, phenolic and epoxy resins and products from petroleum pitch. They have significant amounts of open micropores, which tend to become closed when heated at increasingly higher temperatures. Some hard carbons have been shown to consist of single graphene sheets.

Figure 18 shows a typical first discharge/charge curve of graphite. It has a theoretical capacity of 372 Ah/kg that corresponds to an insertion of one lithium per six carbon atoms ($x = 1$ in Li_xC_6). It shows a significant amount of irreversible capacity during the first discharge/charge cycle due to unwanted, irreversible side reactions with the electrolyte. These side reactions can lead to disintegration of the carbon anode and a decrease in cycle efficiency. More importantly, using electrolytes that consist of propylene carbonate (PC), natural graphite cannot be charged because it leads to an evolution of gas at around 1 V. This is a major drawback of the graphite anode. However, using electrolytes that consist of other solvents such as ethylene carbonate (EC) and diethyl carbonate (DEC), the side reactions are suppressed, and it can be cycled without much difficulty.

The hard carbons generally have higher capacity than graphite. Several models have been proposed to account for the increased capacity: adsorption of lithium on both sides of the single graphene sheets, accommodation of extra lithium into nanometer size cavities, and storage of additional lithium at the edges and surfaces are some of the explanations. Generally, the capacity of carbon materials has been found to increase as the fraction of single layer material increases. However, the hard carbons have a sloping discharge profile between 0 and 1 V, unlike graphite, which has a nearly flat discharge profile between 0 and 0.3 V (Fig. 18). As a result, for lithium-ion cells that use the same cathode, the hard carbons will lead to a slightly lower cell voltage compared to graphite.

Structural modifications to improve the performance of carbon anodes are being pursued by various groups. In

this regard, texture control, surface modification by mild oxidation, and incorporation of other elements such as B, O, Si, and P have been studied. For example, carbon fibers that have different textures and compositions such as $C_{1-y-z}Si_yO_z$ and BC_2N have been investigated.

Other Anodes

The irreversible capacity loss encountered in carbon anodes has motivated the search for other anode hosts. In this regard, a few materials have drawn attention. $Li_4Ti_5O_{12}$ that has a cubic spinel structure accommodates three additional lithium ions per formula unit into the empty 16c sites and has negligible change in volume ($<0.1\%$) at a flat discharge voltage of around 1.5 V (62). However, the higher voltage (1.5 V) and lower capacity (160 Ah/kg) compared to that of carbon make it less attractive. SnO_2 is another candidate that has reversible capacity as high as 600 Ah/kg at 0 to 2 V, but it exhibits high capacity loss during the first cycle (63). More recently, some intermetallic compounds such as Cu_6Sn_5 that have an NiAs-type structure have shown promise (64), and further development of these systems is in progress.

However, all of the preceding candidates and carbon anodes do not contain lithium, and they require that the cathodes be in the lithiated (discharged) form. In this regard, anode hosts based on some lithiated transition-metal nitrides (59) and intermetallic compounds (65) are appealing. For example, $Li_{7-x}MnN_4$ and $Li_{3-x}FeN_2$ exhibit capacities of, respectively, 210 and 150 Ah/kg and a flat discharge voltage of around 1.2 V. $Li_{2.6-x}Co_{0.4}N$ and $Li_{2.6-x}Cu_{0.4}N$ whose structures are similar to that of Li_3N exhibit a much higher capacity of around 470 Ah/kg at 0.2 to 1 V; the formation of an amorphous phase during the initial stages in the latter systems leads to high capacity. Two lithium per formula unit can be reversibly extracted from Li_2CuSn (65). Further development work is necessary to assess the full potential of these lithium-containing anodes. If successful, they have the possibility of being coupled with some of the lithium-free cathodes such as vanadium oxides. However, the resulting lithium-ion cells may have a lower cell voltage compared to the currently available lithium-ion cells fabricated using $LiCoO_2$ cathodes and carbon anodes.

CONCLUDING REMARKS

After providing an introduction to the concepts of electrochemical cells and batteries, lithium-ion batteries that offer much higher energy density compared to other rechargeable systems were discussed. Lithium-ion batteries have become commercial due to the development of smart electrode (cathode and anode) host materials. State-of-the-art lithiumion cells are made by using a transition-metal oxide cathode such as a layered lithium cobalt oxide and a carbon anode. However, only 50 to 65% of the theoretical capacity of the currently available oxide hosts can be practically used. Chemical instability, structural instability, or safety concerns in the charged state limit the usable capacity, depending on the cathode host. The

challenge is to develop smart cathode hosts in which at least one lithium ion per transition-metal ion can be reversibly extracted/inserted and at the same time reduce the material cost; such a cathode can nearly double the energy density. There are also possibilities for increasing the capacity of anode hosts by focusing on amorphous materials. An alternate approach is to develop cells that use lithium-containing anodes and lithium-free cathodes. This strategy will allow using some of the already known high capacity cathodes such as vanadium oxides that have better chemical stability and safety characteristics. From the view points of safety, cycle, and shelf life, cathodes that have a lower voltage (3 to 4 V) but have a significantly increased capacity may be desirable for future applications.

ACKNOWLEDGMENTS

The author thanks the Welch Foundation (Grant F-1254) for financial assistance and Dr. A. M. Kannan, Seungdon Choi, and Ramanan Chebiam for their assistance with the figures.

BIBLIOGRAPHY

1. C.A. Vincent and B. Scrosati, *Modern Batteries: An Introduction to Electrochemical Power Sources*. Edward Arnold, London, 1984, p. 16.
2. D. Linden, ed., *Handbook of Batteries*. 2e, McGraw-Hill, NY, 1995.
3. J.O. Besenhard, ed., *Handbook of Battery Materials*. Wiley-VCH, Weinheim, 1999.
4. P.J. Gellings and H.J.M. Bouwmeester, *The CRC Handbook of Solid State Electrochemistry*, CRC, Boca Raton, FL, 1997.
5. D.R. Lide, ed., *CRC Handbook of Chemistry and Physics*. 81 e, CRC, Boca Raton, FL, 2000, p. 8.21.
6. J.E. Huheey, *Inorganic Chemistry: Principles of Structure and Reactivity*. Harper & Row, NY, 1972, p. 255.
7. Lithium Battery Energy Storage (LIBES) Publication, Technological Research Association, Tokyo, 1994.
8. M.S. Whittingham and A.J. Jacobson, *Intercalation Chemistry*. Academic Press, NY, 1982.
9. J.P. Gabano, *Lithium Batteries*. Academic Press, London, 1983.
10. H.V. Venkatesetty, *Lithium Battery Technology*. Wiley, NY, 1984.
11. G. Pistoia, ed., *Lithium Batteries: New Materials, Developments and Perspectives*, Vol. 5. Elsevier, Amsterdam, 1994.
12. C. Julien and G.A. Nazri, *Solid State Batteries: Materials Design and Optimization*. Kluwer, Boston, 1994.
13. M. Wakihara and O.Yamamoto, ed., *Lithium Ion Batteries: Fundamentals and Performance*. Wiley-VCH, Weinheim, 1998.
14. K. Mizushima, P.C. Jones, P.J. Wiseman, and J.B. Goodenough, *Mat. Res. Bull.* **15**: 783 (1980).
15. J.B. Goodenough, K. Mizushima, and T. Takeda, *Jpn. J. Appl. Phys.* **19**: 305 (1983).
16. M.M. Thackeray, W.I.F. David, P.G. Bruce, and J.B. Goodenough, *Mat. Res. Bull.* **18**: 461 (1983).
17. M.K. Aydinol and G. Ceder, *J. Electrochem. Soc.* **144**: 3832 (1997).

18. J.N. Reimers and J.R. Dahn, *J. Electrochem. Soc.* **139**: 2091 (1992).
19. A. Manthiram, R.V. Chebiam, and F. Prado, *Chem. Mater.* **13**: 2951 (2001), p. 269; R.V. Chebiam, A.M. Kannan, F. Prado, and A. Manthiram, *Electrochem. Commun.* **3**: 624 (2001).
20. L.A. Montoro, M. Abbate, and J.M. Rosolen, *Electrochem. Solid State Lett.* **3**: 410 (2000).
21. A. Hightower, J. Graetz, C.C. Ahn, P. Rez, and B. Fultz, *198th Meet. Electrochem. Soc. Phoenix, AZ, October 22–27, 2000*, Abstract No. 177.
22. G.G. Amatucci, J.M. Tarascon, and L.C. Klein, *J. Electrochem. Soc.* **143**: 1114 (1996).
23. G. Dutta, A. Manthiram, and J.B. Goodenough, *J. Solid State Chem.* **96**: 123 (1992).
24. A. Hirano, R. Kanno, Y. Kawamoto, Y. Tekeda, K. Yamamura, M. Takano, K. Ohyama, M. Ohashi, and Y. Yamaguchi, *Solid State Ionics* **78**: 123 (1995).
25. I. Nakai, K. Takahashi, Y. Shiraiishi, T. Nakagome, and F. Nishikawa, *J. Solid State Chem.* **140**: 145 (1998).
26. W. Li and J. Curie, *J. Electrochem. Soc.* **144**: 2773 (1997).
27. X.Q. Yang, X. Sun, and J. McBreen, *Electrochem. Commun.* **2**: 100 (2000).
28. L. Croguennec, C. Pouillier, and C. Delmas, *J. Electrochem. Soc.* **147**: 1314 (2000).
29. A.N. Mansour, X.Q. Yang, X. Sun, J. McBreen, L. Croguennec, and C. Delmas, *J. Electrochem. Soc.* **147**: 2104 (2000).
30. M. Balasubramanian, X. Sun, X.Q. Yang, and J. McBreen, *J. Electrochem. Soc.* **147**: 2903 (2000).
31. R.V. Chebiam, F. Prado, and A. Manthiram, *J. Electrochem. Soc.* **148**: A49 (2001).
32. J. E. Huheey, *Inorganic Chemistry: Principles of Structure and Reactivity*. Harper & Row, NY, 1972, p. 294.
33. M.M. Thackeray, *Prog. Solid State Chem.* **25**: 1 (1997).
34. A. Manthiram and J. Kim, *Chem. Mat.* **10**: 2895 (1998).
35. A. Manthiram and J. Kim, *Recent Res. Dev. Electrochem.* **2**: 31 (1999).
36. S.J. Wen, T.J. Richardson, L. Ma, K.A. Striebel, P.N. Ross, and E.J. Cairns, *J. Electrochem. Soc.* **143**: L136 (1996).
37. M.M. Thackeray, Y. Shao-Horn, A.J. Kahaian, K.D. Kepler, E. Skinner, J.T. Vaughey, and S.A. Hackney, *Electrochem. Solid State Lett.* **1**: 7 (1998).
38. M.M. Thackeray, A. de Kock, M.H. Rossouw, D.C. Liles, D. Hoge, and R. Bittihn, *J. Electrochem. Soc.* **139**: 363 (1992).
39. R.J. Gummow, A. de Kock, and M.M. Thackeray, *Solid State Ionics* **69**: 59 (1994).
40. J. Kim and A. Manthiram, *J. Electrochem. Soc.* **145**: L53 (1998).
41. S. Choi and A. Manthiram, *J. Electrochem. Soc.* **147**: 1623 (2000).
42. R.J. Gummow, D.C. Liles, and M.M. Thackeray, *Mat. Res. Bull.* **28**: 1249 (1993).
43. A.R. Armstrong and P.G. Bruce, *Nature* **381**: 499 (1996).
44. I.J. Davidson, R.S. McMillan, and J.J. Murray, *J. Power Sources* **54**: 205 (1995).
45. M.M. Doeff, T.J. Richardson, and L. Kepley, *J. Electrochem. Soc.* **143**: 2507 (1996).
46. Y.U. Jeong and A. Manthiram, *Electrochem. Solid State Lett.* **2**: 421 (1999).
47. J. Kim and A. Manthiram, *Nature* **390**: 265 (1997).
48. J. Kim and A. Manthiram, *Electrochem. Solid State Lett.* **2**: 55 (1999).
49. C. Sigala, D. Guyomard, A. Verbaere, Y. Piffard, and M. Tournoux, *Solid State Ionics* **81**: 167 (1995).
50. H. Kawai, M. Nagata, H. Takamoto, and A.R. West, *Electrochem. Solid State Lett.* **1**: 212 (1998).
51. Q. Zhong, A. Bonakdarpour, M. Zhang, Y. Gao, and J.R. Dahn, *J. Electrochem. Soc.* **144**: 205 (1997).
52. Y. Ein-Eli, W.F. Howard, Jr., S. H. Lu, S. Mukerjee, J. McBreen, J.T. Vaughey, and M.M. Thackeray, *J. Electrochem. Soc.* **145**: 1238 (1998).
53. L.A. de Picciotto, M.M. Thackeray, W.I.F. David, P.G. Bruce, and J.B. Goodenough, *Mat. Res. Bull.* **19**: 1497 (1984).
54. L.A. de Picciotto, M.M. Thackeray, *Mat. Res. Bull.* **20**: 1409 (1985).
55. C. Tsang and A. Manthiram, *J. Electrochem. Soc.* **144**: 520 (1997).
56. C. Lampe-Onnerud, J.O. Thomas, M. Hardgrave, and S. Yde-Andersen, *J. Electrochem. Soc.* **142**: 3648 (1995).
57. A. Manthiram and J. B. Goodenough, *J. Power Sources* **26**: 403 (1989).
58. A.K. Padhi, K.S. Nanjundaswamy, and J.B. Goodenough, *J. Electrochem. Soc.* **144**: 1188 (1997).
59. N. Imanishi, Y. Takeda, and O. Yamamoto, in *Lithium Ion Batteries: Fundamentals and Performance*, M. Wakihara and O. Yamamoto, eds., Wiley-VCH, Weinheim, 1998, p. 98.
60. M. Winter and J.O. Besenhard, in *Lithium Ion Batteries: Fundamentals and Performance*, M. Wakihara and O. Yamamoto, eds., Wiley-VCH, Weinheim, 1998, p. 127.
61. R. Kanno, Y. Takeda, T. Ichikawa, K. Nakanishi, and O. Yamamoto, *J. Power Sources* **26**: 535 (1989).
62. D.W. Murphy, R.J. Cava, S.M. Zahurak, and A. Santoro, *Solid State Ionics* **9 & 10**: 413 (1983).
63. W. Liu, X. Huang, Z. Wang, H. Li, and L. Chen, *J. Electrochem. Soc.* **145**: 59 (1998).
64. K.D. Kepler, J.T. Vaughey, and M.M. Thackeray, *Electrochem. Solid State Lett.* **2**: 307 (1999).
65. J.T. Vaughey, K.D. Kepler, R. Benedek, and M.M. Thackeray, *Electrochem. Commun.* **1**: 517 (1999).

BIOMEDICAL APPLICATIONS

JINHAO QIU
MAMI TANAKA
Tohoku University Sendai
Japan

INTRODUCTION

Biomedical applications of smart materials can be divided into three categories: (1) implants and stents, such as bone plates and marrow needles; (2) surgical and dental instruments, devices, and fixtures, such as orthodontic fixtures and biopsy forceps; and (3) devices and instruments for medical checkups, such as ultrasonic devices. The applications of the first category require strict biocompatibility of a material because it is implanted in the body for long periods. Among many traditional materials, including metals, alloys, and ceramics, that are available commercially, only a limited number are currently used as prostheses or biomaterials in medicine and dentistry. The applications in the second category require excellent mechanical

characteristics as well as biocompatibility. The third category is used mainly for transducers.

Among smart materials, the Ti–Ni shape-memory alloy (SMA) has attracted the most attention for biomedical applications in the first and second categories due to its excellent biocompatibility and mechanical characteristics. Research on biomedical applications of the SMA started in the 1970s with animal experiments initially, followed by clinical tests. The first example of a successful biomedical application of the SMA was a bone plate, which was used to repair broken bones. Now, many medical and dental applications of SMAs are available, and many new applications are being developed. On the other hand, piezoelectric materials have been widely used as transducers for medical ultrasonic devices due to their sensor function that uses piezoelectricity and the actuator function that uses inverse piezoelectricity.

In this article, the properties of SMAs for biomedical applications are discussed next, followed by some clinical examples. Recent examples of biomedical applications of SMAs are summarized there after, and finally the recent examples of biomedical applications of piezoelectric materials are summarized.

PROPERTIES OF SMAS FOR BIOMEDICAL APPLICATIONS

The properties of SMAs that are important and have led to its wide acceptance in biomedical applications are discussed in this section. Of these properties, biocompatibility, which simply means the ability of a material to be accepted by the body, is the most important, especially for implants. The other important properties include superelasticity, the shape-memory effect, hysteresis, and fatigue resistance. The properties of SMAs for biomedical applications are discussed in detail in (1).

Biocompatibility

The biocompatibility of a material is its most important property if it is used as prostheses or biomaterials in medicine and dentistry. Biocompatibility means that the material is nontoxic during the implanted period. Because all materials generate a “foreign body reaction” when implanted in the body, the degree of biocompatibility is related to the extent of this reaction. Due to the rigorous demands on material properties for biocompatibility, only these three metallic materials were qualified for use as implant materials: Fe–Cr–Ni, Co–Cr and Ti–Al–V before SMA. Investigations were carried out by many researchers on the biocompatibility of Ti–Ni (2), and an extensive review can be found in (1). The results of these studies show that Ti–Ni has superior corrosion resistance due to the formation of a passive titanium oxide layer (TiO_2) similar to that found on Ti alloys. This oxide layer increases the stability of the surface layers by protecting the bulk material from corrosion and creates a physical and chemical barrier to Ni oxidation.

In *in vitro* dissolution studies, Bishara et al. (3) found that Ti–Ni appliances release an average of $13.05 \mu\text{g/day}$ Ni in saliva, which is significantly lower than the estimated

average dietary intake of $200\text{--}300 \mu\text{g/day}$. In addition, the measured nickel blood levels of orthodontic patients who have Ni–Ti appliances show no significant increase during a 5-month period.

Shape-Memory Effect

Predeformed SMAs can remember their original shapes before deformation and can recover the shape when heated if the plastic deformation takes place in the martensitic phase (4,5). Shape recovery is the result of transformation from the low-temperature martensitic phase to the high-temperature austenitic phase when it is heated. The shape-memory effect makes it easy to deploy an SMA appliance in the body and makes it possible to create a prestress after deployment, when necessary. SMA appliances are first in a compact state during deployment and then restored to their expanded shape by heating. If the phase transformation temperature of an SMA is below body temperature, shape recovery can easily be induced by the heat of the body. When the phase transformation temperature is higher than the body temperature, SMA appliances are usually heated by warm salt water or a high frequency magnetic field.

In recent studies, the shape-memory effect has also been used for actuator functions in medical applications such as a urethral valve and artificial sphincter which are discussed later.

Superelasticity

SMAs exhibit superelasticity when they are in the austenitic phase (1,5). Figure 1 shows the typical superelastic stress–strain curve (solid line) compared with the stress–strain curve of stainless steel (dashed line). As shown in the figure, an important feature of superelastic materials is that they exhibit constant loading and unloading stresses across a wide range of strain. As shown in Fig. 1, the

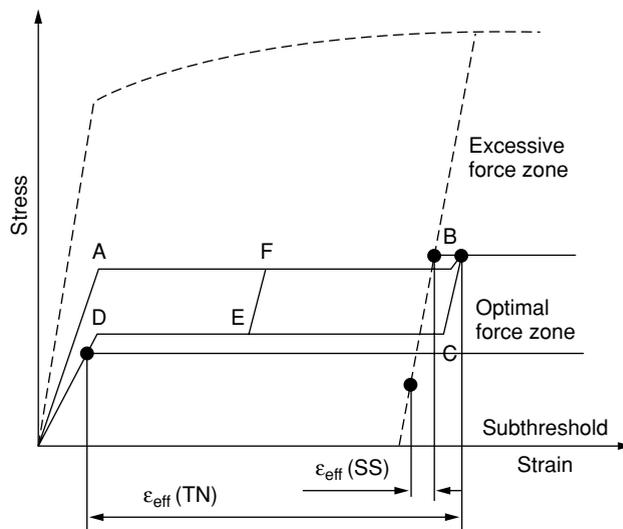


Figure 1. Typical stress–strain curve of superelastic materials and stainless steel. The superelastic materials exhibit constant unloading stress over a wide range of strain.

effective strain range $\varepsilon_{\text{eff}}(\text{TN})$ of Ti–Ni that corresponds to an optimal force zone is much larger than the $\varepsilon_{\text{eff}}(\text{SS})$ of stainless steel. Hence, a superelastic device can provide constant pressure even if the pressed part recedes by a limited amount during the installed period. On the other hand, the pressure exerted by an appliance made from stainless steel will drop drastically if the pressed part deforms, so that performance deteriorates. The orthodontic arch wire that is presented as an example of an application in the next section was the first product to use this property. Another example of applying this property is superelastic eyeglass frames (6). These eyeglass frames have become very popular in the United States, Europe, and Japan and are available in almost every optician's store. These frames can be twisted a full 180° , but more importantly the frames press against the head with a constant and comfortable stress. Not only is "fit" less important, but small bends and twists that may develop do not cause discomfort to the wearer.

The superelasticity of SMAs also makes it easy to deploy SMA stents. Stents made from stainless steel are expanded against the vessel wall by plastic deformation caused by inflating a balloon placed inside the stent. Ti–Ni stents, on the other hand, are self-expanding. More details can be found later.

Hysteresis of SMA

As shown in Fig. 1, superelastic SMA exhibits a hysteretic stress–strain relationship; the stress from A to B in the loading phase and the stress from C to D in the unloading are different. Hysteresis is usually regarded as a drawback in traditional engineering applications, but it is useful in biomedical applications. If the SMA is set at some stress–strain state, E, for example, upon unloading during deployment, it should provide a light, constant force against the organ wall, even under a certain amount of further strain release (e.g., from E to D). On the other hand, it would generate a large resistive force to crushing if it is compressed in the opposite direction because it takes the loading path from E to F. Hence, the SMA material exhibits a biased stiffness at point E, which is very important in designing a SMA stent. Because the stress at the loading phase from A to B and the stress in the unloading phase from C to D depend on the material composition of the SMA, the desirable stress–strain curve can be obtained by optimizing the material composition.

Anti-Kinking Properties

As shown in Fig. 1, the stress of stainless steel remains nearly constant in the plastic region. This means that a small increase in stress in the plastic region could lead to a drastic increase in strain or failure of a medical appliance made from stainless steel (1). On the other hand, the stiffness of superelastic Ti–Ni increases drastically after point B at the end of the loading plateau. The increase in stiffness would prevent the local strain in the high strain areas from further increasing and partition the strain in the areas of lower strain. Hence, strain localization is prevented by creating a more uniform strain than could be realized by using a conventional material.

EXAMPLES OF BIOMEDICAL APPLICATIONS

Orthopedic

Marrow Needles. Figures 2 and 3 show two types of marrow needles that are used in the repair of a broken thigh bone (4,6–8). When a stainless steel Kunster marrow needle is used, blood flow inside the bone can be blocked, and recovery can be delayed. It also has the drawback of low torsional strength. On the other hand, a Kunster marrow needle of SMA can be inserted into the bone in its initial straight shape and transformed to a curved shape by heating, as shown in Fig. 2. Hence, the SMA Kunster marrow

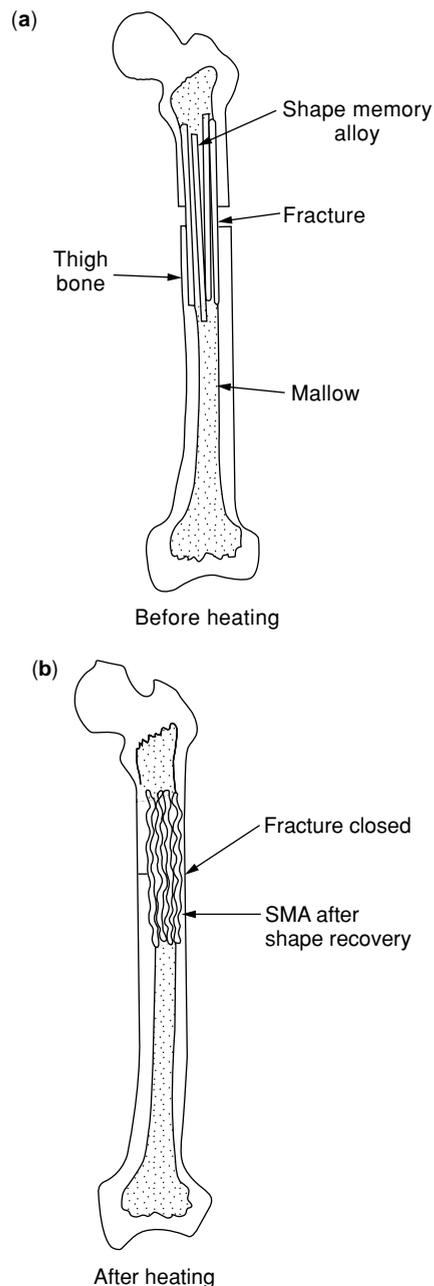


Figure 2. Kunster marrow needle (6).

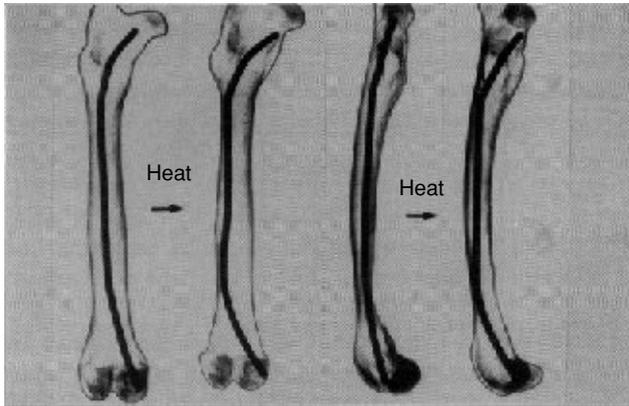


Figure 3. Marrow needles before and after heating (5,8).

needle can avoid the disadvantages of the stainless steel needle. The SMA Kunster marrow needle can also provide a compressive force on the fracture surfaces.

The marrow needle shown in Fig. 3 has a complicated shape for purposes of reinforcement, which makes it difficult to insert in the broken bone. Using the shape-memory effect, insertion can be greatly improved, as shown in the figure, without losing the reinforcing function, because the needles can be inserted in a simpler shape and the necessary size and shape are recovered by heating the needle in the marrow.

Currently available joint prostheses are made of bone cement to be fixed in the bone. Stress acting on the joint prosthesis is quite intense and severe: three to six times the body weight of the patient under nominal action, and the stress is cycled up to 10^6 times. Conventional bone cement causes several inconveniences: gradual loosening after implantation, resultant infection, and other complications. A prosthetic joint made of Ti-Ni SMA was developed to avoid such problems. High wear resistance is also another advantage of the Ti-Ni prosthetic joint.

Bone Staple and Bone Plate. The bone staple shown in Fig. 4 and the bone plate in Fig. 5 are used to fix broken bones (4-6). As shown in Fig. 4, a bone staple made of SMA can be inserted at low temperature in the holes opened in the bone, and then heated by the body temperature, it recovers its original shape to provide a compressive force on the surfaces of the broken bone. Bone plates are attached by screws for fixing broken bones. Bone plates made of Ti-Ni SMA are more effective in connecting broken bones than bone plates made of conventional material because SMA

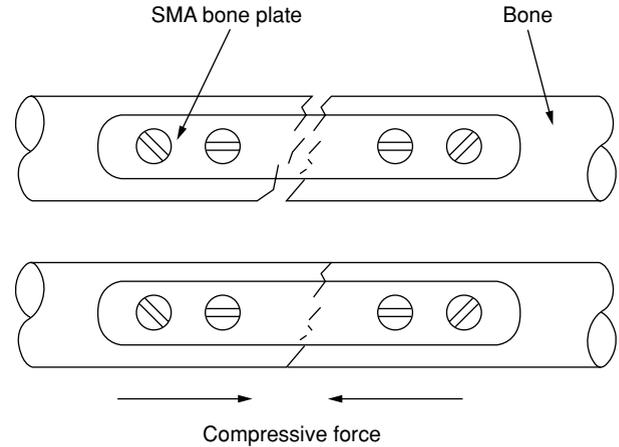


Figure 5. Bone plate used to fix broken bones (4-6).

bone plates provide compressive force on the fracture surface of the broken bones as well as repair, as shown in Fig. 5. Healing proceeds faster under uniform compressive force.

Dental Applications

Orthodontic Fixtures. Due to its superelasticity, Ti-Ni is used in many applications in dentistry. It is obvious that superelasticity gives the orthodontists better mechanical characteristics compared to conventional elastic materials such as stainless steel. Figure 6a,b shows a clinical example of orthodontic treatment using a superelastic Ti-Ni arch wire (5,9). When fixtures made of conventional elastic material such as stainless steel are used, the reforming force drops, and the fixture loosens due to movement of the teeth. Hence, the fixture must be replaced several times before the treatment is finished. An SMA fixture maintains a constant reforming force in a wide range of teeth movement due to its superelasticity, so that no replacement is required after the initial installation. Clinical results also showed faster movement of the teeth and a shorter chair time compared with stainless steel wire.

Tooth-root Prosthesis. Among several methods that restore the masticatory function of patients missing more than one tooth, a tooth-root prosthesis is considered the method that creates the most natural masticatory function. Blade-type implants made of Ti-Ni SMA, as shown in Fig. 7, have been used in Japan (5,6). The open angle of the blade is used to ensure tight initial fixation and to avoid accidental sinking during mastication. But to make the insertion operation easy, the tooth-root prosthesis is

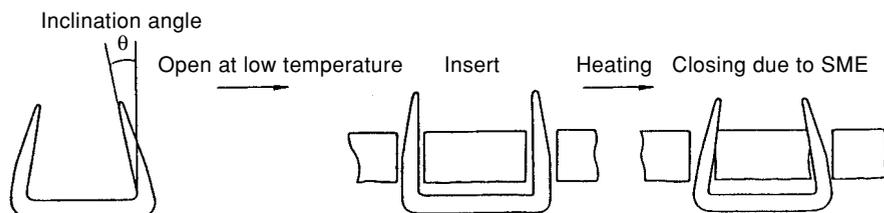
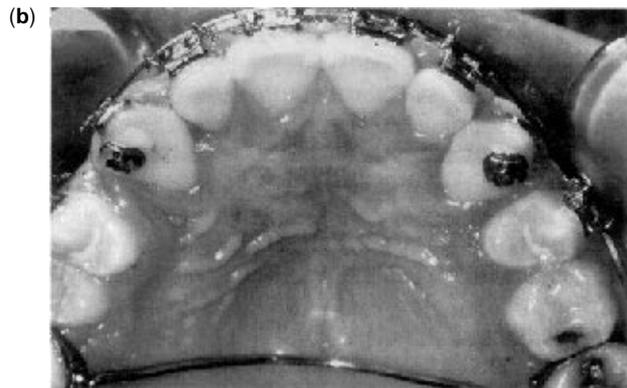


Figure 4. Bone staple used to fix broken bones (4,6).



Misaligned teeth before treatment



Normally aligned teeth after the first stage of treatment

Figure 6. Orthodontic treatment using a superelastic arch wire: (a) misaligned teeth before treatment; (b) normally aligned teeth after the first stage of treatment (5,9).

implanted in the jawbone as a flat shape and then the opened shape is changed by heating. Fig. 8 shows an X-ray photograph of the implanted tooth-root prosthesis. More than 5,000 clinical examples of SMA tooth-root prostheses have been reported.

Partial Denture. The key to a partial denture is the development of an attachment for connecting the partial

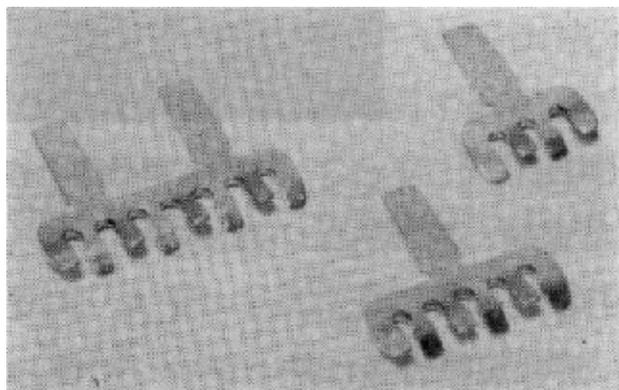


Figure 7. Tooth-root prosthesis (5,6).

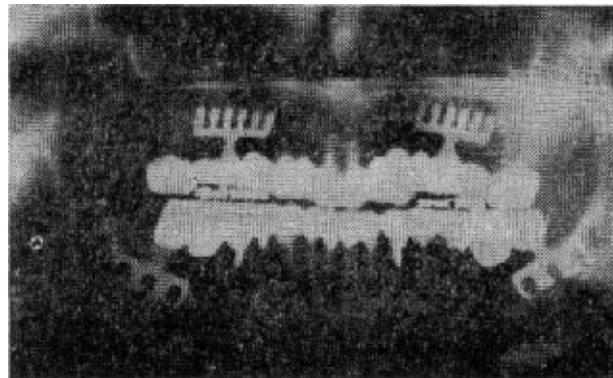


Figure 8. X-ray photo of implanted tooth-root prosthesis (6).

denture to the retained teeth. Clasps have been conventionally used for about a century as the attachment for a partial denture. One of the drawbacks of clasps made of conventional elastic materials is loosening during use; this can be improved by replacing the elastic materials by a superelastic Ti–Ni alloy (5,10). Another drawback of clasps is aesthetics because they are visible in the teeth alignment. To solve the problem, the size of the attachment must be smaller than the width of the teeth so that it can be embedded completely in the teeth. A precision attachment using a small screw has recently become available, but it has to be designed and fabricated very precisely so that it lacks flexibility to follow the change in the setting condition during long-term use due to the shape change of the jawbone. Because of its flexibility, this problem can be solved by using an attachment made of SMA.

The SMA attachment consists of two parts: a fixed part that is made of a conventional dental porcelain-fusible cast alloy and is attached to the full cast crown on the anchor teeth and a movable part that is made of Ti–Ni SMA and is fixed on the side of the partial denture. Examples of movable and fixed parts are shown in Fig. 9.

Surgical Instruments

Since superelastic tubing became available in the early to mid-1990s, a variety of catheter products and other endovascular devices using Ti–Ni has appeared on the market. Early applications of Ti–Ni were retrieval baskets that have Ti–Ni kink-resistant shafts, as well as a superelastic basket to retrieve stones from kidneys, bladders, and bile ducts. An interesting example is the intraaortic balloon pump (IABP) used in cardiac assist procedures (Fig. 10). The use of Ni–Ti allowed a reduction in the size of the device compared with the polymer tube designs and increased the flexibility and kink resistance compared with stainless steel tube designs (1).

Biopsy forceps made from stainless steel are very delicate instruments that can be destroyed by even very slight mishandling. Ti–Ni instruments, on the other hand, can handle serious bending without buckling, kinking, or permanent deformation. Figure 11 shows a 1.5-mm biopsy forceps that consists of thin wall Ti–Ni tubing and a Ti–Ni actuator wire inside. Together they can be bent around a

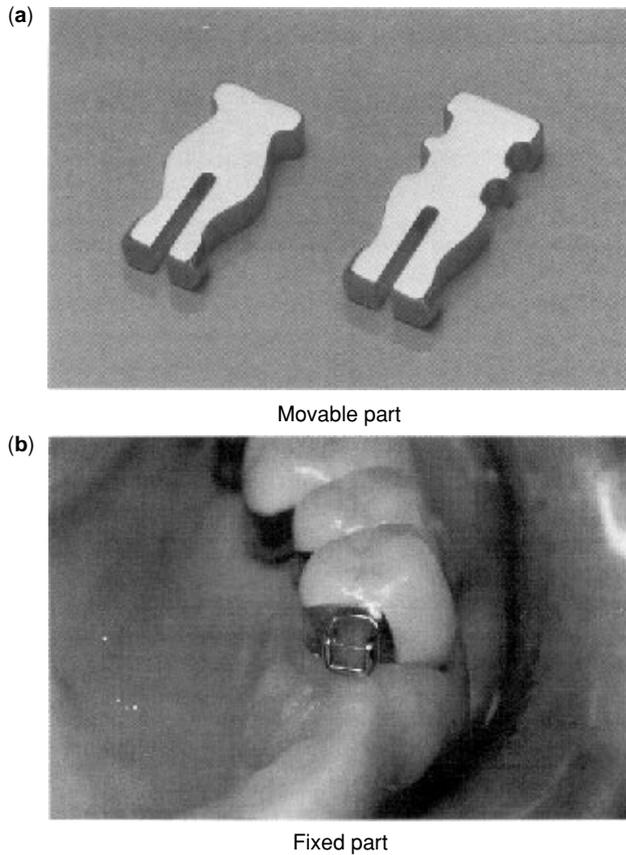


Figure 9. Shape-memory alloy attachment for partial denture; (a) movable part; (b) fixed part (4,5).

radius of less than 3 cm without kinking, and still allow for the opening and closing of the distal grasper jaws without increased resistance. The instrument continues to operate smoothly even while bent around tortuous paths.

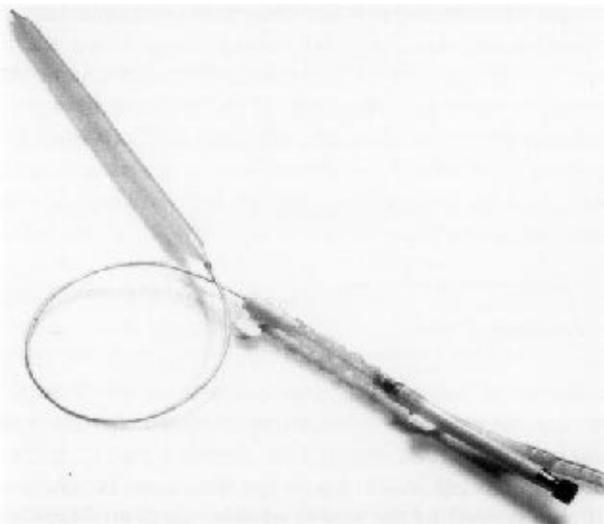


Figure 10. The Arrow interaortic balloon pump uses a Nitinol™ tube to pressurize the balloon (1).

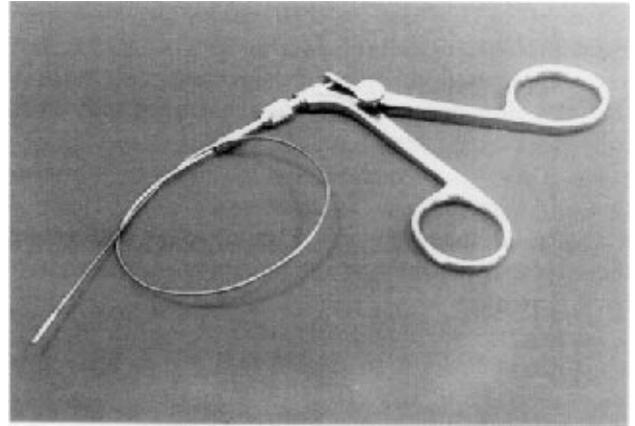


Figure 11. Ninitol™ tubing that has an internal actuating wire allows this 0.8-mm diameter grasper to operate while tied in a knot (1).

Stent

The term stent is used for devices that are used to scaffold or brace the inside circumference of tubular passages or lumens, such as the esophagus biliary duct, and most importantly, a host of blood vessels, including coronary, carotid, iliac, aorta, and femoral arteries (Fig. 12) (1). Stenting in the cardiovascular system is most often used as a follow-up to balloon angioplasty, a procedure in which a balloon is placed in the diseased vessel and expanded to reopen a clogged lumen. Ballooning provides immediate improvement in blood flow, but 30% of the patients have restenosed within a year and need further treatment. The placement of a stent immediately after angioplasty, it has been shown, significantly decreases the propensity for restenosis. Stents are also used to support grafts, for example, in treating aneurysms (Fig. 13).

Most stents today are stainless steel and are expanded against a vessel wall by plastic deformation caused by inflating a balloon placed inside the stent. Ti–Ni stents, on the other hand, are self-expanding—they are shape-set to the open configuration, compressed into a catheter, then

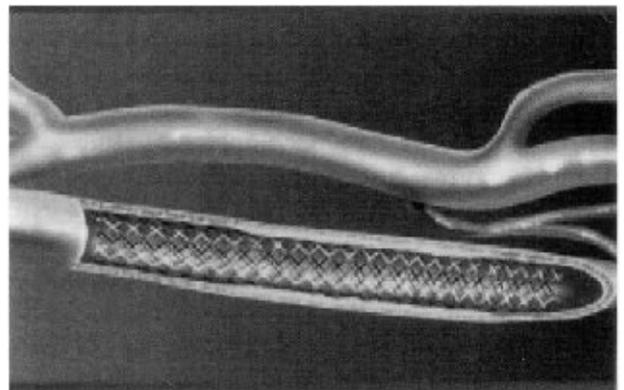


Figure 12. A stent that maintains vessel patency and blood flow to the brain is portrayed in a cutaway view of the internal carotid artery (1).

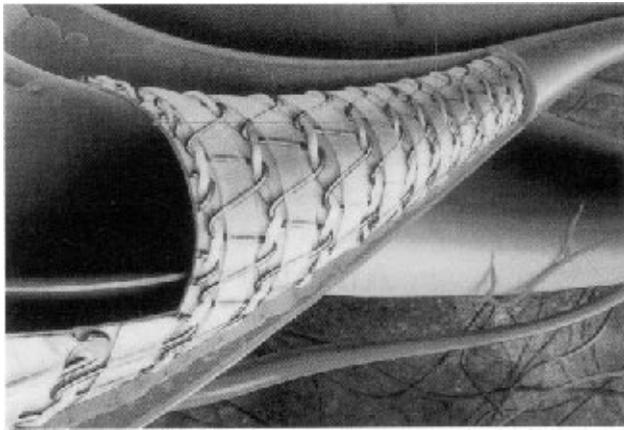


Figure 13. Stentgrafts used to exclude aneurysms, to provide an artificial replacement for injured vessels, or prevent restenosis after angioplasty (1).

pushed out of the catheter and allowed to expand against a vessel wall. Typically, the manufactured stent's outer diameter is about 10% greater than the vessel's diameter to ensure that the stent anchors firmly in place. The flexibility of Ti-Ni is about 10–20 times greater than stainless steel, and it can bear a reversible strain as high as 10%. Ni-Ti stents are made from knitted or welded wire, laser-cut or photoetched sheet, and laser-cut tubing. The preferred devices are laser-cut tubing, thus avoiding overlaps and welds (Fig. 14).

CURRENT BIOMEDICAL APPLICATIONS OF SMA

Artificial Urethral Valve

Urinary incontinence is the involuntary discharge of urine caused by weakness of the urinary canal sphincter muscles due to aging and expansion of the prostate gland. However, the difference in ages, the sex of patients, and the various causes of the disorder make it difficult to treat the disease simply by drugs or surgery. In this section, an artificial urethral valve system driven by an SMA actuator is introduced (11).

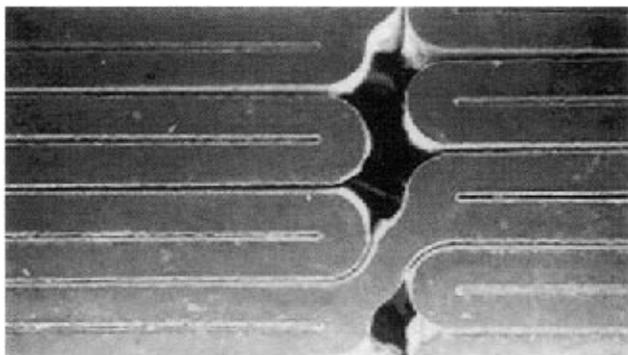


Figure 14. Stents made from laser-cut tubing (1).

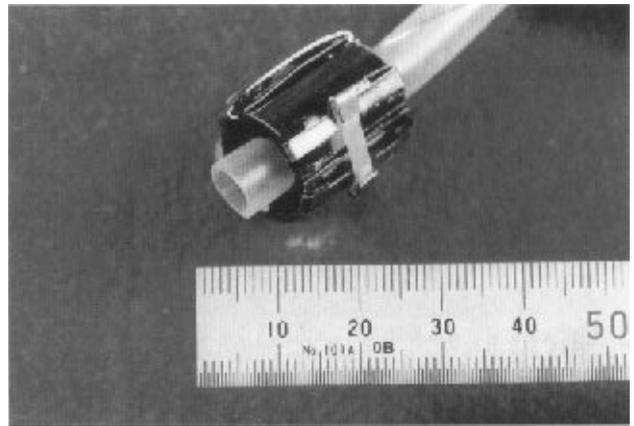


Figure 15. Artificial urethral valve.

Urethral Valve. The artificial urethral valve should be compact and should have no protrusions when it is implanted in the lower abdominal region. In addition, it should be attachable onto various sizes of urethrae. A compact urethral cylindrical valve is presented in Figs. 15 and 16. The valve is 15 mm across and 20 mm long. It is composed of two semicircular stainless steel shells 0.2 mm thick and a 0.2-mm circular-arc Nitinol™ plate. The shells and the Nitinol™ plate are fixed together by stainless steel clamps. Further, a cylindrical sponge rubber filling is placed inside the valve to effect uniform contact between the valve and the urinary canal. In the normal state, the valve presses on the canal, so that it is choked. To free the canal, the valve is opened by actuating the SMA element. The SMA plate, which is cylindrical at body temperature, flattens as heat is increased, and the valve, which is closed by the force of the bias spring in the normal state, is opened to release the choked urethra and allow urinary flow. To heat the SMA, a Nichrome™ wire, insulated by a polyimide membrane, was placed on the surface of the nitinol plate.

Transcutaneous Energy Transformer System. The energy to drive an in-dwelled valve should be supplied from outside the body. A transcutaneous energy transformer system will be effective for this purpose (12,13). The system consists of two induction coils that transmit electrical energy wirelessly from the primary to the secondary coil. In this study, the primary coil was 70 mm in diameter and had 12 turns, and the secondary coil was 60 mm across and had eight turns. The coils are spirals formed of twisted wires that consist of 20 lengths of copper wire 0.2 mm in diameter.

Experimental Setup. Figure 17 is the schematic of an animal experiment that used the urethrae of male dogs of average weight 12 kg, whose thickness is similar to that of human male urethrae. The urinary canal uncovered by the cut was equipped with the SMA valve and then loaded with water at a hydrostatic pressure of 75 cmH₂O, which is comparable to human abdominal muscle pressure. The temperature of the valve and the flow rate of the water passing through the canal were measured. First, the valve

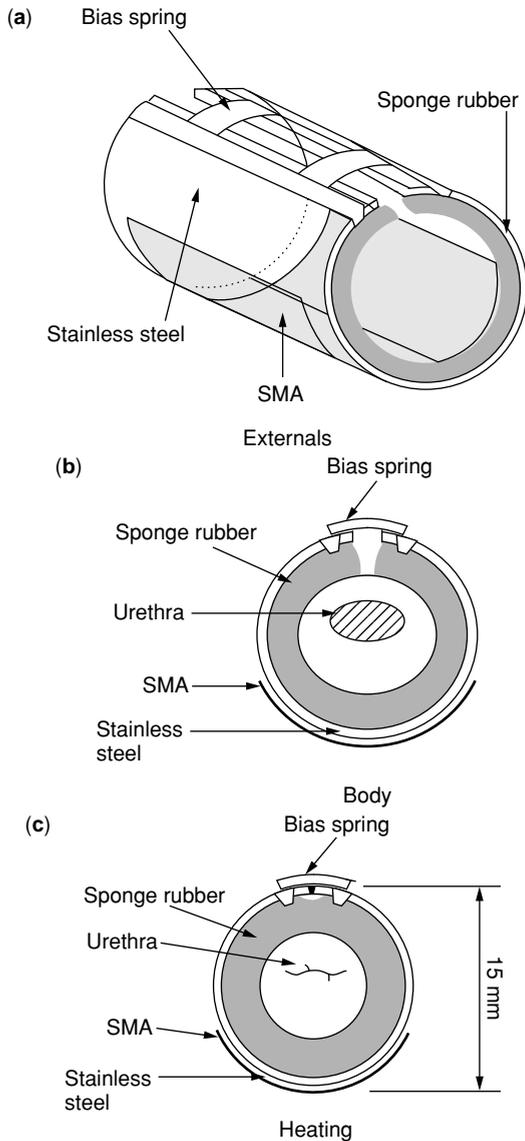


Figure 16. Geometry of urethral valve: (a) externals; (b) body; (c) heating.

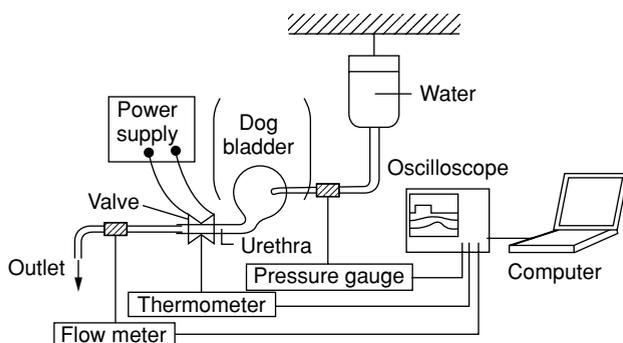


Figure 17. Schema of animal experiment.

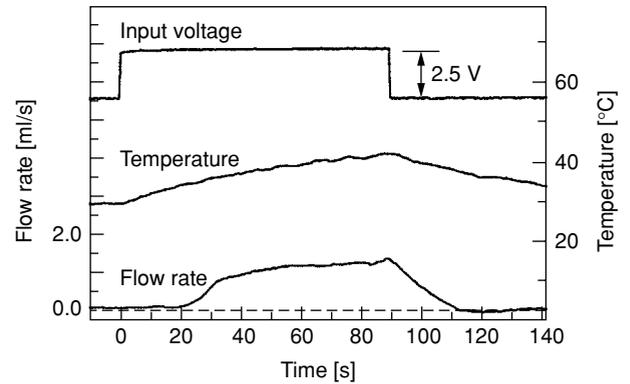


Figure 18. Time variations of applied direct current, valve temperature, and flow rate of water passing through the valve.

was heated by direct current from the transcutaneous energy transformer system and then it was left to cool naturally by cutting the electric power to the heating wire. In the energy transformer system, the secondary coil was placed under the abdominal wall, and the primary coil on the skin face-to-face with the secondary coil. The capacity of the condenser to enhance the resonant effect of the circuit was set at $C_2 = 1.0 \mu\text{H}$. The frequency of the input current, which should coincide with the resonant frequency of the circuit, was 90 kHz. Further, the peak-to-peak amplitude of the current was set at 25 V.

Results and Discussion. Figure 18 shows the opening/closing functions of the valve when it was driven by the direct current heating of the SMA. The input current to the heating wire on the SMA, the temperature of the valve, and the flow rate of water passing out through the urethra are plotted as functions of time. The valve started to open, and the water began to flow 20 seconds after electric current was supplied. The current was cut off when the temperature of the valve reached 45°C . It is seen that the valve completely intercepted the water flow 30 seconds after the electric current was cut off. Thus, the valve possesses the necessary function of an artificial urethral valve.

Figure 19 illustrates the motor functions of the valve when it was driven by the transcutaneous energy transformer system.

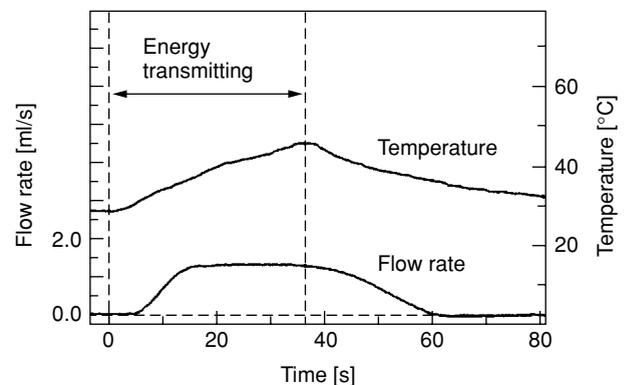


Figure 19. Opening/closing functions of valve driven by transcutaneous energy transformer system.

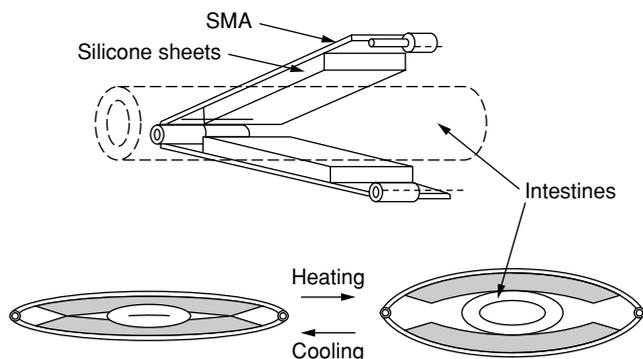


Figure 20. Schematic of the artificial sphincter using SMA.

transformer system. The distance between the primary and secondary coils was 3–4 mm, and the induced current to the heating wire was 7.6 V in the peak-to-peak amplitude. The figure shows that the valve started to open and released the choked canal 6 seconds after the electric current was supplied. The current was cut off when the temperature of the valve reached 45°C, which was 35 seconds after the current supply was started. It is shown that the water flow was intercepted 25 seconds after the current was cut off. These motor functions of the valve showed that the induction coil system worked well as a transcutaneous energy transformer for heating the artificial urethral valve.

Artificial Sphincter

Similar to the urethral valve, the development of an artificial sphincter is also required for the medical treatment of patients who have fecal incontinence due to a colostomy, a congenitally anorectal malformation, or surgical operations for anorectal diseases. The lack of an anal sphincter is the main reason for the problem. In this section, an artificial sphincter using a SMA actuator was one of the solutions of these problems, as introduced in (14).

Mechanism of Artificial Sphincter. A schematic drawing of the proposed artificial sphincter is shown in Fig. 20. The actuator consists of two SMA plates joined by two hinges and the heating coils attached to the SMA plates. The SMA plates are 70 mm long. The width and the thickness are 18 and 0.7 mm, respectively. The material used for the SMA plates, Ti51at%Ni, is known to exhibit an all-round shape-memory effect (ARSME); it reverses shape to the “memorized” shape in its martensitic phase. Because the highest temperature for the complete reverse transformation might reach 55°C, thermally insulated materials such as cork sheets and sponge rubber sheets cover the outer and inner sides of the SMA plates, respectively. When electric power is applied to the coils for heating, the reverse transformation occurs in the SMA plates, accompanied by shape changes from a flat shape to an arc, the restrained shape during annealing. The shape change results in a gap between the two SMA plates to open the intestines. After switching off the electric power, the shape of the SMA plates recovers by natural cooling, and the intestines are closed again.

Animal Experiments. Animal experiments have been conducted to examine the fundamental functions and the biocompatibility of the actuator. The experiments were carried out on a pig that had the same dimensions of intestines as those of the human body.

Preliminary tests of the fundamental functions of the actuator were conducted. As illustrated in Fig. 21, an artificial anus was made in the pigs abdomen using a rectum after the resection of its colons. Then the artificial sphincter was installed around the rectum located between the abdominal wall and the peritoneum (see Fig. 22). The inside pressure of the rectum generated by the artificial sphincter was measured first. A pressure sensor was inserted from the artificial anus into the intestine and then moved out. The pressure exhibited a rise of 50 mmHg in the region clipped by the actuator, corresponding to the width of the SMA plates: 18 mm (see Fig. 23). The pressure tests were also carried out when the inner part of the rectum was

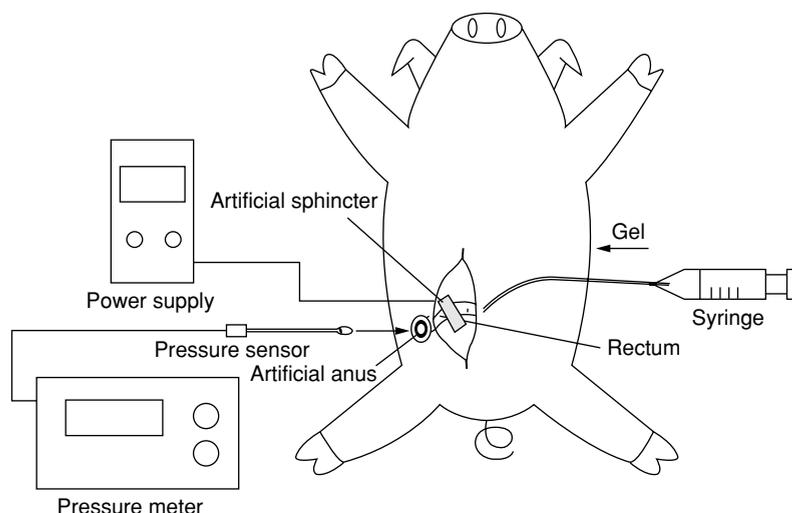


Figure 21. Schema of animal experiment.

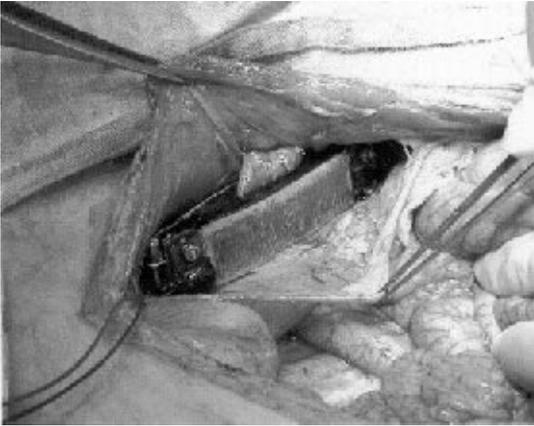


Figure 22. A photograph of the artificial sphincter installed in the rectum of a pig.

subjected to pressure. The inner pressure of the rectum was generated by pouring gel into the intestines. Figure 24 shows the pressure change measured during the test. As seen in Fig. 24, the pressure was increased to 75 mmHg without any leak of the gel, and then decreased by opening the artificial sphincter using electric power. Discharge of the gel from the artificial anus was observed.

A clinical test on a living pig has also been carried out. The artificial sphincter enables a controlled bowel movement of the pig. From the dissecting examination after six days of experiments, neither infections nor burn scars were observed in the body around the artificial sphincter. Furthermore, the artificial anus was in good condition; this suggested that the pressure due to the artificial sphincter was tolerable. Additionally, in these experiments,

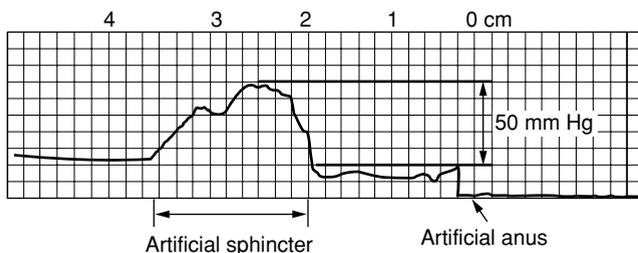


Figure 23. Pressure distribution in an intestine clipped by the artificial sphincter.

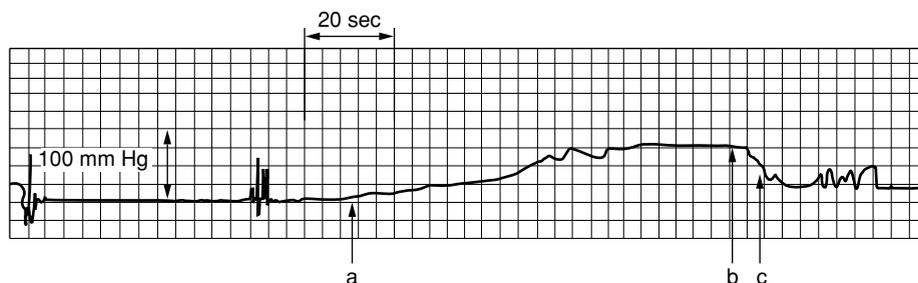


Figure 24. Pressure change in the rectum: (a) start pouring gel into the rectum; (b) switch on the input power; (c) discharge of gel was observed.

electric power was supplied through a pair of current lead wires that penetrated the body. The use of an inductive power transmission for the power supplement of the actuator could lead to a complete implantation-type artificial sphincter based on further research.

CURRENT BIOMEDICAL APPLICATIONS OF PIEZOELECTRIC MATERIALS

Due to its piezoelectricity, a piezoelectric material has a sensor function that can convert a mechanical signal to an electric signal. Electrical voltage generated by mechanical stress in piezoelectric materials decays very fast due to charge dissipation. The voltage signal takes the form of a very brief potential wave at the onset of the applied force and a similar brief wave at termination. It increases as force is applied but drops to zero when the force remains constant. There is no response during the stationary plateau of the applied stimulus. Voltage drops to a negative peak as the pressure is removed and subsequently decays to zero (15). The response is quite similar to the response of the Pacinian corpuscle in the human skin (16), one of the sensory receptors in the dermis. PVDF (polyvinylidene fluoride) piezofilm is suitable for uses in the biomedical field, because it is very flexible and sensitive to the fast variation of stress or strain. In this section, several recent studies of medical applications of PVDF film are introduced.

Active Palpation Sensor for Detecting Prostatic Cancer and Hypertrophy

Prostatic carcinoma and hypertrophy are examined in general by rectal palpation where the doctor's index finger is used as a probe together, in most cases, with ultrasonic tomography. The two lesions found in this study are diagnosed by noticing their morphological features. Prostatic hypertrophy is a symmetrical enlargement of the prostate glands; the stiffness varies from soft to hard. Prostatic cancer, on the other hand, is a hard asymmetrical uneven tumor. Palpation depends on the tactile perception of the forefinger, which is said to be ambiguous, subjective, and much affected by the physician's experiences. Hence, the development of a palpation sensor for detecting prostatic cancer and hypertrophy is important.

Palpation Sensor and Measurement System. The geometry of an active palpation sensor and tip probe are presented

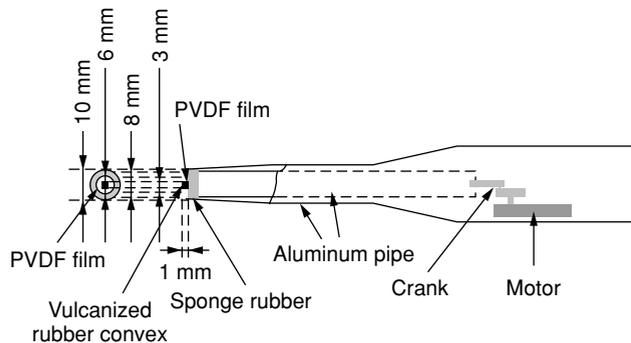


Figure 25. Active palpation sensor that has a recessed sensor head.

in Figs. 25 and 26. The tip probe is mounted on a linear z-translation aluminum bar. It fits into a cylindrical outer aluminum shell 15 mm in diameter and is driven by a dc micromotor and crank mechanism. The drive mechanism is essentially the same as that of an electric toothbrush. The probe is positioned so that its face is to the prostate gland, and, it is oscillated at about 50 Hz at a constant peak-to-peak amplitude of 2 mm. The probe is an assembly of layered media.

The base is a thin aluminum circular plate 10 mm in diameter, on which a cylindrical rubber sponge 8 mm in diameter and 4 mm thick, a PVDF piezopolymer film 6 mm across and 28 μm thick as the sensory receptor, and a thin acetate film as a protect cover for the piezopolymer film are stacked in sequence. Furthermore, a convex layer of vulcanized rubber 3 mm across was placed on the surface of the acetate film to enhance the sensitivity of the sensor (17).

In the experiment, the sensor head is pressed sinusoidally against the object, and the output signal from the piezopolymer film is collected for 100 ms, sent to a digital storage oscilloscope for every sampling time of 0.2 ms, and further forwarded to a personal computer via a GP-IB interface as 500 eight-bit data for processing.

Signal Processing. The output voltage from the piezopolymer film is proportional to the rate of the strain induced in the film, which means that the maximum amplitude of the signal from the sensor is rather superposed by noises from the measuring system. Bearing this fact in mind, the following data analysis can be done by using the absolute output signal of the sensor integrated across the period of data collection:

$$I = \sum_{n=1}^N |V(n)|, \quad (1)$$

Table 1. Average μ , Minimum I_{\min} , Maximum I_{\max} , and Standard Deviation σ of Output

	μ	I_{\min}	I_{\max}	σ
Elastic soft or elastic firm	0.93	0.8	1.0	0.11
Elastic firm	1.72	1.0	3.4	0.75
Hard	16.0	11.5	18.8	2.40

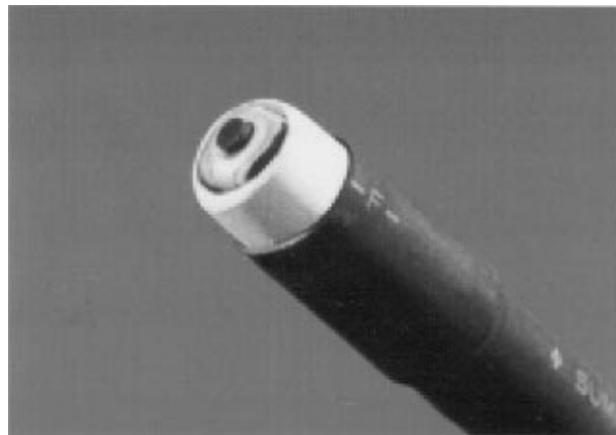


Figure 26. Geometry of a sensory receptor.

where I is the integrated output signal and N ($= 500$) is the total number of data collected in a period of 100 ms.

Clinical Test. Rectal palpation was done by using the sensor to verify the discrimination function of the sensor. First, the doctor examined the stiffness of the prostate gland of a person by using his own index finger. Next, the sensor protected by a medical rubber glove was inserted into the subject's rectum. The relative position of the sensor to the prostate gland was monitored by the ultrasonic diagnostic, and the sensor was placed face to face to the prostate gland and driven to generate sensor output. The examinees in this time were eight in total. One 74-year-old person was suffering from a carcinoma. Four persons of average age 77 were patients who had prostatic hypertrophy, and two persons (average age 52) suffered from prostatitis. The last person diagnosed had no definite lesions on his prostate gland. The stiffness of the prostate glands diagnosed this time was "elastic soft or elastic firm," "definitely elastic firm," and "hard." The results obtained are presented in Table 1, which shows the average, the minimum and maximum output, and the standard deviation of the output. It is seen clearly that the output from the prostate gland of the diagnosis "hard" is greater than the output of the gland of "elastic firm."

Haptic Sensor for Monitoring Skin Conditions

Assessing the pharmaceutical action of liniments on skin disease is a matter of importance to dermatologists. This has drawn much attention to the development of objective techniques for measuring the morphological features of skin (18). Evaluating the substantiation of the cosmetic efficacy of toiletries is another of the objective measuring techniques. The features that affect the health appraisal

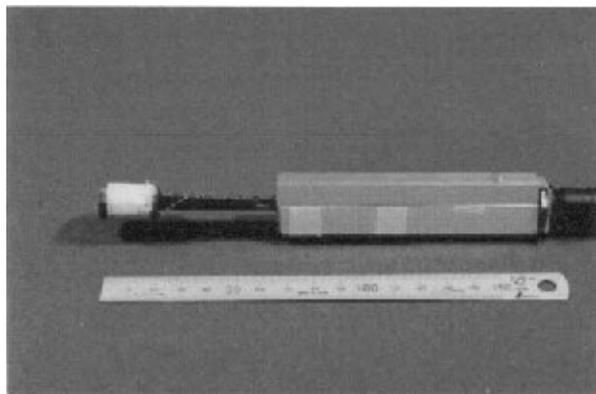


Figure 27. Tribosensor.

and/or the physical beauty of skin are the morphology such as rashes, chaps, or wrinkles. This noninvasive technology has made great advances in the studies of dermatology during the last decade, and several methods have been developed to measure the mechanical properties of the dermis such as measuring transepidermal water loss using an evaporimeter (19) and image processing of a negative replica of the dermis (20). These methods, however, fall in the category of indirect measuring techniques for the dermis. In this section, the development of haptic tribosensors for monitoring skin conditions and distinguishing atrophic and normal healthy skins directly is introduced.

Tribosensor. A tactile sensor for measuring skin surface conditions is presented in Figs. 27 and 28. The sensor is a layered medium, whose construction is analogous to that of the human finger. It is composed of an aluminum shell as the phalanx, sponge rubber 3 mm thick as the digital pulp, a PVDF piezopolymer film 28 μm thick and 12 mm across as the sensory receptor, an acetate film as a protective cover for the piezofilm, and gauze on the surface as the fingerprint that enhances the tactile sensitivity of the sensor. The sensor was attached to the tip of an acrylic elastic beam, and a strain gauge was mounted on the surface of the beam to monitor the force applied to the skin by the sensor.

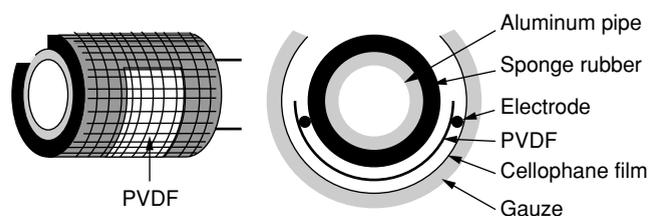


Figure 28. Schematic of PVDF piezofilm sensor.

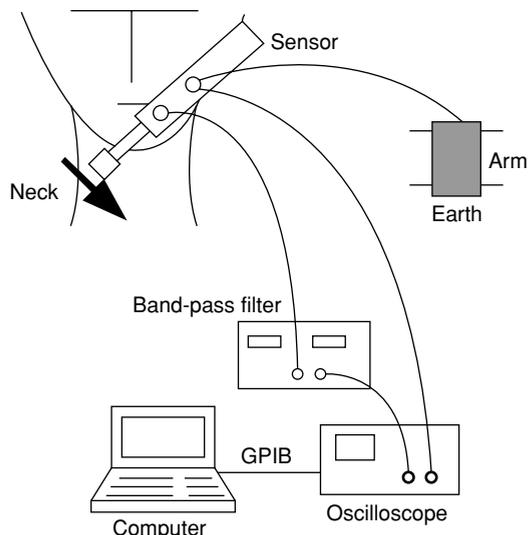


Figure 29. Setup of measuring instruments.

Measuring Apparatus. A measurement system using the tribosensor is presented in Figure 29. The sensor was moved by hand over the sample skin to maintain a constant speed and force. The voltage signal from the PVDF sensory film was sent to a digital storage oscilloscope as eight-bit quantized digital signals of 4096 points and then transmitted to a personal computer via a GPIB board for signal processing. The sampling frequency was held constant at 40 kHz. In the process of measurement, it was necessary to reduce the potential difference between the surface of the skin and the sensor to minimize the overlap of noises on the sensor signal. To this end, the subject and the sensor were grounded by fitting a grounding conductor around the wrist. Still, some noises from the power sources, which were due to outside sources, were observed around 50 Hz and 100 MHz, respectively. Thus, a band-pass filter that had cutoff frequencies of 70 Hz and 20kHz was inserted after the sensor to remove the noise effect.

Signal Processing and Identification of Skins. Dermatitis is a factor that affects the skin condition. The neck skin of subject A, who was suffering from mossybacked atopic dermatitis, was compared with the healthy skins of subjects B–E. The sensor was moved by hand over a prescribed region of skin to maintain a constant speed and force. Hence, a method of identifying sample skins employing signal processing and neural network-based training was introduced. First, the variance was calculated as an index to extract the features of the collected data:

$$V = \frac{1}{N} \sum_{j=1}^N (x(j) - \bar{x})^2, \quad (2)$$

Table 2. Segmented Frequency Range for Power Spectrum Integration

Level	1	2	3	4
Frequency range (Hz)	29–88	98–586	596–1191	1201–19990

Table 3. Rate of Correct Answer for Discrimination of Atopic and Normal Skins

Input Parameter	P_i		P_i, V	
Cycle of Training	10^3	10^4	10^3	10^4
A	0%	0%	100%	100%
B	100%	100%	100%	100%
C	100%	100%	100%	100%
D	100%	100%	100%	100%
E	100%	87%	100%	100%

where, $x(j)$ is the j th quantitized digital signal, \bar{x} is the average of $x(j)$, and N is the total number of digital signals.

Next, FFT analysis was introduced to extract the features of the collected data. Careful reading of the spectra obtained led to an understanding that there were some differences in the distribution profiles of spectra among the subjects. In the light of this, the power spectrum was segmented into several frequency ranges, and the ratio of the power in an individual frequency range to the power of the whole frequency range was calculated and used as the second index that described the characters of the sample skin:

$$(R_s)_i = (S)_i/S, \quad (S)_i = \int_i P(f)df, \quad S = \sum_i (S)_i, \quad (3)$$

where $P(f)$ is the power spectrum density, S is the total power distributed across the frequency range considered, and $(S)_i$ is the power at frequency level i . Here, the level $i(=1-4)$ stands for the range of frequencies given in Table 2.

Discrimination of Skin that has Atopic Dermatitis. The neural network was trained to recognize atopic and normal healthy skins. The neural network employed was a hierarchical network; the training method used was instructor-assisted training, in which the correct answers were provided as an aid to the training; the training algorithm employed was back-propagation. Two networks were examined. One was the network that used the four power ratios $(R_s)_i$, $i = 1-4$, as the cell input, and the other used the variance V in addition to $(R_s)_i$. Five sets of data were obtained for each skin. The neural network had an input sensory layer of four or five cells, an intermediate association layer that had the same number of cells as the input layer, and an output response layer that had a single cell. The correct output value was set to one for the atopic skin, and zero was set for the healthy skin. Training was repeated by using the gradient descent method until the error function became sufficiently small, up to 10^3 and 10^4 cycles. The networks were constructed separately three times for three sets of randomly selected initial values of synaptic weights. After completing the training of 12 individual networks, the recognition experiments were performed on the atopic skin and the healthy skin using the newly obtained five sets of data for each subject. The results obtained are

shown in Table 3. Here, the percentage denotes the rate of the correct answer averaged across the three networks of randomly selected initial synaptic weights. Atopic dermatitis cannot be identified without the input of the variance, and the recognition was perfect when the variance was introduced as the input.

BIBLIOGRAPHY

1. T. Duerig, A. Pelton, and D. Stockel, *Mater. Sci. Eng.* A293-275: 194-161 (1999).
2. S. Trigwell and G. Selvaduray, *Proc. 2nd Int. Conf. Shape Memory and Superelastic Technol.* Pacific Grove, CA, 1994, pp. 383-388.
3. C. Trepanier et al., *J. Biomed. Mater. Res.* **48**: 165 (1999).
4. K. Tanaka, Tobuse, and S. Miyazaki, *Mechanical Properties of Shape Memory Alloys.* Yokendo, 1993.
5. K. Otsuka and C.M. Wayman, *Shape Memory Materials.* Cambridge University Press, 1998.
6. Y. Suzuki, *Topics on Shape Memory Alloys.* Nikkan Kogyo Shimbun, 1988.
7. Ishikawa, Kinashi, and Miwa, *Collections of SMA Applications.* Kogyo Chosakai, Tokyo, 1987.
8. H. Ohnishi, *Artif. Organs* **12**: 862 (1983).
9. R. Sachdeva and S. Miyazaki, *Proc. MRS Int. Meet Adv. Mater.* **9**: 605 (1989).
10. S. Miyazaki, S. Fukutsuji, and M. Taira, *Proc. ICOMAT-92,* Monterey, CA, 1993, p. 1235.
11. M. Tanaka, K. Hirano, H. Goto, T. Namima, K. Uchi, Z.W. Jiang, H. Matsuki, Y. Tanahashi, S. Orikasaand, and S. Chonan *Bio-Med. Mater. Eng.* **9**: 97-112 (1999).
12. J.C. Schuder, H.E. Stephenson, Jr., and J.F. Townsent, *IRE Internet. Conv. Rec.* **Pt9-9**: 119-126 (1961).
13. H. Matsuki, M. Shiiki, K. Murakami et al., Flexible transcuteaneous transformer for artificial heart system, *IEEE Trans. Magn.* **26**: 1548-1550 (1990).
14. T. Takagi, Y. Luo, S. Hara, T. Yambe, S. Amae, and M.W.H. Nakamura *J. Soc. Adv. Sci.* (in press).
15. G. Harsanyi, *Sensing in Polymer Films in Sensor Applications,* Technomic, Lancaster, Basel, 1995, p. 97.
16. G.M. Shepherd, in *Neurobiology.* 3e., Oxford University Press, New York, Oxford, 1994, p. 272.
17. S. Chonan, Z.W. Jiang, M. Tanaka, T. Kato, and M. Kamei, *Int. J. Appl. Elect. Mech.* **9**: 25-38 (1998).
18. H. Tagami, *Fragrance J.* **10**: 11-15 (1993).
19. T. Yamamura, Measuring methods of skin surface water content, *Fragrance J.* **10**: 35-41 (1993).
20. M. Takahashi, *Fragrance J.* **10**: 16-26 (1993).

BIOMEDICAL SENSING

CHRISTOPHER S. BRAZEL
University of Alabama
Tuscaloosa, Alabama

INTRODUCTION

In recent years, biomedical diagnostics research has led to the development of simple, less invasive, and more accurate evaluative techniques. Much of this is due to biosensor technology, which has played an important role in material improvements used to sense, carry signals, and respond to signals. Biosensor devices can measure micromolar or even smaller quantities of biological substances, including chlorides, glucose, lactose, and urea. Pietro (1) defines biosensors as “any discrete sensing device that relies on a biologically derived component as an integral part of its detection mechanism,” although sensors that are used to monitor biological conditions are usually also included. Medical applications abound, ranging from diagnostic tests for home testing to *in vivo* monitoring of vital conditions and using feedback to control drug delivery or send amplified signals of a change in patient health. Smart materials are frequently combined with biological components to create systems that respond to environmental conditions, such as temperature, pH, concentration of particular analytes, or even light. These smart materials are typically polymeric, and most have the common characteristic that changing environmental condition cause a thermodynamic change between hydrophilic and hydrophobic states, as detailed later. pH paper strips may be one of the simplest sensing devices; they relay a physiological condition through a colorimetric response, and because they are a form of dry chemistry, they simplify detection and diagnostics, which is especially important for home diagnostic kits.

This article presents the use of intelligent polymers in sensing systems, that range from traditional biosensors to advanced self-responsive systems. This review of biomedical diagnostics using smart polymers examines some of the medical applications for intelligent polymers used in electrode-based systems and in bioconjugate systems.

Sensing devices are categorized by the mechanisms of signal detection and response, and examples of systems that monitor analytes, such as glucose, are addressed. Additional modes of response beyond the traditional electronic signal as well as synthesis techniques and examples of medical diagnostics using smart polymers, will be discussed.

Smart materials sense their environment, judge the magnitude of changes, and respond to obtain the most thermodynamically favorable state. This response can cause changes such as surface modification (from hydrophobic to hydrophilic or vice versa), swelling or shrinking of gels, enzyme solubility (if covalently attached to a phase-separating polymer), and binding of polymer and proteins (especially in reversible ionic interactions) (2). These reversible phenomena are useful for biosensors in protecting biological materials, collecting and concentrating analytes, and responding to stimuli to transduce a signal or deliver a drug (3). Smart materials act as on/off switches and can be barriers for an enzyme to protect it from the body's immune system and from harmful solutes or pH conditions. The ability of polymers to phase separate can also aid in separating and preserving enzymes or antibodies to be reused in future diagnostic tests.

The use of intelligent polymer systems in sensors provides the possibilities of combining sensing, transduction of signals, and response in the same independent device and the possibility of controlling biological events based on the signal by delivery of drugs or other means. These three behaviors (sensing, transduction, and response) are characteristic of biosensors, although the traditional response is often electronic (Fig. 1).

Sensing and response processes in smart materials are reversible, and after the hydrogen ions diffuse from the polymer, it returns to its normal unswollen state, and drug release stops. This creates a positive feedback mechanism whereby the device senses abnormal biological or biomedical events and treats the diseased state only when the triggering molecule is present. An example of the structure of a feedback drug delivery biosensor is shown in Fig. 2, where a silicon chip is used as a platform for the biosensor (4). This is similar to the current research of the National Science Foundation's “Lab on a Chip” technology program, where chemical moieties are analyzed on a molecular scale. In the scheme shown, drug reservoirs are kept

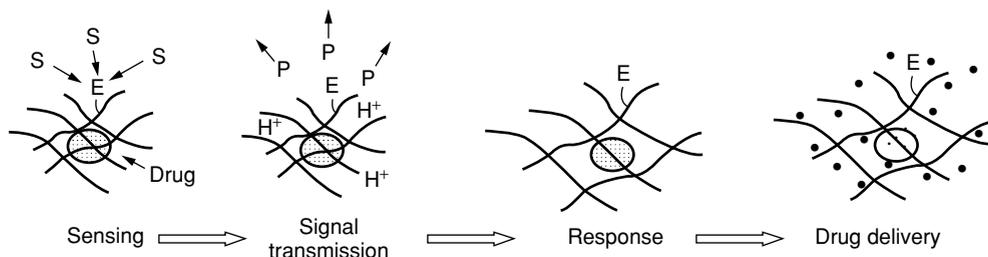


Figure 1. Steps involved in biomedical sensing using intelligent polymers. First, the substrate (S) diffuses to an enzyme immobilized in the polymeric device and reacts; the signal that a particular substrate is present is then transmitted to the surrounding polymer by diffusion of the products (P, H^+) of the enzymatic reaction, such as hydrogen ions. The hydrogen ions, in turn, elicit a swelling response from a smart material, such as pH-sensitive poly(acrylic acid), which can then deliver an embedded drug by diffusion.

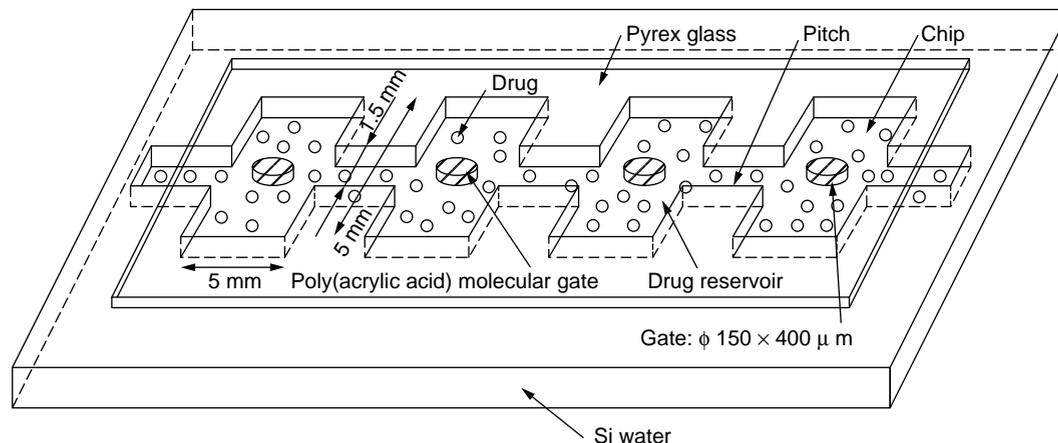


Figure 2. Model scheme for the structure of a sensor device [reprinted with permission from (4); copyright 1996 The Controlled Release Society, Inc.].

behind molecular gates made of pH-sensitive poly(acrylic acid) to yield a sensor that provides a drug delivery feedback response to changing pH.

Traditional biosensors include electrochemical sensors that have been under development since the early 1970s and in which enzymes, antibodies, chemoreceptors, and cellular tissue are used as catalysts for reactions that create electrical signals (5). An example of an electrochemical sensor is shown in Fig. 3, where an enzyme is deposited directly onto a metal conducting electrode and a polymer or protein is used to immobilize the enzyme. This

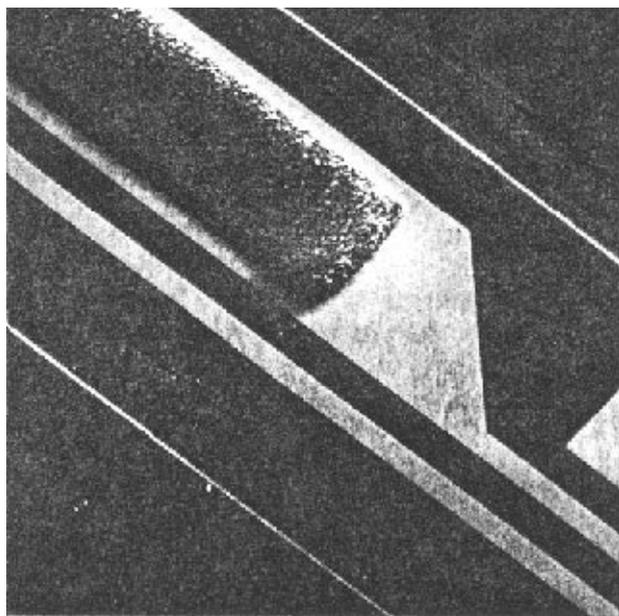


Figure 3. SEM photograph (200x) of a portion of the working electrode that had glucose oxidase and albumin electrodeposited and cross-linked onto the exposed surface. The visible metal layer surrounding the layer of glucose oxidase and albumin is the conductor underlying the insulating layer [reprinted with permission from (6); copyright 1994 American Chemical Society].

particular device was designed for subcutaneous implantation to monitor glucose concentrations in diabetic patients (6). The magnitude of the sensor response is proportional to the analyte concentrations, but the response also depends on the diffusion of the analyte to the enzyme or receptor, the kinetics of the enzyme reaction, and the diffusion of reaction products (such as O_2 or H_2O_2) to an electrode. Biosensor technologies include biochemical sensors, enzymatic sensors, cellular sensors, sensors for redox reactions, antigen/antibody interactions, and other materials that provide recognition surfaces (7). They typically include a specific reaction site (sensor), a mode of transporting the signal (transducer), and a signal measurement or feedback mechanism (responder). These sensing materials can involve antibody/antigen (8) or enzyme/substrate interactions, where the interaction is highly specific and a singular target analyte can be recognized (9). Organelles, whole cells, or tissue can also be used as the sensor (10). Spichiger-Keller (11) divides the types of sensors by the recognition processes used (Table 1). Biosensor devices are one class of intelligent materials, but they can also be combined with smart polymers (such as those sensitive to environmental pH or temperature) to aid in signal collection through specific interactions with the analyte, signal transduction, or the response mechanism (as depicted in Fig. 1).

Because of the complex nature of biosensors, it is important to understand the requirements of sensing in general, and then apply those guidelines to the additional requirements needed to use the sensor in a biological environment. Complications can arise when multiple functional materials are combined into the sensor device. Several important characteristics of biosensors ensure accurate, reproducible, and specific results. Diamond (12) summarized some of the important considerations in Table 2. Many of the characteristics, such as reproducibility of results, robustness, and proportional signal output, are common to process control theory. In addition to the requirements of Table 2, materials (especially the surfaces exposed to biological environments) must be biocompatible, the device should be as noninvasive as possible for medical applications,

Table 1. Classification According to the Type of Recognition Process^a

Sensor Type	Reacting Pairs	Chemical Reaction Recognition Process
Chemical sensors in the strict sense	Host–guest; ligand–analyte; carrier–ion; ion, neutral species, and gas sensors	Complexation association, addition, typical equilibrium reactions
Oxide semiconductor sensors	Inorganic metal oxide; layer–reactive gases	Absorption, reduction, oxidation
Enzymatic sensors	Active site–substrate–cosubstrate mediated sensing reactions; active site–mediator–electrode	Metabolic turnover, typical steady state, Kinetic reactions
Immunochemical sensors	Antibody (catalytic antibody)–antigen; antibody–antigenic protein–hapten	Affinity, association, Equilibrium reactions
Receptrodes, living organs, bilayers hybridised as abzymes, etc.	Receptor–substrate	Association, affinity, metabolic turnover

^a Reprinted with permission from Wiley-VCH and the author (11). Copyright 1998 Wiley-VCH.

Table 2. Ideal Characteristics of a Sensor^a

Characteristic	Comments
Signal output should be proportional or bear a simple mathematical relationship to the amount of the species in the sample	This is becoming less important because of on-device electronics and integration of complex signal processing options to produce so-called smart sensors.
No hysteresis	The sensor signal should return to baseline after responding to the analyte.
Fast response times	Slow response times arising from multiple sensing membranes or sluggish exchange kinetics can seriously limit the range of possible application and prevent use in real-time monitoring situations.
Good signal-to-noise(S/N) characteristics	The S/N ratio determines the limit of detection; can be improved by using the sensor in flow analysis rather than for steady-state measurements; S/N ratio can also be improved by filter or impedance conversion circuitry built into the device(“smart” sensor).
Selective	Without adequate selectivity, the user can not confidently relate the signal obtained to the target species concentration.
Sensitive	<i>Sensitivity</i> is defined as the change in signal per unit change in concentration (i.e., the slope of the calibration curve); this determines the ability of the device to discriminate accurately and precisely between small differences in analyte concentration.

^a(From (12) copyright © 1998. Reprinted by permission of John Wiley & Sons, Inc.

and biosensors must have adequate lifetimes, especially if implanted. Immobilization is often necessary in enzymatic or immunosensors, so that the active material can be kept near the electrode or other transduction and feedback device.

The ideal biosensor characteristics listed in Table 2 provide some of the driving forces for research in this field. Hysteresis limits the effective lifetime of a sensing device; because of repetitive cycling, the sensor's response becomes less reproducible. Poshossian et al. (13) report a penicillin biosensor based on a pH-sensitive gel that has a hysteresis of less than 4 mV (less than 0.4% of the signal) and a usable lifetime of at least one year; perhaps more importantly, they demonstrated that by using an absorptive technique to immobilize penicillinase, the sensor could be regenerated by desorption and resorption of fresh enzyme. To achieve faster responses in immobilized enzyme-based systems, macroporous gels have been employed to reduce any potential hindrances due to diffusion (14). The limits of sensitivity are also important in creating sensing devices. The minimum detection level has dropped as low as the nanomolar level (15), but the range of analyte concentration across which the biosensor is useful must also match the system being monitored.

Although the functioning of a biosensor must be accurate and robust, the largest barrier to successful implementation of a biomedical sensing device is ensuring that it is readily usable and simple for patient compliance. The development of assays that extract metabolites across the skin (16–18) or those that sample biological fluids that are more readily collected, such as urine, sweat, or saliva (19), are examples.

Compared to all sensor markets, the medical and biological fields have become leading drivers for new research in recent years and have great potential for improving pharmaceutical processing, medical diagnostics, and patient treatment (12). The driving forces for future development of biomedical sensing devices include improving and diversifying recognition mechanisms, developing new materials for immobilization that meet the stringent requirements of biocompatibility, developing materials that do not use enzymes or biological components (using techniques such as enzyme mimics and molecular imprinting), improving the flexibility of design (especially through dry chemistry and removing the requirement for electrolytic fluids in the sensor), discovery and development of new sensor molecules that are highly specific to diseased states, and improving signal processing and reproducibility (12). Many biosensors have a linear response to concentrations across only a narrow window, so one continuing focus will be to develop sensors that are more robust and have proportional responses across a wider range of analyte concentrations.

A variety of biological components have been used in sensing mechanisms for biosensors; the majority is based on enzymes because enzymatic reactions are highly specific, occur at low analyte concentrations, and can differentiate between enantiomers and compounds of similar structures. Particular enzymes can also be selected to catalyze any of a range of medically significant biochemical reactions. Because obtaining highly purified enzymes or

proteins is often expensive, McCormack et al. (9) suggest using whole cells in the substrate recognition step. Cell-based sensors would be more adaptable and resilient compared to proteins or enzymes and may lead to biosensor products that have longer lifetimes; cells can carry out more complex reactions by using multiple enzymes and metabolic pathways to produce a product that signals a response from the biosensor. On the negative side, cell-based biosensors are much more complex, and the direct cause-effect (reactant–product) relationship of enzymatic reactions becomes more difficult to define. One example where a cell-based sensor would be preferred is in monitoring the products of cell metabolism to determine the availability of nutrient supply. If carbon or nitrogen sources are in short supply, secondary metabolites may be formed, which could be monitored; alternatively, if the oxygen supply is short, partial metabolites, such as lactates, may build up in the tissue near the biosensor. This would be potentially useful in monitoring cells used for tissue engineering or implanted organs to verify that the region is becoming vascularized and not rejected by the host. It is also possible to use multiple cells or even plant or mammalian tissue in combination with sensors to monitor the production of highly specific metabolites.

MEDICAL, THERAPEUTIC, AND DIAGNOSTIC APPLICATIONS OF BIOSENSORS

Biosensors have made it possible to reduce human health care costs by using available at-home test kits so that patients can monitor their own glucose levels, pregnancy hormones, and cholesterol (20). Most of these kits can monitor easily collected biological specimens, such as saliva or urine, or require training the patient to collect blood samples (for glucose monitoring). Several *ex vivo* diagnostic tests, including glucose and cholesterol screenings, have been made possible through biosensors that incorporate enzyme assays into dry chemistry electrodes (21). In many of these tests, polymers are included with the biological sensing component along with a simple readout, such as an indicating dye. Biosensors are used in many biological systems, including control of reactions in biological reactors for producing pharmaceutical agents (22), measuring analytes in biological samples, such as blood or urine (23), and monitoring health information *in vivo* (16).

As an example of biosensors used with smart materials, Mizutani et al. (24) developed enzyme electrodes based on oxidoreductase enzymes for monitoring lactates and ethanol. Their device consisted of a stimuli-responsive “smart” polymer system of poly(L-lysine)/poly(4-styrenesulfonate) to which enzymes are bound by ionic interactions with the polymers. The ionic polymer is involved in both enzyme immobilization and in screening solutes such as L-ascorbic acid and uric acid, which interfere with signal transduction if they are near the electrode. Lactate levels in sour milk and human serum were tested using the biosensor and were compared to conventional test kit methods (Table 3). Results must compare favorably in accuracy and reproducibility to consider biosensor devices feasible.

Table 3. Comparison of Results Obtained for L-Lactic Acid in Human Sera and Sour Milk by Different Methods^a

Sample	L-Lactate Concentration Proposed Method	(mM) F-kit method
Serum 1	2.03	2.02
Serum 2	1.44	1.36
Serum 3	2.80	2.86
Serum 4	1.75	1.67
Sour milk 1	63.8	61.4
Sour milk 2	74.6	73.0
Sour milk 3	53.3	53.6
Sour milk 4	87.7	86.1
Sour milk 5	72.4	74.4

^aReprinted with permission from. Copyright 1996 American Chemical Society.

Medical applications of biosensors cover a wide range of analytes (Table 4) and medical conditions (Table 5). Diabetes monitoring and treatment is the primary thrust of research and product development in current medical sensing technology, but as the understanding of molecular biochemistry advances, there are possibilities of developing economically feasible sensors based any of these conditions and more.

Some of the enzymes of potential use in biomedical sensors are listed in Table 6. Much of biosensor development has focused on oxidoreductases because changes in oxidized states of chemicals cause electron flow that can be detected by using electrodes.

Biomaterials that respond to environmental changes and fractures to self-repair are considered smart materials that sense biological events and give feedback by releasing healing chemicals. In these systems, the "sensing" is not as sophisticated as in traditional biosensors, but the use of self-repairing structures in biomaterials makes them able to sense stresses and cracks and respond *in vivo* without surgical procedures to replace or reset the implant. Dry (30) described techniques to improve biomaterial performance by using smart materials that self-heal upon mechanical erosion. This is particularly important in developing load-bearing biomechanical materials for replacing bones and joints. Hastings (33) also cites the use of smart materials that self-repair upon shear and fracture or release drugs or hormones in conjunction with biomaterial implants to reduce inflammatory response. These devices typically use encapsulation to hold the active ingredient, and shear or pressure are used to break the capsule wall and trigger release.

Another area where biosensing is done nontraditionally is targeted drug delivery. Some of the same interactions, especially using chemoreceptors, can be used to design drug

Table 4. Examples of Chemical Analytes Subject to Biosensing

Lactate	Pyruvate	Glucose
Fructose	Galactose	CO ₂
O ₂	Ascorbic acid	Cholesterol
Urea/uric acid		

Table 5. Examples of Biomedical Applications of Sensing Materials

Diabetes monitoring and treatment (25)
Detection of viruses/toxins (26, 27)
Monitoring metabolic substrate use (biological oxygen demand) (9)
Monitoring metabolic products (such as lactates or partial oxidation metabolites) (28)
Determining the efficiency of dialysis and filtering (in organs or <i>ex vivo</i>) (19)
Biopharmaceutical production and testing (e.g., monitoring cell activity in fermenters) (22, 29)
Repair of fractured tissue or bone (30)
Targeted drug delivery (such as cardiovascular or gastrointestinal) (31, 32)

carrier surfaces so that a drug is delivered to a specific area. Yang and Robinson (34) used glycoproteins to bind drug delivery vehicles to a selected site. These surface derivatives may be used to anchor controlled release devices or biosensors to a particular type of cell in the body. Smart polymers used in biosensors for drug delivery are detailed later in this article.

Monitoring lactate and pyruvate levels by using a biosensor can indicate when secondary and partial oxidation products are formed *in vivo* and indicate when the supply of nutrients or oxygen to tissue is insufficient (35). These enzymatic electrochemical sensors are used in extracorporeal evaluation of blood in patients who have an artificial pancreases; and the lifetime of the sensor is 30 days/300 assays. The results from the biosensor correlated well with spectrophotometric analysis of blood serum.

Many biomedical sensing devices have been proposed and constructed for monitoring blood glucose levels, and some of these approaches are detailed later in this article. The reader is referred to Campanella and Tomassetti (36) for a review of biosensors for clinical and pharmaceutical analysis.

POLYMERS AS ELECTRODE COATINGS AND BIOSENSOR MEDIATORS

To date, most research on biosensors has focused on devices that consist of a metal layer to conduct electrons as the signal, an enzyme or antibody to sense the presence of a particular analyte, and a membrane to immobilize the enzyme and also possibly aid in transducing the signal to the electrode. A common approach used to design glucose sensors, as demonstrated by Johnson et al. (6), is based on glucose oxidase, which is immobilized between an outer membrane permeable to glucose and oxygen and a platinum electrode to reduce the formed hydrogen peroxide and transmit electronic signals (Fig. 4). The design of the membrane is crucial to operation because the pore size must allow rapid diffusion of glucose and oxygen and yet retain the enzyme.

Polymer gel coatings on electrodes serve multiple purposes. Polymer gels form semipermeable membranes to

Table 6. Enzymes Used in Biosensors

Enzyme	Medical Use
Lactate dehydrogenase	Test for metabolic acidosis
Alcohol dehydrogenase, alcohol oxidase	Test for blood alcohol content
Monoamine oxidase	Drug testing—amphetamines or cocaine
Glucose-6-phosphate dehydrogenase	Test for diabetes
Glucose oxidase	Test for high/low glucose concentration in blood
Isocitrate dehydrogenase	Test for normal metabolism (citric acid cycle)
Lactate oxidase, pyruvate oxidase	Detect partial oxidation metabolism—sports medicine
Catalase	Used in conjunction with glucose oxidase to drive reaction toward gluconic acid product by consuming hydrogen peroxide
Ascorbate oxidase	Monitoring of vitamin levels
Urease	Detoxification of blood
Cholesterol oxidase	Detect cholesterol levels

control substrate diffusion and immobilize an enzyme near an electrode. Because hydrophilic polymers, such as poly(vinyl alcohol), poly(lactic acid-co-glycolic acid), and poly(ethylene glycol) have densities and compositions similar to those of natural tissue and interact minimally with the immune system, they are used to make biosensor surfaces more biocompatible. The definition of biocompatibility has been redefined in recent years from totally inert when exposed to living tissue to actively integrating with the biological components (37). Some of the important characteristics of biocompatible materials include thromboresistance, infection resistance, and minimal effects on blood flow and nutrient supply (38), but the physical aspects of biocompatibility may change for the intended application, so as to obtain the desired host–material interaction (39). The surface characteristics of implanted biomaterials are often modified to create interfaces that have low platelet adhesion by derivatizing polymers with heparin or poly(ethylene glycol) (40,41). Interfacial properties are important for biocompatibility, but the polymer matrix can also act as a signal mediator to transmit electrons from the reaction site to the electrode, especially if the polymer is doped with conducting materials (42,43). Electron transfer in many polymer systems is hindered by the insulating nature of polymers, so several materials have been used to dope or modify the electrical properties of polymers used in immobilization. The use of naturally occurring redox chemicals, such as FAD/FADH₂ and NAD/NADH, has improved signal transduction and response time for some biosensors

(5). Good mediators should have reversible kinetics so that they are ready to transmit multiple signals; they should be independent of environmental factors such as temperature or pH, be stable, and be retained easily near the surface of the electrode. Some small molecules, such as quinone, methylene blue, thionine, and ferrocene, have been used because they diffuse easily through porous polymers (5). These mediators have been successfully applied to such commercial products as home glucose meters. Usmani (44) and Chen et al. (45) used osmium-containing poly(4-vinyl pyridine), ferrocene-modified polysiloxanes, quinones, and other organic salts to aid in signaling and charge transport.

Polymers selected for use in biosensors must meet many requirements: they must not interfere with the chemistry of the reaction in the sensor and must also be neutral and stabilizing for biological components used in the sensor, such as enzymes (21). Many smart polymers are ionic to accommodate pH-sensitive transitions, but ionic interactions between the polymer and biological substances may cause interference with or fouling of the sensor device (46,47). The method used for synthesis may also leave residual monomer, cross-linking agent, or solvent trapped within the polymer gel, which can be detrimental in sensitive measurements. So care must be taken to purify polymers used in biosensors or to select techniques that do not use any potentially harmful or interfering chemicals. Particle size and porosity must also be considered because the diffusion path and surface area exposed to a solution being monitored may greatly affect the amplitude of the response.

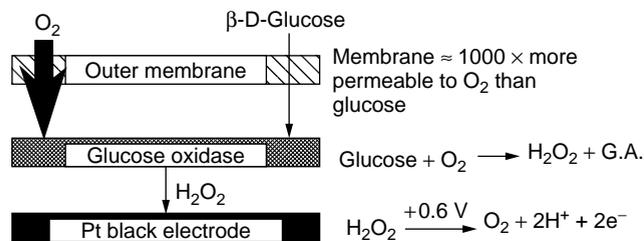


Figure 4. Block schematic of a hydrogen peroxide-based electroenzymatic glucose sensor that has a differentially permeable outer membrane layer [reprinted with permission from (6); copyright 1994 American Chemical Society].

IMMOBILIZATION TECHNIQUES AND MATERIALS

Enzymes are often used in biomedical sensing devices to catalyze reactions that are specific for certain substrates. The products formed by enzymes embedded in smart polymer networks in turn elicit a response, such as swelling to release drugs for treatment, markers for detection, or phase separation to change the hydrophilic/hydrophobic balance of the local environment. Oxidation–reduction enzymes lend themselves well to combination with pH-sensitive polymers because hydrogen ions can be used

Table 7. Physical and Chemical Methods of Immobilization

Entrapment within a polymeric gel
Microencapsulation in a semiporous capsule
Cross-linking the enzyme
Covalent linkages
Chemisorption
Physical adsorption to polymer surfaces
Chemical binding using macromonomeric techniques or pendent chain chemistry

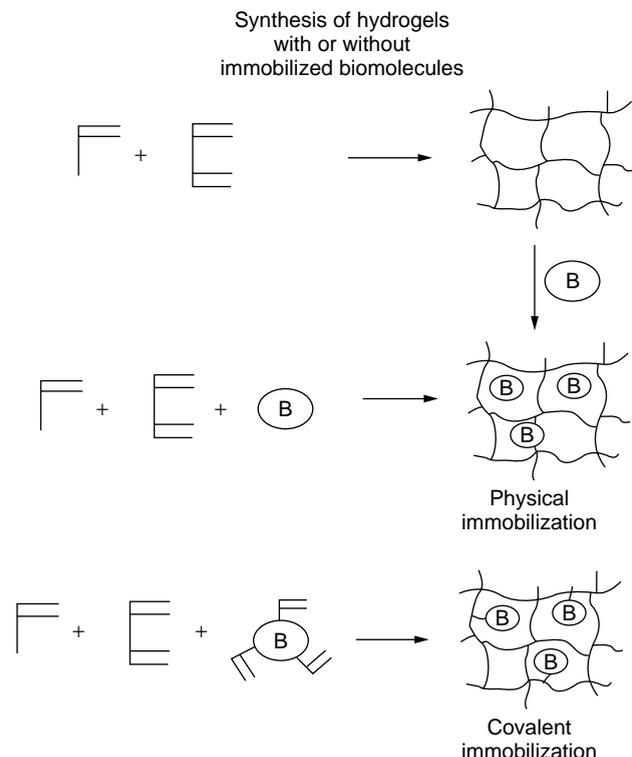


Figure 5. Synthesis of hydrogels with or without immobilized biomolecules [reprinted from (48); copyright 1987 with permission from Elsevier Science].

or released as a result of enzymatic reactions. Hoffman (48) presents three possible methods of immobilizing an enzyme in a hydrogel: equilibrium partitioning, cross-linking in the presence of active agent, or derivatizing the enzyme with a reactive double bond to form covalent linkages the polymer gels (Fig. 5).

Enzymes are typically large protein molecules that have molecular weights in the tens to hundreds of thousands and can be combined with polymer gels and networks by a number of methods. The methods described before by Hoffman (48) can be used, as can other methods developed recently (49–52). Platé, Valuev, and co-workers used a macromonomer reaction for incorporating enzymes into polymer networks. These systems are formed by first forming a reactive derivative of the target enzyme by attaching a polymerizable double bond, such as an acrylate, to the enzyme by reacting it with acryloyl chloride (Fig. 6).

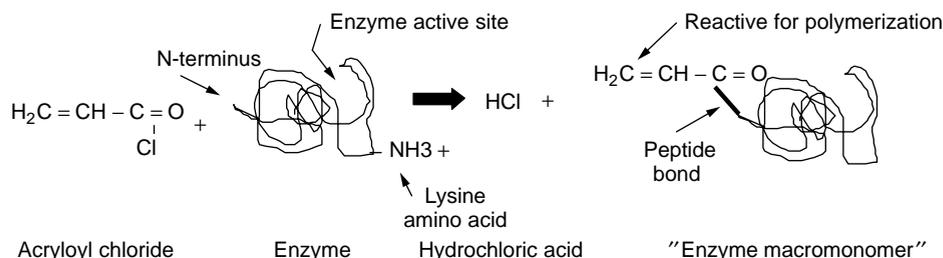


Figure 6. Macromonomer procedure for preparing derivatized enzymes that have an active double bond for polymerization.

Acryloyl chloride is particularly reactive to basic groups, such as amino organic groups found commonly in enzymes on the side groups of amino acids, such as lysine, histidine, and arginine. The reaction forms an acrylate linkage through one of these enzyme subgroups, which is reactive to free radical polymerizations to form cross-linked gels. As long as the solutions used to form the macromonomers and polymer gels do not destroy the folded structure of the enzyme and the double bond reaction does not take place in the enzyme active site, the result is a polymer that has an active enzyme covalently attached and immobilized. The activity of the enzyme is typically reduced somewhat during the derivatization procedure, depending on the environmental pH and temperature, as well as the effect of the polymerization chemicals on the protein configuration. Because the active site of an enzyme is only a small portion of the protein's molecular structure, attachment of the acrylate double bond within the active site occurs in only a small fraction of enzyme molecules, although inhibition of enzyme kinetics may occur due to linkages anywhere on the enzyme. The resulting systems synthesized by the macromonomer technique include a biological chemical, such as an enzyme or protein, dispersed throughout a polymer network but chemically immobilized within the membrane.

In addition to the methods mentioned before, enzymes and antibodies (8) can be entrapped by physical or chemical methods (Table 7). Immobilization normally uses polymers to restrict diffusion or to serve as matrices for adsorbing or binding enzymes. A variety of polymers, both synthetic and naturally occurring, have been used for chemical linkages to enzymes or in gelled forms to control diffusion (Table 8). Some standard microencapsulation techniques that can be used for immobilization include spray drying, rotary atomization, coextrusion, fluid bed coating, solvent evaporation, and emulsion and suspension polymerization (53).

Table 8. Materials Used in Immobilizing Enzymes and Biological Components^a

Synthetic Polymers	Natural Materials
Polyacrylamides	Agarose
Poly(maleic anhydrides)	Dextrans
Poly(meth)acrylic acids	Cellulose
Acrylates	Glass
Poly(vinyl alcohol)	Collagen
Dacron TM	Alumina
Nylon	Polysaccharides
Polystyrenes	Polypeptides

^aCompiled from 23,54,55.

Immobilization provides several advantages to biosensors. Immobilization can improve enzyme stability and preserve its biological providing activity by a nondenaturing environment (55) and preventing the loss of enzyme to the surrounding fluid; this allows using the device to multiple times. The polymer matrices used can be designed to control the diffusion of the substrate to the enzyme, and conducting polymers can transduce redox charges from enzyme active sites to electrode surfaces (19). Immobilization is also used to protect enzymes from denaturing proteins and helps avoid extreme pH and chemical microenvironments for the biomolecule (56). Enzyme stability is crucial for commercial viability because destabilization can result in false biosensor readings.

The sterility of medical devices is also of utmost importance in developing biosensors. This issue can lead to selecting particular methods of immobilization even though the activity of the enzyme and the structure of the membrane may not be ideal. Some alternatives to chemical cross-linking of polymers include radiative cross-linking (57), thermal gelation, or the use of ionic polymers, such as alginates, that do not require monomer or organic solvents. Preservation of the biological activity is paramount in ensuring the robustness of the biosensing device; organic solvents, extreme pH or temperature, and radiation can affect the structure and thus the activity of proteins and enzymes.

Pazur et al. (58) and Gibson and Woodward (55) studied the activity of alcohol oxidase stabilized by saccharides and stored in dry form. The activity was monitored as a function of time, and certain sugars, especially cellobiose, inositol, and trehalose, maintained the enzyme at high activity levels, and the resulting mixtures also demonstrated higher activity in methanol oxidation (Fig. 7). Enzyme activity can be altered in many ways. The geometric structure of the polypeptides that make up enzymes must be kept intact, so that the active site of the enzyme retains its configuration. Enzymatic structure can be altered by pH, temperature, organic solvents, and shear stresses as well as the presence of other chemicals in solution. These chemicals can act as inhibitors or activators to alter the rates of substrate/enzyme reaction. The reaction mechanisms are shown in Scheme 1.

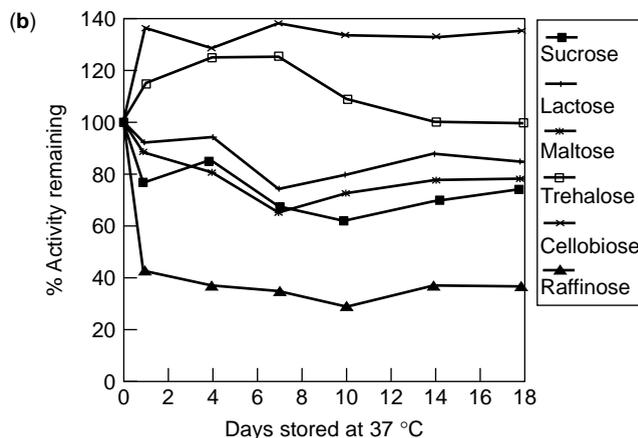


Figure 7. Alcohol oxidase stabilization by disaccharides. Sugars were added to the enzyme at concentrations of 1–10% immediately before drying. The solutions were dried in shallow dishes at 30°C under a vacuum and then harvested, ground to a powder, and stored in vials at 37°C. Enzymatic activity was assayed with methanol as a substrate using an oxygen electrode and a colorimetric assay [reprinted with permission from (55); copyright 1992 American Chemical Society].



Scheme 1. Enzyme (E) reaction mechanisms for substrate (S) conversion involve the formation of an enzyme–substrate complex at the active site, which catalyzes the conversion of substrate(s) to product(s) (P). The presence of an inhibitor (I) chemical can reduce reaction velocities because fewer enzymes are available for complexation with the substrate. Alternatively, enzyme activators (A) help to stabilize the enzyme conformation and may make the active site more accessible to forming the substrate–enzyme complex (EAS, when the enzyme is complexed with both the activator and substrate).

Enzyme activity can also be reduced due to diffusional limitations imposed by the polymer network on interactions between the enzyme and its substrate. The Damköhler number, Da , as defined following, provides a measure of the importance of diffusional resistance to normal enzyme kinetic behavior:

$$Da = \frac{v_{\max}}{\left(\frac{D_{es}}{\delta^2}\right) s_0} \quad (1)$$

In Eq. (1), v_{\max} is the maximum reaction velocity for the enzyme–substrate reaction, D_{es} is the effective diffusion coefficient for diffusion of the substrate through the polymer membrane, taking tortuosity and porosity into account; δ is the distance required for the solute to diffuse through the membrane; and s_0 is the substrate concentration in the bulk solution (59). For Da values greater than 1, the enzyme–substrate reaction proceeds as usual, but for Da much less than 1, the matrix is a significant hindrance to substrate diffusion and lowers the observed reaction rates.

SMART POLYMERS FOR IMMOBILIZATION AND BIOCONJUGATE MATERIALS

The recent research in smart polymers has led to intriguing materials for applications ranging from separations and surface modifications to drug delivery and self-contained biomedical sensing devices. In addition to the materials listed in Table 7, many researchers have used smart polymers to immobilize enzymes. The devices formed by this combination afford possibilities of protecting the enzyme from temperature or pH swings or acting as part of a biosensor response mechanism. Chen et al. (60) immobilized α -amylase by esterifying the side chains of temperature-sensitive poly(*N*-isopropylacrylamide) and poly(*N*-acryloxysuccinimide) gels to improve the thermal stability of the enzymes by preventing structural changes and exposure to solvent as the immobilized enzyme is exposed to high temperatures. There was no significant loss in enzymatic activity after immobilization when exposed to high temperatures because the temperature-sensitive phase separation of the polymer network shielded the amylase from solutions above the polymer lower critical solution temperature (LCST). Above this temperature (in aqueous solutions, at approximately 30°C for polymers based on *N*-isopropylacrylamide), the gel collapses, reducing the amount of water surrounding the enzyme, and the polymer collapses providing stabilization to the enzyme structure.

In addition to protecting enzymes from thermal denaturation, thermally reversible gels can also be used to control rates of reaction. For example, the temperature-dependence of α -chymotrypsin activity was affected by the polymer behavior when immobilized by physical entrapment in poly(*N*-isopropylacrylamide-*co*-hydroxyethyl methacrylate) gels (61). Figure 8 shows the effect of temperature on free enzymatic activity versus enzyme bound in the temperature-sensitive gel. The maximum reaction rate for the immobilized enzyme was observed below the gel's LCST, at 30°C, and decreasing activity observed at higher temperatures was attributed to the polymer which

caused a change in enzyme conformation and diffusion limitations on the substrate, which must pass through the collapsed porous structure.

Smart polymers have also been used to influence binding and interactions between polymers and solutes, which can be controlled thermodynamically using pH, electrical current, temperature, or solution ionic strength as modulators for reversible processes in separations (62). These materials can be used to improve or modulate the sensitivity of biosensors by concentrating very low levels of analytes near the electrode or sensing device to amplify the signal (63). Schild (64) and Ogata (65) used phase-separating polymers to collect a particular analyte (Figure 9), such as immunoglobulin G (IgG). This method can aid in concentrating the analyte for detection, such as in antigen capture fluorescence immunoassays (48), where IgG concentrations can be correlated with fluorescent intensity. This method would be ideal for use in combination with optical detectors. As shown in Fig. 9, antibodies bound or otherwise immobilized to temperature-sensitive polymers can be designed to select and attract antigens.

BIOSENSOR OPERATION

In Table 1, biosensors are categorized by recognition processes. They can also be classified by the response mechanism used. Biosensor responses can be categorized as chemical such as production of dissolved oxygen or hydrogen ions, electrical, optical (using fluorescent or colored molecules), or mechanical (7). Colorimetric dyes, pH indicators, and bioluminescent molecules create optical responses that can be quantified spectrophotometrically; electrodes are used to deliver electronic signals that can be interpreted by computerized or simplified controllers; and smart polymers can change shape to give an optical response, swell, or shrink to control diffusion of a drug for release (Table 9). Bioluminescent molecules, such as luciferin, can be used in sensors, where light is emitted through metabolic production of adenosine triphosphate (ATP), so they are often used to detect bacteria in food and pharmaceutical production (20,66). Bioluminescence can also be used to monitor uptake of ATP due to the presence of glucose in sufficient quantity to begin the glycolytic metabolic pathway at the first reaction to convert glucose to glucose-6-phosphate; this results in reduced luminescent intensity (20).

Intelligent materials are used in biosensors in three main areas: signal detection, transmission of signal to a measuring electrode or the response element, and controlling the feedback response to the signal. Intelligent materials are those that respond to changes in the neighboring environment, and the response can be of different types. Polymer swelling can lead to physical work, such as shutting off a valve or making contact between the sensor and a secondary component. Swelling is also used in controlled drug delivery, whereby diffusion of embedded drugs controlled by the porosity of the cross-linked polymer carrier can be triggered by environmental changes that indicate the need for therapeutic agents. The material may provide a pathway for electron transport by using semiconducting

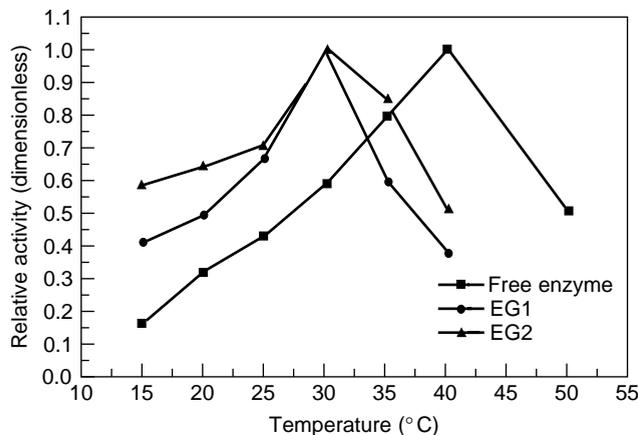


Figure 8. The variation of relative activities of the free enzyme and enzyme-gel matrix vs. temperature [from (61); copyright © 1998; reprinted by permission of John Wiley & Sons, Inc].

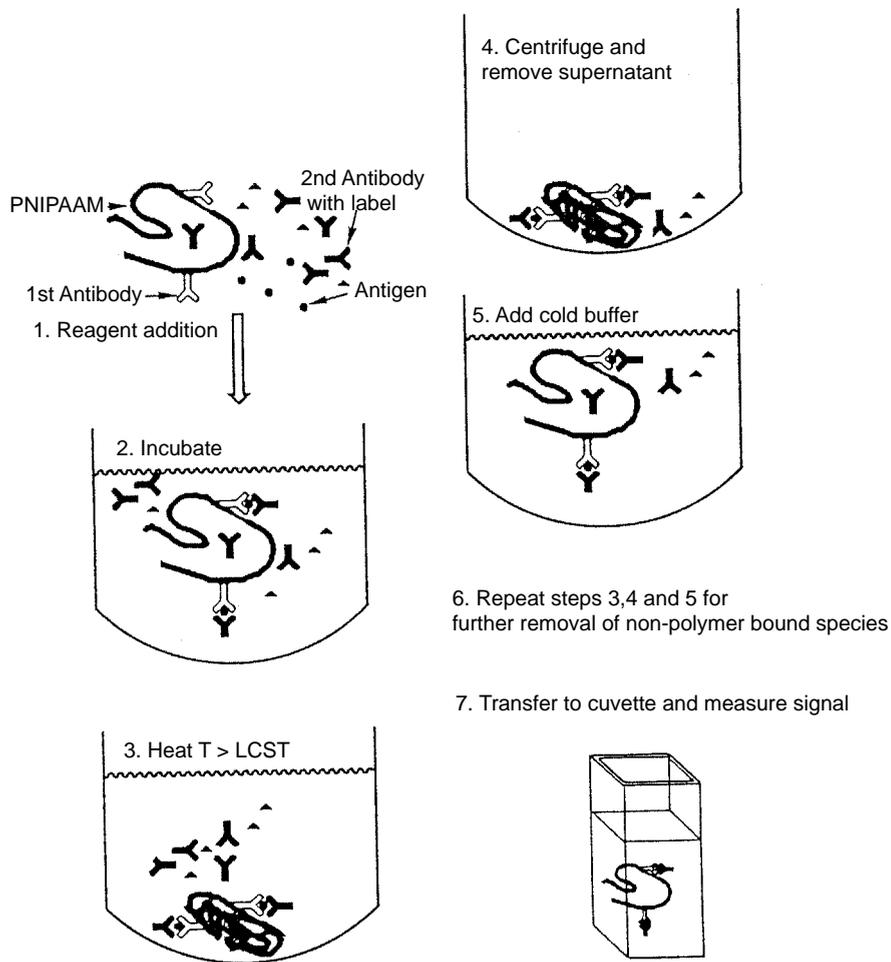


Figure 9. Immunoassay scheme for PNI-PAAm [reprinted from (64); copyright 1992, with permission from Elsevier Science].

polymers, such as polypyrrole (67), polyaniline, or poly(vinyl pyridine) (68). Schuhmann (69) showed ways to improve electron transfer mediation in immobilized enzyme systems by using polypyrrole or polyazulen to facilitate electron hopping, similar to semiconductor materials, so that the signal could be transferred from the enzyme to the electrode (Fig. 10). Many intelligent polymers behave according to two-component thermodynamics and phases separate from a homogeneous solution as the polymer turns from a primarily hydrophilic to a hydrophobic entity.

The signal can be transduced by using amperometric or potentiometric methods, field effect transistors, piezoelectric crystals, thermistors or optoelectronic systems (10), or by using closed-loop systems within the device where feedback is sent by using a smart polymer. Amplification

of the signal is important, especially when the target molecule is in very small concentrations. Skaife and Abbott (70) used liquid crystals to measure the binding of IgG, where the concentration ranged from 1–100 nM.

GLUCOSE SENSORS

Much of the recent literature on biomedical sensing has focused on methodologies for detecting glucose levels in diabetic patients to sense the need for insulin release without requiring self-diagnosis through needle sticks. Diabetes is a highly prevalent disease in the United States, and many researchers have been searching for methods of mimicking a naturally functioning pancreas through tissue engineering, encapsulating pancreatic islets of Langerhans cells, creating insulin pumps that use electronic glucose sensors, or developing responsive polymers to detect high glucose levels and deliver insulin. As of 1996, the blood glucose monitoring market in the United States amounted to approximately \$750 million per year and was growing at a rate of 10% per year (21). There is great potential and motivation to study more convenient, reproducible, and cheaper methods of determining blood glucose, and because of the recent intensity of research in this area, many novel glucose monitors, such as the Glucoprocessor(R)

Table 9. Smart Material Responses to Analyte Detection

Releasing drugs or chemicals by diffusion through enlarging pores or squeezing out of shrinking pores
Acting as a mechanical valve by reversibly swelling
Reversible adhesion
Completing a circuit by changing electrical properties or shape
Visual response by changing from transparent to opaque or changing shape
Trapping molecules to separate or concentrate

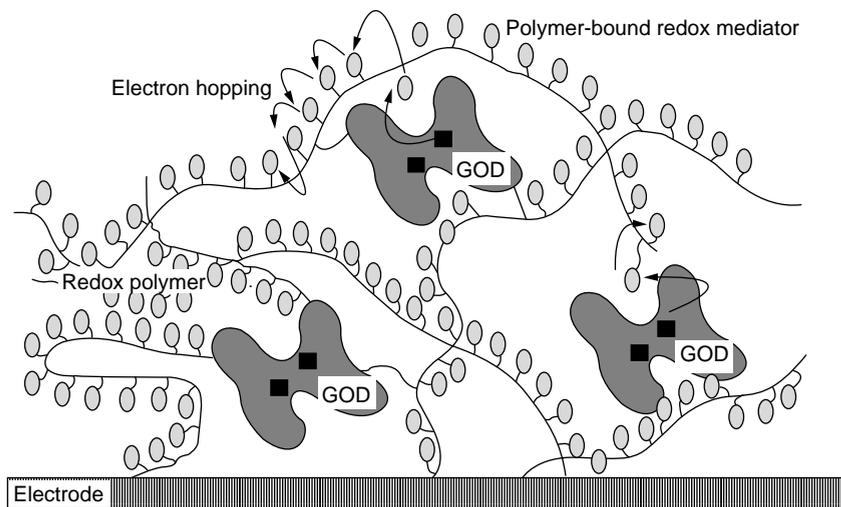


Figure 10. Electron transfer in a redox polymer via electron hopping between adjacent redox centers [reprinted with permission from (69); copyright 1994 American Chemical Society].

marketed by Solea-Tacussel in France as early as 1988 (10), Bayer's Glucometer[®] DEX[®] Diabetes Care System, and Roche Diagnostics' Accu-Chek[™] line, have reached consumers.

Glucose sensors include a range of devices. Amperometric biosensors use immobilized glucose oxidase, an enzyme that converts glucose to gluconic acid and hydrogen peroxide. Hydrogen peroxide is measured by electrodes and can be correlated with the glucose levels, even at micromolar concentrations (6, 71–73). Subcutaneous insulin pumps are either user-controlled or contain a glucose-sensing system that creates a closed-loop feedback insulin delivery system to provide a simplified method for monitoring and treating hyperglycemia (21). Diabetic glucose is typically monitored by sampling a small quantity of blood and using a chemical test kit; biosensors have been developed to analyze samples extracted across the skin by electroporation or iontophoresis (16). Several methods using smart materials for insulin delivery incorporate glucose-sensitive enzymes or chemical linkages that are disrupted by glucose (74,75).

Martin et al. (76) studied ultrathin film composite glucose sensors based on a glucose-permeable membrane to immobilize a solution of glucose oxidase, ferrocene as a mediator, and amperometric electrodes. Transduction of signals from the reaction is rapid because all diffusion is in the aqueous phase. They found a linear relationship between glucose concentration and current in the range of 2–22 mM glucose (Fig. 11).

Schuhmann et al. (69) demonstrated that adding electron transport functional groups to polymers (such as the material shown in Fig. 10) using conducting polymers (β -amino(polypyrrole), poly(4-aminophenyl)azulen, or poly(*N*-(4-aminophenyl)-2,2'-dithienyl) pyrrole), the amperometric response to glucose could be modified (Fig. 12). The poly(*N*-(4-aminophenyl)-2,2'-dithienyl) pyrrole conduction mediator showed the greatest proportional amperage response to glucose, but the linear relationship ended around 5 mM glucose concentration. The other two mediator-conducting polymers provided nearly linear current-concentration curves across the range from 1–14 mM glucose. Fortier et al. (77) verified the same

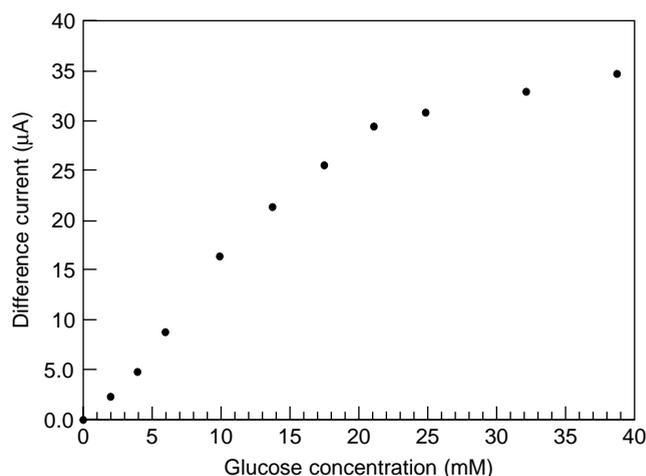


Figure 11. Calibration curve for the prototype ultrathin film composite glucose sensor. The regression coefficient for the linear region (up to 22 mM) is 0.995 [reprinted with permission from (76); copyright 1994 American Chemical Society].

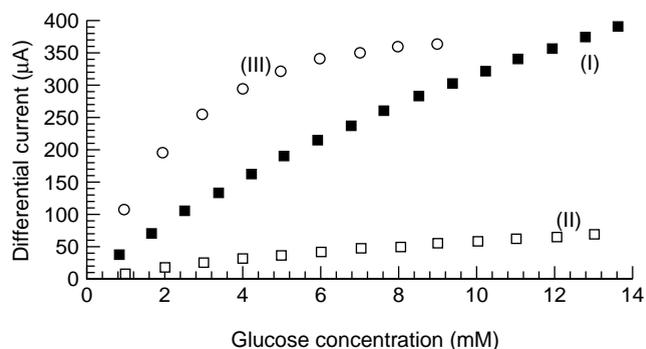


Figure 12. Calibration graphs for glucose obtained from amperometric enzyme electrodes where glucose oxidase is covalently bound to different conducting polymers. I. β -amino(polypyrrole); II. poly-(4-aminophenyl) azulen; III. poly[N-(4-aminophenyl)-2,2'-dithienyl]pyrrole [reprinted with permission from (69); copyright 1994 American Chemical Society].

general behavior of glucose oxidase electrodes, using a catalytic current from 0–200 microamps generated from 0–20 mM glucose solutions. Sung et al. (78) also used polypyrrole composites with polyanionic-GOx conjugates to improve electrical conduction within the hydrogel to achieve a correlation between glucose concentration and current. These results may provide a rationale for selecting particular components based on the glucose concentration range of interest.

These devices are designed for subcutaneous implantation, but biomedical diagnostic sensors can also be valuable for use *ex vivo* or on the surface of the skin to monitor glucose levels. A unique use of biosensors in detecting glucose is a method developed by Berner et al. (16), where glucose can be monitored without sampling blood. Glucose is iontophoretically extracted across the skin, a noninvasive method that may be much more convenient for diabetic patients. A smart material is used to sense glucose, using glucose oxidase to create hydrogen peroxide, which is determined electrochemically by a platinum electrode. Tierney et al. (15) developed a hydrogel coating for an electrode that incorporates glucose oxidase to make it possible to sense glucose levels electro-osmotically extracted from the skin as a diagnostic indicator for diabetic patients. Micromolar concentrations of H_2O_2 can be correlated with electrical signals of the order of several hundred nanoamps. The glucose sensors were also tested for response time to a small (2 nM) increase in glucose concentration (Fig. 13). The biosensor's response was an increase in current of $1.5 \mu A$ in about 30 seconds. Tamada et al. (79) also used the transdermal extraction technique to monitor glucose concentrations in blood.

Intelligent polymers have been used in a number of insulin delivery systems. Podual et al. (75) offer a different method for detecting glucose. Their method is based on glucose oxidase, as those mentioned before, to cause the reaction to produce gluconic acid, but the response is due to the localized increased concentration of acid (decreased pH) within the polymer membrane which causes pH-sensitive polybasic hydrogels to swell and release embedded insulin. Catalase, a second enzyme in the gels developed by Podual et al., was used to help drive the reaction toward

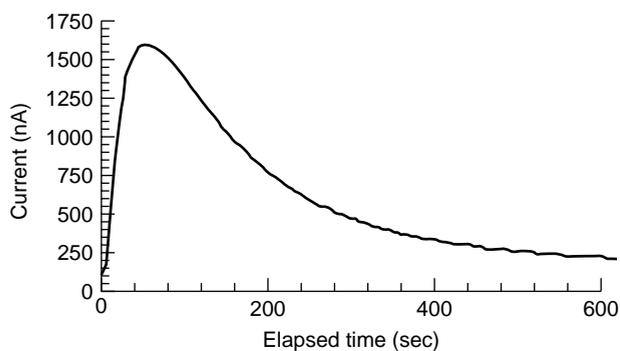


Figure 13. Biosensor response to 2-nmol glucose spike [reprinted with permission from (15); copyright 1998 The Controlled Release Society, Inc.].

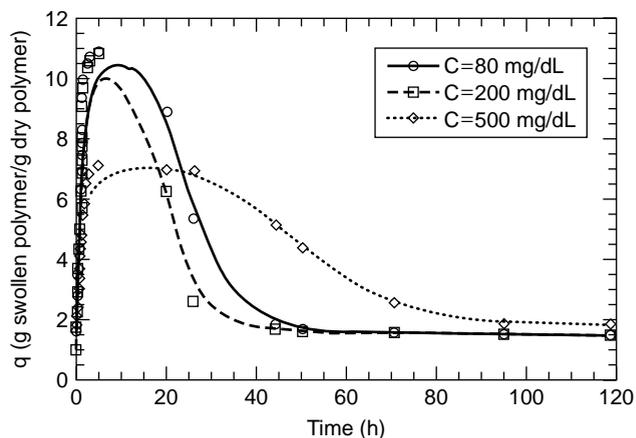


Figure 14. Weight swelling ratios of glucose-sensitive P(MAA-g-EG) hydrogels as a function of time in response to swelling in glucose solutions at $37^\circ C$ [reprinted with permission from (80); copyright 1997 American Chemical Society].

the formation of more gluconic acid by removing the hydrogen peroxide produced. Hassan et al. (80) used gels based on this technique to demonstrate that pH-sensitive hydrogels that contain glucose oxidase incorporated by Valuev's macromonomer technique are responsive to glucose concentrations, swell rapidly as glucose is converted to gluconic acid, and return to normal swelling states as the hydrogen ions dissipate from the localized area within the gel (Fig. 14).

Another method for glucose-sensitive insulin release was proposed by Kim et al. (74) and Okano and Yoshida (81), where glycosylated insulin is attached to concavalin A and encapsulated in a semipermeable membrane that allows glucose to diffuse in. Concavalin A prefers to bind to glucose, and insulin is released in proportion to the glucose that enters the capsule.

Okano (82) also demonstrated a concept using smart polymers to sense the presence of glucose. Insulin is encapsulated inside a composite polymer membrane that is cross-linked with boric acid (Fig. 15). Glucose disrupts the borate cross-links and opens small pores in the membrane, which allow the encapsulated insulin to diffuse through the network until the glucose level returns to normal. This method may require some improvements to function *in vivo* because the pH and the buffering effect of physiological fluids may disrupt the release mechanism.

Yuk et al. (83) showed that the pH- and temperature-responsive copolymers P(DMAEMA-co-EAAM) could be pressed into tablets that include glucose oxidase and insulin; an observed pulsatile release of insulin occurs, and release is turned on at glucose concentrations of 5.0 g/L and off when it dropped to 0.5 g/L (Fig. 16). This method makes it possible to combine sterilizable polymers and known amounts of insulin and glucose oxidase without the complications of forming hydrogel systems, which generally require a solvent. This method also may prevent degradation or inactivation of biological components, including both the enzyme and insulin, during the immobilization step.

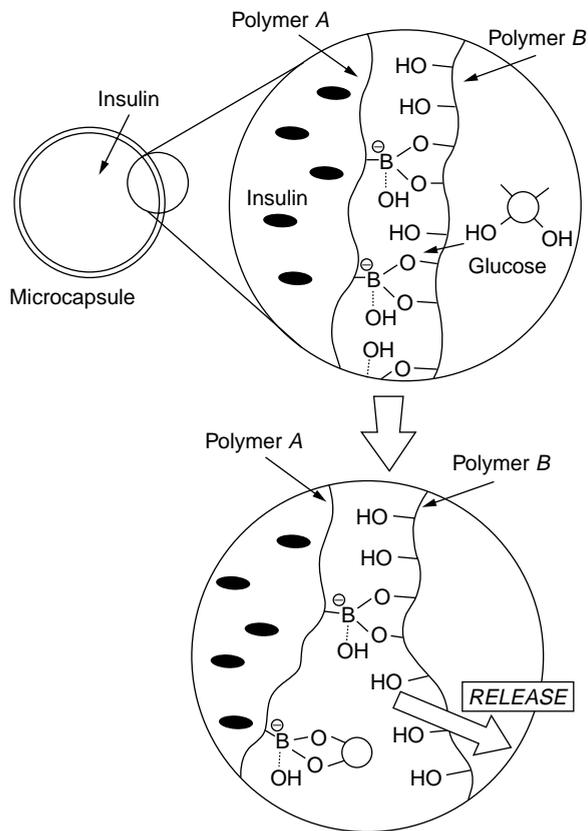


Figure 15. Concept of glucose-sensitive insulin release system using PVA/poly(NVP-co-PBA) complex system [reprinted with permission from (82) Figure 16; copyright 1993 Springer-Verlag GmbH & Co. KG].

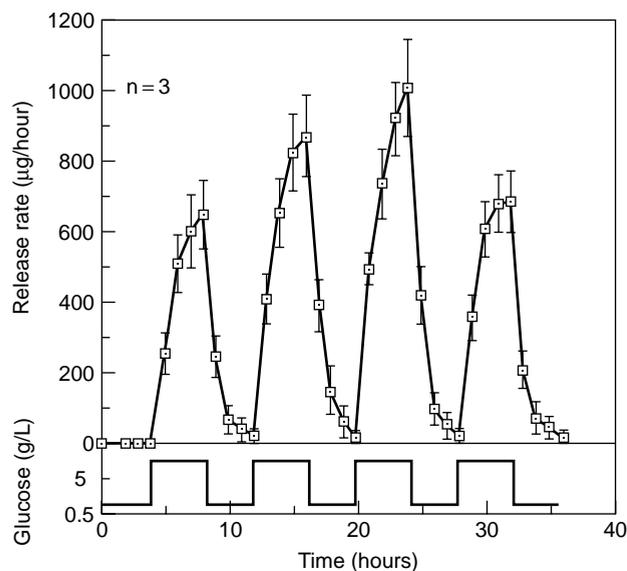


Figure 16. Insulin release from the insulin-loaded matrix in response to alternating change of glucose concentration [reprinted with permission from (83); copyright 1997 American Chemical Society].

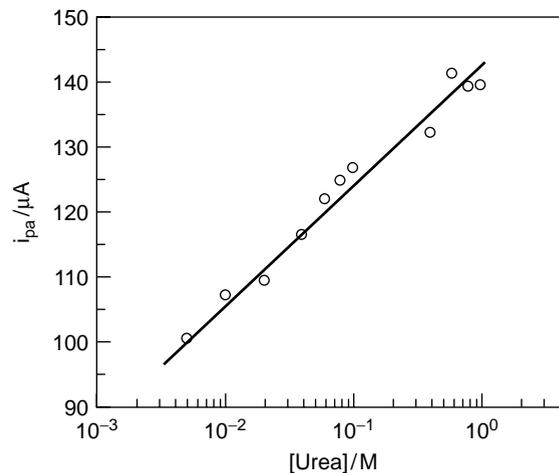


Figure 17. Urea-dependent changes in the cathodic peak current of cyclic voltammograms [reprinted with permission from (88); copyright 1994 American Chemical Society].

OTHER ANALYTES FOR BIOLOGICAL SENSING

There is tremendous potential for biosensors based on analytes besides glucose. In addition to the chemicals listed in Table 4, biosensors have been developed that respond to or measure pH, chloride levels, magnesium (18), bilirubin, blood gases (84), triglycerides (21), creatinine, and various saccharides (23,85). For example, Karube and Sode (85) used microbial detectors to determine fish freshness by immobilizing CO_2 -using bacteria to measure metabolic rates in fish. They also developed microbial immobilization methods for determining creatinine levels in kidney dialysis and enzymatic immobilization methods to determine hypoxanthine concentrations. Maeda et al. (86) formed block copolymers of poly(styrene-co-acrylonitrile) with poly(L-glutamate) which respond to Ca^{2+} and urea and produce smart materials that give a linear current response to urea concentration (Fig. 17), because the poly(L-glutamate) changes conformation in response to high urea concentrations.

Sirkar and Pishko (87) showed that hydrogel biosensors can detect galactose and lactose by incorporating their respective oxidases into polymer networks. Galactose sensing is useful in monitoring liver response to sepsis (88), and lactose monitoring can be used in sports medicine, myocardial infarction, and pulmonary edema to determine lack of oxygen supply to tissue (89). Sirkar and Pishko's biosensors produce a nanoamp range current proportional to galactose concentrations, but also noted that oxygen in the air surrounding the device reduces the response, due to enzyme inhibition by O_2 .

MODES OF RESPONSE IN SMART POLYMERS

As shown in Fig. 1, the use of smart materials in conjunction with sensing devices makes it possible to have a closed-loop device that responds by pH-stimulated drug

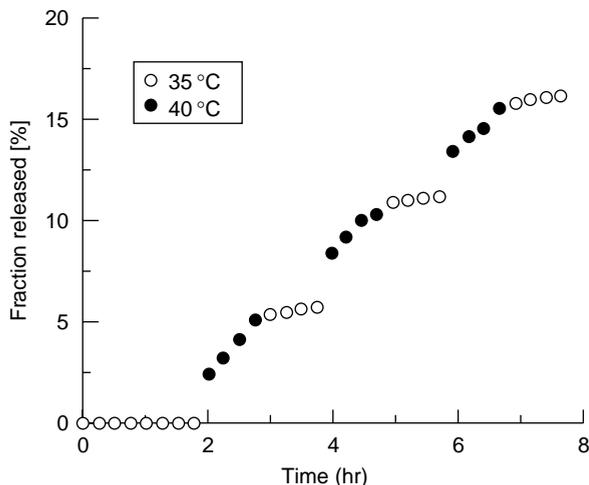


Figure 18. An “on-off” release profile of acetaminophen from NiPAAm/AAC gel in response to a temperature change between 35 and 40°C [reprinted from (93); copyright 1997, with permission from Elsevier Science].

delivery. Kaetsu et al. (4) immobilized acetylcholine esterase and glucose oxidase in poly(acrylic acid) to achieve a biosensor that has a pH-sensitive drug delivery feedback mechanism, and they showed that the device can deliver drugs in response to elevated substrate concentrations.

Ichikawa and Fukumori (90) developed temperature-sensitive networks that have small temperature-responsive beads made of poly(*N*-isopropylacrylamide) dispersed in ethylcellulose that allow the pores to open as temperature increased for drug delivery. Drug delivery based on pH- and temperature-responsive materials has been extensively researched (64,91–95), and is reviewed elsewhere within this article. Using materials that act similarly to amperometric biosensors, Guiseppi-Elie et al. (96) showed that electroconductive gels synthesized from polyaniline/polypyrrole could respond to an electric charge for direct delivery of peptides. These drug delivery systems do not respond directly to an increase in a particular molecular concentration (except for $[H^+]$ in pH-responsive systems), but they do sense changes in biological conditions that may occur naturally, such as pH-gradients in the gastrointestinal tract, pH changes due to the coagulation cascade, or temperature changes in tissue which is necrosed or nutrient-starved. Gutowska et al. (93) showed an on-off acetaminophen delivery system as a function of

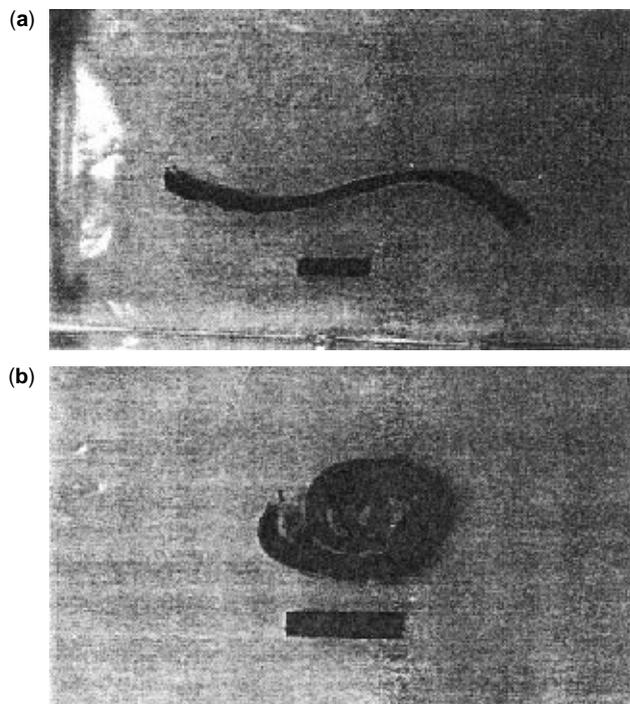


Figure 19. (a) A long bi-gel strip that has one PAAM side modulated by NIPA gel. At room temperature, the bi-gel is straight. (b) When the sample temperature is raised to 39°C, the gel becomes a spiral [From (Vol); copyright 1997; reprinted by permission of John Wiley & Sons, Inc.].

temperature for a gel made from *N*-isopropylacrylamide and acrylic acid (Fig. 18).

Targeted drug delivery devices also fit in the category of materials for biomedical sensing because they are designed to detect cellular or other biological surfaces or chemicals in the body to trigger release. Enteric delivery systems (97) are a well-used version of targeted systems. Other targeted systems rely on surface modification, so that the carrier recognizes the target tissue or biomolecules (98).

A unique response for diagnostic tests using biosensors and smart polymers is the change of device shape observed when asymmetrical gels are exposed to variations in temperature or pH. Zhang et al. (99) and Hu et al. (100) showed that PNIPAAm asymmetrical gels bend due to differential swelling capacities to change shape as temperature changes (Fig. 19). These gels have potential applications based on visual observation of analyte level changes and also on the possibility of controlling electrical conduction



Figure 20. Method of forming molecularly imprinted surfaces on polymers. (a) Target molecules (such as proteins or antibodies) are placed at the surface of a prepolymer solution. This solution is then polymerized or cross-linked in the presence of the target molecule at the surface (b), creating a geometry on the surface of the polymer which is complementary to the target molecule and can fit together like a jigsaw puzzle after surface extraction (c).

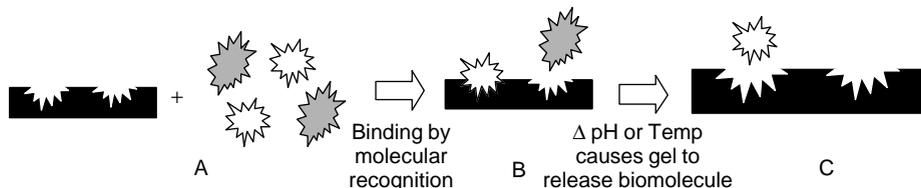


Figure 21. Mechanism of selective interaction between a molecularly imprinted polymer surface and biological molecules. The surface is designed to match the three-dimensional geometry of proteins or other biomolecules and prevent interactions with nontarget molecules. If the imprinted polymer is an environmentally sensitive material, a change in pH, temperature, or other conditions will provide a transition mechanism to attract and release the target molecule.

by alternately completing and disconnecting molecular switches if the asymmetric gels are formed from semiconducting polymers.

MOLECULAR IMPRINTING

Molecular imprinting is a relatively new research area that may have applications in biomedical sensing. Polymer templates are formed, as shown in Fig. 20, where an imprint of the target molecule is placed on the surface of the polymer or gel to enhance interactions and binding between the surfaces and mimic the complementary geometries of enzymes and substrates. These materials can easily be mass-produced and would be much cheaper than their biosensor counterparts, which use enzymes or cells for detection. The combination of smart materials and molecularly imprinted surfaces may make binding interactions reversible because the geometrically complementary site will match only under specific physiological conditions (Fig. 21).

A recent example of the development of molecularly imprinted polymers is reported by Sreenivasan (102), who created cholesterol-recognition sites in radiation-polymerized poly(2-hydroxyethyl methacrylate) which increased the affinity for molecular interactions between the polymer and cholesterol, so that it was easier to detect very small quantities of “imprinted” analytes. Arnold et al. (103) reported the use of molecularly imprinted polymers in combination with fiber-optic luminescence to create highly sensitive chemical sensors. Several other researchers have reported methods of synthesizing imprinted devices (104) and advantages of this technique (105). This field shows much promise for making synthetic biosensors that do not rely on often expensive enzymes or microbes for detection.

POSSIBILITIES FOR FUTURE DEVELOPMENT

Based on rapid advances in thin film chemistry, design of microchips, and tissue and cellular engineering, it is quite foreseeable that biological sensors can be made more reproducibly to detect analytes and disease states using complex enzyme or cellular reactions. Intelligent polymers will be used in these devices in the feedback loop—either in conducting a signal or in responding to treat the abnormality

directly—through combination with drug reservoirs, for example. Some of the challenges still lying ahead for the developing materials for biomedical sensing include improvements in biocompatibility, lifetime of the sensor, minimizing signal drift, maximizing sensitivity, developing the capability to recalibrate sensors *in vivo*, and using materials that are sterile or sterilizable.

BIBLIOGRAPHY

1. W. Pietro, in *Encyclopedia of Chemical Technology*, 4th ed., J.I. Kroschwitz, ed., Wiley, NY, 1992, pp. 208–221.
2. A. Kikuchi and T. Okano, in *Biorelated Polymers and Gels: Controlled Release and Applications in Biomedical Engineering*, T. Okano, ed., Academic Press, NY, 1998, pp. 1–28.
3. P.S. Stayton, A.S. Hoffman, Z. Ding, T. Shimobooji, C. Lackey, C. Cheung, N. Murthy, R. To, and O. Press, *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol.*, (1999), p. 95.
4. I. Kaetsu, S. Ueta, K. Uchida, and K. Sutani, *Proc. Int. Symp. Continuous Release Bioactive Mater.* **23**: 773–774 (1996).
5. J.F. Cassidy, A.P. Doherty, and J.G. Vos, in *Principles of Chemical and Biological Sensors*, D. Diamond, ed., Wiley, NY, 1998 pp. 73–132.
6. K.W. Johnson, D.J. Allen, J.J. Mastrototaro, R.J. Morff, and R.S. Nevin, in *Diagnostic Biosensor Polymers*, A.M. Usmani and N. Akmal, eds., American Chemical Society, Washington, DC, (1994), pp. 84–95.
7. E.A.H. Hall, in *Biosensors and Chemical Sensors: Optimizing Performance through Polymeric Materials*, P.G. Edelman and J. Wang, eds., American Chemical Society, Washington, DC, 1992, pp. 1–14.
8. I. Karube and M. Gotoh, in *Analytical Uses of Immobilized Biological Compounds for Detection, Medical and Industrial Uses*, G.G. Guilbault and M. Mascini, eds., Reidel, Boston, 1988, pp. 267–279.
9. T. McCormack, G. Keating, A. Killard, B.M. Manning, and R.O’Kennedy, in *Principles of Chemical and Biological Sensors*, D. Diamond, ed., J. Wiley, NY, 1998, pp. 133–194.
10. P.R. Coulet, in *Analytical Uses of Immobilized Biological Compounds for Detection, Medical and Industrial Uses*, G.G. Guilbault and M. Mascini, eds., Reidel, Boston, 1988, pp. 319–327.
11. U.E. Spichiger-Keller, *Chemical Sensors and Biosensors for Medical and Biological Applications*. Wiley-VCH, NY, 1998.
12. D. Diamond, “Overview,” in *Principles of Chemical and Biological Sensors*, D. Diamond, ed., Wiley, NY, 1998, pp. 1–18.

13. A. Poshossian, M.J. Schoning, P. Schroth, A. Simonis, and H. Luth, *Sensors and Actuators B: Chem.* **76**: (2001) 519–526.
14. O. Okay, *Prog. Polym. Sci.* **25**: (2000) 711–779.
15. M.J. Tierney, J.A. Tamada, N. Parris, R.O. Potts, and B. Berner, *Proc. Int. Symp. Continuous Release Bioactive Mater.* **25**: 69–70 (1998).
16. B. Berner, P. Joshi, A. Ewing, P. Vijaykumar, W. Abraham, A. Zhang, and P. Plante, *Proc. Int. Symp. Continuous Release of Bioactive Materials*, (1998), **25**, pp. 740–741.
17. V. Merino, A. Lopez, D. Hochstrasser, and R.H. Guy, *J. Controlled Release*, **61**: (1999) 65–69.
18. M.J. Tierney, J.A. Tamada, R.O. Potts, R.C. Eastman, K. Pitzer, N.R. Ackerman, and S.J. Fermi, *Ann. Med.* **32**: (2000) 632–641.
19. J.A. Hubbell, S.P. Massia, P.D. Drumheller, C.B. Herbert, and A.W. Lyckman, *Polym. Mater. Sci. Eng. Proc.* **68**: 30–31 (1992).
20. J.D. Andrade, C.-Y. Wang, D.-L. Min, R. Scheer, R. Stewart, J.-I. Sohn, and P. Triolo, *Proc. Int. Symp. Continuous Release Bioactive Mater.* (1996) **23**: pp. 4–5.
21. N. Akmal and A.M. Usmani, in *Polymers in Sensors: Theory and Practice*, N. Akmal and A.M. Usmani, eds., American Chemical Society, Washington, DC, (1996) pp. 2–23.
22. Tothill, I.E., J.D. Newman, S.F. White, and A.P.F. Turner, “Monitoring of the Glucose Concentration During Microbial Fermentation using a Novel Mass-Produced Biosensor Suitable for On-line Use,” *Enz. Microb. Tech.* **20** (1997) 590–596.
23. A. Manzoni, in *Analytical Uses of Immobilized Biological Compounds for Detection, Medical and Industrial Uses*, G.G. Guilbault and M. Mascini, eds., Reidel, Boston, 1988, pp. 71–82.
24. F. Mizutani, S. Yabuki, Y. Sato, and Y. Hirata, in *Polymers in Sensors: Theory and Practice*, N. Akmal and A.M. Usmani, eds., American Chemical Society, Washington, DC, (1996), pp. 46–56.
25. Wilkins, E. and P. Atanasov, “Glucose Monitoring: State of the Art and Future Possibilities,” *Med. Eng. Phys.* **18** (1996) 273–288.
26. Lee, W.E. and H.G. Thompson, “Detection of Newcastle Disease Virus Using an Evanescent Wave Immuno-based Biosensor,” *Can. J. Chem.* **74** (1996) 707–712.
27. Boiarski, A.A., J.R. Busch, L.S. Miller, A.W. Zulich, and J. Burans, “Integrated Optic Immunoassay for Virus Detection,” *Proceed. SPIE* (1995) 576–584.
28. Casimiri, V. and C. Burstein, “Co-immobilized L-lactate Oxidase and L-lactate Dehydrogenase on a Film Mounted on Oxygen Electrode for Highly Sensitive L-lactate Determination,” *Biosens. Bioelect.* **11** (1996) 783–789.
29. Loibner, A.P., O. Doblhoff-Dier, N. Zach, K. Bayer, H. Hatinger, C. Lobmaier, T. Schalkhammer, and F. Pittner, “Automated Glucose Measurement with Microstructured Thin-layer Biosensors for the Control of Fermentation Processes,” *Sensors Actuators B: Chem.* **B19** (1994) 603–606.
30. C.M. Dry, *Eur. Conf. Smart Struct. Mater.* (1992), **1**: pp. 407–411.
31. Hunt, C.A., R.D. MacGregor, and R.A. Siegel, “Engineering Targeted In Vivo Drug Delivery. I. The Physiological and Physicochemical Principles Governing Opportunities and Limitations,” *Pharm. Res.* **3** (1986) 333–344.
32. Torchilin, V.P., “Targeting of Drugs and Drug Carriers within the Cardiovascular System,” *Adv. Drug Deliv. Rev.* **17** (1995) 75–101.
33. G.W. Hastings, *Eur. Conf. Smart Struct. Mater. Proc. SPIE* **1**: 357–362 (1992).
34. X. Yang and J.R. Robinson, in *Biorelated Polymers and Gels: Controlled Release and Applications in Biomedical Engineering*, T. Okano, ed., Academic Press, NY, 1998, pp. 135–192.
35. M. Mascini, in *Analytical Uses of Immobilized Biological Compounds for Detection*, G.G. Guilbault and M. Mascini, eds., *Medical and Industrial Uses*, Reidel, Boston, 1988, pp. 153–167.
36. L. Campanella and M. Tomassetti, in *Analytical Uses of Immobilized Biological Compounds for Detection*, G.G. Guilbault and M. Mascini, eds., *Medical and Industrial Uses*, Reidel, Boston, 1988, pp. 169–176.
37. B.D. Ratner, *J. Biomed. Mater. Res.* **27**: (1993) 837–850.
38. M.N. Helmus *MRS Bull.* 33–38 (1991).
39. H. Park and K. Park, *Pharm. Res.* **13**: (1996) 1770–1776.
40. C.D. Ebert, E.S. Lee, J. Deneris, and S.W. Kim, in *Biomaterials: Interfacial Phenomena and Applications*, N.A.P.S.L. Cooper, A.S. Hoffman and B.D. Ratner, eds., American Chemical Society, Washington, DC, 1982, pp. 161–176.
41. J.S. Tan, D.E. Butterfield, C.L. Voycheck, K.D. Caldwell, and J.T. Li, *Biomaterials* **14**: (1993) 823–833.
42. A. Coury, J. Perrault, and D. Michels, *New Mater. Process.* **3**: 411–415 (1985).
43. P. Calvert, M. Oner, J. Burdon, P. Rieke, and K. Farmer, *Proc. SPIE* 354–361 (1993).
44. A.M. Usmani, in *Diagnostic Biosensor Polymers*, A.M. Usmani and N. Akmal, eds., American Chemical Society, Washington, DC, (1994), pp. 2–20.
45. J.W. Chen, D. Belanger, and G. Fortier, in *Biosensors and Chemical Sensors: Optimizing Performance through Polymeric Materials*, P.G. Edelman and J. Wang, eds., American Chemical Society, Washington, DC, 1992, pp. 31–39.
46. N. Wisniewski, B. Klitzman, and W.M. Reichert, *Proc. Int. Symp. Controlled Release Bioactive Mater.*, **25**: 73–74 (1998).
47. K. Ishihara, N. Nakabayashi, M. Sakakida, K. Nishida, and M. Shichiri, in *Polymers in Sensors: Theory and Practice*, N. Akmal and A.M. Usmani, eds., American Chemical Society Washington, DC, (1996) pp. 24–33.
48. A. Hoffman, *J. Controlled Release* **6**: 297–305 (1987).
49. Hubbell, J.A., S.P. Massia, P.D. Drumheller, C.B. Herbert, and A.W. Lyckman, “Bioactive and Cell-Type Selective Polymers Obtained by Peptide Grafting,” *Polym. Mater. Sci. Engin. Proceed* **68** (1992) 30–31.
50. S. Luo and D.R. Walt, *Anal. Chem.* **61**: (1989) 1069–1072.
51. L.I. Valuev, O.N. Zefirova, I.V. Obydenova, N.A. Plate, *J. Bioactive Compatible. Polym.* **9**: (1994) 55–65.
52. N.A. Platé, and L.I. Valuev, *Polym. Prepr.* **33**: (1992) 939.
53. P.B. Deasy, *Microencapsulation and Related Drug Processes*. Marcel Dekker, NY, 1984.
54. A. Singh, M.A. Markowitz, L.-I. Tsao, and J. Deschamps, in *Diagnostic Biosensor Polymers*, A.M. Usmani and N. Akmal, eds., American Chemical Society, Washington, DC, (1994), pp. 252–263.
55. T.D. Gibson, and J.R. Woodward, in *Biosensors and Chemical Sensors: Optimizing Performance through Polymeric Materials*, P.G. Edelman and J. Wang, eds., American Chemical Society, Washington, DC, 1992, pp. 40–55.
56. Y. Agi and D.R. Walt, *J. Polym. Sci. A Polym. Chem.* **35**: 2105–2110 (1997).

57. A.K. Piskin, in *Analytical Uses of Immobilized Biological Compounds for Detection, Medical and Industrial Uses*, G.G. Guilbault and M. Mascini, eds., Reidel, Boston, 1988, pp. 17–23.
58. J.H. Pazar, H.R. Knull, and P.L. Simpson, *Biochim. Biophys. Res. Commun.* **40**: (1970) 110–116.
59. J.E. Bailey and D.F. Ollis, *Biochemical Engineering Fundamentals*. 2e. McGraw-Hill, NY, 1986, pp. 204–211.
60. J.-P. Chen, D.-H. Chu, and Y.-M. Sun, *J. Chem. Tech. Biotechnol.* **69**: 421–428 (1997).
61. H. Cicek, and A. Tuncel, *J. Polym. Sci. A. Polym. Chem.* **36**: 543–552 (1998).
62. Y.H. Bae and I.C. Kwon, in *Biorelated Polymers and Gels: Controlled Release and Applications in Biomedical Engineering*, T. Okano, ed., Academic Press, NY, 1998, pp. 93–134.
63. A. Hoffman, in *Polymers in Medicine III*, C. Migliaresi et al., eds., Elsevier Amsterdam, (1988), pp. 161–167.
64. H.G. Schild, *Prog. Polym. Sci.* **17**: (1992) 163–249.
65. N. Ogata, *J. Controlled Release*, **48**: (1997) 149–155.
66. S. Ren and P.D. Frymier, *AIChE Extended Abstracts, 1999 Annu. Meet.* 15004.
67. N.C. Foulds, and C.R. Lowe, *Anal. Chem.* **60**: 2473–2478 (1988).
68. P.D. Hale, L.I. Boguslavsky, T.A. Skotheim, L.F. Liu, H.S. Lee, H.I. Karan, H.L. Lan, and Y. Okamoto, in *Biosensors and Chemical Sensors: Optimizing Performance through Polymeric Materials*, P.G. Edelman and J. Wang, eds., American Chemical Society, Washington, DC, 1992, pp. 1–14.
69. W. Schuhmann, in *Diagnostic Biosensor Polymers*, A.M. Usmani and N. Akmal, eds., American Chemical Society, Washington, DC, (1994), pp. 110–123.
70. J.J. Skaife, and N. Abbott, *Proc. AIChE Nat. Meet.* (1999), Session 15D09.
71. F. Mizutani, S. Yabuki, and T. Katsura, in *Diagnostic Biosensor Polymers*, A.M. Usmani and N. Akmal, eds., American Chemical Society, Washington, DC, (1994), pp. 41–46.
72. B. Linke, W. Kerner, M. Kiwit, M. Pishko, and A. Heller, *Biosensors & Bioelectronics* **9**: (1994) 151–158.
73. L. Sirkar and M.V. Pishko, *Proc. Int. Symp. Controlled Release Bioactive Mater.*, **25**: 120–121 (1998).
74. S.W. Kim, C.M. Pai, K. Makino, L.A. Seminoff, D.L. Holmberg, J.M. Gleeson, D.E. Wilson, and E.J. Mack, *J. Controlled Release*, **11**: (1990) 193–204.
75. K. Podual, F.J. Doyle, and N.A. Peppas, *Polymer* **41**: (2000) 3975–3983.
76. C.R. Martin, B. Ballarin, C. Brumlik, and D.R. Lawson, in *Diagnostic Biosensor Polymers*, A.M. Usmani and N. Akmal, eds., American Chemical Society, Washington, DC, (1994), pp. 158–168.
77. G. Fortier, J.W. Chen, and D. Belanger, in *Biosensors and Chemical Sensors: Optimizing Performance Through Polymeric Materials*, P.G. Edelman and J. Wang, eds., American Chemical Society, Washington, DC, 1992, pp. 22–30.
78. W.J. Sung, and Y.H. Bae, *Proc. Int. Symp. Continuous Release Bioactive Mater.* **25**: 742–743 (1998).
79. J.A. Tamada, N.J.V. Bohannon, and R.O. Potts, *Proc. Int. Symp. Controlled Release Bioactive Mater.* **23**: 190–191 (1996).
80. C.M. Hassan, F.J. Doyle, and N.A. Peppas, *Macromolecules* **30**: 6166–6173 (1997).
81. T. Okano, and R. Yoshida, in *Biomedical Applications of Polymeric Materials*, T. Tsuruta et al., eds., CRC, Ann Arbor, (1993) pp. 407–428.
82. T. Okano, *Adv. Polym. Sci.* **110**: (1993) 179–197.
83. S.H. Yuk, S.H. Cho, and S.H. Lee, *Macromolecules* **30**: (1997) 6856–6859.
84. D.R. Walt T.A. Dickinson, B.G. Healey, J.S. Kauer, and J. White, *Proc. SPIE* **7**: 111–116 (1995).
85. I. Karube and K. Sode, in *Analytical Uses of Immobilized Biological Compounds for Detection, Medical and Industrial Uses*, G.G. Guilbault and M. Mascini, eds., Reidel, Boston, 1988, pp. 115–130.
86. M. Maeda, K. Makano, and M. Takagi, in *Diagnostic Biosensor Polymers*, A.M. Usmani and N. Akmal, eds., American Chemical Society, Washington, DC, (1994), pp. 238–251.
87. K. Sirkar, and M.V. Pishko, *Anal. Chem.* **70**: (1998) 2888–2894.
88. M.S. Dahn, et al. *Surgery* **107**: 295–301 (1997).
89. G. Kenausis, Q. Chen, and A. Heller, *Anal. Chem.* **69**: (1997) 1054–1060.
90. H. Ichikawa and Y. Fukumori, *J. Controlled Release*, **63**: (2000) 107–119.
91. C.S. Brazel and N.A. Peppas, *J. Controlled Release*, **39**: 57–64 (1996).
92. S.K. Vakkalanka, C.S. Brazel, and N.A. Peppas, *J. Biomater. Sci. Polym. Educ.*, **8**: (1996) 119–129.
93. A. Gutowska, J.S. Bark, I.C. Kwon, Y.H. Bae, Y. Cha, and S.W. Kim, *J. Controlled Release*, **48**: (1997) 141–148.
94. L.M. Schwarte, K. Podual, and N.A. Peppas, in *Tailored Polymeric Materials for Controlled Delivery Systems*, I. McCulloch and S.W. Shalaby, eds., ACS Symposium Series **709**, (1997), pp. 56–66.
95. A.B. Scranton, B. Rangarajan, and J. Klier, *Adv. Polym. Sci.* **122**: (1995) 1–53.
96. A. Guiseppi-Elie, A.M. Wilson, and A.S. Sujdak, in *Tailored Polymeric Materials for Controlled Delivery Systems*, I. McCulloch and S.W. Shalaby, eds., ACS Symposium Series **709**: (1997), pp. 185–202.
97. L.-C. Dong and A.S. Hoffman, *J. Controlled Release*, **15**: 141–152 (1991).
98. Y. Nagasaki and K. Kataoka, in *Tailored Polymeric Materials for Controlled Delivery Systems*, I. McCulloch and S.W. Shalaby, eds., ACS Symposium Series **709**, (1997), pp. 105–116.
99. X. Zhang, Y. Li, Z. Hu, and C.L. Littler, *J. Chem. Phys.* **102**: (1995) 551–555.
100. Z. Hu, X. Zhang, and Y. Li, *Science* **269**: 525–527 (1995).
101. Y. Li, Z. Hu and Y. Chen, *J. Appl. Polym. Sci.* **63**: (1997) 1173–1178.
102. K. Sreenivasan, *Polym. Int.* **42**: (1997) 169–172.
103. B.R. Arnold, A.C. Euler, A.L. Jenkins, U.M. Uy, and G.M. Murray, *Johns Hopkins APL Tech. Dig.* **20**: 190–198 (1999).
104. A. Katz and M.E. Davis, *Macromolecules*, **32**: (1999) 4113–4121.
105. S.A. Piletsky, E.V. Piletskaya, T.L. Panasyuk, A.V. El'skaya, R. Levi, I. Karube, and G. Wulff, *Macromolecules* **31**: (1998) 2137–2140.

BIOMIMETIC ELECTROMAGNETIC DEVICES

MORLEY O. STONE
 RAJESH R. NAIK
 Wright-Patterson Air Force Base
 Dayton, Ohio

INTRODUCTION

Literally, the term *biomimetics* means to imitate life. Practically, biomimetics is an interdisciplinary effort aimed at understanding biological principles and then applying those principles to improve existing technology. This approach can mean changing a design to match a biological pattern, or it can mean actually using biological materials, for example, proteins, to improve performance (1). Biomimetics had its earliest and strongest footholds in materials science, and it is rapidly spreading to areas such as electromagnetic sensors and computer science.

The area of biomimetics covered in this article applies to sensing electromagnetic (EM) radiation. Volumes have been written regarding how higher organisms perceive visible light, and thus, this will be largely ignored here. Mostly, coverage is limited to biological sensing of EM radiation on either side of the visible, the ultraviolet, and the infrared. From a biological sensing perspective, spectral regions are defined as follows: near-ultraviolet wavelengths (λ) from 200–400 nm and infrared wavelengths from 0.75–15 μm (Fig. 1). Operationally, both the ultraviolet and infrared extend over larger wavelength intervals. This article outlines some of the electromagnetic detection/sensitive systems in biology.

BIOLOGICAL ULTRAVIOLET AND VISIBLE SYSTEMS

The electromagnetic spectrum extends from gamma and X rays of wavelengths less than 0.1 nm through ultraviolet, visible, infrared, radio, and electric waves. The solar spectrum of radiation that reaches the earth's surface ranges from 300–900 nm; radiation in the ultraviolet (200–300 nm) is mostly absorbed by the ozone layer in the upper atmosphere. Light that passes through the atmosphere reaches the earth's surface and also enters large water bodies. As penetration increases, the extremes of the visible spectrum are absorbed, allowing a narrow radiation of

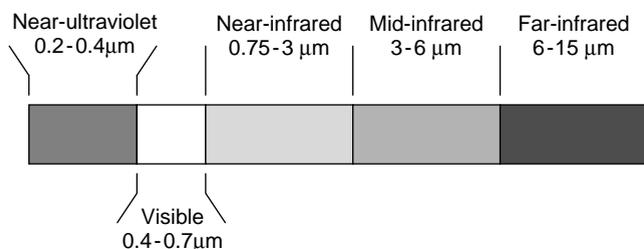


Figure 1. Chart of EM radiation definitions used in this article.

blue-green light (500 nm) to penetrate at depths greater than 100 meters. Radiation in the near ultraviolet, from 300–500 nm, is important in photobiological responses that include phototropism, phototaxis, and vision. For example, the spectral sensitivity of many insects and some birds is from 360–380 nm, and for invertebrates, it is from 500–600 nm. Radiation from 600–700 nm is important in photosynthesis, and radiation from 660 nm into the near-infrared is important for plant and animal growth, flowering of plants, and sexual cycles in animals. Nucleic acids and proteins absorb ultraviolet radiation at 260 nm and 280 nm, respectively. Absorption at these wavelengths produces damaging effects including mutations and cell death. Fortunately, organisms can repair the damaging effects of ultraviolet light on their genetic material by activating UV repair mechanisms.

Insect Vision

The discovery of visual sensitivity in insects dates back to the 1800s, but substantial proof of the visual sensitivity of insects was provided in the last 20 years or so. The spectral range visible to insects extends from the ultraviolet through the red. Studies of the behavioral response of insects suggest that insects have wavelength-dependent behaviors. The ability of an organism to discriminate differences in wavelength distribution and use this information to direct its behavioral response within a given environmental setting enables it to select its food source, avoid detection by prey, and identify potential mates (2).

The compound eyes of insects are made of eye facets, or ommatidia. The number of ommatidia varies from one species to another: about 10,000 in the eyes of dragonflies, 5500 in worker honeybees, 800 in *Drosophila*, and one in ants. Each ommatidium is a complete eye that consists of an optical system and the photoreceptor, which converts light energy to electric energy. The optical system consists of the corneal lens and the crystalline cone that transmits the image to the photoreceptors. The compound eye of the common fruit fly *Drosophila melanogaster* is composed of about 800 ommatidia (Fig. 2); each ommatidium is approximately 10 μm in diameter and 100 μm long. The ommatidium consists of a corneal lens, a crystalline cone, retinula cells, and a sheath of pigment cells that extend across the entire length of the rhabdom (2). The rhabdom consists of eight individual rhabdomeres (R1 to R8), but as seen in Fig. 2D, only R1–R7 are visible; R8 is not visible because it lies underneath R7 (1). The ommatidia of honeybees consist of nine rhabdomeres (3). The rhabdomeres contain unique photopigments (opsins) that have characteristic spectral properties.

Image formation depends on the optical properties of the corneal lens and crystalline cone that aids in maximizing the amount and quality of light and focuses that light onto the rhabdomeres. The chromoproteins (rhodopsins) within the rhabdomeres interact with visible photons and in turn convert light energy to electrical energy (4). The rhabdom acts as a light guide or waveguide, mainly because of its long narrow cylindrical geometry. The length of the rhabdom increases the probability that an entering photon of visible light will interact with the visual pigment.

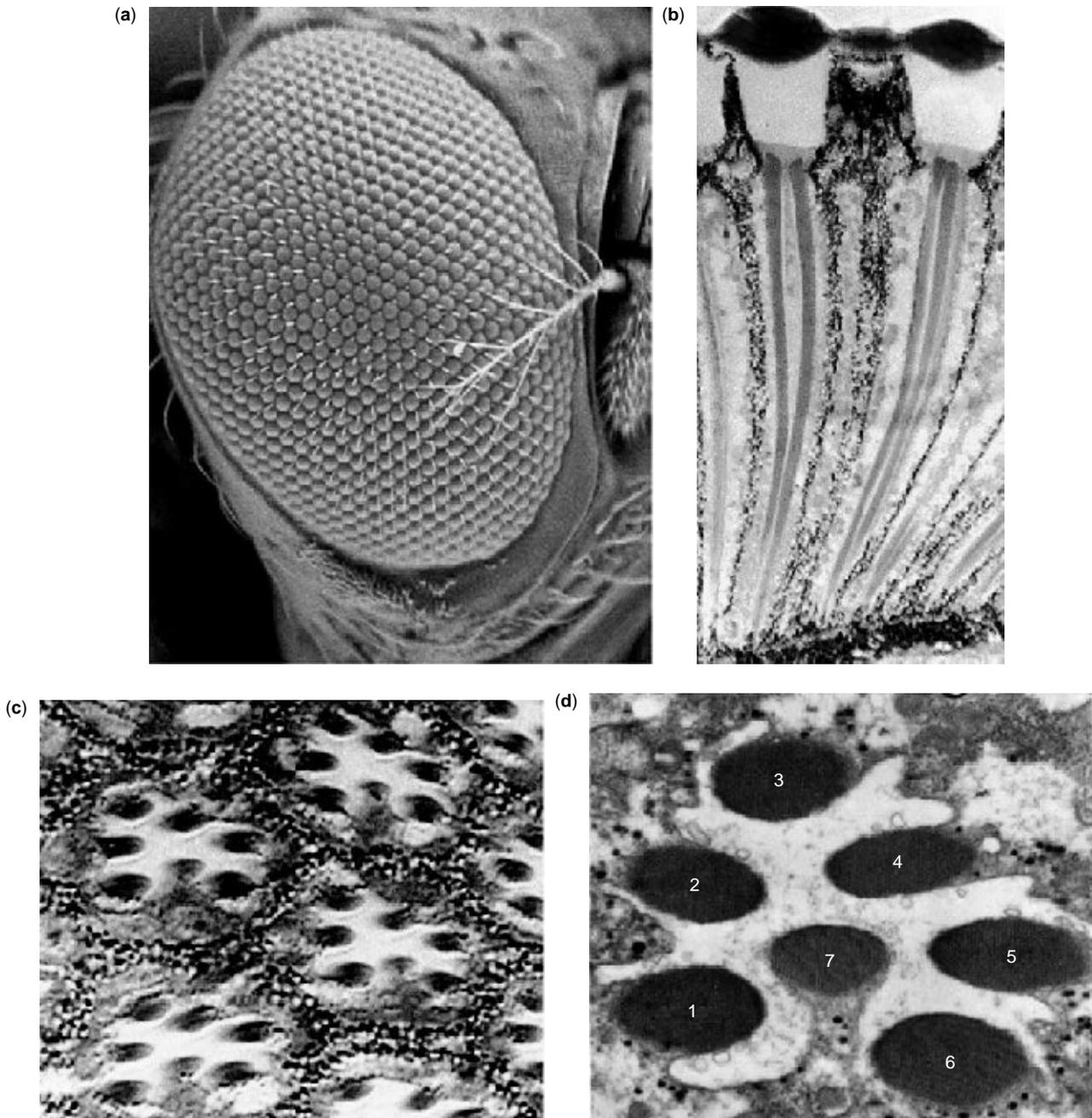


Figure 2. (A) Scanning electron micrograph of the adult eye; (B) longitudinal section through the eye showing the corneal lens, crystalline cone, rhabdom and the pigment sheath; (C) cross section through the ommatidium; and (D) cross section through the rhabdom to show the orientation of the photoreceptors (courtesy of John Archie Pollock).

Light that enters the rhabdom at a certain angle to its long axis is totally reflected and contained within the rhabdom, whereas light that enters at oblique angles is lost.

Rhabdomeres whose diameters ($0.5 \mu\text{m}$) are similar to the wavelength of light in the visible spectrum function as waveguides. The rhabdomere can transmit or guide this electromagnetic energy within its small cross-sectional area. The small cross-sectional area of rhabdomeres prevents light from being uniformly distributed which causes interference. This, in turn, causes the light to

propagate in patterns known as modes (dielectric waveguide). The effect of confining the photopigment within a rhabdomere of small cross section, it is believed, causes a shift from its visible absorptive peak to lower wavelengths and increases the UV peak absorption (5). In other words, the physical properties (size and shape) of rhabdomeres affect their spectral, polarization, angular, and absolute sensitivity. For example, in *Drosophila*, photoreceptor R7 that has a smaller diameter filters out a high proportion of the ultraviolet and blue light, whereas R8 (beneath R7)

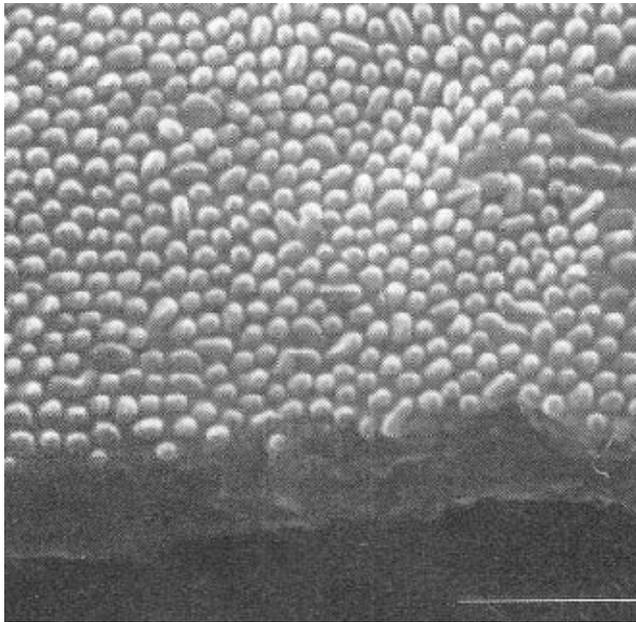


Figure 3. A scanning electron micrograph of a transparent insect wing (courtesy H. Ghiradella).

receives longer wavelengths. In bees, the rhabdomeres that are short and have a larger cross-sectional area mediate polarized vision (3). The sky appears bright blue because sunlight is scattered by molecules in the atmosphere (Rayleigh scattering), and as a result the light becomes polarized. The polarization pattern of skylight offers insects a reference for orientation. The ommatidia in the dorsal rim of compound eyes in insects are used as a polarized light detector (3). Rhabdomeres generally fall into three classes that are maximally sensitive to light: 350 nm (ultraviolet), 440 nm (blue), and 540 nm (green) (6).

Optical engineers have attempted to develop imaging systems that function like the eyes of animals. Although replicating the visual system of an eye is probably a very arduous task, imaging systems have been developed that combine light sensors, photocells, microchips, and cameras. Nature has developed optical systems in animals through millions of years of evolution, so much can be learned by studying animal or insect eye architecture. The compound eyes of insects have been quite informative in designing imaging devices. A multiaperture lens that has an array of glass rods arranged in a hemisphere was developed by Zinter in 1987. The multiaperture lens consists of hundreds of rod elements that have a graded index of refraction, and each rod or optic element acts as a single lens. Each lens transmits a small portion of the image, and at the focal point, each rod produces an overlapping image. The images are then transmitted via optic-fiber bundles, and the superimposition of the image creates an intensified image. This type of device has advantages such as detecting objects in low light and a wide field of view.

Antireflective Coatings

The surfaces of compound eyes of numerous insects appear smooth, but in certain insects, the front surface of the corneas are completely covered by protuberances known

as “corneal nipples” (7). These corneal nipples are an antireflective device in the broadband wavelengths from the near-UV to red (8). It is believed that these structures function to match the impedance of air to that of the lens cuticle; thereby they increase the transmission of light and decrease the reflection from the corneal surface across a broad wavelength range. The nipple array gives the insect two important advantages: greater camouflage and increased visibility.

Apart from insect eyes, insect wings also, it is known, function as antireflective devices. The transparent wing of a hawkmoth *Cephonodes hylas* functions as an antireflective device (9). It was shown that the wing reduces light reflectance by about 29–48% across a broad wavelength range (200–800 nm). An examination of the wing surface reveals the presence of nanosized protuberances similar to the corneal nipples in insect eyes (Fig. 3). The wing protuberances, like the corneal nipple, also function to match the impedance between the cuticle and air. The individual protuberances are not detectable under visible light because their small size limits visible light diffraction. Each single protuberance functions as an antireflective device. By packing the protuberances closely in two-dimensional space, nature has optimized the most efficient way for the wing to function as an antireflective device. The transparent wing is difficult to distinguish from the background and provides good camouflage to the insect. Interestingly, wings coated by a thin layer of nanosized gold particles (8 nm) reduced the metallic reflection of the gold particles presumably by changing the surface metal from a reflective to a dark absorptive one (9). The optical properties of the nanostructures on insect wings can be used as a model for designing new optical devices that consist of nanosized structures (Fig. 4).

Butterfly Wings

The wings of butterflies are adorned by beautiful patterns and colors. Some wings are uniformly colored, whereas others reflect yellow, orange, red, green, blue, violet, or black. The colors of butterfly wings as well as insect eyes are generated by interference, diffraction, and scattering. The nature of the wing surface structure leads to the absorption of certain wavelengths and the reflection of other wavelengths of light; in addition, some wavelengths may even be transmitted. The rainbow-like display of colors on butterfly wings is caused by iridescence, due to the reflection from multiple thin-film interference filters on the wing scales (10). For example, the metallic iridescence of *Morpho* butterflies is an interference color attributable to the structural features of its wings. Interference colors result from the reflection of light from a series of superimposed structures separated by distances equal to the wavelengths of light. The wing scale of the butterflies consists of a flat basal plate that carries a large number of vanes (or ridges) that run parallel to the length of the scale (Fig. 5). Each vane consists of a series of obliquely horizontal and vertical lamellae oriented lengthwise on the scale and their spacing is comparable to the bands of reflection. In *Morpho* wing scales, the horizontal lamellae are approximately 185 nm apart and are stacked in an alternate fashion to each other. *Morpho* butterfly wings produce a metallic blue

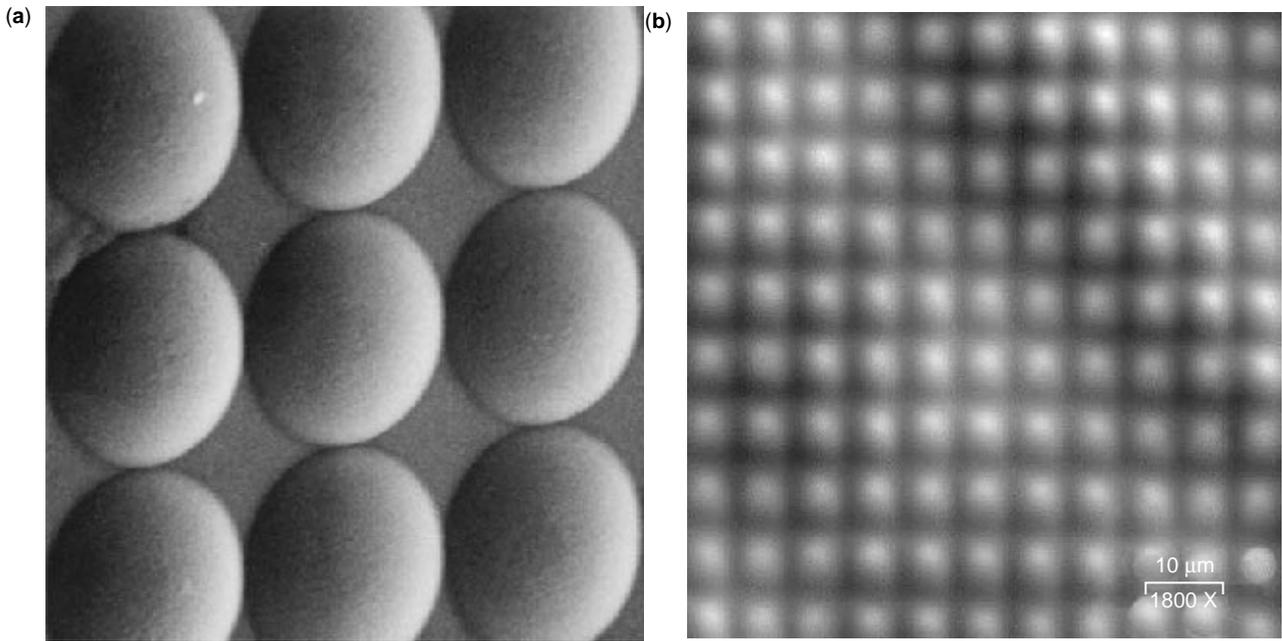


Figure 4. The fine nipples function as an antireflective coating. (a) Microlens array that functions to focus light onto the active area of a sensor; (b) subwavelength structure that functions as an antireflective surface (courtesy SY Technology, Inc.).

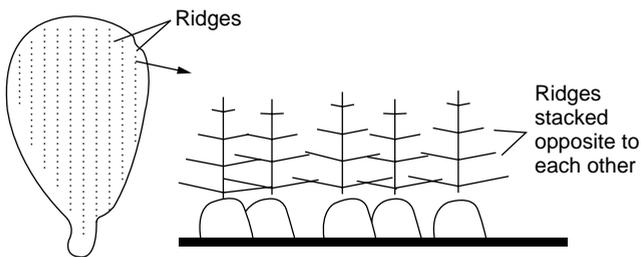


Figure 5. Butterfly wing scale structure.

interference color. *Morpho* butterflies can be seen from low-flying aircraft due to the iridescent colors of their wings and are visible from a quarter of a mile (11).

The male *Eurema* wing scales reflect light whose peak wavelength is 350 nm. The lamellae in the *Eurema* wing scales are stacked opposite each other and separated by an equal spacing of 83 nm. Ghiradella et al. (1972) demonstrated ultraviolet reflection from wing surfaces when the wing was tilted by 20° with respect to the incident light. The flapping of butterfly wings caused a change in hue proportional to the angle of the wing. A change in reflectance, abrupt intensity change, and a strong ultraviolet component that contrasts with the background may serve as long-range communication signals between insects. As mentioned earlier, insect compound eyes are maximally sensitive to UV, and hence, UV patterns caused by this reflection may serve as a potential source of communicative signals. Furthermore, vegetation generally absorbs UV wavelengths in this region, and this may serve to maximize color contrast with respect to insect vision. Most vertebrates do not see UV, so this spectral region may serve as a private channel of communication among insects.

BIOLOGICAL INFRARED DETECTION

The ability of biological organisms to sense infrared radiation has been studied in snakes, beetles, moths, bacteria, numerous other organisms, and even subcellular organelles. In much of this study, there is overlap and confusion between differentiating a process as infrared photon detection versus thermal detection. This is evidenced by the fact that literature comprising the infrared sensing area is quite limited; however, literature in the thermoreception area is voluminous. This section treats infrared and thermal reception as indistinguishable processes from a biological standpoint, and the terms are used interchangeably. Additionally, this treatment extends to three experimental infrared/thermal systems: snake, beetle, and bacteria.

Before delving into the specifics of biological infrared detection, a brief tutorial on blackbody radiation and thermal sources of IR is needed. Warm objects, such as mammals, emit energy in the infrared part of the electromagnetic spectrum. Table 1 lists three different temperatures: 300 K is listed as a background temperature, 310 K is listed to represent 37°C prey, and the 1000 K listing is representative of a forest fire. The 310 K listing and the 1000 K listing pertain to the snake model and beetle model of infrared

Table 1. List of Blackbody Infrared Emission Values For Objects at Various Temperatures

Temperature (K)	λ_{\max} (μm)	Total Flux (W/cm^2)	Bandpass Flux (W/cm^2)	
			3–5 Microns	8–12 Microns
300	9.66	4.59e-02	5.84e-04	1.21e-02
310	9.35	5.23e-02	8.31e-04	1.42e-02
1000	2.90	5.67	2.04	5.04e-01

reception, respectively, because each system tries to detect objects at these temperatures.

At first glance, the main difference between snake-based and beetle-based infrared detection is the wavelength region of peak intensity (λ_{max}). Cooler objects, for example, mammals at 37°C, emit maximally in the far-IR, in the 8–12 μm atmospheric transmission window. As an object becomes hotter, for example, a forest fire at $\sim 750^\circ\text{C}$, the λ_{max} shifts to shorter wavelengths that place it in the 3–5 μm atmospheric transmission window. Roughly two orders of magnitude more total flux come from a 1000 K object compared to that from a 310 K object. An object at 310 K emits 27% of its total flux in the 8–12 μm bandpass and 1.6% in the 3–5 μm bandpass. Alternatively, an object at 1000 K emits 8.9% of its total flux in the 8–12 μm bandpass and 36% in the 3–5 μm bandpass. This discussion of infrared emitting objects and which is the better emitter is important to keep in mind as we discuss biological infrared detectors.

Bacterial Thermoreception

Cellular processes are influenced by temperature, and therefore, cells must possess temperature-sensing devices that allow for the cell's survival in response to thermal changes. Virtually all organisms show some kind of response to an increase or decrease in temperature, but sensing mechanisms are not well understood. When bacterial cells are shifted to higher temperatures, a set of proteins known as "heat-shock" proteins are induced. These proteins include molecular chaperones that assist in re-folding proteins that aggregate at higher temperatures as well as proteases that degrade grossly misfolded proteins (12,13). Changes in temperatures can also be sensed by a set of coiled-coil proteins called methyl-accepting proteins (MCPs), that regulate the swimming behavior of the bacterium *Escherichia coli* (14). Coiled-coil proteins are formed when a bundle of two or more alpha-helices are wound into a superhelix (Fig. 6) (15). The MCPs can be reversibly methylated at four or five glutamate residues (16). Methylation and demethylation, it is presumed, is the trigger that dictates the response during temperature

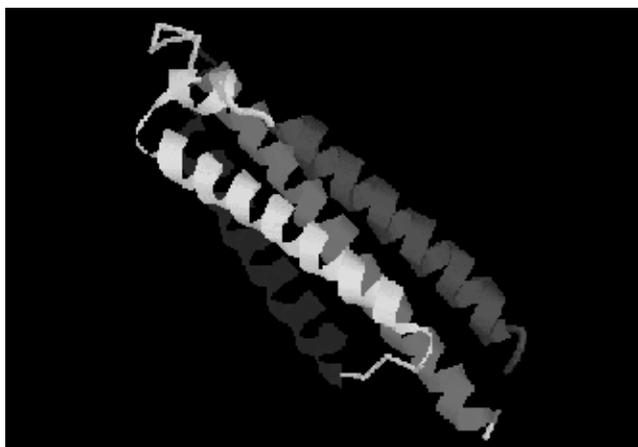


Figure 6. A cartoon showing the coiled-coil structure of MCP-II from *Escherichia coli*.

changes. The mechanism through which MCPs sense temperature is still not fully understood. In *Salmonella*, a coiled-coil protein known as TlpA has been identified as a thermosensing protein (17). TlpA regulates the transcription of genes by binding to sequence-specific regions on the DNA molecule. At low temperatures ($<37^\circ\text{C}$), TlpA interacts with another molecule of TlpA to form a functional (dimeric) molecule. As the temperature increases, TlpA dissociates from itself and becomes nonfunctional. However, the unwinding of TlpA helices is highly reversible, and a downshift in temperature leads once again to the formation of functional dimers. Because TlpA is not irreversibly denatured, it serves as an active thermosensing device. The fact that the denaturation and renaturation process is rapid allows cells to adapt quickly to changes in temperature. As shown in Fig. 7, the change in the structure of TlpA was measured by circular dichroic spectroscopy as a function of temperature. We observed that the thermal unfolding–folding is reversible and the protein displayed 100% recovery. To date, of all the proteins tested by us, TlpA exhibits the highest degree of reversibility with respect to this thermal unfolding transition. It is likely that TlpA, as well as MCPs, represent an adaptation of the coiled-coil motif as a temperature sensor by coupling its folding and unfolding to temperature cues. In addition, the ability of short synthetic coiled-coil peptides to undergo rapid thermal denaturation and renaturation (Naik and Stone, unpublished observations), suggests that the coiled-coil motif would be a model for designing new peptide-based thermosensing devices.

Snake Infrared Reception

The longest and best studied system of biological infrared sensing is the snake system. Snakes from two families, Crotalidae (pit vipers) and Boidae (boas and pythons), can sense infrared radiation by using specialized organs. In the crotalines, two infrared pit organs are positioned on either side of the head between the eyes and upper jaw. In boids, an array of infrared pit organs line the upper and lower jaw, and the number of pit organs is species specific. The ability of these organs to detect thermal energy was first

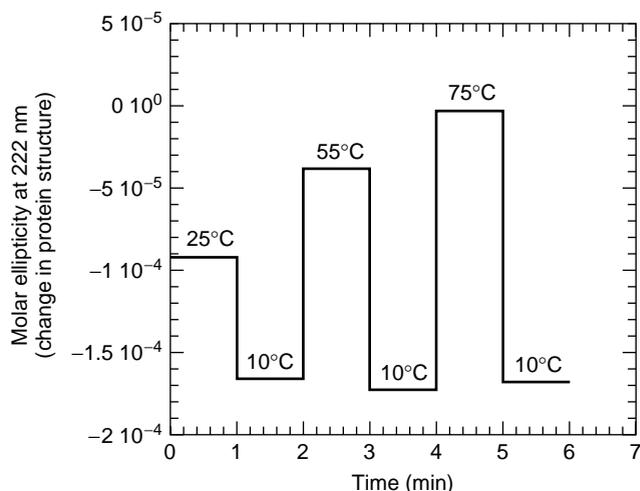


Figure 7. Reversibility of the thermal unfolding of TlpA.

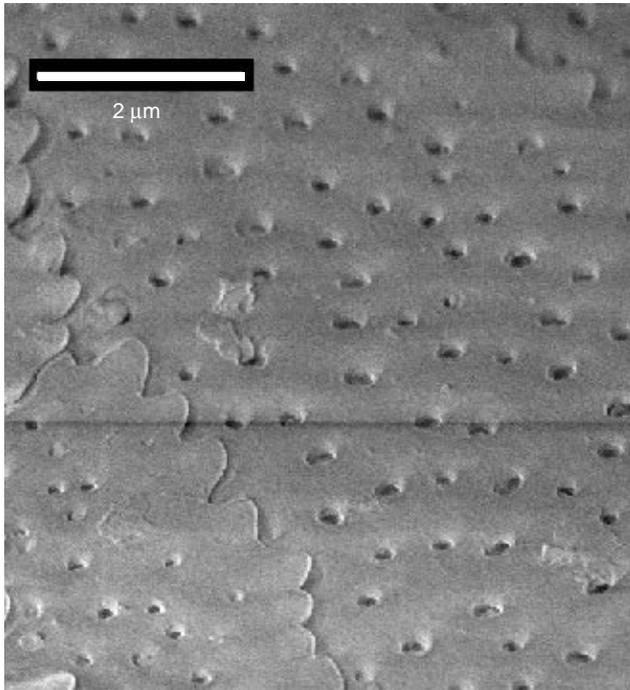


Figure 8. SEM micrograph of IR pit organ surface.

described by Noble and Schmidt in the 1930s (18). Bullock and co-workers at UCLA further defined this area by their electrophysiological studies in the 1950s. His publications from this period continue as the referenced sources for the stated sensitivity of 0.003°C for crotaline infrared pit organs (19,20). Hartline continued to further the study of thermoreception in snakes throughout the 1970s, and he wrote a wonderful review article for the layperson in 1982 (21). For more than three decades, the center of snake infrared research has been in Japan based on the work of Terashima and Goris. Recently, this group published a book that compiles their research papers from this past decade (22).

Much of this previous body of work has been electrophysiological and descriptive using electron microscopy techniques. We recently published a detailed examination of the morphology of Boidae infrared pits using both atomic force microscopy (AFM) and scanning electron microscopy (SEM) (23). Our results were consistent with the earlier results of Amemiya et al. (24). In both publications, the function of the unique surface morphology that covers the infrared pit organs was speculated about (see Fig. 8). This speculation centered on the hypothesis that unwanted wavelengths of light, that is, visible, were being scattered and desired wavelengths of light, that is, infrared, were being preferentially transmitted.

To prove the speculation about visible light, we conducted a series of spectroscopy experiments to test the spectral properties of infrared pit scales compared to other parts of the snake (Fig. 9). This data suggested that the IR pit organ surface microstructure indirectly aids infrared detection by scattering unwanted visible wavelengths of light. Using various samples and repeated measurements, there was consistently more than a fourfold reduction in the amount of transmitted visible light. This loss of

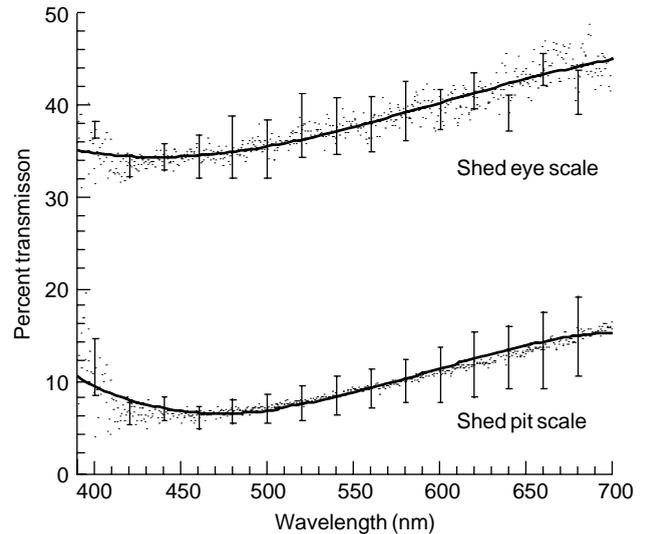


Figure 9. Fiber-optic spectrophotometry, visible wavelengths.

transmission was attributed to scatter due to measurements using a helium–neon laser at 632 nm and a silicon detector. Shed IR pit skin transmission dropped faster as a function of detector distance compared to eye scale transmission; this indicated an increased scattering angle and limited sample absorption. The increased visible light scatter can be accounted for by using a simple Rayleigh model of scatter and incorporating the micropit dimensions of different snakes (23).

This difference in skin surface morphology as a function of location on the snake is a wonderful example of evolved tissue engineering. These unique dimensions are confined to a few square millimeters within the IR pit organ. From the standpoint of chemical composition, there is no difference, as indicated by FT-IR analysis (Fig. 10). The FT-IR spectra from shed IR pit skin and shed spectacle (eye) skin are identical to the amide bands of keratin that dominate the absorbance profile. Interestingly, note that regions of high skin transmission correspond to regions of high atmospheric transmission (3–5 and 8–12 microns).

As mentioned previously, the sensitivity of crotaline (pit viper) infrared detection, widely stated as 0.003°C , refers to the seminal work by Bullock and co-workers (20). However, this value was never measured directly but rather extrapolated from calculated assumptions. Furthermore, the measured values were determined as water was running over the pits—a conductive mode rather than a radiant mechanism of heat transfer. The function of prey detection has been studied extensively for these sensors (25). Bearing this function in mind, we attempted to examine the phenomenon of snake infrared reception in the context of the thermal radiative transfer among the sensor, prey, and background.

The actual molecular mechanism for infrared pit organ function is an active area of research in our group and others. Several models were proposed by de Cock Buning, and based on his work, we sought to construct a radiative transfer model that would measure the radiant flux of a biological object as a function of distance (26–28). De Cock Buning (27) presented thresholds and corresponding

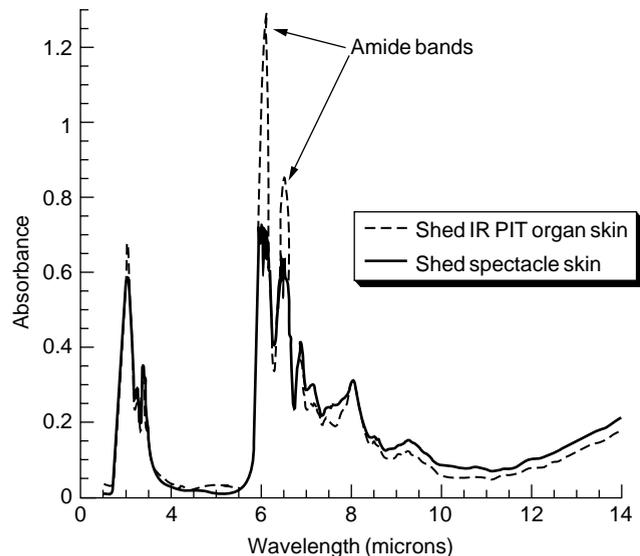


Figure 10. FT-IR analysis of shed crocotaline skin.

detection ranges, but this analysis did not take into account the form factor relationships between emitter and detector and ignored the effect of the thermal background from the soil and atmosphere. The output from our model is the change in radiant flux (ΔQ) at the infrared pit organ as a 37°C object is moved. When this value becomes negative, the object (prey) no longer has a thermal signature greater than the background—essentially, it becomes invisible from an infrared, or thermal perspective. What was surprising in this analysis was how quickly the ΔQ value became negative, indicating extremely short detection distances of the order of <4 cm. The specifics of this model have been published elsewhere (29).

This modeling result raises very probing questions about the function of the infrared pit organs and suggests limited function in long-range prey detection. Instead, we agree with the speculation of Theodoratus et al. that IR sensing may be playing a role in strike orientation (25). Interestingly, when we apply this same type of modeling analysis to the beetle infrared system, we agree with the sensing distance quoted by Schmitz that a 10-hectare fire can be sensed at a distance of 12 km (see later) (30). This limited detection distance also raises questions as to how this research can contribute to IR sensor technology. Approaches that we are taking to increase the efficiency of biological thermal detection are covered in Biomimetic Applications.

Beetle (*Melanophila acuminata*) Infrared Sensing

The Buprestid family of beetles encompasses the genus *Melanophila*; for almost sixty years, research has shown that it is attracted to fires and smoke (31). Evans published the first scientific analysis of *Melanophila acuminata*'s response to specific infrared wavelengths (32). This early work documented *Melanophila*'s ability to detect forest fires at extreme distances. A current estimate is that it can detect a 10-hectare fire at a distance of 12 km, but distances as long as 50 km were proposed in the early literature

(30,31). The obvious question is why these particular beetles are attracted to forest fires. The answer is that many insects are drawn to forest fires because the burnt trees lack a natural defense against insect larvae and *M. acuminata* is the best characterized insect in this regard.

Estimated forest fire temperature is between 500 and 1000°C . We have chosen 1000 K as an approximate median, and as mentioned in the introduction to this section, an object at this temperature would emit radiation maximally around $3\ \mu\text{m}$. Therefore, this family of beetles responds to a fundamentally different part of the infrared spectrum, the $3\text{--}5\ \mu\text{m}$ atmospheric transmission window, compared to snake infrared reception at longer infrared wavelengths. Another distinction between the two systems is that snake IR reception is definitely thermal and may or may not involve mechanoreception; however, it is most likely that *Melanophila* IR reception is based on mechanoreception. In a recent report, Gronenberg and Schmitz analyzed the neurons from *M. acuminata* IR sensilla and postulated that they evolved from mechanosensory ancestors (33).

The IR pit organ of *M. acuminata* is a small structure that measures approximately $450 \times 200\ \mu\text{m}$. *Melanophila* possess two such organs under the second set of mesothoracic legs. Each organ is comprised of 70–100 spherical sensilla that are approximately $15\ \mu\text{m}$ in diameter each (Fig. 11). The only other components of the IR organ are

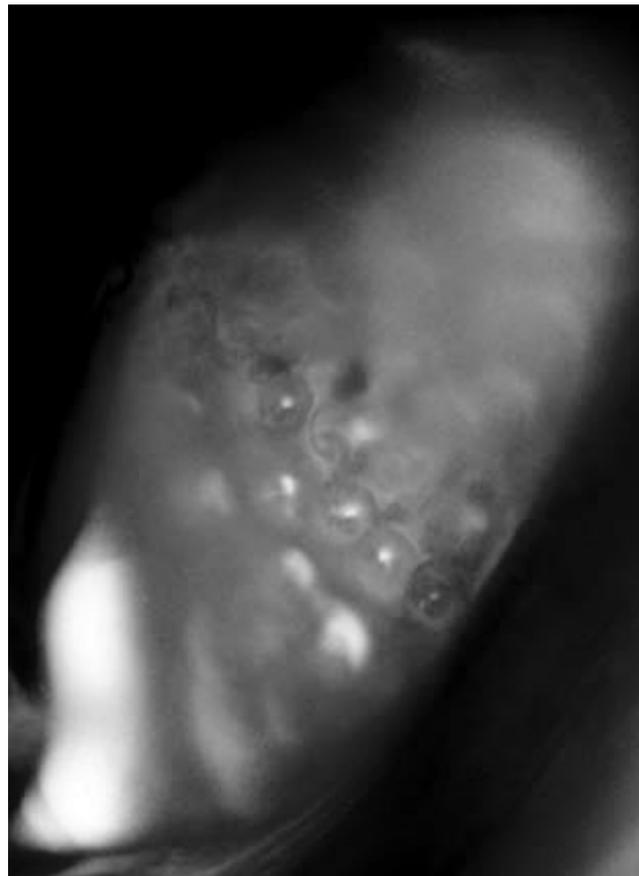


Figure 11. Optical microscope image of *Melanophila acuminata* IR organ.

numerous wax glands that, it is hypothesized, keep the organ free of dust and dirt (34). The chemical composition of insect cuticle (chitin) and the secreted wax contain numerous C–H, N–H, and O–H chemical bonds that respond via stretch and vibrational resonances in this 3–5 μm wavelength range. Upon absorption of 3- μm radiation from a sufficiently hot thermal source, for example, a forest fire, the sensillum are thought to expand approximately 1 nm. This minute expansion is sufficient to trigger the firing of a mechanoreceptor at the base of each sensillum.

From an application standpoint, the IR pit organs of *M. acuminata* are a much more attractive target for biomimetic EM sensing. The known mechanical nature of the organ and the unique morphology of the sensilla make attractive targets for replication in a biosensor. In fact, we have begun to view this organ as nature's equivalent to the Golay detector developed in the 1950s. The following is a definition from Hudson's book on infrared detectors (35): "(An absorber) is heated by the incident radiation, which in turn heats the gas in the chamber. The resulting increase in pressure is observed optically by the deflection of a small flexible mirror." If one were to replace the small flexible mirror by a mechanosensitive neuron, this definition of a Golay cell's operation becomes very similar to the way *Melanophila's* sensilla detect IR radiation. Particulars regarding this type of biosensor development are covered in the next section.

ELECTROMAGNETIC APPLICATIONS OF BIOMIMETIC RESEARCH

When examining the landscape of biomimetics, the application is obvious in many areas, and many of these applications are defense-related. The study of fish swimming has obvious tie-ins to underwater locomotion and naval interests (36). Much of the work in structural biomimetics has Army interest due to the potential of producing next-generation, lightweight armor based on naturally occurring, biological composites (37,38). From a commercial standpoint, few biomimetic results have been as exciting as the recent successes in the biocatalyzed formation of silica and silica polymerization (39,40). A major portion of our economy, especially the technology sector, is based on manipulating silicon. It is easy to see why the ability to manipulate this element under benign, ambient conditions by using enzymes has many people excited.

Sensing electromagnetic radiation is of particular interest in aviation because of the increasing distances over which sensors operate. The ability to detect EM in the infrared without cryogenics has been an important technology driver because of increased sensor reliability and reduced payloads. The latter are becoming more important as space migration dominates defense and commercial interests. Against this backdrop, it is easy to see why biomimetics, and in particular biomimetic EM sensing, has been a growing part of research in many funding agencies.

We already discussed in the first section (Biological Ultraviolet and Visible Systems) how nature evolved incredibly intricate coatings and patterns to reflect, absorb, and transmit light. The complexity of these natural

coatings has made replicating them a challenge. Many of the curved surfaces involved in biological coatings, for example, the hawkmoth's corneal nipples and *Melanophila's* domed IR sensilla, would require gray-scale lithography, which at present is not a "standard" technique in micro- and nano fabrication. However, the 15- μm domed structure of each IR sensillum is giant compared to the feature sizes currently being produced by the microprocessor industry. Commercial companies are currently engaged in applying advanced lithographic procedures to replicate biological surfaces, and many of these lithographic techniques are being applied to nonstandard, that is, nonsilicon, materials like germanium (41).

The ability of insect structures like hairs or microscopic spines to gather electromagnetic radiation was postulated by Callahan (42). In that publication, insect antennae are considered dielectric waveguides that work in the infrared. Similarity is drawn between this biological structure and a drawing of an electromagnetic wave energy converter (EWEC) that was patented through NASA (US Pat. No. 3,760,257) for converting microwave EM energy into electrical energy (42).

In our own research, replicating the surface structure of boid and crotaline infrared pit organs has been a top priority. We feel that the replicating this surface structure would be an important advancement in optical coatings for infrared optics. The micropits of the IR pit organ are approximately 300 nm in diameter and the scale ridges are spaced at 3.5 μm . This latter dimension has implications in the infrared, and the former dimension has visible light consequences, as mentioned earlier. In recent publications, we have reported successful holographic duplication of snake scale structure in a photopolymer matrix (43,44). In using a holographic approach, light is used to record the fine details of a biological surface. By combining this "reading" beam and a reference beam, the resulting interference pattern can record a multitude of biological information.

Before proceeding from coatings to the application of biomimetics to infrared sensors, we will briefly review the state of artificial or man-made sensors. For further reference, *Infrared System Engineering* by Richard Hudson is an excellent source of information (35). Broadly, infrared sensors fall into two categories: thermal and photon (quantum) detectors. On the thermal side are thermocouples, thermopiles, bolometers, and pneumatic (Golay) detectors. The microbolometer format currently dominates this class of noncooled IR detectors in state-of-the-art detectors. On the photon detector side are photoconductive, photovoltaic (p-n junction), and electromagnetic detectors. In general, this class of detectors is cryogenically cooled and made from semiconductor materials. In comparing these two classes of detectors, speed has always been a big differentiator; thermal detectors respond relatively slowly times (ms) versus photon detectors (μs).

From a biological perspective, it is clear that biological infrared sensing is thermal. This conclusion arises from the fact that the pit organs are not made of semiconductors and that IR photons in the mid- to far-IR region of the EM spectrum simply lack sufficient energy to catalyze or trigger a conventional biochemical reaction. This is graphically represented in Fig. 12; the energy required to move an

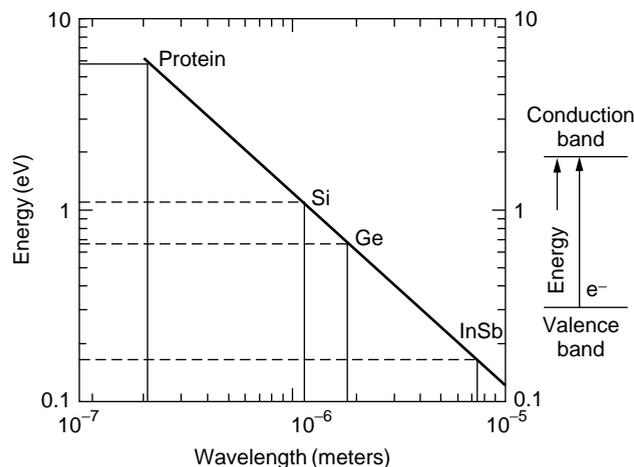


Figure 12. Graph of energy (quanta) as a function of wavelength for various detector materials.

electron into the conduction band is plotted on the y axis and the wavelength on the x axis. In a photon detector, arriving quantized energy displaces electrons from the valence band to the conduction band. In semiconductors like InSb (indium antimonide), the forbidden energy gap is small, so that the energy contained in a mid- to long-IR photon is still sufficient to move an electron across this barrier. A material such as silicon has a larger forbidden energy gap, so that a photon past $\sim 1 \mu\text{m}$ in wavelength no longer possesses enough energy to move an electron into the conduction band (Fig. 12). If one extends this treatment to a generalized protein, the main intrinsic absorption at 220 nm (via the amide bond) correlates with an energy gap that can be bridged only by high-energy photons outside of the infrared region of the EM spectrum. Even highly conjugated biological chromophores, for example, chlorophylls, cannot use light that is beyond the very near-infrared.

After reaching this conclusion that biological infrared sensing is thermal, how then does one apply this knowledge to new detector strategies? To compete with an artificial inorganic detector that directly converts a photon to an electron, one needs to make the biological thermal process more efficient. In a biological system, infrared photons are absorbed in the form of bond vibrational and stretch resonant frequencies inherent in the chemical structure of the tissue. This molecular motion is eventually dissipated as thermal energy on a very minute scale. We believe that this is enough of a thermal change to alter the dynamic ionic concentration gradient maintained in the terminal nerve masses of the IR pit organ that eventually leads to a change in the neuronal firing rate. This change in neuronal firing rate is interpreted by the brain as either “hot” (increased rate) or “cold” (decreased rate). A successful biomimetic approach would simplify this process by engineering the “trigger” in this process, the original IR absorbing biological macromolecule.

A model for this engineering process is the aforementioned bacterial thermoproteins. The ability to manipulate bacterial genes easily and produce the desired recombinant proteins via fermentation make this a model system. To increase the efficiency of this biological system, we are exploring ways of optically sensing thermally induced

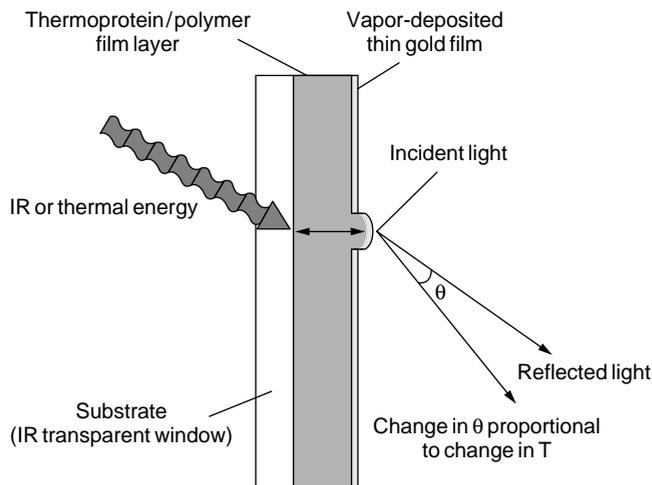


Figure 13. Schematic drawing of a thermal sensor based on a thermoprotein.

changes in protein structure. The use of circularly polarized light in circular dichroic (CD) spectroscopy is a common laboratory technique that optically records changes in protein secondary structure. A recent publication examined temperature-induced changes in polymer hydrogel swelling behavior using synthetic coiled-coil domains and CD spectroscopy (45). We are examining similar sensing concepts, as shown in Fig. 13. A critical step in the maturation of biomimetics for EM sensing will be meshing traditional synthetic polymer synthesis and processing with biochemistry and molecular biology.

There is a growing awareness of the contribution biomimetics can make to numerous well-established research areas, of which electromagnetic sensing is a small part. The highly interdisciplinary nature of biomimetic work makes it difficult for a single research group to be successful unless it truly spans several departments. The work is not only interdisciplinary, but additionally, few areas span basic, fundamental science to applied research as completely as biomimetics. Bearing this grand challenge in mind, there are still undreamed advances that can be made by imitating nature’s optimization that has occurred across millions of years.

ACKNOWLEDGMENTS

Our work is sponsored by the Air Force Office of Scientific Research (AFOSR) through the “Biological Infrared Sensing Initiative.” We thank Laura Sowards and Bryan Jones for help in preparing this manuscript.

BIBLIOGRAPHY

1. C.S. Zuker, *Proc. Natl. Acad. Sci. USA* **93**: 571–576 (1996).
2. Chapman, *The Insects: Structure and Functions*. Cambridge University Press, Cambridge, 1998.
3. T. Labhart, and E.P. Meyer, *Microsc. Res. Tech.* **47**: 368–379 (1999).
4. D.P. Smith, M.A. Starnes, and C.S. Zuker, *Annu. Rev. Cell Biol.* **7**: 161–190 (1991).

5. A.W. Snyder, in *The Compound Eye and Vision of Insects*, G.A. Horridge, ed., Oxford University Press, London, 1975.
6. R. Menzel, D.F. Ventura, H. Hertel, J.M. de Souza, and U. Greggers, *J. Comp. Physiol. A* **108**: 165–177 (1986).
7. C.G. Bernhard, and W.H. Miller, *Acta. Physiol. Scand.* **56**: 385–386 (1962).
8. C.G. Bernhard, W.H. Miller, and A.R. Moller, *Acta. Physiol. Scand.* **63**: 1–79 (1965).
9. A. Yoshida, M. Motoyama, A. Kosaku, and K. Miyamoto, *Zool. Sci.* **14**: 737–741 (1997).
10. H. Ghiradella, *J. Morphol.* **202**: 69–88 (1989).
11. R.E. Silberghied, in *The Biology of Butterflies*, R.I. Vane-Wright and P.R. Ackery, eds., Academic Press, London, 1985.
12. F.U. Hartl, *Science* **381**: 571–580 (1996).
13. T. Yura, H. Nagai, and H. Mori, *Annu. Rev. Microbiol.* **47**: 321–350 (1993).
14. B.L. Stoddard, H.P. Biemann, and D.S. Koshlan, Jr., *Cold Spring Harbor Symp. Quant. Biol.* **57**: 1–15 (1992).
15. A. Lupas, *Trends Biochem. Sci.* **21**: 375–382 (1996).
16. T. Nara, I. Kawagishi, S.-O. Nishiyama, M. Homma, and Y. Imae, *J. Biol. Chem.* **271**: 17932–17936 (1996).
17. R. Hurme, K.D. Berndt, S.J. Normark, and M. Rhen, *Cell* **90**: 55–64 (1997).
18. G.K. Noble, and A. Schmidt, *Proc. Am. Philos. Soc.* **77**: 263–288 (1937).
19. T.H. Bullock, and R.B. Cowles, *Science* **115**: 541–543 (1952).
20. T.H. Bullock, and F.P.J. Diecke, *J. Physiol.* **134**: 47–87 (1956).
21. E. Newman, and P. Hartline, *Sci. Am.* **246**: 98–107 (1982).
22. S.-I. Terashima, and R.C. Goris, *Infrared Receptors and the Trigeminal Sensory System*. Harwood Academic Press, Amsterdam, 1998.
23. A.L. Campbell, T.J. Bunning, M.O. Stone, D. Church, and M.S. Grace, *J. Struct. Biol.* **126**: 105–120 (1999).
24. F. Amemiya, T. Ushiki, R.C. Goris, Y. Atobe, and T. Kusunoki, *Anat. Rec.* **246**: 135–146 (1996).
25. D.H. Theodoratus, D. Chiszar, and H.M. Smith, *Psychol. Rec.* **47**: 461–472 (1997).
26. T. de Cock Buning, *Acta Biotheoretica* **34**: 193–206 (1985).
27. T. de Cock Buning, *Am. Zool.* **23**: 363–375 (1983).
28. T. de Cock Buning, S. Terashima, and R.C. Goris, *Cell. Mol. Neurobiol.* **1**(1): 69–85 (1981).
29. B.S. Jones, W.F. Lynn, and M.O. Stone, *J. Theo. Bio.* **209**: 201–211 (2001).
30. H. Schmitz, and H. Bleckmann, *J. Comp. Physiol. A* **182**: 647–657 (1998).
31. E.G. Linsley, *J. Econ. Entomol.* **36**: 341–342 (1943).
32. W.G. Evans, *Nature* **202**: 211 (1964).
33. W. Gronenberg, and H. Schmitz, *Cell. Tissue Res.* **297**: 311–318 (1999).
34. W.G. Evans, *Ann. Entomol. Soc. Am.* **59**: 873–877 (1966).
35. R.D. Hudson, *Infrared System Engineering*. J Wiley, NY, 1969.
36. M. Sfakiotakis, D.M. Lane, and J.B.C. Davies, *IEEE J. Oceanic Eng.* **24**(2): 237–252 (1999).
37. X. Shen, A.M. Belcher, P.K. Hansma, G.D. Stucky, and D.E. Morse, *J. Biol. Chem.* **272**: 32472–32481 (1997).
38. M. Sarikaya, *Micros. Res. Tech.* **27**: 360–375 (1994).
39. N. Kroger, R. Deutzmann, and M. Sumper, *Science* **286**: 1129–1132 (1999).
40. J.N. Cha, K. Shimizu, Y. Zhou, S.C. Christiansen, B.F. Chmelka, G.D. Stucky, and D.E. Morse, *Proc. Natl. Acad. Sci. USA* **96**: 361–365 (1999).
41. M. Banish, and R.F. Crouse, SY Technology, Huntsville, AL, 1999.
42. P.S. Callahan, *Tuning in to Nature: Solar energy, Infrared radiation, and the Insect Communication System*. The Devin-Adair Company, Old Greenwich, CT, 1975.
43. S. M. Kirkpatrick, J.W. Baur, C.M. Clark, L.R. Denny, D.W. Tomlin, B.A. Reinhardt, R. Kannan, and M.O. Stone, *Appl. Phys. A* **69**: 461–464 (1999).
44. M.O. Stone, J.W. Baur, L.A. Sowards, and S.M. Kirkpatrick, *Photonics West, SPIE Proc.* San Jose, 2000.
45. C. Wang, R.J. Stewart, and J. Kopecek, *Nature* **397**: 417–420 (1999).

BIOSENSORS, POROUS SILICON

ANDREAS JANSHOFF

Johannes-Gutenberg-Universität
Mainz, Germany

CLAUDIA STEINEM

Universität Regensburg
Regensburg, Germany

INTRODUCTION

Biosensors consist of a biologically active layer that responding to an analyte in solution and a powerful transducer that transforms and amplifies the reaction into a measurable signal. Biosensors can constantly measure the presence, absence, or concentration of specific organic or inorganic substances in short response time and ultimately at low cost. They are used commercially in health care, biotechnological process control, agriculture, veterinary medicine, defense, and environmental pollution monitoring. A common requirement of all of these applications is on-site chemical information—preferably in real time—on some dynamic or rapidly evolving process. Most biosensors are based on molecular events as they take place at the cellular membrane or inside the cell involving enzyme cascades. Their perceived advantages over existing technologies include the ability to monitor broad or narrow spectra of analytes continuously in real time, and their weakness is the instability of the biological molecules outside their natural environment, which results in a restricted lifetime for the device. The challenge is to find a matrix for biomolecules that provides high compatibility of the material with biological substances, low-cost fabrication, and special optical and electrical properties to generate a signal that measures the interaction between analytes in solution and the receptive biological layer. It is also desirable that it be compatible with conventional microfabrication techniques to miniaturize the device or to build individually addressable arrays.

The high surface area in conjunction with its unique optical and electrical properties and its compatibility with silicon microelectronics fabrication techniques has led to the proposal that porous silicon may be a suitable material for building sensor devices. Several different transducer

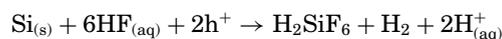
schemes have evolved based on thin film interference, capacitance changes, and the photoluminescent properties of porous silicon.

HISTORICAL OVERVIEW

Porous silicon is not a newly discovered material. Ulhir reported 45 years ago that porous silicon (PSi) is generated during the electropolishing of silicon under anodic polarization in a hydrofluoric acid containing electrolyte if the current density falls short of a critical value (1). Since its first discovery, the material has been studied extensively because it was considered suitable for electronic applications (local insulation, gettering of impurities, sacrificial layers, etc.). However, the impact of PSi increased far more than expected in 1990 when Canham unexpectedly discovered a red bright photoluminescence from PSi at room temperature (2). The emission of visible light from PSi produced a sensation because the energy gap of silicon (1.1 eV) corresponds to the infrared region and does not explain the occurrence of photoluminescence in the visible regime. Within months after this observation, several labs reportedly detected visible light emission from PSi by passing an electric current through it (electroluminescence) (3). This was a vital discovery because any optoelectronic device that might use PSi will probably operate by conventional electroluminescence. Inspired by this unique property of PSi, the efforts of the scientific community during the last 10 years led to much useful information about aspects of PSi formation and its physical and chemical properties. Despite these efforts, several aspects of PSi formation and even some of the physical and chemical properties are still a matter of discussion.

POROUS SILICON FORMATION

PSi layers can be prepared chemically or electrochemically (4). The electrochemical route starting from boron (p-type) or phosphorus (n-type) doped silicon is mostly employed. For most electrochemical preparations of PSi (2–6), single-crystal silicon [(100)- or (111)-oriented wafers] is anodized in an aqueous or ethanolic HF solution under constant current conditions. The exact dissolution chemistry of silicon is still in question, although it is generally accepted that holes are required in the initial oxidation steps. This means that for n-type material, significant dissolution occurs only under illumination, high electric fields, or other hole-generating mechanisms. A couple of facts have been gathered about the course of pore formation: (1) hydrogen gas evolves in a 2:1 atomic ratio to silicon; (2) current efficiencies have been measured at approximately two electrons per dissolved silicon atom and (3) the final, stable end product for silicon in HF is H_2SiF_6 (4,5). Though the reaction mechanism is still unclear and several different mechanistic variants for the anodization of silicon surfaces have been proposed, a simplified sum equation can be written for the dissolution process:



One mechanistic model presented by Lehmann and Gösele comprises an entirely surface-bound oxidation scheme of hole capture and subsequent electron injection to produce the divalent silicon oxidation state (7). The silicon surface continuously vacillates between hydride and fluoride coverage at each pair of electron/hole exchanges. It appears that, despite the thermodynamic stability of the Si–F bond, it does not remain on the silicon surface in any stable, readily measured form. The present consensus is that hydrogen exists on the silicon surface in at least two different forms, Si–H and Si–H₂ and possibly a third form, Si–H₃. For both n- and p-type silicon, low current densities are essential in PSi formation (Fig. 1).

Low current densities ensure a sufficient amount of HF molecules (or F[−] ions) at the silicon–electrolyte interface. Because holes from the bulk silicon phase reach the bottoms of the pores first, silicon at the pore bottoms is preferentially dissolved. This is, however, a very simple explanation. Several other aspects of pore propagation are discussed in the literature, such as image force effects, hole diffusion, crystallography, charge transfer, quantum confinement, and surface tension (5). Higher current densities result in an excess of holes at the silicon–electrolyte interface, and the corrosion reaction becomes limited by diffusion of HF molecules (or F[−] ions). This leads to a preferential reaction of the upper parts of the silicon surface that results in smooth electropolishing (Fig. 1). Because electropolishing does not occur in organic solutions, it appears to depend on the formation of an oxide layer atop the silicon surface.

The formation of pores results from the complex interplay of chemical kinetics, charge distribution, and differing crystal face reactivities, so it is obvious that the issue of PSi films comprises rather different porous structures ranging from those holding micron-sized pores to sponge-like layers that contain nanometer-sized pores. Pore structures and dimensions are determined by a large number of preparative conditions: doping level and type, crystal orientation, composition of electrolyte, construction of the electrolytic cell, anodization regime, sample preconditioning, and postanodization processing (5). In fact, samples produced by different research groups are hardly comparable, even if the preparative conditions are apparently the same. No wonder great controversy exists over the mechanism of PSi formation.

CHARACTERIZATION OF POROUS SILICON

The body of knowledge about PSi formation has been obtained from current–voltage characteristics, as described earlier (5). Besides electrochemistry, several other methods have been employed to study the morphology of PSi. Among them, transmission electron microscopy (TEM) has contributed a large amount of information about structural details on individual pore propagation and silicon microcrystals because it is the only method to visualize microporous silicon directly (4). Scanning electron microscopy (SEM) is used mainly for macroporous silicon obtained from n-type or heavily doped p-type silicon etched at high current densities. Scanning probe techniques such as atomic force microscopy (AFM) are especially useful for

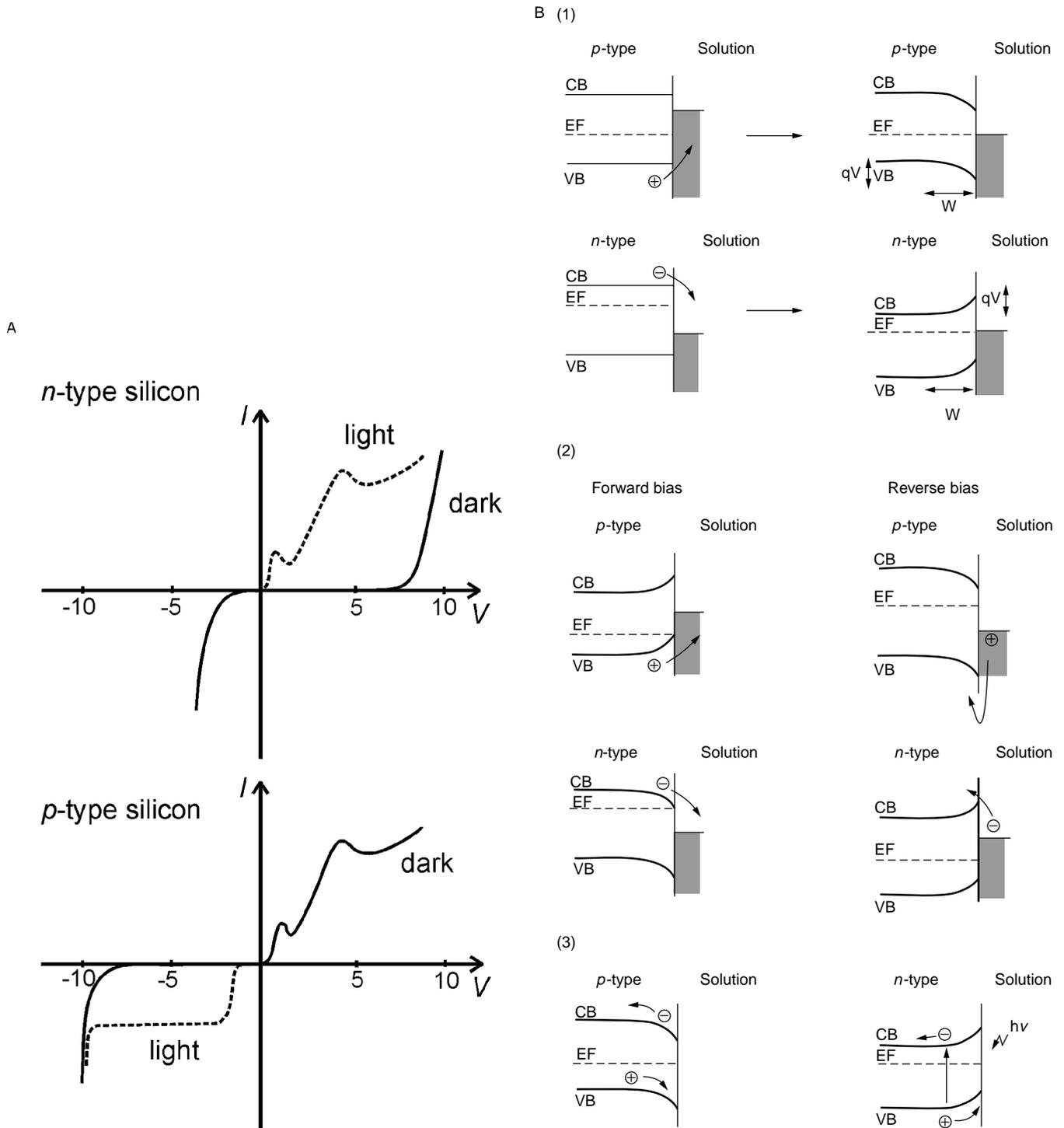


Figure 1. (a) Typical current-voltage relationships for n - and p -type silicon. The solid line is the dark response, and the dashed line indicates the response under illumination. The first (lower) current peak corresponds to a surface anodic oxide formed during and required for electropolishing. The second (higher) current peak marks the beginning of stable current oscillations and the possible formation of a second type of anodic oxide (5). (b) (1) The semiconductor-electrolyte interface before (left) and in thermal equilibrium (right), (2) at forward and reverse bias, and (3) during anodic etching. n -type PSi has to be illuminated to provide holes for the etching process. CB: conduction band, EF: Fermi energy, VB: valence band, W : width of the depletion layer.

detecting topographical features in conjunction with material properties such as friction, elasticity, conductance, and energy dissipation. Quantitative data about porosity and poreradii distribution may be inferred from low-temperature adsorption and desorption of gases. The most prominent technique, the BET (Brunauer–Emmett–Teller) method, is based on measuring the gas volume adsorbed by a material as a function of pressure; the BJH (Barret–Joyner–Halendra) method uses the Kelvin equation to infer the pore radius from gas condensation inside the pores (8). Simple gravimetric analysis and profilometer measurements of pore nucleation and propagation have provided valuable information about the anodization of silicon (9). Optical properties and morphological details are studied by spectroscopic techniques such as UV-vis, Raman and IR spectroscopy, as well as spectroscopic ellipsometry (10). Ellipsometry reveals information about porosity and the dielectric function of the material and is particularly useful for determining changes in the refractive index and thickness of the material. Details of pore morphology can also be obtained from X-ray crystallography measurements, as demonstrated by grazing angle experiments using X rays and synchrotron radiation.

Key parameters that describe the overall properties of porous material are porosity and pore radius, which depend mainly on the composition and temperature of the electrolyte, the dopant concentration, and the current density (5). Pore sizes can vary over a wide range from macropores (pores >50 nm wide) and mesopores (2–50 nm) to micropores (<2 nm). Generally, an increase in pore radius accompanies an increase in the anodization potential or current density for both n- and p-type silicon. At low current densities, the pores are randomly oriented and filamentlike. In contrast, the pores “pipe” at high current densities close to the electropolishing regime. The effect of dopant concentration on pore morphology is well explored. The pore diameters and interpore spacings of lightly doped p-type silicon are between 1 and 5 nm and exhibit a networklike appearance. Increasing the dopant concentration results in forming clear channels that have larger pore diameters and directed pore growth. Although the n-type silicon is more complex, increased dopant concentration is characterized by decreasing pore diameter and interpore spacing. The pore diameters in n-type PSi are considerably larger than those of the p-type silicon at low dopant concentration (3,5). Electrochemical etching of lightly doped n-type silicon wafers in the dark results in forming low porosity materials that exhibit macropores whose radii are in the micrometer range. Under illumination, much larger porosities can be obtained and micro- to macropores are found. Using p-type silicon of low resistivity, the porous texture is always thin, and the pore size distribution is in the 1 to 5-nm regime.

The results of a systematic study of porous layers formed in heavily doped substrates has been published by Herino (11). Generally, the porosity increases as HF concentration decreases in p-type silicon, whereas the influence of the HF concentration on the pore size of the n-type is not very pronounced. The specific surface area is in the range of 180–230 m²/cm² in p-type silicon and 90–230 m²/cm² in n-type silicon and is not very sensitive to the forming parameters.

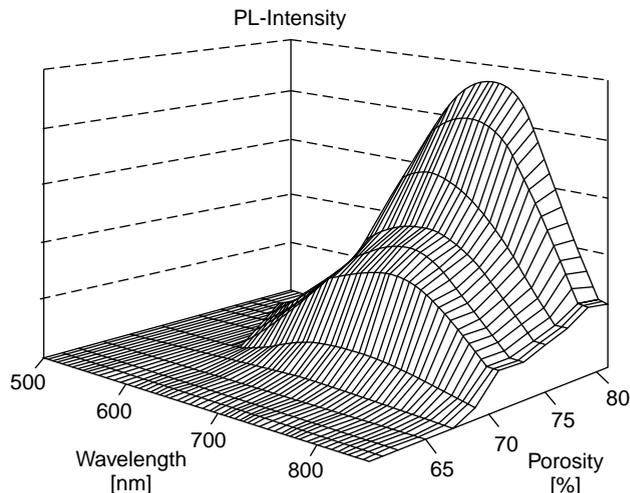


Figure 2. Photoluminescent spectra of lightly doped p-type PSi layers of various porosities. The layer is about 1 μm thick, and the specific resistivity of the silicon 0.2 $\Omega\text{ cm}$ in all cases [reprinted with permission from (10)].

OPTICAL PROPERTIES OF POROUS SILICON

The demand for visible light-emitting devices made entirely from silicon is enormous because silicon is the dominant material for electronic and optical devices such as waveguides, detectors, and modulators. However, silicon is an indirect semiconductor, and thus light emission is inefficient. A direct photon transition at the energy of the minimum band gap does not meet the requirement of conservation of momentum in silicon. Therefore, electrons at the minimum of the conduction band need a significant amount of time to receive the necessary momentum transfer to recombine with holes in the valence band. Consequently, nonradiative recombination reduces the quantum efficiency considerably and results in emission of weak infrared wavelength light due to its small indirect band gap of 1.1 eV (12). In 1990, Canham announced the discovery of photoluminescence from PSi electrochemically etched at room temperature (2). Figure 2 shows typical photoluminescent spectra of p-type PSi that depend on porosity.

Tunable photoluminescence from anodically etched silicon is expected to have great impact on the development of optoelectronics, filters, chemical and biological sensors, and optical data storage, to name just a few applications. The mechanism of luminescence, however, is still a matter of controversial discussions. Available models can be grouped into four classes: those based on quantum confinement alone, nanocrystal surface states, specific defects or molecules, and structural disordered phases (13). Experimental data, however, are most consistent with the so-called smart quantum confinement model that comprises the quantum confinement model, including contributions from surface states (14). The general features of light emission from PSi may be explained in terms of reduced nonradiative recombination, as deduced from time development of photoluminescent intensity after short laser pulse excitation. The rather slow decay provides evidence that reduction of nonradiative recombination, rather than an increased amount of radiative

transitions, is the reason for the enhanced quantum efficiency, compared to bulk silicon. Significant light emission is observed only for microporous silicon, and the band gap widens (1.4–2.2 eV) as crystal size decreases, essentially identical to the particle in the box phenomenon in quantum mechanics. The increased path length of electrons in larger crystals renders recombination with surface defects or other mechanisms very likely. Consequently, light emission from larger structures is poor, whereas bright luminescence occurs in microporous material and is accompanied by a shift to higher photon energies from the near IR to the visible region. There is a correlation between porosity as an indirect measure of particle size and emissive energy. The smallest particle size is obtained from lightly doped p-type PSi etched at low current densities. Moreover, evidence for nanocrystallinity of the porous material from ESR analysis, TEM measurements, and phonon-assisted luminescence strongly support the quantum confinement model (13). All three primary colors were obtained, and the consequences are important for future display applications (13,15). Many chemical sensors based on PSi use luminescent reduction and thus provide a transducing mechanism for quantifying adsorption of analytes on the surface. Examples of chemical sensors employing photoluminescent reduction are given later.

Because recent biosensor developments are based on the dielectric function of PSi films, it is instructive to review briefly the reflectance and transmission of PSi layers and emphasize interference patterns and suitable effective medium approximations. The dielectric function $\varepsilon(\omega)$ connects the dielectric displacement D to the electric field vector E (12). The polarization P represents the part of D that arises due to polarization of the dielectric material induced by an external electrical field. The total polarizability of matter is usually separated into three parts: the electronic, ionic, and dipolar. The dielectric constant at optical wavelengths (UV-vis) arises almost entirely from electronic polarizability, and the dipolar and ionic contributions are small at high frequencies (Fig. 3).

The dielectric function is not a constant but depends strongly on the frequency of the external electrical field. The frequency dependence of the dielectric function arises from relaxation processes, vibrations of the electronic system and atomic cores, that are accompanied by macroscopic polarization. At certain wavelengths, however, it is reasonable to assume a constant value. The dielectric function of a solid can be inferred from measuring the reflectivity, refractive index, and absorbance, all functions that are accessible by optical spectroscopy. The real and imaginary function of the dielectric function can be accessed from reflectivity measurements. The refractive index $n(\omega)$ and the extinction coefficient $K(\omega)$ are related to the reflectivity $r(\omega)$ at normal incidence in vacuum by (16)

$$r(\omega) = \frac{n + iK - 1}{n + iK + 1}. \quad (1)$$

By definition, n and K are related to the dielectric function by

$$\sqrt{\varepsilon(\omega)} = n + iK. \quad (2)$$

The measured quantity is the reflectance $R(\omega)$, which is related to the reflectivity $r(\omega)$ by its complex product:

$$R(\omega) = \frac{E_{\text{ref}}^* E_{\text{ref}}}{E_{\text{inc}}^* E_{\text{inc}}} = r^* r, \quad (3)$$

where E_{ref} is the electric field vector of the reflected light and E_{inc} that of the incoming light. The following description of thin film interference provides the necessary foundation to understand the functioning of most popular biosensors based on the shift of interference fringes that arise from reflections at thin transparent PSi layers. Because PSi can be described as a film of a particular dielectric function different from that of bulk silicon, it is instructive to look at wave propagation in thin films on solid supports. Assumption of transparency due to the high porosity of PSi simplifies the treatment. At the interface between two media that have different refractive indices (n_1 and n_2), an incident wave is partially transmitted in the medium and reflected (Fig. 4).

This follows from the boundary conditions for electric and magnetic fields. Reflectivity and transmission coefficients can be obtained from the Fresnel equations which can be simplified for normal incidence and ideal transparent media by taking $K = 0$.

In the visible range, the dielectric function of PSi may be described by an effective medium approximation (EMA). Porous silicon consists of two media, the pore filling and the pore walls. The geometry of the pores determines the way the dielectric functions of these two media can be combined to give an effective dielectric function between that of silicon and the pore filling medium. The following section briefly summarizes the most prominent effective

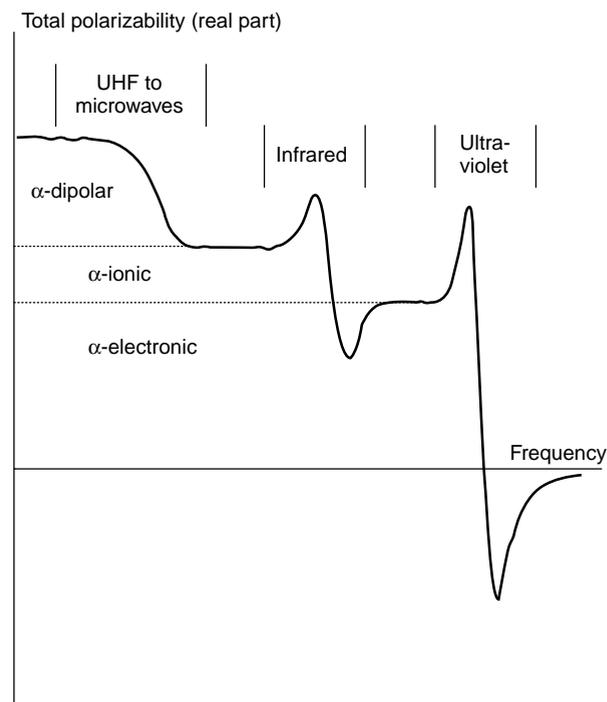


Figure 3. Frequency dependence of the different contributions to polarizability (12).

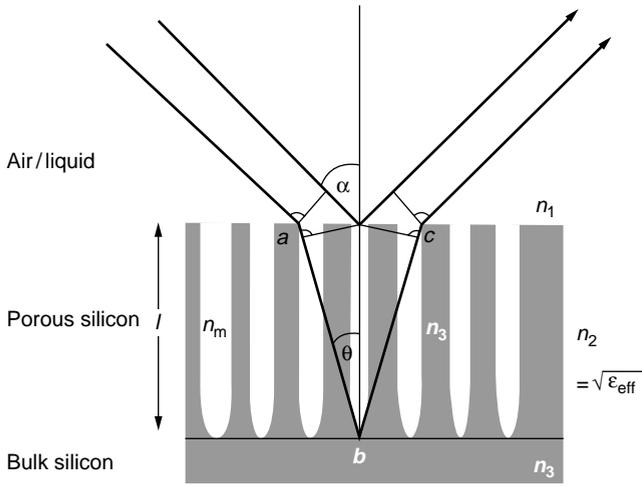


Figure 4. Concept of thin film interference. The incoming light is reflected at the PSi surface whose thickness is l and combines with the light beam reflected from bulk silicon to form an interference pattern. The path difference between the two rays is $\delta = \overline{abc} = 2n_2l\cos(\alpha)$. n_1 : refractive index of the upper medium; n_m : refractive index of the pore filling medium; n_3 : refractive index of bulk silicon; and n_2 : refractive index of the porous layer that has a pore filling.

medium theories (10), that are applicable to PSi films. The Maxwell-Garnett approach for two media is given by

$$\frac{\varepsilon_{\text{eff}} - \varepsilon_m}{\varepsilon_{\text{eff}} + 2\varepsilon_m} = (1 - p) \frac{\varepsilon_3 - \varepsilon_m}{\varepsilon_3 + 2\varepsilon_m}, \quad (4)$$

where p is the porosity of the material, ε_{eff} the effective dielectric function, ε_3 the dielectric function of bulk silicon, generally the host material, and ε_m denotes the dielectric function of the medium inside the pores—air or liquid in most sensors. The simple Maxwell-Garnett model is a good approximation for highly porous, spherical particles at a large distance from each other. It is seldom applied to PSi. The Bruggeman approximation is most frequently used to describe the effective dielectric function of two or more different media:

$$p \frac{\varepsilon_m - \varepsilon_{\text{eff}}}{\varepsilon_m + 2\varepsilon_{\text{eff}}} + (1 - p) \frac{\varepsilon_3 - \varepsilon_{\text{eff}}}{\varepsilon_3 + 2\varepsilon_m} = 0. \quad (5)$$

A concept for highly porous solids is provided by the Looyenga model that also involves one parameter, the porosity, to describe the microtopology of the material:

$$\varepsilon_{\text{eff}}^{1/3} = (1 - p)\varepsilon_3^{1/3} + p\varepsilon_m^{1/3}. \quad (6)$$

Theiss and co-workers developed a more realistic approximation taking into account the strength of percolation and a broadening parameter of resonances (10). This three-parameter approach is a good compromise between the general model from Bergmann that has a normalized distribution function $g(n, p)$ of so-called geometrical resonances and the simple one-parameter approaches. Porous silicon multilayers or superlattices may serve as a material for interferometric devices that lead to a number

of different applications in the design of Fabry–Perot interferential filters, distributed Bragg reflectors, and interferometric biosensors. Illumination of the porous matrix by white light leads to a characteristic interference pattern in the reflectance spectrum. Assuming smooth surfaces and a negligible absorption coefficient, one may infer the effective optical thickness n_2l , where l is the thickness of the layer and n_2 the refractive index of the effective medium from the reflectance spectrum that displays interference fringes due to alternating constructive and destructive interference of the light reflected from the top and bottom of the porous layer. Assuming an incident angle of 0° , the reflectance R of a thin PSi layer is given by

$$R = \frac{r_1^2 + 2r_1r_2\cos\left(\frac{4\pi}{\lambda}n_2l\right) + r_2^2}{1 + 2r_1r_2\cos\left(\frac{4\pi}{\lambda}n_2l\right) + r_1^2r_2^2}, \quad (7)$$

where r_1 is the reflectivity at the interface between the medium on top of the film (n_1) and the porous layer (n_2) and r_2 is the reflectivity at the interface PSi (n_2) and bulk silicon (n_3):

$$r_1 = \frac{n_1 - n_2}{n_1 + n_2},$$

$$r_2 = \frac{n_2 - n_3}{n_2 + n_3}. \quad (8)$$

Constructive interference occurs if the path difference fulfills $\delta = 2n_2l$ and is a multiple of the wavelength $m\lambda = 2n_2l$ ($m = 1, 2, 3, \dots$) that marks the distance between the interference maxima. It is instructive to consider the minimum and maximum reflectance at normal incidence. Evaluation of R_{min} and R_{max} permits one to obtain the effective refractive index of the PSi layer from interference measurements without knowing the thickness of the layer ($n_3 > n_2 > n_1$):

$$R_{\text{max}} = \frac{(n_3 - n_1)^2}{(n_3 + n_1)^2},$$

$$R_{\text{min}} = \frac{(n_2^2 - n_1n_3)^2}{(n_2^2 + n_1n_3)^2}. \quad (9)$$

For $n_2 > n_3 > n_1$, the expressions for R_{min} and R_{max} are vice versa, but this does not apply for PSi. Because the refractive index of the porous matrix is smaller than that of bulk silicon, one has to take into account that the reflectivity of the whole system is smaller than that of bulk silicon. Lower n_2 , however, results in steeper fringes, as shown in Fig. 5. Correction terms for finite roughness at both interfaces of the PSi layer were introduced by Léron del et al. (17). Theiss and co-workers established a theory accounting for thickness variations (10).

The preceding treatment considers merely two-beam interference. However, multiple internal reflections occur in transparent films that give rise to sharper maxima than sinusoidal curves. An exiting wave, either in reflection or transmission, will combine the waves that have corresponding phase increments at each stage. It can be shown that the intensity of the transmitted light (a geometric series) gives a Lorentzian function. Multiple PSi layers

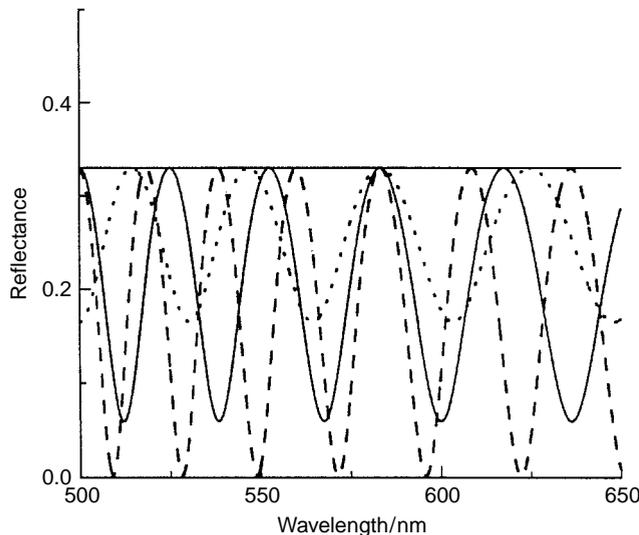


Figure 5. Simulated reflectance spectra of PSi exposed to air. The layer was $3.5 \mu\text{m}$ thick, the refractive index of bulk silicon 3.7, and the effective refractive index of the porous layer was chosen as 1.25 (dotted), 1.5 (solid), and 2 (dashed line). The maximum of the spectrum is limited by the differences between the refractive index of the outer medium and that of bulk silicon, and the fringe visibility increases as the refractive index of the PSi rises.

require more cumbersome mathematics but are of great commercial interest in designing Bragg reflectors characterized by an alternating sequence of layers of low (L) and high (H) porosity [air–HLHL ($\times n$) HLHL–PSi] so called random PSi multilayers (10,18).

For successful comparison between experimental and simulated spectra, it is important to find reasonable referencing. For instance, a simple measurement gives the intensity $I(\omega) = R(\omega)I_0(\omega)$, in which $I_0(\omega)$ contains all spectral features of the incident light source. Therefore, the reflectance $R(\omega)$ can be obtained only if $I_0(\omega)$ is measured as accurately as possible by using highly reflective reference samples such as metal-coated smooth surfaces of known reflectivity. Reflectance spectra of bulk silicon can be described very well by a constant and real dielectric function where $\epsilon_\infty = 11.7$ and a Drude contribution from the absorption of free carriers that depends on the doping level (10). In transmission spectra of thick silicon wafers, typical absorption bands that arise from carbon (610 cm^{-1}) and oxygen (1105 cm^{-1}) impurities occur, as well as multiphonon excitations. The dielectric function of freshly prepared PSi is governed by Si–H vibrations, which can be modeled by harmonic analysis, assuming a Gaussian distribution of resonant frequencies. Stretching, scissors, and wagging modes are found. A comparison with spectra obtained from Si–H-terminated amorphous PSi samples shows significant difference in resonant frequencies and bandwidths. Although silicon (lightly doped) is sufficiently transparent in the IR region, accurate conversion from transmission to absorbance is not possible because the reflectivity of bare silicon as the reference and PSi differ significantly from each other. This is due to the lower effective refractive index of PSi compared to bulk silicon and multiple beam interference within the porous

layer. Interference patterns in the infrared show up as a baseline. Therefore, reflectance rather than transmission techniques are recommended to cope with this problem for routine measurements.

Optical reflection spectroscopy in the UV-vis has been employed to investigate the electronic band structure because direct transitions occur that contribute significantly to the dielectric function of PSi. Transmission is usually too weak near the interband transitions, and the penetration depth of UV light is small. Within the reflectance spectrum, there is a clear distinction between the low-frequency region of transparent PSi that gives rise to the formation of interference fringes and the high-frequency part (UV) where no radiation is reflected from the interface between PSi and bulk silicon. In the UV region, a broad peak is detected due to the vast number of dipole-allowed transitions that arise from the complex microstructure of PSi (Fig. 6). The peak broadens as porosity increases and thus gives rise to the assumption that quantum size effects play a key role.

Effective modeling of the dielectric function remains to be elucidated, although introducing a sufficient number of extended oscillatory terms provides good agreement with experimental data. Once a model for the dielectric function of the pore walls has been found, EMA theories need to be employed to ensure the “right averaging” between the dielectric function of the pore-filling and the silicon pillars.

FUNCTIONALIZATION OF POROUS SILICON SURFACES

Any chemical or biochemical sensor is based on a highly specific receptive layer. These layers are best prepared by chemical reactions of the PSi surface. A large number of mild chemical reactions have recently been developed (Scheme 1) to modify PSi surfaces for optical and sensor applications (19). The formation of PSi by anodic dissolution of crystalline silicon in a HF-based electrolyte leads to a hydride-terminated silicon surface that is the starting point for a variety of modifying procedures.

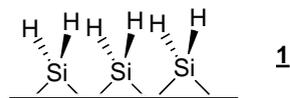
Si–H-terminated Surfaces

A PSi surface obtained by an anodic or chemical etch of crystalline silicon using HF comprises Si–H_x ($x = 1, 2, 3$) bonds that can be readily oxidized and hydrolyzed, single-bonded Si–Si groups, and Si atoms that have free valences (dangling bonds). In the last few years, new reaction schemes have been developed based on hydride-terminated PSi layers.

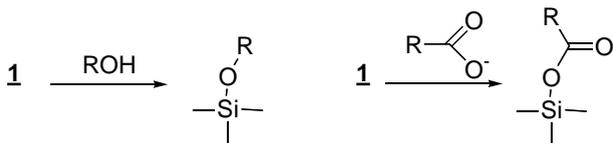
Formation of Si–O–C-Bonds

Chemical reactions at the surface of electronic materials can be very different from the corresponding solution-phase transformations. In particular, the electronic structure of the semiconductor provides a source of electrons and holes that can be used to induce surface reactions. For example, for a nucleophilic attack on n-type PSi, the surface is brought under positive potential control. This is called the reverse-bias condition, where the applied potential adds to the band bending potential, thereby increasing the

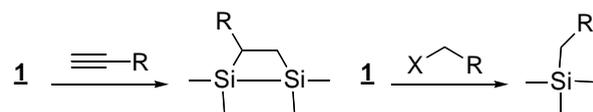
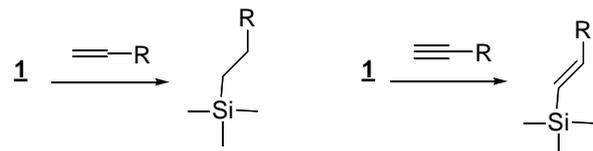
Scheme 1

Reactions of Si-H_x terminated surfaces

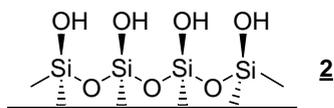
Formation of Si-O-C bonds



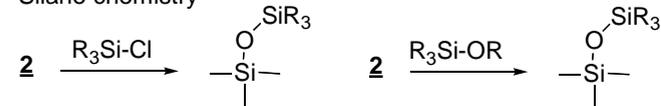
Formation of Si-C bonds



Reactions of Si-OH terminated surfaces



Silane chemistry



Reactions of Si-X terminated surfaces

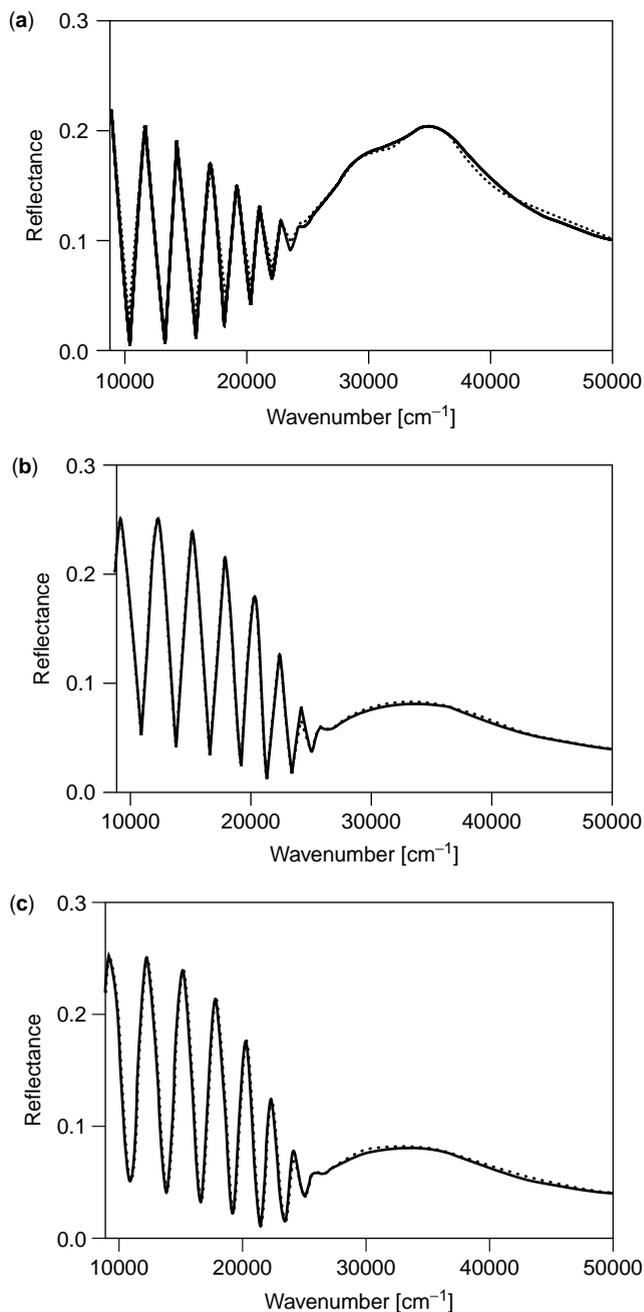
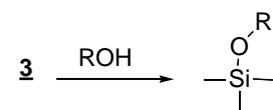
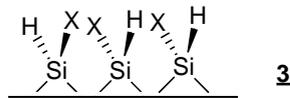


Figure 6. Measured (solid lines) and calculated (dotted lines) reflectance spectra of A 61%, B 71%, and C 79% porosity layers 1 μm thick [reprinted with permission from (10)].

barrier height. As a result, in n-type silicon, an excess of positive charges in the semiconductor renders the silicon surface susceptible to nucleophilic attack. Nucleophiles such as H_2O and CH_3OH react with this surface of reverse-biased PSi (20). Less active nucleophiles such as trifluoroacetic, acetic, and formic acids, however, react only with the silicon surface and produce a silyl-ester-modified surface upon irradiation (20–22). Because this kind of reaction takes place only under illumination, porous silicon can be photopatterned by illuminating the surface through a mask during the derivatization procedure. Esterification changes the chemical and physical properties of

silicon surfaces. The ester-modified surface is hydrophilic as opposed to the hydrophobic, native, hydride-terminated surface. Reaction of an ester-modified surface with organomethoxysilanes results in replacing the ester by organosilanes, and they do not react with hydride-terminated surfaces. Reactions that are not based on photo- or electrochemical methods were carried out by Laibinis and co-workers using alcohols under modest heat that resulted in the forming Si–O–C bonds (23,24). The major disadvantage of a surface modification based on Si–O–C bonds is their limited stability in aqueous solution. The Si–O bond in this case is readily hydrolyzed and limits the applicability of these functionalized surfaces to sensor devices.

Formation of Si–C Bonds

Since Canham and co-workers discovered the photoluminescence of PSi in 1990, modification and characterization of photoluminescent PSi surfaces has become an area of intense interest. However, hydride-terminated silicon oxidizes slowly in air, often resulting in the loss of photo- and electroluminescent properties. Many studies addressed the chemical properties of H-terminated silicon surfaces to protect PSi from losing its luminescent properties and to prepare PSi that is chemically stable. The attachment of species to silicon surfaces by forming Si–C bonds (25) provides greater stability to oxidation. Methyl groups were grafted on PSi surfaces by an anodic electrochemical stimulus using CH_3MgBr or CH_3Li (26,27). Without photo- or electrochemical methods that often proceed by oxidizing the substrate, it is possible to derivatize PSi surfaces by using a variety of Grignard (23,24,28) or aryllithium and alkylolithium reagents (29,30) at room temperature. Si–Li species on the surface are readily hydrolyzed by water resulting in considerable surface oxidation and thus, loss of photoluminescence. However, surface-bound lithium can also be replaced by H- or acyl species that reduce the rate of air oxidation. Greater stability to hydrolysis and oxidation can be obtained by using hydrosilylation reactions applicable to a wide range of different PSi samples, independent of doping and pore morphology. Hydrosilylation of native hydride-terminated PSi can be induced by Lewis acids (31–35). Insertion of alkenes and alkynes into surface Si–H groups yields alkyl or alkenyl termination, respectively. Robins et al. (36) reported on electrochemical grafting of terminal alkynes. Cathodic electrografting attaches alkynes directly to the surface, whereas anodic electrografting yields an alkyl surface. Hydrophobic surfaces capped by a monolayer of long alkyl chains are dramatically stabilized under chemically demanding conditions, such as basic solutions, compared to nonfunctionalized PSi. High surface coverage and short reaction time were achieved by an electrochemical method based on the reductive electrolysis of alkyl iodide, alkyl bromide, and benzyl bromide (37).

Silicon OH-Terminated Surfaces

Besides direct use of Si–H-terminated surfaces as obtained from a HF etch, the PSi surface can first be oxidized, resulting in an OH-terminated surface that has a layer of SiO_2 underneath. Several methods have been employed

that partly transform silicon into silica (19). The extent of oxidation depends on the procedure. Porous silicon that has very stable photoluminescent properties can be generated by rapid thermal oxidation in O_2 at high temperature (38). Then, the surface is then coated by a thick silicon oxide layer and is stable in air indefinitely. However, only a few OH groups are exposed. Other techniques have been evolved, including chemical oxidation using reagents such as hydrogen peroxide, nitric oxide, and ozone, leading to an OH-terminated PSi surface.

Silane Chemistry

Traditionally, almost all of the chemistry of silicon at moderate temperatures and pressures is based on OH-terminated surfaces. Common examples are the use of substituted chloro- and alkoxy silanes to form self-assembled monolayers of organosilanes (39,40). Trichloro- and trialkoxy silanes react with OH groups on the PSi surface but also cross-react with themselves to form an organosilane network, depending on the conditions. A sole reaction of organosilanes with OH groups on the surface can be accomplished by using monofunctional instead of trifunctional silanes. However, the surface coverage in this case depends strongly on the number of available OH groups, which in turn is determined by the oxidation procedure.

Si-Halide-Terminated Surfaces

A variation of traditional silane chemistry starting with Si–OH species is the formation of reactive Si halides on the PSi surface. Exposing a Si–H-terminated surface to halogen vapor breaks Si–Si bonds and creates Si halide species (41,42). The Si halide surface is then exposed to a silanol or an alcohol to generate Si–OR surfaces. The halogenation route avoids the need to generate a silicon oxide surface before derivatization. Exposure to air leads to oxidation of only the outer layer of Si atoms and leaves a large number of OH groups behind. However, in contrast to other oxidation procedures, Si–H groups are still present on the surface and make PSi susceptible to hydrolysis and further oxidation. A different strategy was followed by Lewis and co-workers (43), who functionalized silicon with an alkyl layer by first chlorinating the H-terminated silicon surface with PCl_5 . The Si–Cl surface is then treated with an alkyl Grignard or alkylolithium reagent to generate the surface-bound alkyl species.

POROUS SILICON CHEMOSENSORS

Most chemical sensors based on PSi use the material's unique property to emit light efficiently at room temperature. Reversible reduction of photoluminescence due to the specific or nonspecific adsorption of analytes from vapor to the porous matrix renders PSi a fast responding sensor for many vapors and, if suitably functionalized, for adsorbents in liquids. A typical photoluminescent spectrum of PSi usually has a bandwidth of 200 nm and the wavelength of maximum emission varies from 500–900 nm. Time-resolved spectroscopy revealed half-lives of the order of several tens of microseconds at the high wavelength

limit of the spectrum and 5 μs at the blue end. The decay is indicative of a distributed number of emission lifetimes rather than a single one. This is readily explained by an ensemble of different quantum structures of varying sizes that give rise to a broad emission spectrum that has a distribution of lifetimes. It is well known that surface contamination leads to reduced quantum efficiency, thus resulting in decreased emission intensity. Any covalently bound compound may act as a surface defect, if its orbital energies are within the band gap, that results in nonradiative recombination. Fortunately, the energies of the Si–H and Si–O bonds, which are among the most stable bonds of silicon, do not lie within the band gap. Most likely, chemical binding of a species to PSi adds a nonradiative trap but does not change the spectral features of the photoluminescent spectrum. If shifts are observed, they may arise from differences in the photoluminescent lifetimes that range from nanoseconds in the blue to microseconds in the red. Thus, the red part of the spectrum may be more strongly reduced than the blue part, leading to a slight blue shift of the overall spectrum.

Sailor and co-workers reported on PSi chemical sensors that detect vapors by partially reversible photoluminescent reduction. They found that the visible photoluminescence of n-type PSi is quenched by nitric oxide to detection limits of 2 ppm and that of nitrogen oxide to 70 ppb (44). At a partial pressure in the millitorr range, photoluminescent reduction is partly reversible. Recovery from nitrogen oxide occurs on a timescale of minutes. Reversible quenching for both nitric oxide and nitrogen dioxide fits a Stern–Vollmer kinetic model in the low concentration regime, and it deviates at higher partial pressures; a permanent loss of photoluminescence due to oxidation occurs. Interestingly, no significant quenching was observable for nitrous oxide and carbon dioxide and only minor quenching for carbon monoxide and oxygen. A PSi-based NO_x sensor, which is used for monitoring NO concentrations in industrial processes and pollution control, can be used to detect both small and large amounts of NO_x that can overload conventional sensors based on SnO_2 .

Using a similar approach, Content et al. (45) detected explosives such as 2,4-dinitrotoluene (DNT), 2,4,6-trinitrotoluene (TNT), and nitrobenzene in an air stream by the quenching PSi photoluminescence. Detection limits of 500 ppb, 2 ppb, and 1 ppb were observed for nitrobenzene, DNT, and TNT, respectively. Combined with a second transduction mode—Fabry–Perot optical interference—Letant et al. (46) developed an electronic artificial nose based on PSi surfaces that discriminated among solvent vapors, ethyl esters, and perfumes. Discrimination index obtained by PSi sensors have been as good as those obtained from metal oxide sensors.

Zhang et al. (47) reported on the successful functionalization of p-type PSi using calixarene carboxylic derivatives. They described a method for depositing a uniform film of calixarene derivatives varying in ring size that is stable in aqueous and heptane solutions. The authors showed that photoluminescent reduction due to the addition of copper(II) in aqueous solution depends on the ring size and enables one to determine the binding constant from a Stern–Vollmer plot. A concentration of 1504 M^{-1} for calix[8]-COOH-coated PSi versus 128 M^{-1} for

calix[4]-COOH-coated PSi was found for copper(II) ions dissolved in water.

Besides the reduction of photoluminescence, other transducing properties of PSi have been used to design chemical sensors. Recently, Letant and Sailor (48) described the design of a chemical HF vapor sensor based on detecting the effective refractive index. The authors report on the dissolution of SiO_x species upon exposure to wet HF vapor that was detected by a decreased effective optical thickness.

Tobias and co-workers (49) reported a 440% increase in capacitance in response to a humidity change from 0 to 100% using an aluminum contact to p-type PSi (Schottky-barrier sensor). This sandwich structure, in which PSi is located between the Al film and the bulk silicon, serves as the dielectric sensor matrix that responds to the condensation of vapor inside the pores. The capillary condensation is readily described by the Kelvin equation for closed-end capillaries:

$$\ln \frac{p}{p_0} = \frac{-2\gamma M \cos \theta}{\rho r RT}, \quad (10)$$

where p is the effective vapor pressure, p_0 the standard vapor pressure, γ the surface tension, M the molecular weight, θ the contact angle, ρ the density of the liquid, r the pore radius, R the gas constant, and T the temperature.

BIOSENSOR APPLICATIONS OF POROUS SILICON

Although silicon technology has a lot to offer in miniaturized intelligent devices, the range of applications has been limited to those where the electronic chip is almost isolated from a biological environment primarily because aqueous media rapidly destroy silicon—it has not been considered biocompatible. However, it was demonstrated that PSi could be designed to be more compatible than bulk silicon with biological environments due to recent developments in surface derivatization (50). This implies that this particular material might be well suited for developing biosensor devices based on silicon technology. Several physical properties of PSi have been employed to detect analytes (signal readout) in solution.

Photoluminescent Transduction

When enhanced photo- and electroluminescence were discovered, PSi also excited great interest among scientists working with biological sensors for detecting a biological analyte fast and at very low concentrations. Starodub and co-workers (51–53) exploited photoluminescence to monitor binding of human myoglobin to mouse monoclonal antibodies. They used PSi samples whose pore size was 10 to 100 nm obtained by laser-beam-treating and chemically etching monocrystalline p-type silicon (specific resistance: $10 \Omega \text{ cm}$). The PSi was functionalized by physisorption of mouse monoclonal antihuman antibodies on the passivated PSi surface that can binds human myoglobin. The physisorption itself induced only little change in the photoluminescent intensity. The influence of nonspecific adsorption on photoluminescent intensity was verified by using bovine serum albumin, rabbit IgG, and sheep

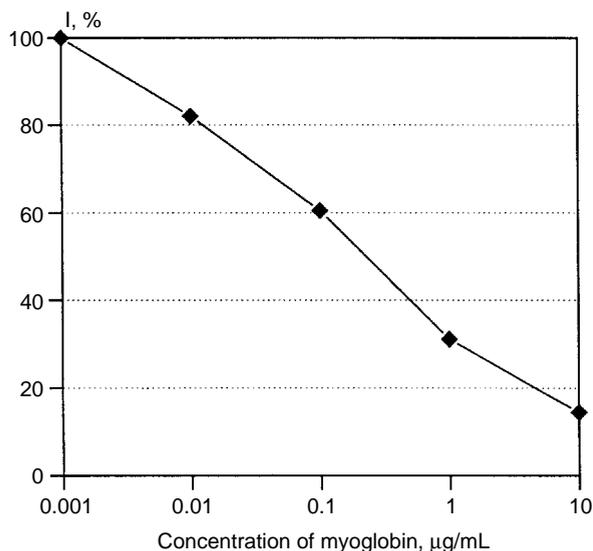


Figure 7. Changes of the photoluminescent intensity (I) upon immersion of the PSi sample antibodies in a myoglobin solution. The silicon surface was functionalized by using monoclonal anti-human antibodies by physisorption. Different myoglobin concentrations were added, and the photoluminescence was monitored [reprinted with permission from (53)].

antirabbit IgG. No change was detected within 2–2.5 h. Adding human myoglobin, however, resulted in a large decrease in photoluminescent intensity (Fig. 7). The origin of the photoluminescent decrease is discussed in terms of dehydrogenation of the PSi surface after formation of a specific immune complex. Hydrogen is released from Si–H bonds and subsequently captured by the immune complex. The sensitivity of the sensor is about 10 ng/mL myoglobin, and the overall detectable concentration regime ranges from 10 ng/mL to 10 $\mu\text{g/mL}$ in buffer solution.

To demonstrate the effectiveness of their biosensor, the concentration of myoglobin in human serum of patients suffering from heart failure determined by the PSi sensor was compared to results from a standard ELISA (enzyme-linked immunosorbent assay) test: in all three cases, the two techniques gave almost the same results; the difference was less than 5%. The overall time, however, taken by the ELISA test (at least 3 h) is significantly greater than for the PSi sensor (15–30 min). Unfortunately, the PSi biosensor cannot be reused. It was found that after the first cycle—including binding and release of myoglobin, which was done by lowering the pH resulting in destruction of the antigen–antibody complex—the photoluminescent intensity is decreased by 50% of the initial value. The authors discuss such a decrease in photoluminescent intensity in terms of a possible destruction of the PSi surface or incomplete removal of the immune complex from the surface. Despite this drawback, the approach offers a simple and cheap technique of preparation and operation combined with high specificity and sensitivity.

Electrochemical Transduction

The use of the electrical characteristics of PSi is a different approach. One advantage of an electrochemical sensor

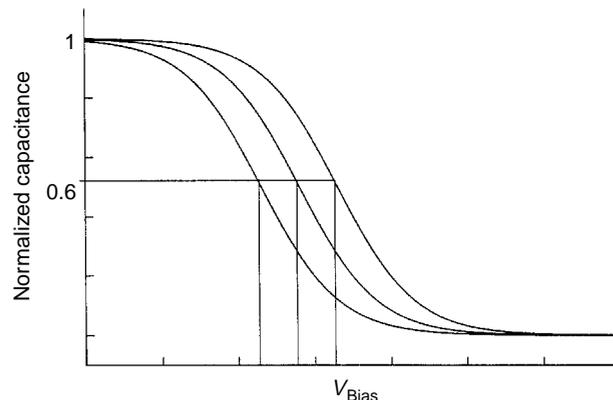
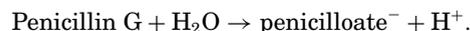


Figure 8. Schematic drawing of typical behavior of C – V curves for different H^+ concentrations. As the pH of the solution increases, the C – V curves are shifted to larger values along the x axis (54).

based on PSi compared to well-established silicon microelectronics such as ion sensitive field-effect transistors (ISFETs) is the high surface area, which allows for higher sensitivity but uses a smaller active area. Lüth and co-workers investigated PSi as a substrate material for potentiometric biosensors operating in aqueous solution. The principle of this device is a shift of the capacitance (C)–voltage (V) curve upon pH shifts (Fig. 8).

The shape of the C – V curve for p-type silicon can be explained as follows: at negative voltage, an accumulation of holes occurs at the interface, and as a result, the measured differential capacitance is close to that of the SiO_2 layer. As the negative voltage is reduced, a depletion region that acts as a dielectric in series with the SiO_2 layer is formed near the silicon surface, leading to a decrease in overall capacitance. The parallel shift of the C – V curve is caused by the flat-band voltage shift toward positive values as pH decreases. It can be explained by the presence of Si–OH groups at the surface of hydrated SiO_2 , described in site-binding theory. The ionization state of the silanol groups changes by varying the pH, and the resulting surface charge affects the depletion layer at the Si/SiO₂ interface. Thus, the performance of the sensor is strongly affected by the response of the oxide-covered PSi to pH change. Any reaction that changes the pH close to the silica surface can be measured by monitoring corresponding C – V curves. Lüth and co-workers developed a biosensor that detects penicillin by the following enzyme-catalyzed reaction (54–56):



The enzyme that catalyzes this reaction is penicillinase, also termed β -lactamase from *Bacillus cereus*. Porous silicon samples were prepared from p- or n-type by anodic etching followed by thermal annealing in an oxygen atmosphere to generate a SiO_2 layer, which protects the PSi surface from corrosion in aqueous solution. The bioactive compound, penicillinase, is immobilized via physisorption on the PSi surface. This mild immobilization method requires no additional reagents and does not affect the activity of the enzyme. Due to the holes within the porous

material, fast leaching out of the sensor compound was prevented. A penicillin G (benzylpenicillin) concentration in the range of 0.1–100 mmol/L can be monitored by a linear potentiometric response from 0.5 to 20 mM and a sensitivity of about 40 mV. Experiments performed using n-type PSi indicated even higher sensitivities of about 50 mV. To enhance pH sensitivity of the biosensor to penicillin, Lüth and co-workers deposited Si_3N_4 by plasma-enhanced chemical vapor deposition. Calibration curves indicate a pH sensitivity of 54 mV per decade that is close to the theoretical Nernstian slope of 59.1 mV/pH (57). An example of a constant capacitance measurement of a penicillinase-covered PSi surface is given in Fig. 9. The penicillin concentration is varied between 0.01 and 1 mM, and the voltage change is monitored on-line (Fig. 9a). The calibration curve (sensor signal vs. penicillin concentration) is almost linear in the concentration range of 0.01 to 0.75 mM penicillin G (Fig. 9b). The mean sensitivity is 138 ± 10 mV/mM in a concentration range of 0.025 to 0.25 mM penicillin for the first 20 days of operation.

Using a different approach, Al_2O_3 was deposited as a pH-sensitive material, which was characterized by long-term stability, stability to corrosion, and very little drift compared to the Si_3N_4 layer. The pH sensitivity was 55 mV/pH (58). To improve the biosensor further, it might be desirable to immobilize the enzyme molecules covalently to the surface via cross-linkers. Penicillinase was bound by *N*-5-azido-2-nitrobenzoyloxysuccinimide to a planar Si_3N_4 surface of the sensor. This sensor was stable for 250 days (58). The sensor needs to be miniaturized to realize capacitive microsenors. For this purpose, a multisensor array was established by coating the silicon wafer with polyimide as a passivation material that forms a microelectrode array (57).

Optical Transduction—Interferometry

Sensitive label-free biosensors are highly desirable for applications in high-throughput screening and diagnostics.

Optical transduction mechanisms such as interferometric and surface-plasmon-related methods offer several advantages, most notably label-free analyte sensing, which simplifies sample preparation. Ghadiri, Sailor, and co-workers established several biosensor surfaces based on detecting changes in the interference patterns of thin PSi layers (59–63). In a comprehensive study, Ghadiri and co-workers developed a sensor surface that detects streptavidin binding to biotin by interferometry. Several requirements had to be considered to design a proper sensor surface. The prerequisite for using PSi as an optical interferometric biosensor is to adjust the size and the geometric shape of the pores by choosing appropriate etching parameters. The pore size has to be large enough to allow proteins to enter the pores freely but small enough to retain optical reflectivity of the PSi surface. Moreover, it is necessary for the material to be mechanically stable in aqueous solutions to provide reproducible and predictable binding signals. Janshoff et al. (59) extensively studied various parameters in the fabrication of PSi. p-type silicon that has resistivities of 0.1–10 Ω cm etched in aqueous or ethanolic HF solutions generally displays a network of micropores, rather than the desired well-defined cylindrical meso- or macropores. However, the pore size of p-type PSi can be increased by increasing the concentration of the dopant and decreasing the aqueous HF concentration. On the other hand, low current densities result in random orientation of highly interconnected filament-like micropores. Large, cylindrically shaped pores can be obtained when higher current densities are applied near the electropolishing region. By anodizing heavily doped (10^{-3} Ω cm) p-type silicon (100) in ethanolic HF solution at ambient temperatures, Janshoff et al. (59) predictively fabricated PSi layers that had cylindrically shaped structures and tunable pore diameters in the range of 5 to 1200 nm, as deduced from scanning force microscopy images (Fig. 10).

Using low current densities (150 mA/cm²), pores are scarcely visible, and the relatively flat surface is dominated by a distinct hillock structure. As the current densities

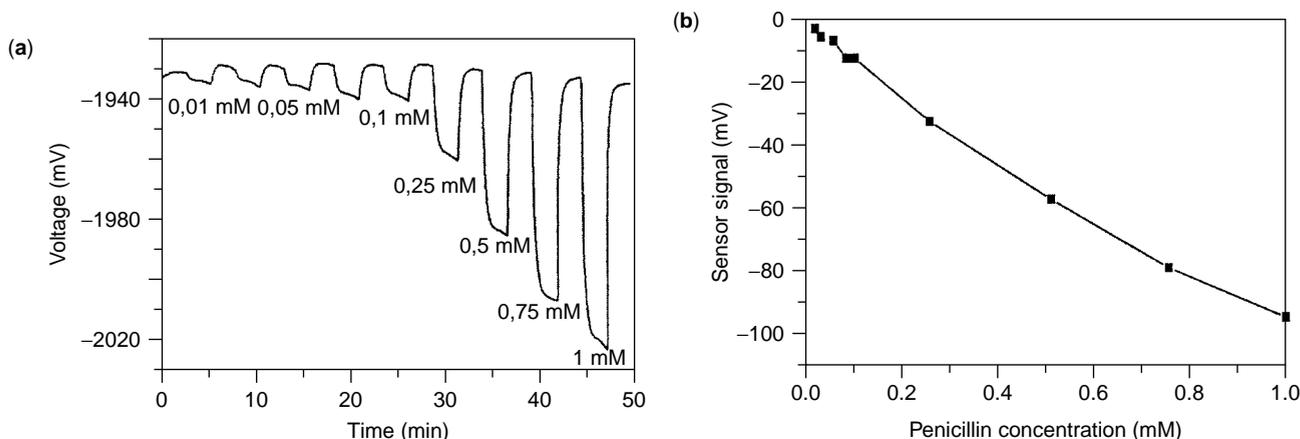


Figure 9. (a) Typical constant capacitance measurement. The enzyme penicillinase is immobilized by physisorption onto a Si_3N_4 -covered PSi surface. Different concentrations of penicillin G sodium salt were added, and the change in voltage was monitored on-line. (b) Corresponding calibration curve that exhibits a wide linear range from 0.01 to 0.75 mM penicillin [reprinted with permission from (55)].

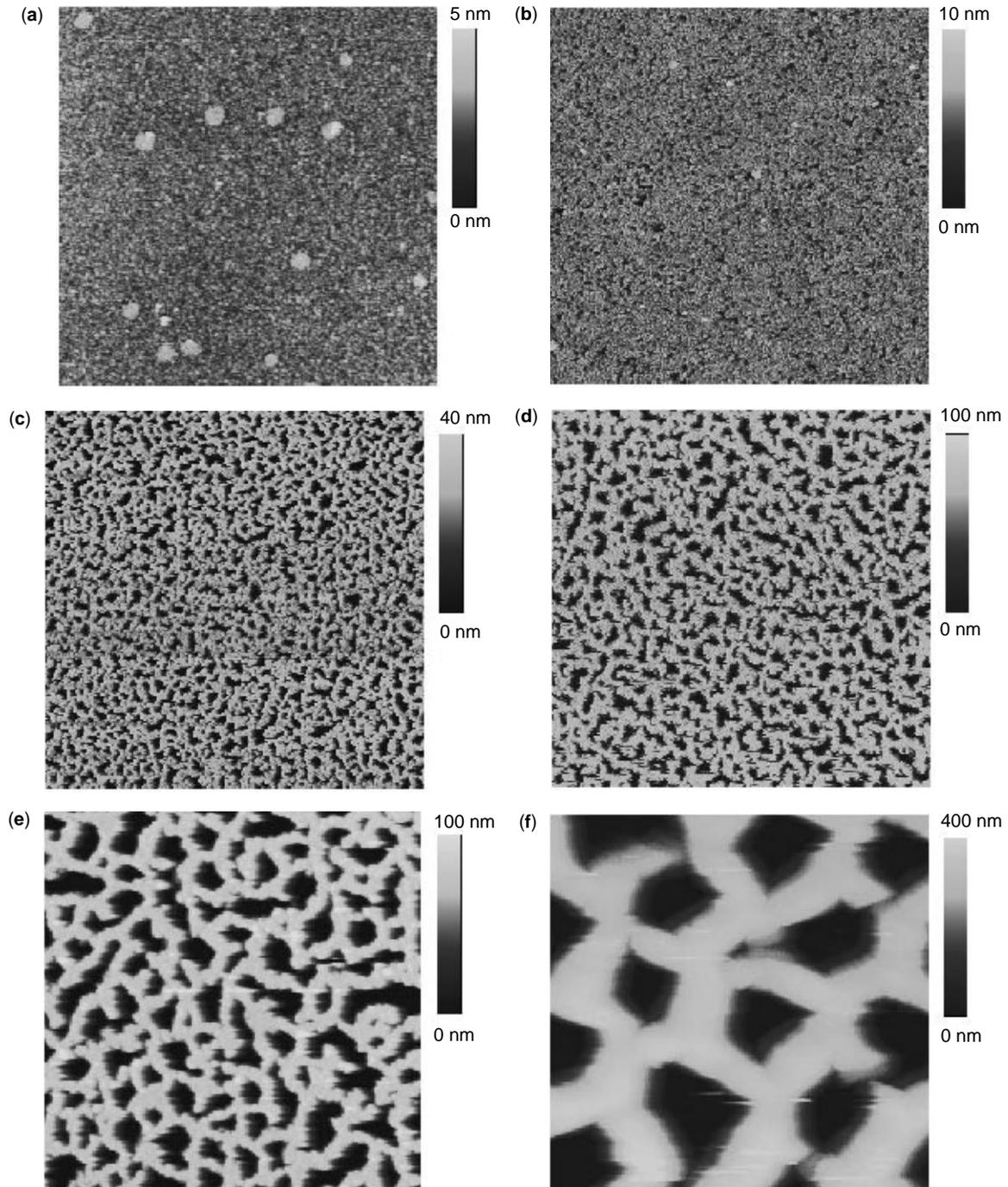


Figure 10. SFM images (tapping mode) of porous p-silicon layers freshly etched at different current densities. (a) $1.5 \times 1.5 \mu\text{m}^2$ image etched at 150 mA/cm^2 ; all following images are $5 \times 5 \mu\text{m}^2$ (b) etched at 295 mA/cm^2 ; (c) 370 mA/cm^2 ; (d) 440 mA/cm^2 ; (e) 515 mA/cm^2 ; and (f) at 600 mA/cm^2 . The dopant concentration ($1 \text{ m}\Omega \text{ cm}$) and anodizing solution (37% ethanolic HF) were the same for all samples. All samples were etched at a constant charge of 4.5 C/cm^2 [reprinted with permission from (59)].

are increased, larger pore sizes can be obtained. The pore radius depends approximately exponentially on the current density. The surface porosity of silicon layers, calculated from SFM images by integrating the number of pixels, increases slightly from 27% (330 mA/cm^2) to 30% (410 mA/cm^2) and finally up to 40% by applying

densities $>440 \text{ mA/cm}^2$. The interference or fringe patterns obtained from these PSi layers anodized at different current densities are presented in Fig. 11.

Fabry–Perot fringes using visible light illumination were observed on samples prepared at current densities between 150 and 600 mA/cm^2 . Anodization of p-type

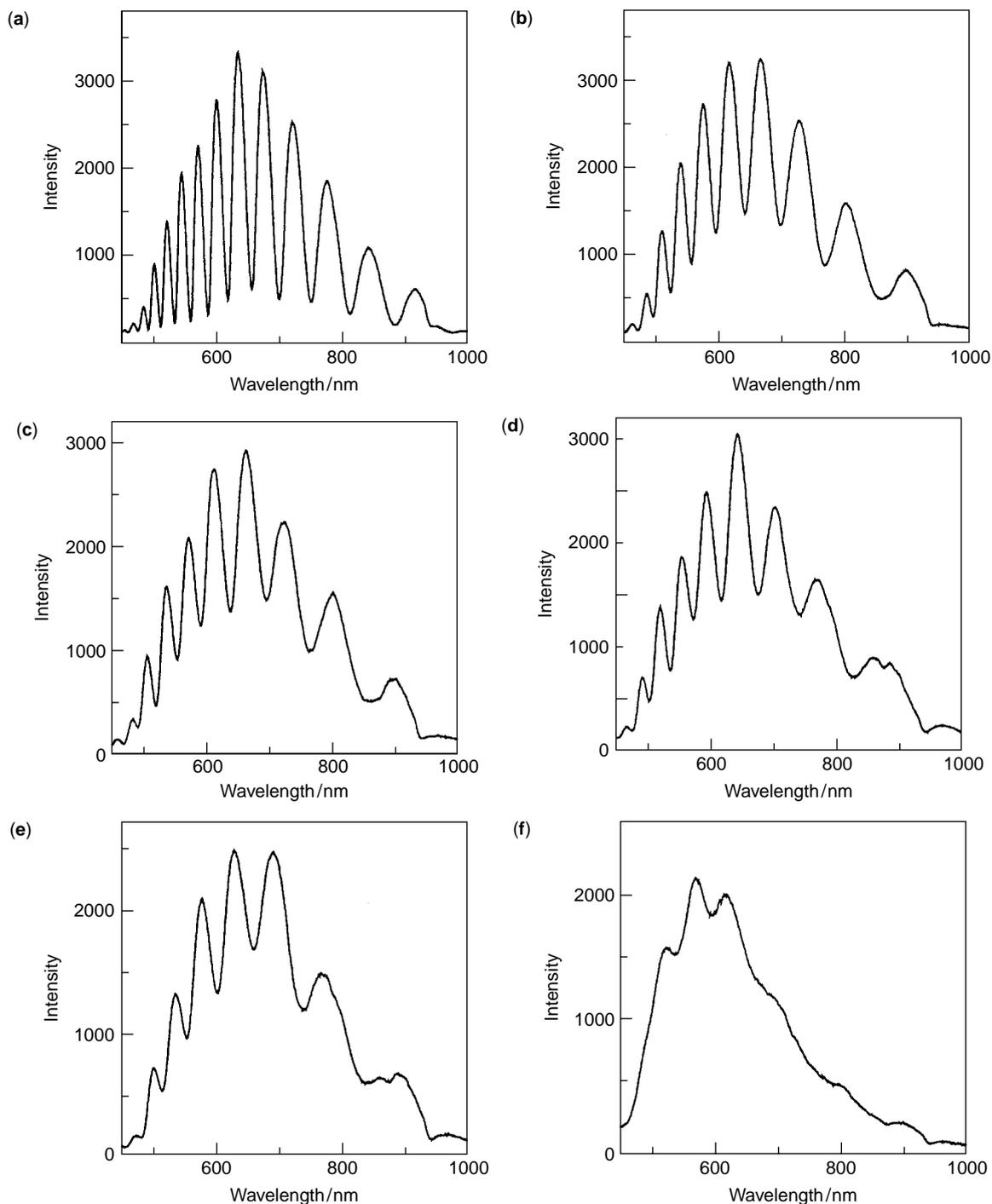


Figure 11. Interference fringe patterns of p-type PSi etched at different current densities. All samples were etched at a constant charge of 4.5 C/cm^2 . The spectra were taken in the center of the chip. (a) 150 mA/cm^2 ; (b) 295 mA/cm^2 ; (c) 370 mA/cm^2 ; (d) 440 mA/cm^2 ; (e) 515 mA/cm^2 ; (f) 600 mA/cm^2 [reprinted with permission from (59)].

silicon at a current density of 600 mA/cm^2 resulted in an obvious matte surface that had a barely discernible fringe pattern due to insufficient reflectivity of the upper PSi layer. Electropolishing occurs at a current density higher than 700 mA/cm^2 . The number of fringes in the observed wavelength range depends on the porosity and the thickness of the porous layer. Samples approximately 3000 nm

thick typically display 9 to 12 fringes in the wavelength region of $500\text{--}1000 \text{ nm}$, depending on the effective refractive index. The higher the current density, the fewer fringes are observed, consistent with the observation that higher current densities lead to greater porosities. To determine the porosity p and thickness l of the porous layers, the pores were filled with organic solvents of different refractive

index n , and the effective optical thickness was determined from interferometric reflectance spectra. Different EMAs were applied to the data to obtain the porosity and thickness of the porous layer simultaneously. The parameters p and l for each sample were determined from the fit to a plot of $n_{\text{eff}}l$ versus n . Independent of the EMA used, the estimated porosity of PSi increases with increasing current density in close agreement with experimental observation. According to the theory of Looyenga, the porosities of the samples etched at different current densities were in the range of 64–90% in good agreement with gravimetric measurements which yielded a porosity of $80 \pm 5\%$ for PSi samples etched at 150 mA/cm^2 and $90 \pm 5\%$ for samples etched at 400 mA/cm^2 .

Because freshly etched, hydride-terminated PSi readily suffers oxidative and hydrolytic corrosion, the surface needs to be oxidized and functionalized to stabilize it. Stability was proven by measuring the effective optical thickness (EOT) as a function of time (Fig. 12).

The observed decrease in EOT is caused by oxidation and dissolution of the PSi. The conversion of silicon to silica results in a decrease in the effective refractive index of the PSi layer, leading to the observed blue shift of the interference fringes. Furthermore, dissolution of the porous layer can lead to a decrease in thickness of the layer, which would also result in a decrease in the effective optical thickness, and therefore to a shift of the spectrum to shorter wavelengths. Ghadiri and co-workers found that ozonolysis followed by capping using a long-chain alkoxy silane linker (Scheme 2a) stabilized the surface sufficiently for sensing in aqueous media. Ozonolysis was the preferred oxidation route because a larger number of Si–OH groups are generated compared to thermal oxidation through which binding of the alkoxy silane linker occurred. A monoalkoxy instead of a trialkoxy silane was used to prevent the formation of cross-linking reactions, which might result in clogging the pores by silane polymers. By tethering a biotin molecule to the end of the linker, streptavidin can bind to the chemically modified PSi surface. To eliminate nonspecific binding further and to space the binding sites apart to reduce crowding on the surface, Sailor and co-workers (60) synthesized a linker molecule containing bovine serum albumin (BSA) (Scheme 2b).

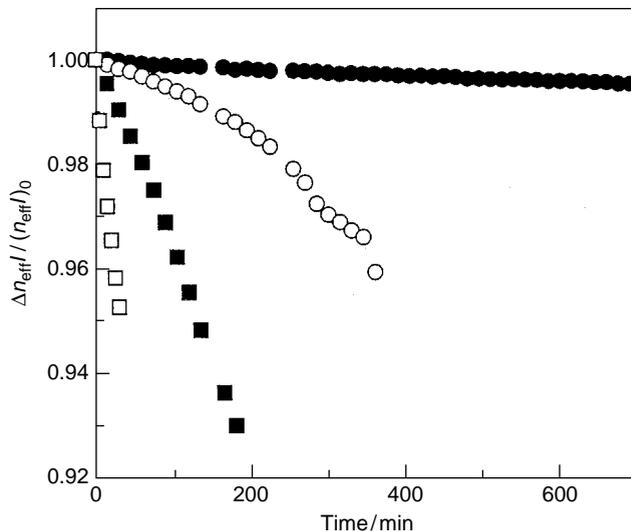
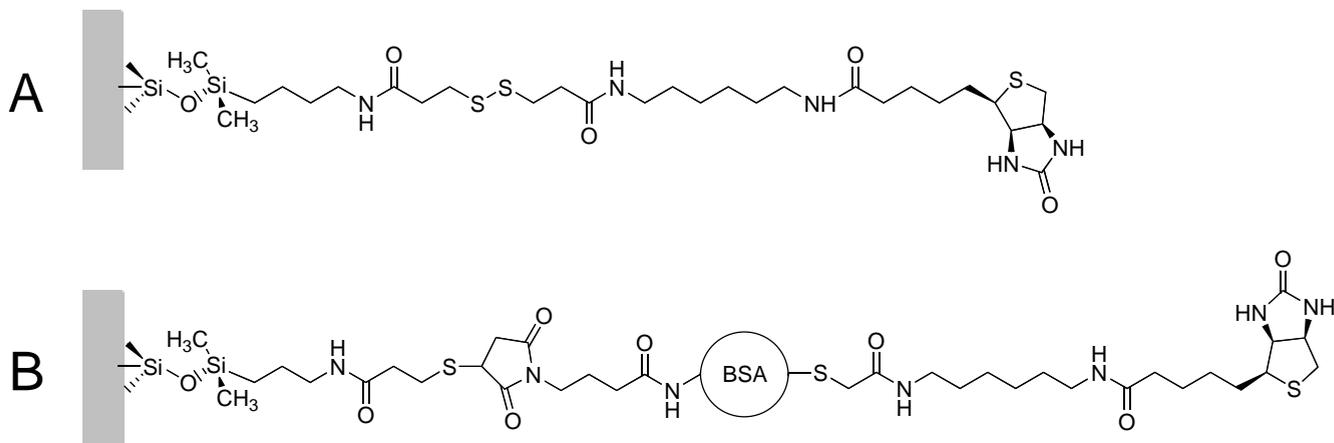


Figure 12. Stability of various surface-derivatized PSi samples in 10% (v/v) EtOH in PBS buffer, pH 7.4, presented as the normalized relative effective optical thickness change (normalized EOT) as a function of time. The slopes of $n_{\text{eff}}l/t$ are given in brackets (\square). Hydride-terminated PSi sample (6 nm/min); (\blacksquare) ozone oxidized sample (2 nm/min); (\circ) thermally oxidized (400°C , 1h), (1 nm/min); (\bullet) ozone oxidized PSi wafer functionalized by using (2-pyridyldithiopropionamido) butyldimethylmethoxysilane (0.05 nm/min) [reprinted with permission from (59)].

The accessibility of the porous matrix to biological molecules was probed by exposure of a concentrated BSA solution in PBS buffer to an ozone-oxidized PSi sample functionalized by 2-pyridyldithio(propionamido) dimethylmonomethoxysilane and pretreated by using with BSA to inhibit nonspecific adsorption to the silicon surface. The expected shift in EOT of about 10–30 nm, considering the volume of the pores and the refractive index of the aqueous BSA solution, was reached within 2–3 min and confirmed that proteins can enter and fill the porous matrix within a reasonable timescale. Although the observed shift is mainly due to the bulk effect of the protein solution, the slower rate of recovery after rinsing the sample with buffer suggests that some proteins were physisorbed on

Scheme 2



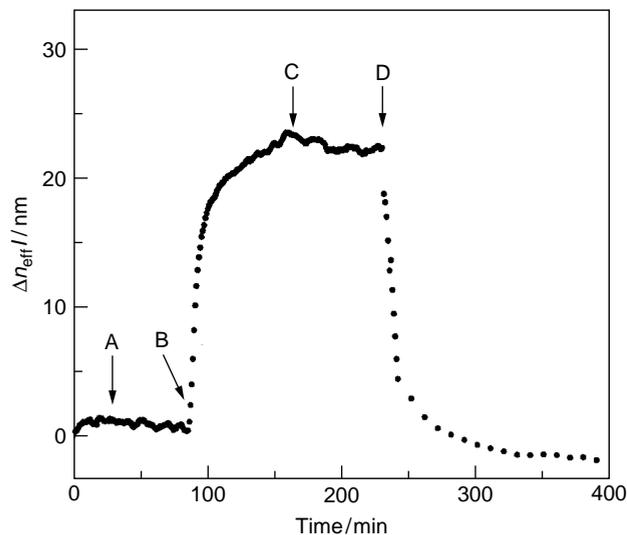


Figure 13. Time course of the EOT (n_{eff}) of a p-type PSi chip etched at 440 mA/cm^2 , oxidized by ozone for 20 min, and functionalized as shown in Scheme 2 a. The arrow labeled **A** identifies the addition of $10 \mu\text{M}$ streptavidin preincubated in 1 mM biotin dissolved in PBS buffer, pH 7.4 (control); **B** addition of $10 \mu\text{M}$ streptavidin without biotin (washing cycles in between); **C** washing cycles with buffer; **D** addition of dithiothreitol, which was used to reduce the disulfide bridge and therefore release the bound protein-linker complex. The sample was mounted in a flow cell using a constant flow rate of 0.5 mL/min [reprinted with permission from (59)].

the silicon walls. Using an ethanol-water mixture instead of the protein solution results in a rectangular signal response upon adding the mixture and rinsing with water.

Specific binding of streptavidin to the biotin-functionalized PSi matrix was measured by monitoring the changes in EOT time-resolved in a PBS buffer containing 0.1% TritonTM to minimize nonspecific adsorption (Fig.13).

As expected, specific binding of streptavidin to the biotin-derivatized porous layer resulted in an increase in the measured effective optical thickness. The change in the EOT is due to binding of proteins that have a higher refractive index ($n_{\text{protein}} = 1.42$) than the water ($n_{\text{water}} = 1.33$) in the pores and is in direct quantitative agreement with what was expected from effective medium approximations. The overall change in the EOT (Δn_{eff}) after 80 min was 23 nm. In a control experiment, in which all streptavidin binding sites were deactivated by saturating them with biotin in solution, a change in EOT was not observed, suggesting that there is little or no nonspecific protein adsorption to the PSi matrix. Rinsing the surface with buffer after the protein has bound does not alter the EOT significantly. However, because the biotin recognition element is linked to the surface via a disulfide bond, the protein-ligand complex could be released from the surface by adding dithiothreitol to the bulk phase. The initial red shift of 23 nm upon binding streptavidin to the biotinylated PSi can be completely reversed and provides further support for the interpretation that the observed red shift is due to specific binding of the protein to the functionalized surface. Moreover, the reversible linkage of the proteins via disulfide bridges to the surface offers the possibility of reusing the functionalized PSi chips for further binding experiments. Sailor and co-workers bound protein A to the PSi surface through the BSA-containing linker (60,61). Streptavidin binds to the biotin-terminated linker and adds three accessible free biotin-binding sites to the surface (Fig. 14).

Adding a solution of biotinylated protein A results in attaching it to the surface. This prefunctionalized surface can be used for binding studies of aqueous human IgG. The observed change in EOT for binding IgG required several minutes to reach a steady-state value, presumably due to slow diffusion of this large molecule into the pores of the PSi film. The proteinA/IgG complex was partly dissociated by rinsing with buffer and completely dissociated by a pH switch to a low pH. Protonation of the binding sites on

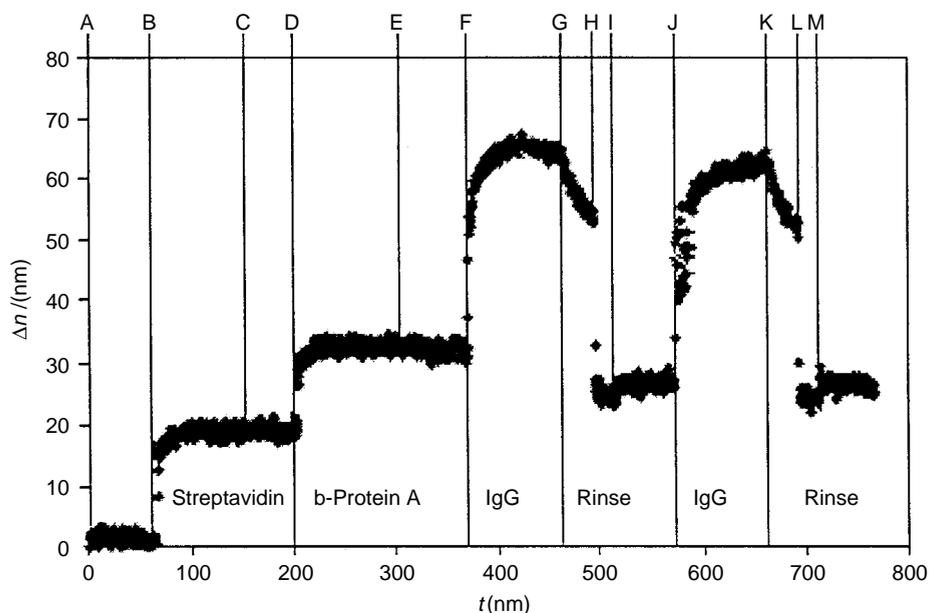


Figure 14. Binding curve (change in EOT) on a PSi surface functionalized as shown in Scheme 2b. Sequential addition of streptavidin (1 mg/mL), biotinylated protein A (2.5 mg/mL), and human IgG (2.5 mg/mL). Reversible binding of IgG was demonstrated by binding of IgG followed by a pH-induced release and a second binding of IgG to the immobilized protein A layer [reprinted with permission from (60)].

protein A by decreasing the pH of the solution releases IgG from protein A. A second binding of IgG after its release can be demonstrated that shows the reproducibility of the method. The incorporation of BSA in the linker offered two advantages. Due to the increased hydrophilicity of the chemically modified PSi, surface nonspecific adsorption was not observed, and the addition of detergent in the buffer was no longer necessary. A second reason for incorporating BSA in the linker was to separate binding sites in the PSi films. Sailor and co-workers (61) found that without BSA the sensor did not scale with the mass of analyte, as was expected, assuming the same refractive index for all proteins investigated. Larger analytes were consistently underestimated, indicating crowding of binding sites at the surface. The insertion of BSA in the linker avoided crowding and thus, the sensor scaled with the analyte mass above 20 kDa (60).

Optical Transduction—Ellipsometry

Optical biosensing is usually based on the interaction of light with biomolecules. Techniques such as surface plasmon resonance and ellipsometry have focused mostly on interactions on a macromolecular scale, for example, antigen–antibody and nucleic acid interactions. The optical detection of small molecules (0.2–2 kDa) that have biological receptors is much more difficult due to their small change in EOT. Mandenius and co-workers (64) demonstrated the advantage of using oxidized PSi as a surface enlargement for binding small receptor molecules such as biotin or small peptides. They used p-type silicon that had (111) orientation and a resistivity of 0.01–0.02 Ω cm. The samples were thermally oxidized to stabilize the porous structure. The PSi surface was functionalized by using streptavidin, either physisorbed on the silica surface or cross-linked via glutardialdehyde. Streptavidin adsorption monitored by ellipsometry showed a 10-fold larger response compared to a planar surface. However, the rate of adsorption was one order of magnitude lower, probably due to the long diffusion time of the protein within the pores. Theoretically, the refractive index and the thickness of a thin layer can be calculated from the measured parameters ψ (the ratio of the amplitude change of light polarized parallel and perpendicular to the plane of incidence) and Δ (the phase shift). For PSi, however, the microstructure of the porous layer is very complicated, and a simple optical model that allowing quantifying film thickness and surface concentration is not straightforward to define. Therefore, Madenius and co-workers used changes in ψ and Δ as a direct measure of analyte binding without quantification. Using this setup, they detected binding of biotin and an oligopeptide in a concentration range of 2–40 μ M and a response time of 30 s for the oligopeptide at a concentration of 40 μ M.

CONCLUSIONS

Porous silicon based biosensors may add a new dimension to conventional technologies due to their unique optical and electronic properties. Tunable properties such as pore size,

porosity, dielectric function, and thickness render porous silicon a versatile matrix for biological compounds that act as the receptive layer for molecular recognition of analytes in solution. Interferometry has been successfully employed to detect changes in the effective optical thickness upon adsorption of molecules on the pore walls. The large surface area of porous silicon that displays a spongelike appearance or exhibits ordered cylindrical pores provides a quasi three-dimensional space that increases the signal-to-noise ratio of many transducing principles.

BIBLIOGRAPHY

1. A. Uhler, *Bell. Syst. Tech. J.* **35**: 333–347 (1956).
2. L. Canham, *Appl. Phys. Lett.* **57**: 1046–1048 (1990).
3. V. Parkhutik, *Solid-State Electron.* **43**: 1121–1141 (1999).
4. M.I.J. Beale, J.D. Benjamin, M.J. Uren, N.G. Chew, and A.G. Cullis, *J. Cryst. Growth* **73**: 622–636 (1985).
5. R.L. Smith and S.D. Collins, *J. Appl. Phys.* **71**: R1–R22 (1992).
6. P.C. Searson, *Advances in Electrochemical Sciences and Engineering*. VCH, Mannheim, 1994.
7. V. Lehmann and U. Gösele, *Appl. Phys. Lett.* **58**: 856–858 (1991).
8. S.J. Gregg and K.S.W. Sing, *Adsorption, Surface Area and Porosity*. Academic Press, NY, 1982.
9. D. Brumhead, L.T. Canham, D.M. Seekings, and P.J. Tufton, *Electrochim. Acta* **38**: 191–197 (1993).
10. W. Theiss, *Surf. Sci. Rep.* **29**: 91–192 (1997).
11. R. Herino, in *Properties of Porous Silicon, Vol. Datareview Ser. No. 18*, L.T. Canham, ed., London, 1997, pp. 89–96.
12. C. Kittel, *Introduction to Solid State Physics*. J. Wiley, NY, 1996.
13. L.T. Canham, in *Properties of Porous Silicon, Vol. Datareview Ser. No. 18*, L.T. Canham, ed., London, 1997, pp. 249–255.
14. F. Kozłowski, P. Steiner, and W. Lang, *J. Luminescence* **57**: 163–167 (1993).
15. H. Mizuno, H. Koyama, and N. Koshida, *Appl. Phys. Lett.* **69**: 3779–3781 (1996).
16. S.G. Lipson, H. Lipson, and D.S. Tannhauser, *Optical Physics*. Cambridge University Press, Cambridge, 1995.
17. G. L erondel and R. Romestain, *Thin Solid Films* **297**: 114–117 (1997).
18. L. Pavesi and P. Dubos, *Semicond. Sci. Technol.* **12**: 570–575 (1997).
19. J.H. Song and M.J. Sailor, *Comments Inorg. Chem.* **21**: 69–84 (1999).
20. E.J. Lee, T.W. Bitner, A.P. Hall, and M.J. Sailor, *J. Vac. Sci. Technol.* **B 14**: 2850–2854 (1996).
21. E.J. Lee, J.S. Ha, and M.J. Sailor, *J. Am. Chem. Soc.* **117**: 8295–8296 (1995).
22. J.E. Lee, T.W. Bitner, J.S. Ha, M.J. Shane, and M.J. Sailor, *J. Am. Chem. Soc.* **118**: 5375–5382 (1996).
23. N.Y. Kim and P.E. Laibinis, *Mater. Res. Soc. Symp. Proc.* **536**: 167–172 (1999).
24. N.Y. Kim and P.E. Laibinis, *ACS Symp. Ser.* **727**: 157–168 (1999).
25. J.M. Buriak, *Chem. Commun.* 1051–1060 (1999).
26. T. Dubois, F. Ozanam, and J.-N. Chazaviel, *Proc. Electrochem. Soc.* **97**: 296–310 (1997).

27. C. Vieillard, M. Warntjes, F. Ozanam, and J.-N. Chazalviel, *Proc. Electrochem. Soc.* **95**: 250–258 (1996).
28. N.Y. Kim and P.E. Laibinis, *J. Am. Chem. Soc.* **120**: 4516–4517 (1998).
29. J.H. Song and M.J. Sailor, *J. Am. Chem. Soc.* **120**: 2376–2381 (1998).
30. J.H. Song and M.J. Sailor, *Inorg. Chem.* **38**: 1498–1503 (1999).
31. J.M. Holland, M.P. Stewart, M.J. Allen, and J.M. Buriak, *J. Solid State Chem.* **147**: 251–258 (1999).
32. J.M. Buriak and M.J. Allen, *J. Am. Chem. Soc.* **120**: 1339–1340 (1998).
33. J.M. Buriak, M.P. Stewart, T.W. Geders, M.J. Allen, H.C. Choi, J. Smith, D. Raftery, and L.T. Canham, *J. Am. Chem. Soc.* **121**: 11491–11502 (1999).
34. J.M. Buriak, M.P. Stewart, and M.J. Allen, *Mater. Res. Soc. Symp. Proc.* **536**: 173–178 (1999).
35. M.P. Stewart and J.M. Buriak, *Angew. Chem. Int. Ed.* **37**: 3257–3260 (1998).
36. E.G. Robins, M.P. Stewart, and J.M. Buriak, *Chem. Commun.* 2479–2480 (1999).
37. C. Gurtner, A.W. Wun, and M.J. Sailor, *Angew. Chem. Int. Ed.* **38**: 1966–1968 (1999).
38. V. Petrova-Koch, T. Muschik, A. Kux, B.K. Meyer, F. Koch, and V. Lehmann, *Appl. Phys. Lett.* **61**: 943–945 (1992).
39. V.M. Dubin, C. Vieillard, F. Ozanam, and J.-N. Chazalviel, *Phys. Status Solidi B* **190**: 47–52 (1995).
40. R.C. Anderson, R.S. Müller, and C.W. Tobias, *J. Electrochem. Soc.* **149**: 1393–1395 (1993).
41. J.M. Lauerhaas and M.J. Sailor, *Science* **261**: 1567–1568 (1993).
42. J.M. Lauerhaas and M.J. Sailor, *Mat. Res. Soc. Symp. Proc.* **298**: 259–263 (1993).
43. A. Bansal, X. Li, I. Lauermann, N.S. Lewis, S.I. Yi, and W.H. Weinberg, *J. Am. Chem. Soc.* **118**: 7225–7226 (1996).
44. J. Harper and M.J. Sailor, *Anal. Chem.* **68**: 3713–3717 (1996).
45. S. Content, W.C. Trogler, and M.J. Sailor, *Chem. Eur. J.* **6**: 2205–2213 (2000).
46. S.E. Letant, S. Content, T.T. Tan, F. Zenhausern, and M.J. Sailor, *Sensors and Actuators B* **69**: 193–198 (2000).
47. L. Zhang, J.L. Coffey, J. Wang, C.D. Gutsche, J.-J. Chen, and O. Zhyhan, *J. Am. Chem. Soc.* **118**: 12840–12841 (1996).
48. S.E. Letant and M.J. Sailor, *Adv. Mater.* **12**: 355–359 (2000).
49. R.C. Anderson, R.S. Muller, and C.W. Tobias, *Sensors and Actuators A21–A23*: 835–839 (1990).
50. M.P. Stewart and J.M. Buriak, *Adv. Mater.* **12**: 859–869 (2000).
51. N.F. Starodub, L.L. Fedorenko, V.M. Starodub, S.P. Dikij, and S.V. Svechnikov, *Sensors and Actuators B* **35–36**: 44–47 (1996).
52. N.F. Starodub and V.M. Starodub, *Proc. Electrochem. Soc.* **98**: 185–191 (1998).
53. V.M. Starodub, L.L. Fedorenko, A.P. Sisetkiy, and N.F. Starodub, *Sensors and Actuators B* **58**: 409–414 (1999).
54. M. Thust, M.J. Schöning, S. Frohnhoff, R. Arens-Fischer, P. Kordos, and H. Lüth, *Meas. Sci. Technol.* **7**: 26–29 (1996).
55. M. Thust, M.J. Schöning, P. Schroth, U. Malkoc, C.I. Dicker, A. Steffen, P. Kordos, and H. Lüth, *J. Mol. Catal. B: Enzym.* **7**: 77–83 (1999).
56. H. Lüth, M. Steffen, P. Kordos, and M.J. Schöning, *Mater. Sci. Eng.* **B 69–70**: 104–108 (2000).
57. M.J. Schöning, F. Ronkel, M. Crott, M. Thust, J.W. Schultze, P. Kordos, and H. Lüth, *Electrochim. Acta* **42**: 3185–3193 (1997).
58. H. Lüth, M.J. Schöning, and M. Thust, *Phys. Bl.* **53**: 423–427 (1997).
59. A. Janshoff, K.-P.S. Dancil, C. Steinem, D.P. Greiner, V.S.-Y. Lin, C. Gurtner, K. Motesharei, M.J. Sailor, and M.R. Ghadiri, *J. Am. Chem. Soc.* **120**: 12108–12116 (1998).
60. K.-P.S. Dancil, D.P. Greiner, and M.J. Sailor, *J. Am. Chem. Soc.* **121**: 7925–7930 (1999).
61. K.-P.S. Dancil, D.P. Greiner, and M.J. Sailor, *Mater. Res. Soc. Symp. Proc.* **536**: 557–562 (1999).
62. V.-Y. Lin, K. Motesharei, K.S. Dancil, M.J. Sailor, and M.R. Ghadiri, *Science* **278**: 840–842 (1997).
63. A.M. Tinsley-Brown, L.T. Canham, M. Hollings, M.H. Anderson, C.L. Reeves, T.I. Cox, S. Nicklin, D.J. Squirrel, E. Perkins, A. Hutchinson, M.J. Sailor, and A. Wun, *Phys. Stat. Sol. A* **182**: 547–553 (2000).
64. D. van Noort, S. Welin-Klintström, H. Arwin, S. Zangoie, I. Lundström, and C.-F. Mandenius, *Biosensors and Bioelectronics* **13**: 439–449 (1998).

CERAMICS, PIEZOELECTRIC AND ELECTROSTRICTIVE

ANDREI KHOLKIN
BAHRAM JADIDIAN
AHMAD SAFARI
Rutgers University
Piscataway, NJ

INTRODUCTION

In a rapidly developing world, the use of smart materials becomes increasingly important when executing sophisticated functions within a designed device. In a common definition (1), smart materials differ from ordinary materials because they can perform two or several functions, sometimes with a useful correlation or feedback mechanism between them. For piezoelectric or electrostrictive materials, this means that the same component may be used for both sensor and actuator functions. Piezoelectric/electrostrictive sensors convert a mechanical variable (displacement or force) into a measurable electrical quantity by the piezoelectric/electrostrictive effect. Alternately, the actuator converts an electrical signal into a useful displacement or force. Typically, the term transducer is used to describe a component that serves actuator (transmitting) and sensor (receiving) functions. Because piezoelectrics and electrostrictors inherently possess both direct (sensor) and converse (actuator) effects, they can be considered smart materials. The degree of smartness can vary in piezoelectric/electrostrictive materials. A merely smart material (only sensor and actuator functions) can often be engineered into a “very smart” tunable device or further, into an “intelligent structure” whose sensor and actuator functions are intercorrelated with an integrated processing chip.

Recent growth in the transducer market has been rapid and, it is predicted will continue on its current pace through the turn of the century. The sensor market alone rose to \$5 billion in 1990, and projections are \$13 billion worldwide by the year 2000 and an 8% annual growth rate during the following decade (2). Piezoelectric/electrostrictive sensors and actuators comprise a significant portion of the transducer market. There is a growing trend due especially to automobile production, active vibration damping, and medical imaging. In this article, the principles of piezoelectric/electrostrictive sensors and actuators are considered along with the properties of the most useful materials and examples of successful devices.

PIEZOELECTRIC AND ELECTROSTRICTIVE EFFECTS IN CERAMIC MATERIALS

Piezoelectricity, first discovered in Rochelle salt by Jacques and Pierre Curie, is the term used to describe the ability of certain crystals to develop an electric charge that is directly

proportional to an applied mechanical stress (Fig. 1a) (3). Piezoelectric crystals also show the converse effect: they deform (strain) proportionally to an applied electric field (Fig. 1b). To exhibit piezoelectricity, a crystal should belong to one of the twenty nonsymmorphic crystallographic classes. An important subgroup of piezoelectric crystals is ferroelectrics, which possess a mean dipole moment per unit cell (spontaneous polarization) that can be reversed by an external electric field. Above a certain temperature (Curie point), most ferroelectrics lose their ferroelectric and piezoelectric properties and become paraelectrics, that is, crystals that have centrosymmetric crystallographic structures do not spontaneously polarize. Electrostriction is a second-order effect that refers to the ability of all materials to deform under an applied electrical field. The phenomenological master equation (in tensor notation) that describes the deformations of an insulating crystal subjected to both an elastic stress and an electrical field is

$$x_{ij} = s_{ijkl} X_{kl} + d_{mij} E_m + M_{mni} E_m E_n, \quad (1)$$

$$i, j, k, l, m, n = 1, 2, 3,$$

where x_{ij} are the components of the elastic strain, s_{ijkl} is the elastic compliance tensor, X_{kl} are the stress components, d_{mij} are the piezoelectric tensor components, M_{mni} are the electrostrictive moduli, and E_m and E_n are the components of the external electrical field. Here, the Einstein summation rule is used for repeating indexes. Typically, the electrostriction term ($\propto E_m E_n$) is more than an order of magnitude smaller than the piezoelectric term in Eq. (1), that is, the electrostrictive deformations are much smaller than the piezoelectric strains. In this case, under zero stress, Eq. (1) simply transforms to

$$x_{ij} \approx d_{mij} E_m, \quad i, j, m = 1, 2, 3. \quad (2a)$$

Eliminating symmetrical components simplifies the relationship in matrix notation (4) expressed as

$$x_i \approx d_{mi} E_m, \quad m = 1, 2, 3, \quad (2b)$$

$$i = 1, 2, 3, 4, 5, 6,$$

where $i = 4, 5,$ and 6 describe the shear strains perpendicular to the crystal axis resulting from application of the electrical field. Equations (2a) and (2b) describe the converse piezoelectric effect where the electrical field induces a change in the dimensions of the sample (Fig. 1b). The piezoelectric effect is absent in centrosymmetric materials, and the elastic strain is due only to electrostriction. In ferroelectric crystals that have a centrosymmetric paraelectric phase, the piezoelectric and electrostriction coefficients can be described in terms of their polarization and relative permittivity. For example, when the electrical field and deformation are along the orthogonal axis in a tetragonal crystal system, longitudinal piezoelectric d_{33} and longitudinal electrostrictive M_{11} coefficients can be

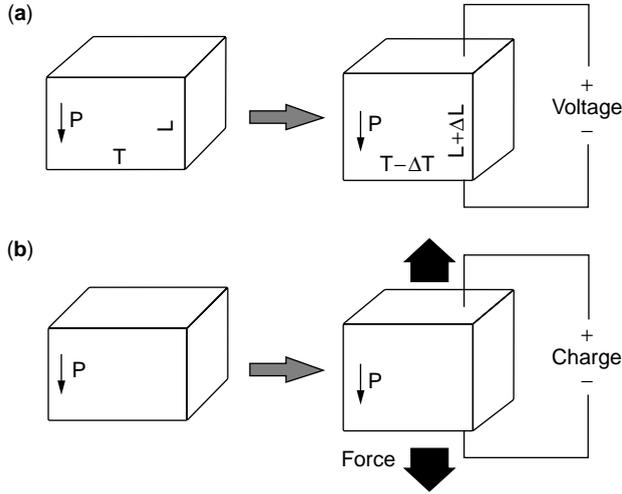


Figure 1. Schematic representations of the direct and converse piezoelectric effect: (a) an electric field applied to the material changes its shape; (b) a stress on the material yields an electric field across it.

described in matrix notation as follows (5):

$$d_{33} = 2Q_{11}\epsilon_0\epsilon_{33}P_3, \quad (3a)$$

$$M_{11} = Q_{11}(\epsilon_0\epsilon_{33})^2, \quad (3b)$$

where ϵ_{33} and P_3 are the relative permittivity and polarization in the polar direction, $\epsilon_0 = 8.854 \times 10^{-12}$ F/m is the permittivity of vacuum, and Q_{11} is the polarization electrostriction coefficient, which couples longitudinal strain and polarization (in matrix notation), as described by the general electrostriction equation,

$$S_3 = Q_{11}P_3^2. \quad (4)$$

In matrix notation, the mathematical definition of the direct piezoelectric effect, where applied elastic stress causes a charge to build on the major surfaces of the piezoelectric crystal, is given by

$$P_i = d_{ij}X_j, \quad i = 1, 2, 3, \quad (5)$$

$$j = 1, 2, 3, 4, 5, 6,$$

where P_i is the component of electrical polarization. In electrostriction (centrosymmetric crystals), no charge appears on the surface of the crystal upon stressing. Therefore, the converse electrostriction effect is simply a change of the inverse relative permittivity under mechanical stress:

$$\Delta \left[\frac{1}{\epsilon_0\epsilon_{33}} \right] = 2Q_{11}X_3. \quad (6)$$

The piezoelectric and electrostrictive effects were described for single crystals in which spontaneous polarization is homogeneous. A technologically important class of materials is piezoelectric and electrostrictive ceramics, that consist of randomly oriented grains, separated by grain boundaries. Ceramics are much less expensive to process than single crystals and typically offer comparable piezoelectric and electrostrictive properties. The piezoelectric effect of individual grains in nonferroelectric

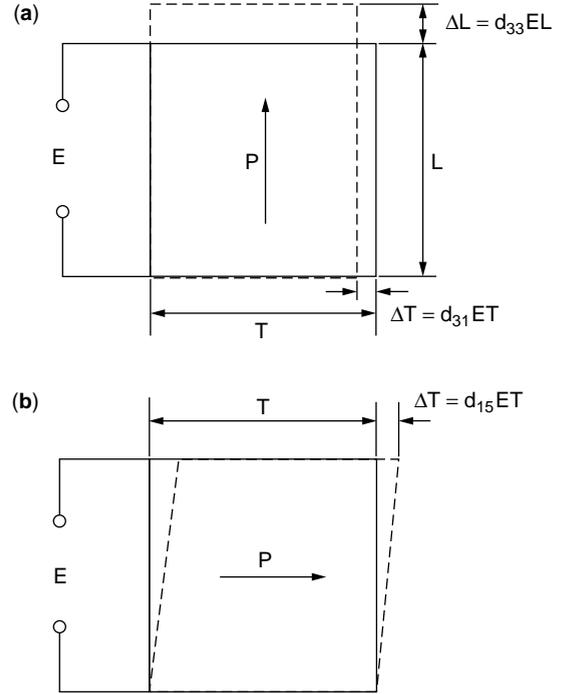


Figure 2. Schematic of the longitudinal (a), transverse (a) and shear deformations (b) of the piezoelectric ceramic material under an applied electric field.

ceramics is canceled by averaging across the entire sample, and the whole structure has a macroscopic center of symmetry that has negligible piezoelectric properties. Only electrostriction can be observed in such ceramics. Sintered ferroelectric ceramics consist of regions that have different orientations of spontaneous polarization—so-called ferroelectric domains. Domains appear when a material is cooled through the Curie point to minimize the electrostatic and elastic energy of the system. Domain boundaries or domain walls are movable in an applied electric field, so the ferroelectric can be poled. For example, domains become oriented in a crystallographic direction closest to the direction of the applied electric field. Typically, poling is performed under high electric field at an elevated temperature to facilitate domain alignment. As a result, an initially centrosymmetric ceramic sample loses its inversion center and becomes piezoelectric (symmetry ∞m). There are three independent piezoelectric coefficients: d_{33} , d_{31} , and d_{15} , which relate longitudinal, transverse, and shear deformations, respectively, to the applied electric field (Fig. 2).

Other material coefficients that are frequently used to characterize the piezoelectric properties of ceramics are the piezoelectric voltage coefficients g_{ij} , which are defined in matrix notation as

$$E_i = g_{ij}X_j, \quad (7)$$

where E_i are components of the electric field that arise from external stresses X_j . The piezoelectric charge d_{ij} and voltage g_{ij} coefficients are related by the following equation:

$$g_{ij} = d_{ij}/(\epsilon_0\epsilon_{ii}). \quad (8)$$

An important property of piezoelectric and electrostrictive transducers is their electromechanical coupling coefficient k defined as

$$k^2 = \text{output mechanical energy/input electrical energy,}$$

or

$$k^2 = \text{output electrical energy/input mechanical energy.} \quad (9)$$

The coupling coefficient represents the efficiency of the piezoelectric material in converting electrical energy into mechanical energy and vice versa. Energy conversion is never complete, so the coupling coefficient is always less than unity.

MEASUREMENTS OF PIEZOELECTRIC AND ELECTROSTRICTIVE EFFECTS

Different means have been developed to characterize the piezoelectric and electrostrictive properties of ceramic materials. The resonance technique involves measuring characteristic resonance frequencies when a suitably shaped specimen is driven by a sinusoidal electric field. To a first approximation, the behavior of a poled ceramic sample close to its fundamental resonance frequency can be represented by an equivalent circuit, as shown in Fig. 3a. The schematic behavior of the reactance of the sample as a function of frequency is shown in Fig. 3b. The equations used to calculate the electromechanical properties are described in the IEEE Standard on piezoelectricity (6). The simplest example of a piezoelectric measurement by the resonance technique can be shown by using a ceramic rod (typically 6 mm in diameter and 15 mm long) poled along its length. The longitudinal coupling coefficient (k_{33}) for this configuration is expressed as a function of the fundamental series and parallel resonance frequencies f_s and f_p ,

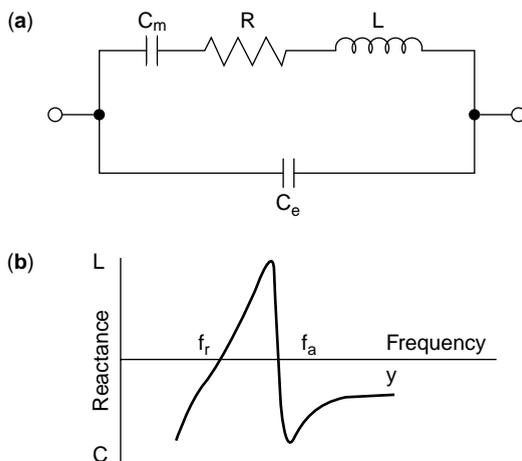


Figure 3. (a) Equivalent circuit of the piezoelectric sample near its fundamental electromechanical resonance (top branch represents the mechanical part and bottom branch represents the electrical part of the circuit); (b) electrical reactance of the sample as a function of frequency.

respectively:

$$k_{33} = (\pi/2)(f_s/f_p) \tan[(\pi/2)(f_p - f_s)/2]. \quad (10)$$

Then, the longitudinal piezoelectric coefficient d_{33} is calculated using k_{33} , the elastic compliance s_{33} , and the low-frequency relative permittivity ϵ_{33} :

$$d_{33} = k_{33}(\epsilon_{33}s_{33})^{1/2}. \quad (11)$$

Similarly, other coupling coefficients and piezoelectric moduli can be derived using different vibration modes of the same ceramic sample. The disadvantage of the resonance technique is that measurements are limited to specific frequencies determined by the electromechanical resonance. Resonance measurements are difficult for electrostrictive samples due to the required application of a strong dc bias field to induce a piezoelectric effect in relaxor ferroelectrics (see next section of the article).

Subresonance techniques are often used to evaluate the piezoelectric properties of ceramic materials at frequencies much lower than their fundamental resonance frequencies. These include the measurement of piezoelectric charge upon the application of a mechanical force (direct piezoelectric effect) and the measurement of electric-field-induced displacement (converse piezoelectric effect) when an electric field is introduced. It has been shown that piezoelectric coefficients obtained by direct and converse piezoelectric effects are thermodynamically equivalent.

The electrostrictive properties of ceramics are easily determined by measuring displacement as a function of the electric field or polarization. Thus the M and Q electrostrictive coefficients can be evaluated according to Eqs. (1) and (4), respectively. As an alternative, Eqs. (3b) and (6) can also be used for electrostriction measurements.

A direct technique is widely used to evaluate the sensor capabilities of piezoelectric and electrostrictive materials at sufficiently low frequencies. Mechanical deformations can be applied in different directions to obtain different components of the piezoelectric and electrostrictive tensors. In the simplest case, metal electrodes are placed on the major surfaces of a piezoelectric sample normal to its poling direction (Fig. 1b). Thus, the charge produced on the electrodes with respect to the mechanical load is proportional to the longitudinal piezoelectric coefficient d_{33} and the force F exerted on the ceramic sample: $Q = d_{33}F$. The charge can be measured by a charge amplifier using an etalon capacitor in the feedback loop. Details of direct piezoelectric measurements can be found in a number of textbooks (7).

Electric-field-induced displacements can be measured by a number of techniques, including strain gauges, linear variable differential transformers (LVDT), the capacitance method, fiber-optic sensors, and laser interferometry. Metal wire strain gauges are the most popular sensors used to measure strain at a resolution of about 10^{-6} m. To perform the measurement, the strain gauge is glued to the ceramic sample, and the resistance of the gauge changes according to its deformation. The resistance variation is measured by a precise potentiometer up to a frequency of several MHz. However, several gauges need to be used to obtain a complete set of piezoelectric and electrostrictive coefficients of the sample.

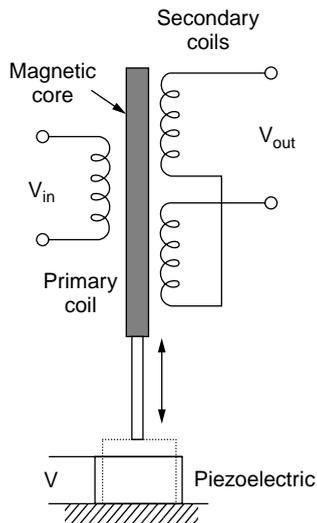


Figure 4. Principle of the linear variable differential transformer (LVDT) used for measuring electric-field-induced deformations in a piezoelectric sample.

Figure 4 illustrates the design of an LVDT. The moving surface of the sample is attached to the magnetic core inserted into the center of the primary and secondary electromagnetic coils. The change of the core position varies the mutual inductance of the coils. An ac current supplies the primary coil, and the signal in the secondary coils is proportional to the displacement of the core. The response speed depends on the frequency of the ac signal and the mechanical resonance of the coil, which typically does not exceed 100 Hz. Generally the resolution is sufficiently high and approaches $\sim 10^{-2}$ – 10^{-1} μm , depending on the number of turns.

The capacitive technique for strain measurements is based on the change of capacitance in a parallel-plate capacitor that has an air gap between two opposite plates. One of the plates is rigidly connected to the moving surface of the sample, and another plate is fixed by the holder. The capacitance change due to the vibration of sample can be measured precisely by a zero-point potentiometer and a lock-in amplifier. Therefore, high resolution (in the \AA range) can be achieved by this technique. The measurement frequency must be much lower than the frequency of the ac input signal, which typically does not exceed 100 Hz.

All of the aforementioned techniques require mechanical contact between the sample and the measurement unit. This, however, limits the resolution and the maximum operating frequency, which prevents accurate measurement of piezoelectric loss (the phase angle between the driving voltage and the displacement). The force exerted on the moving surface of the sample (especially on a thin ceramic film) may damage the sample. Therefore, noncontact measurements are often preferred to determine the electric-field-induced displacement of piezoelectric and electrostrictive materials accurately. Figure 5 shows the operating principle of a Photonic fiber-optic sensor, which can be used to examine the displacement of a flat reflecting

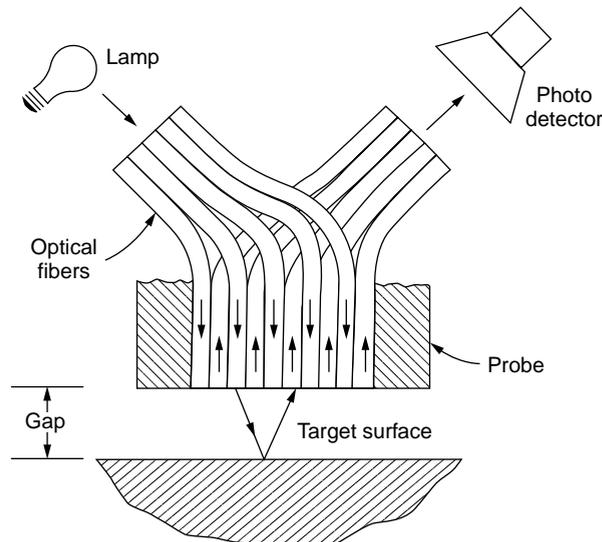


Figure 5. Schematic of the fiber-optic photonic sensor used for nondestructive evaluation of electric-field-induced strains.

surface (8). The sensor head consists of a group of transmitting and receiving optical fibers located in the immediate vicinity of the vibrating surface of sample. The intensity of the reflected light depends on the distance between the moving object and the probe tip. This dependence allows exact determination of displacement in both dc and ac modes. Using a lock-in amplifier to magnify the output signal, which is proportional to the light intensity, a resolution of the order of 1 \AA can be achieved (8). The frequency response is determined by the frequency band of the photodiode and the amplifier (typically of the order of several hundreds of kHz).

Optical interferometry is another technique that allows noncontact accurate measurement of the electric-field-induced displacements. Interferometric methods of measuring small displacements include the homodyne (9), heterodyne (10), and Fabri-Perot (11) techniques. The most common technique is the homodyne interferometer that uses active stabilization of the working point to prevent drift from thermal expansion. When two laser beams of the same wavelength (λ) interfere, the light intensity varies periodically ($\lambda/2$ period) corresponding to the change of optical path length between the two beams. If one of the beams is reflected from the surface of a moving object, the intensity of the output light changes, which can later be translated to the amount of displacement. Using a simple Michelson interferometer (12), a very high resolution of $\sim 10^{-5}$ \AA is achievable. However, the measurements are limited to a narrow frequency range because the sample is attached to a rigid substrate and only the displacement of the front surface of the sample is monitored (12). As a result of this configuration, the errors arising from the bending effect of the sample can be very high, especially in ferroelectric thin films. In response to that, a double beam (Mach-Zehnder) interferometer is used to take into account the difference of the displacements of both major surfaces of the sample (13). The modified version of the double-beam

interferometer, specially adapted to measure thin films, offers resolution as high as 10^{-4} Å in the frequency range of $10\text{--}10^5$ Hz and long-term stability ($<1\%$) (14).

COMMON PIEZOELECTRIC AND ELECTROSTRICTIVE MATERIALS

Single Crystals

A number of single crystals (ferroelectric and nonferroelectric) have demonstrated piezoelectricity. However, nonferroelectric piezoelectric crystals exhibit piezoelectric coefficients much lower than those of ferroelectric crystals. The former are still extensively used in some applications in which either high temperature stability or low loss is required. The most important nonferroelectric piezoelectric crystal is quartz (SiO_2) which has small but very stable piezoelectric properties [e.g., $d_{11} = 2.3$ pC/N, x-cut (15)]. Ferroelectric LiNbO_3 and LiTaO_3 crystals that have high Curie temperatures (1210 and 660°C , respectively) are used mostly in surface acoustic wave (SAW) devices. Recent investigations (15) have shown that rhombohedral single crystals in the $\text{Pb}(\text{Zn}_{1/3}\text{Nb}_{2/3})\text{O}_3\text{--PbTiO}_3$ system have exceptionally large longitudinal piezoelectric ($d_{33} = 2500$ pm/V) and coupling ($k_{33} = 0.94$) coefficients. In addition, ultrahigh strain of 1.7% has been observed in these materials under high electric field. These single crystals are now being intensively investigated and show significant promise for future generations of smart materials.

Piezoelectric and Electrostrictive Ceramics

As indicated earlier, the randomness of the grains in as-prepared polycrystalline ferroelectric ceramics yields nonpiezoelectric centrosymmetric material. Thus “poling” the ceramic (Fig. 6) is required to induce piezoelectricity. Due to symmetry limitations, all of the domains in a ceramic can never be fully aligned along the poling axis. However, the end result is a ceramic whose net polarization along the poling axis has sufficiently high piezoelectric properties.

The largest class of piezoelectric ceramics is made up of mixed oxides that contain corner-sharing octahedra of O^{2-} ions. The most technologically important materials in this class are perovskites that have the general formula ABO_3 , where $\text{A} = \text{Na}, \text{K}, \text{Rb}, \text{Ca}, \text{Sr}, \text{Ba},$ or Pb , and $\text{B} = \text{Ti}, \text{Sn}, \text{Zr}, \text{Nb}, \text{Ta},$ or W . Some piezoelectric ceramics that have this structure are barium titanate (BaTiO_3), lead titanate (PbTiO_3), lead zirconate titanate

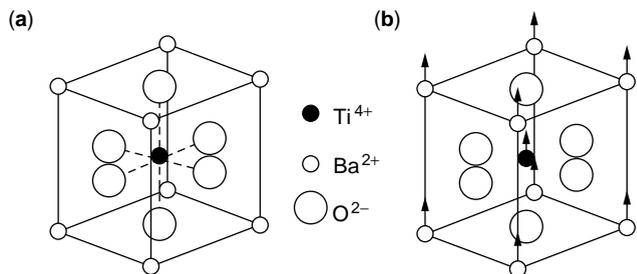


Figure 7. The crystal structure of BaTiO_3 : (a) above the Curie point, the cell is cubic; (b) below the Curie point, the cell is tetragonal, and Ba^{2+} and Ti^{4+} ions are displaced relative to O^{2-} ions.

($\text{PbZr}_x\text{Ti}_{1-x}\text{O}_3$, or PZT), lead lanthanum zirconate titanate ($\text{Pb}_{1-x}\text{La}_x(\text{Zr}_y\text{Ti}_{1-y})_{1-x/4}\text{O}_3$, or PLZT), and lead magnesium niobate ($\text{PbMg}_{1/3}\text{Nb}_{2/3}\text{O}_3$, or PMN).

The piezoelectric effect in BaTiO_3 was first discovered in the 1940s (3), and it became the first recognizable piezoelectric ceramic. The Curie point of BaTiO_3 is about $120\text{--}130^\circ\text{C}$. Above 130°C , a nonpiezoelectric cubic phase is stable, and the center of positive charges (Ba^{2+} and Ti^{4+}) coincides with the center of the negative charge (O^{2-}) (Fig. 7a). When cooled below the Curie point, a tetragonal structure (Fig. 7b) develops where the center of positive charges is displaced relative to the O^{2-} ions. This induces an electric dipole. The piezoelectric coefficients of BaTiO_3 are relatively high: $d_{15} = 270$ and $d_{33} = 190$ pC/N (3), and the coupling coefficient of BaTiO_3 is approximately 0.5. Due to its high dielectric constant, BaTiO_3 is widely used as a capacitor.

Lead titanate (PbTiO_3) first reported to be ferroelectric in 1950 (3), has a structure similar to BaTiO_3 but has a significantly higher Curie point ($T_c = 490^\circ\text{C}$). When cooled through the Curie temperature, the grains go through a cubic to tetragonal phase change that leads to a large strain which causes the ceramic to fracture. Thus, it is difficult to fabricate pure lead titanate in bulk form. This spontaneous strain has been decreased by adding dopants such as Ca, Sr, Ba, Sn, and W. Calcium-doped PbTiO_3 (16) has a relative permittivity of ~ 200 and a longitudinal piezoelectric coefficient (d_{33}) of 65 pC/N. Because of its high piezoelectric coefficient and low relative permittivity, the voltage piezoelectric coefficient of lead titanate ceramic is exceptionally high. Therefore, lead titanate is used in hydrophones and sonobuoys (17).

Lead zirconate titanate (PZT) is a binary solid solution of PbZrO_3 and PbTiO_3 (3). It is an ABO_3 perovskite structure in which Zr^{4+} and Ti^{4+} ions randomly occupy B sites. PZT has a temperature-independent morphotropic phase boundary (MPB) between tetragonal and rhombohedral phases, when the Zr:Ti ratio is 52:48. (Fig. 8). This composition of PZT has efficient poling and excellent piezoelectric properties because of its large number of polarization orientations. At the MPB composition, PZT is usually doped by a variety of ions to form what are known as “hard” and “soft” PZTs (3). Doping PZT with acceptor ions, such as K^+ or Na^+ at the A site, or Fe^{3+} , Al^{3+} , or Mn^{3+} at the B site, creates hard PZT. This doping reduces the piezoelectric properties and makes the PZT

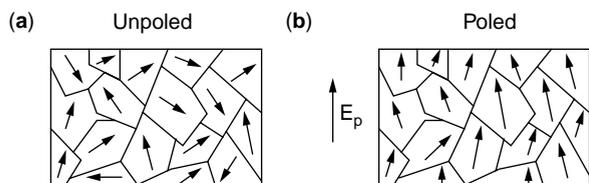


Figure 6. Schematic of the poling process in piezoelectric ceramics: (a) in the absence of an electric field, the domains have random orientation of polarization; (b) the polarization within the domains are aligned in the direction of the electric field.

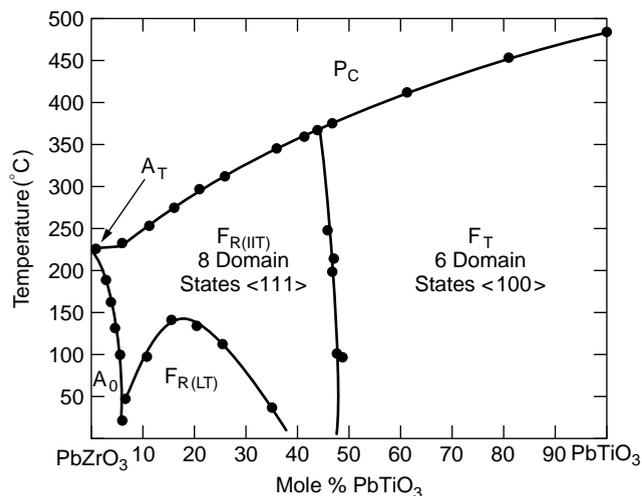


Figure 8. Phase diagram of lead zirconate titanate piezoelectric ceramics (PZT) as a function of mole% PbTiO_3 .

more resistant to poling and depoling. Introducing donor ions such as La^{3+} into the A site, or Nb^{5+} or Sb^{5+} into the B site, makes soft PZT. This doping increases the piezoelectric properties and makes the PZT easier to pole and depole. Table 1 compares the piezoelectric properties of several major piezoelectric ceramics.

Lead magnesium niobate, $\text{PbMg}_{1/3}\text{Nb}_{2/3}\text{O}_3$ (PMN), is a perovskite ceramic known as a relaxor ferroelectric. Unlike normal ferroelectrics, which have well-defined Curie points in their weak-field relative permittivity, relaxor ferroelectrics exhibit a broad transition peak between ferroelectric and paraelectric phases (18). This kind of transition is often referred to as a diffuse phase transition. The distinctive features of relaxor ferroelectrics are their strong frequency dispersion of relative permittivity and a shift of their maximum relative permittivity with frequency. Local inhomogeneity of B site ions (e.g., Mg^{2+} and Nb^{5+}) in the perovskite lattice are the proposed cause of relaxor properties. Relaxors do not possess piezoelectricity without a dc bias field to break the paraelectric cubic phase into the rhombohedral ferroelectric piezoelectric phase. Relaxors have been used as actuators because of their negligible hysteresis and large induced polarization (electrostrictive strain of the order of 10^{-3}). Figure 9 compares the electric-field-induced strains of typical piezoelectric (PZT) and electrostrictive (PMN) ceramics.

Processing of Piezoelectric Ceramics

The electromechanical properties of piezoelectric ceramics are largely influenced by their processing conditions.

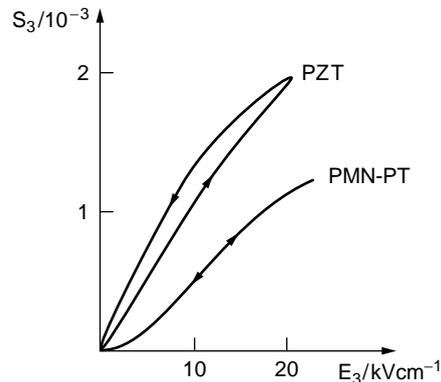


Figure 9. Comparison of the electric-field-induced strain in a typical piezoelectric (PZT) and relaxor (0.9PMN–0.1PT).

Each step of the process must be carefully controlled to yield the best product. Figure 10 is a flowchart of a typical oxide manufacturing process for piezoelectric ceramics. First, high purity raw materials are accurately weighed according to their desired ratio and then are mechanically or chemically mixed. During the calcination step, the solid phases react to form the piezoelectric phase. After calcination, the solid mixture is milled to fine particles. Shaping is accomplished by a variety of ceramic processing techniques, including powder compaction, tape casting, slip casting, and extrusion. During the shaping operation, organic materials are typically added to the ceramic powder to improve its flow and binding characteristics. The organic is then removed in a low-temperature (500–600°C) burnout step. After organic removal, the ceramic structure is fired to an optimum density at an elevated temperature. Lead-containing ceramics (PbTiO_3 , PZT, PMN) are fired in sealed crucibles in an optimized PbO atmosphere to prevent lead loss above 800°C.

PIEZOELECTRIC COMPOSITES

Single-phase piezoelectric/electrostrictive materials are not ideally suited for hydrostatic and ultrasonic applications where ceramic elements radiate and receive acoustic waves. Although d_{33} and d_{31} piezoelectric coefficients are exceptionally high in PZT ceramics, their hydrostatic voltage response is relatively low due to the high dielectric constant and low hydrostatic charge coefficient $d_h = d_{33} + 2d_{31}$. Because $d_{31} \approx -0.4d_{33}$ in PZT ceramics (3), their hydrostatic sensor capabilities are rather low. In

Table 1. Piezoelectric Properties of Major Piezoelectric Ceramics

	Quartz	BaTiO_3	PZT-4	PZT-5	$\text{PbTiO}_3\text{:Sm}$
d_{33} (pC/N)	2.3(x-cut)	190	289	593	65
g_{33} (10^{-3}Vm/N)	58	12.6	26.1	19.7	42
k_t	0.09	0.38	0.51	0.50	0.30
k_p		0.33	0.58	0.65	0.03
ϵ_{33}	5	1700	1300	3400	175
T_c (°C)		130	328	193	355

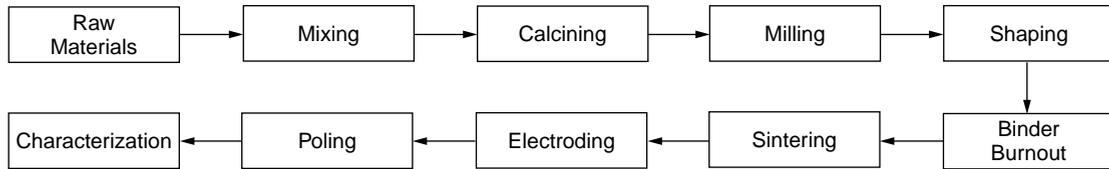


Figure 10. Flowchart for processing piezoelectric ceramics.

addition, the high density of ceramics results in a high acoustic impedance mismatch between the transducer and the medium in which the acoustic waves are propagating. On the other hand, piezoelectric polymers have low density (low impedance), dielectric constant, and piezoelectric coefficients.

In the past three decades, researchers have focused on methods for combining the best characteristics of ceramics and polymers to overcome the aforementioned deficiencies. Integration of a piezoelectric ceramic with a polymer allows tailoring the piezoelectric properties of composites. The mechanical and electrical properties of a composite depend strongly on the characteristics of each phase and the manner in which they are connected. In a diphasic composite, the materials can be oriented in ten different ways in a three-dimensional space (19). The possible connectivity patterns are 0-0, 1-0, 2-0, 3-0, 1-1, 2-1, 3-1, 2-2, 3-2, and 3-3. As a matter of convention, the first and second numbers in the connectivity denote the continuity of the piezoelectric and polymer phases, respectively. Figure 11 shows some of the composites made in the past 30 years (20). The most important connectivity patterns are 0-3, 1-3, 3-3, and 2-2. The 0-3 composites are made of a homogeneous distribution of piezoelectric ceramic particles within a polymer matrix. The primary advantage of these composites is that they can be formed into shapes and still retain their piezoelectricity. However, they cannot be sufficiently poled because the ceramic phase is not self-connected in the poling direction. On the other hand, 3-0

composites that are simply the ceramic matrix containing a low concentration of polymer inclusions or voids can be effectively poled and exhibit hydrostatic properties superior to those of single-phase PZT (20).

In composites of 3-3 connectivity, the piezoceramic and polymer phases are continuous in three dimensions and form two interlocking skeletons. The first composite of 3-3 connectivity was formed by the replamine process using a coral skeleton (21). Another effective method of making 3-3 composites is called BURPS (acronym for burned out plastic spheres) (22) which provides properties similar to the replamine composites. In this process, a mixture of PZT powder and burnable plastic spheres is used to fabricate the PZT/polymer composites. Other techniques, such as relic processing (23) and distorted reticulated ceramics (24) have been developed to fabricate 3-3 composites. Recently, fused deposition modeling (FDM) and fused deposition of ceramics (FDC) have been used to make ladder and 3-D honeycomb composites (25). In the FDM technique, a 3-D plastic mold is prepared and filled with PZT slurry. The FDC process deposits a mixture of PZT and polymer directly in the form of a three-dimensional ladder structure. Either structure is heat treated to burn the organic, sintered, and embedded in epoxy polymer.

The composites most extensively studied and used in transducer applications are those that have 1-3 connectivities. They consist of individual PZT rods or fibers aligned in the direction parallel to poling and embedded in a polymer matrix. The rod diameter, spacing between them,

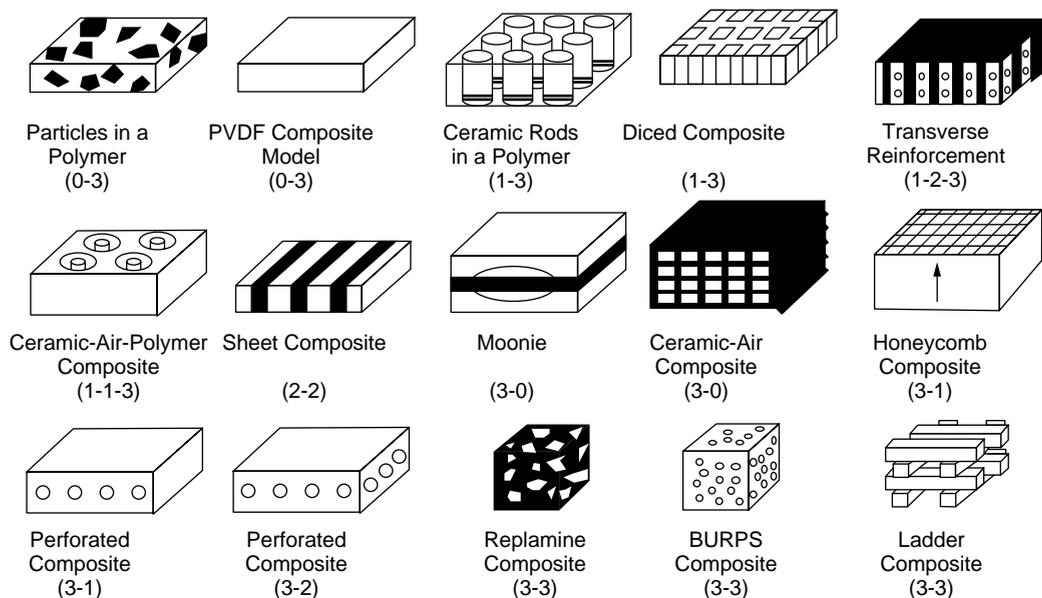


Figure 11. Schematic of various piezoelectric composites of different connectivities.

composite thickness, volume% of PZT, and polymer compliance influence the composite's performance. The most common methods of forming 1–3 composites are the dice and fill technique (26) and injection molding (27). In the former method, the composite is fabricated by dicing deep grooves in perpendicular directions into a solid sintered block of poled PZT. The grooves are backfilled with polymer, and the base is removed via grinding or cutting. In the latter method, a thermoplastic mixture of ceramic powder and organic binder is injected into a cooled mold. The process can be used to form composites that have a variety of rod sizes, shapes, and spacings. This technique has recently been employed by Materials Systems, Inc. to mass produce SmartPanels™ (28).

APPLICATIONS OF PIEZOELECTRIC/ELECTROSTRICTIVE CERAMICS

By directly coupling mechanical and electrical quantities, piezoelectrics and electrostrictives have been extensively used in a variety of electromechanical devices for both sensor and actuator applications. The direct piezoelectric effect is currently being used to generate charge (voltage) in applications such as gas igniters, acoustic pressure sensors, vibration sensors, accelerometers, and hydrophones (29). The best known examples of actuators, which take advantage of the converse effect, are piezoelectric motors, piezoelectrically driven relays, ink-jet heads for printers, noise cancellation systems, VCR head trackers, precise positioners, and deformable mirrors for correcting of optical images (30). Acoustic and ultrasonic vibrations can be generated by piezoelectrics using an ac field at resonance conditions and/or detected by a piezoelectric receiver. Very often, an acoustic sender and receiver are combined in the same piezoelectric devices. Transducers have a variety of applications, including imaging, nondestructive testing, and fish finders (31). At high frequencies, piezoelectric transducers also function as frequency control devices, bulk and surface acoustic wave (SAW) resonators, filters, and delay lines.

Ultrasonic transducers operate in a so-called pulse-echo mode, where a transducer sends an acoustic wave that is reflected from the interfaces and is received by the very same transducer. These echoes vary in intensity according to the type of interface, which may include tissue and bone. Therefore, the ultrasonic image that is created clearly represents the mechanical properties of human tissue. Thus, anatomic structures of different organs can be recognized in real time. A sensitive ultrasonic transducer that generates low-intensity acoustic waves can be one of the safest diagnostic devices for medical imaging. These transducers are usually composed of matching and backing layers and the piezoelectric material itself. The matching layers are added to the transducer to reduce the acoustic impedance mismatch between the imaged object and the transducer, and the backing layers dampen the acoustic backwaves. Composite materials instead of single phase materials are frequently used to increase the performance of transducers (20).

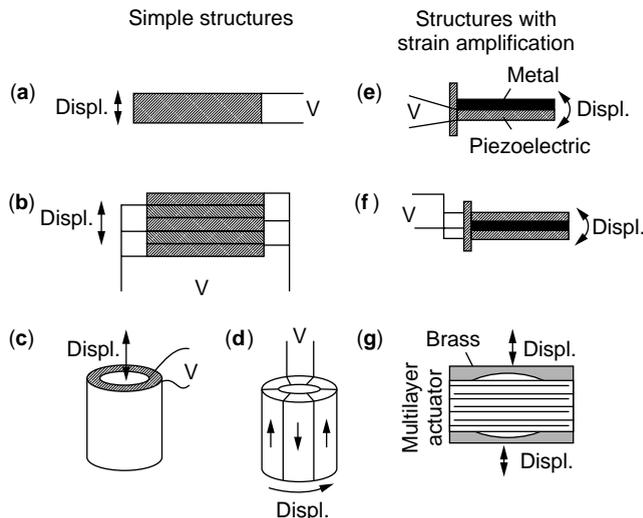


Figure 12. Typical actuator designs: simple structures [(a)–(d)] and structures with strain amplification [(e)–(g)].

When a transducer function is to displace an object, it is called an actuator. It is desirable for an actuator to generate a significant displacement and/or generative force under a moderate electric field. In addition, actuators must have reproducible displacements when precise positioning is important. Thus, electrostrictive materials such as PMN or its solid solution with PT are preferred over PZT materials due to their small hysteresis. Figure 12 shows several possible designs of piezoelectric/electrostrictive actuators. In simple structures, like those shown in Fig. 12a–d, the actuator displacements are solely due to d_{33} , d_{31} , or d_{15} effects of the ceramic rod, plate, or tube. Because strain is limited to 10^{-3} , the typical displacement of a 1-cm long actuator is $\sim 10 \mu\text{m}$. Multilayer actuators (Fig. 12b) use a parallel connection of ceramic plates cemented together. In this case, the displacements of many individual sheets of a piezoelectric ceramic are summed. The advantage of multilayer actuators is their small operating voltage, fast speed, and large generative force. A useful design is the piezoelectric tube (Fig. 12c) which is poled and driven by the voltage applied in a radial direction (through the wall width). The axial response is due to the d_{31} coefficient of the material and is proportional to the length/width ratio. The radial response can be tuned to almost zero by manipulating the geometry of the tube (32). This configuration is beneficial in suppressing unwanted lateral displacements. Another important design is a shear actuator (Fig. 12d) which directly transforms the voltage applied normal to the polarization vector into a pure rotation due to the d_{15} coefficient (33).

As previously indicated, all of the simple structures are based on pure d_{31} , d_{33} , or d_{15} actions, so that displacements are limited to tens of microns. The amplification of strain at the expense of generative force can be achieved by using monomorph and bimorph structures (Fig. 12e–f). These types of actuators produce large displacements (up to several mm) but have low generative force and slow response. Another type of strain amplification can be achieved by flexensional transducers. One of the designs,

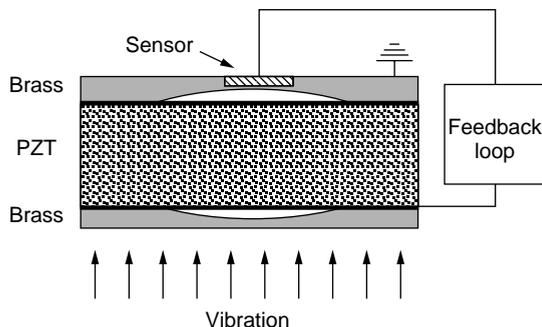


Figure 13. Example of the smart system using a PZT sensor incorporated in the MOONE actuator (36).

called MOONE, is shown in Fig. 12g (34). This type of actuator uses the bending effect of the moon-shaped metallic cap attached to both sides of a multilayer actuator. The d_{31} motion of the actuator is amplified by bending the metallic cap. Other examples of flextensional actuators are RAINBOWs and CYMBALs (not shown in the figure) (35, 36). Flextensional actuators have characteristics intermediate between multilayers and bimorphs and are now extensively used in various actuator applications. An example of a smart structure using flextensional actuator (MOONE) is shown in Fig. 13. The actuator portion of the device consists of the standard MOONE and a small piezoelectric ceramic embedded in the upper cap that serves as a sensor. The sensor detects vibrations normal to the actuator surface and, via a feedback loop, sends a signal of appropriate amplitude and phase to the actuator, so that it effectively cancels the external vibration. Potential applications of the smart structure shown in Fig. 13 include active optical systems, rotor suspension systems, and other noise cancellation devices.

Recent trends toward miniaturization have resulted in extensive use of piezoelectric/electrostrictive materials in microelectromechanical systems (MEMS). Because miniaturization of bulk ceramics is limited, these materials are used in a thin/thick film form. Thin film actuators based on the piezoelectric effect in PZT materials have been demonstrated. They include micromotors (37), acoustic imaging devices (38), components for atomic force microscopes (AFM) (39), and micropumps (40). Figure 14 shows the design of an atomic force microscope using PZT film for both sensing and actuating functions. The excitation ac voltage signal superimposed on the actuation dc voltage is applied to the PZT film deposited on the Si cantilever. The vibrational amplitude, which is sensitive to the atomic force between the tip and investigated surface, is detected by measuring the difference between the cantilever current and the reference current. The feedback system maintains a constant current while scanning in the x, y plane. This system does not require optical registration of the vibration that makes PZT-based AFM compact and it allows the multiprobe systems to be achieved. Because PZT film is very sensitive to vibrations, the vertical resolution of such an AFM approaches that of conventional systems. This electromechanically driven AFM is an excellent example of using piezoelectric ceramic thin films as smart materials.

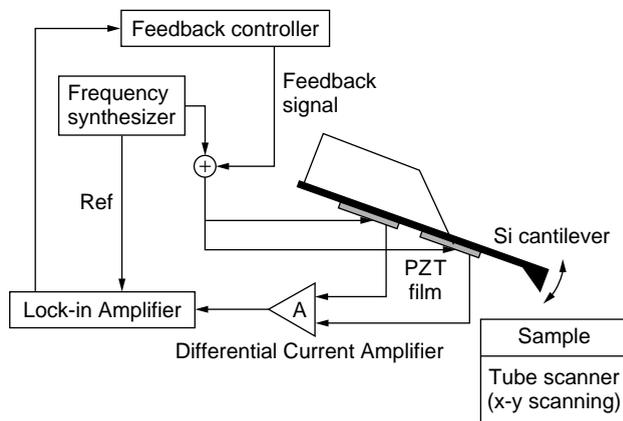


Figure 14. Schematic of the AFM cantilever sensor and actuator based on a PZT thin film.

FUTURE TRENDS

Most piezoelectric/electrostrictive ceramics currently rely on lead oxide based materials due to their excellent properties as sensors and actuators. However, due to increased public awareness of health problems associated with lead and environmental protection policies, future research will be focused on finding lead-free compounds that have piezoelectric properties similar to those of PZT. Relaxor single-crystal materials that have a giant piezoelectric effect will probably find a wide range of applications from composite transducers for medical imaging to microelectromechanical systems. The current trend of miniaturization will continue to give rise to complex sensors and actuators integrated directly on a silicon chip. Further, batch processing will effectively reduce the cost of such devices. In addition, the research will continue toward the development of more resilient piezoelectric/electrostrictive materials used for operation under severe external conditions (temperature, pressure, harsh chemical environments). This will further improve their potential application in space and deep ocean exploration, as well as in noise cancellation in airplanes and helicopters.

BIBLIOGRAPHY

1. R.E. Newnham and G.R. Ruschau, *J. Intelli. Materi., Syst. Struct.* **4**, 289 (1993).
2. L. Ristic, ed., *Sensor Technology and Devices*, Chapter 1. Artech House, Boston, MA, 1994.
3. B. Jaffe, W.R. Cook, Jr., and H. Jaffe, *Piezoelectric Ceramics*, Marietta, R.A.N., OH, 1971.
4. J.F. Nye, *Physical Properties of Crystals*, Oxford University Press, Oxford, UK, 1985.
5. M.J. Haun, E. Furman, S.J. Jang, and L.E. Cross, *IEEE Trans. Ultrason., Ferroelectr. Freq. Control* **36**(4), 393 (1989).
6. IEEE Standards on Piezoelectricity, *IEEE Stand.* **176** (1978).
7. V.F. Janas, A. Safari, A. Bandyopadhyay, and A. Kholkin, in *Measurements, Instrumentation and Sensor Handbook* (J.G. Webster, ed.). CRC Press, Boca Raton, FL.

8. *MTI-2000 FotonicTM Sensor*, Instruction Manual, MTI Instruments Division, Latham, New York.
9. S. Sizgoric and A.A. Gundjian, *Proc. IEEE* **57**, 1312 (1969).
10. R.M. De La Rue, R.F. Humphreys, I.M. Masom, and E.A. Ash, *Proc. IEEE* **119**, 117 (1972).
11. V.E. Bottom, *Rev. Sci. Instrum.* **35**, 374 (1964).
12. Q.M. Zhang, W.Y. Pan, and L.E. Cross, *J. Appl. Phys.* **65**, 2807 (1988).
13. W.Y. Pan and L.E. Cross, *Rev. Sci. Instrum.* **60**, 2701 (1989).
14. A.L. Kholkin, C. Wuethrich, D.V. Taylor, and N. Setter, *Rev. Sci. Instrum.* **67**, 1935 (1996).
15. S.-E. Park and T.R. Shrout, *J. Appl. Phys.* **82**, 1804 (1997).
16. *Piezoelectric Products*, Sensor Technology Limited, Collingwood, Ontario, Canada, 1991.
17. K.M. Rittenmyer and R.Y. Ting, *Ferroelectrics* **110**, 171 (1990).
18. L.E. Cross, in *Ferroelectric Ceramics: Tutorials, Reviews, Theory, Processing and Applications*, (N. Setter and E. Colla, eds.). Birkhaeuser, Basel, 1993.
19. R.E. Newnham, D.P. Skinner, and L.E. Cross, *Mater. Res. Soc. Bull.* **13**, 525 (1978).
20. T.R. Gururaja, A. Safari, R.E. Newnham, and L.E. Cross, in *Electronic Ceramics* (L.M. Levinson, ed.), p. 92. Dekker, New York.
21. D.P. Skinner, R.E. Newnham, and L.E. Cross, *Mater. Res. Soc. Bull.* **13**, 599 (1978).
22. T.R. Shrout, W.A. Schulze, and J.V. Biggers, *Mater. Res. Soc. Bull.* **14**, 1553 (1979).
23. S.M. Ting, V.F. Janas, and A. Safari, *J. Am. Ceram. Soc.* **79**, 1689 (1996).
24. M.J. Creedon and W.A. Schulze, *Ferroelectrics* **153**, 333 (1994).
25. A. Bandyopadhyay, R.K. Panda, V.F. Janas, M.K. Agarwala, S.C. Danforth, and A. Safari, *J. Am. Ceram. Soc.* **80**, 1366 (1997).
26. H.P. Savakus, K.A. Klicker, and R.E. Newnham, *Mater. Res. Soc. Bull.* **16**, 677 (1981).
27. R.L. Gentilman, D.F. Fiore, H.T. Pham, K.W. French, and L.W. Bowen, *Ceram. Transa.* **43**, 239 (1994).
28. D. Fiore, R. Torri, and R. Gentilman, *Proc. 8th U.S.-Jpn. Semin. Dielectr. Piezoelectr. Ceram.*, Plymouth, MA, 1997.
29. A.J. Moulson and J.M. Herbert, *Electroceramics: Materials, Properties, Applications*, Chapman & Hall, London, 1990.
30. K. Uchino, *Piezoelectric Actuators and Ultrasonic Motors*, Kluwer, Boston, MA, 1997.
31. J.M. Herbert, *Ferroelectric Transducers and Sensors*, Gordon & Breach, New York, 1982.
32. Q.M. Zhang, H. Wang, and L.E. Cross, *Proc. SPIE, Smart Struct. Mater.* **1916**, 244 (1993).
33. A.E. Glazounov, Q.M. Zhang, and C. Kim, *Appl. Phys. Lett.* **72**, 2526 (1998).
34. A. Dogan, Q.C. Xu, K. Onitsuka, S. Yoshikawa, K. Uchino, and R.E. Newnham, *Ferroelectrics* **156**, 1 (1994).
35. G. Haertling, *Bull. Am. Ceram. Soc.* **73**, 93 (1994).
36. A. Dogan, K. Uchino, and R.E. Newnham, *IEEE Trans. Ultrason., Ferroelectr. Freq. Control* **44**, 597 (1997).
37. P. Muralt, M. Kohli, T. Maeder, A. Kholkin, K. Brooks, N. Setter, and R. Luthier, *Sensors Actuators* **A48**, 157 (1995).
38. R.E. Newnham, J.F. Fernandez, K.A. Markowski, J.T. Fielding, A. Dogan, and J. Wallis, *Mater. Res. Soc. Proc.* **360**, 33 (1995).
39. J. Bernstein, K. Houston, L. Niles, K. Li, H. Chen, L.E. Cross, and K. Udayakumar, *Proc. 8th Int. Symp. Integr. Ferroelectr.*, Tempe, AZ, 1996.
40. T. Fuji and S. Watanabe, *Appl. Phys. Lett.* **68**, 487 (1996).

CERAMICS, TRANSDUCERS

KENJI UCHINO

YUKIO ITO

The Pennsylvania State University

University Park, PA

INTRODUCTION

Smart Material

Let us start with the “smartness” of a material. Table 1 lists the various effects that relate the input—electric field, magnetic field, stress, heat and light—to the output—charge/current, magnetization, strain, temperature, and light. Conducting and elastic materials that generate current and strain outputs, respectively, from input voltage, or stress (well-known phenomena!) are sometimes called “trivial” materials. Conversely, pyroelectric and piezoelectric materials that generate an electric field from the input of heat or stress (unexpected phenomena!) are called “smart” materials. These off-diagonal couplings have a corresponding converse effect such as electrocaloric and converse piezoelectric effects, and both “sensing” and “actuating” functions can be realized in the same materials. “Intelligent” materials should possess a “drive/control” or “processing” function which is adaptive to the change in environmental conditions, in addition to the actuation and sensing functions. Note that ferroelectric materials exhibit most of these effects, except magnetic-related phenomena. Thus, the ferroelectrics are said to be very “smart” materials.

The “actuator” in a narrow meaning stands for materials or devices that generate mechanical strain (or stress) output. As indicated by the thick columnar border in Table 1, solid state actuators use converse piezoelectric, magnetostriction, elasticity, thermal expansion, or photostriction phenomena. A shape-memory alloy is a kind of thermally expanding material. On the other hand, a “sensor” requires charge/current output in most cases. Thus, conducting/semiconducting, magnetoelectric, piezoelectric, pyroelectric, and photovoltaic materials are used for detecting electric fields, magnetic fields, stress, heat, and light, respectively (see the thin columnar border in Table 1).

In this sense, piezoelectric materials are most popularly used in smart structures and systems because the same material is applicable to both sensors and actuators, in principle. We treat mainly piezoelectric transducers, sensors, and actuators in this article. Even though transducers, in general, are devices that convert input energy to a different energy type of output, the piezoelectric

Table 1. Various Effects in Ferroelectric and Ferromagnetic Materials^a

		Input \rightarrow Material Device \rightarrow Output			
Output Input	Charge Current	Magnetization	Strain	Temperature	Light
Elec. eld	Permittivity Conductivity	Elect.-mag effect	Converse piezo-effect	Elec. caloric effect	Elec.-optic effect
Mag. eld	Mag.-elect. effect	Permeability	Magnetostriction	Mag.caloric effect	Mag.optic effect
Stress	Piezoelectric effect	Piezomag. effect	Elastic constant		Photoelastic effect
Heat	Pyroelectric effect		Thermal expansion	Speci c heat	
Light	Photovoltaic effect		Photostriction		Refractive index

^a Off-diagonal coupling = **Smart Material** Sensor Actuator

“transducer” is often used to denote a device that possesses both sensing and actuating functions, exemplified by underwater sonar.

Piezoelectric Effect

Certain materials produce electric charges on their surfaces as a consequence of applying mechanical stress. When the induced charge is proportional to the mechanical stress, it is called a direct piezoelectric effect and was discovered by J. and P. Curie in 1880. Materials that show this phenomenon also conversely have a geometric strain generated that is proportional to an applied electric field. This is the converse piezoelectric effect. The root of the word “piezo” is the Greek word for “pressure”; hence the original meaning of the word piezoelectricity implied “pressure electricity” (1,2).

Piezoelectric materials couple electrical and mechanical parameters. The material used earliest for its piezoelectric properties was single-crystal quartz. Quartz crystal resonators for frequency control appear today at the heart of clocks and are also used in TVs and computers. Ferroelectric polycrystalline ceramics, such as barium titanate and lead zirconate titanate, exhibit piezoelectricity when electrically poled. Because these ceramics possess significant and stable piezoelectric effects, that is, high electromechanical coupling, they can produce large strains/forces and hence are extensively used as transducers. Piezoelectric polymers, notably polyvinylidene difluoride and its copolymers with trifluoroethylene and piezoelectric composites that combine a piezoelectric ceramic and a passive polymer have been developed and offer high potential. Recently, thin films of piezoelectric materials are being researched due to their potential use in microdevices (sensors and microelectromechanical systems). Piezoelectricity is being extensively used in fabricating various devices such as transducers, sensors, actuators, surface acoustic wave devices, and frequency controls.

We describe the fundamentals of piezoelectric effect first, then present a brief history of piezoelectricity, followed by present day piezoelectric materials that are used, and finally various potential applications of piezoelectric materials are presented.

PIEZOELECTRICITY

Relationship Between Crystal Symmetry and Properties

All crystals can be classified into 32 point groups according to their crystallographic symmetry. These point groups are divided into two classes; one has a center of symmetry, and the other lacks it. There are 21 noncentrosymmetric point groups. Crystals that belong to 20 of these point groups exhibit piezoelectricity. Although cubic class 432 lacks a center of symmetry, it does not permit piezoelectricity. Of these 20 point groups, 10 polar crystal classes contain a unique axis, along which an electric dipole moment is oriented in the unstrained condition.

The pyroelectric effect appears in any material that possesses a polar symmetry axis. The material in this category develops an electric charge on the surface owing to the change in dipole moment as temperature changes. The pyroelectric crystals whose spontaneous polarization are reorientable by applying an electric field of sufficient magnitude (not exceeding the breakdown limit of the crystal) are called ferroelectrics (3,4). Table 2 shows the crystallographic classification of the point groups.

Piezoelectric Coefficients

Materials are deformed by stresses, and the resulting deformations are represented by strains ($\Delta L/L$). When the stress X (force per unit area) causes a proportional strain x ,

$$x = sX, \tag{1}$$

where all quantities are tensors, x and X are second rank, and s is fourth rank. Piezoelectricity creates additional

Table 2. Crystallographic Classification According to Crystal Centrosymmetry and Polarity^a

Polarity	Symmetry	Crystal system										
		Cubic		Hexagonal		Tetragonal		Rhombohedral		Orthorhombic	Monoclinic	Triclinic
Non-polar (22)	Centro (11)	<i>m3m</i>	<i>m3</i>	<i>6/mmm</i>	<i>6/m</i>	<i>4/mmm</i>	<i>4/m</i>	<i>3m</i>	<i>3</i>	<i>mmm</i>	<i>2/m</i>	
	Non-centro (21)	432	23	622	6	422	4	32		222		
		43m		6m2		42m						
Polar (pyroelectric) (10)				6mm	6	4mm	4	3m	3	mm2	2 <i>m</i>	1

^a Piezoelectrics are inside the boldline.

strains by an applied electric field E . The piezoelectric equation is given by

$$x_{ij} = s_{ijkl}X_{kl} + d_{ijk}E_k, \quad (2)$$

where E is the electric field and d is the piezoelectric constant which is a third-rank tensor. This equation can be also expressed in a matrix form such as for a poled ceramic:

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{pmatrix} = \begin{pmatrix} 0 & 0 & d_{31} \\ 0 & 0 & d_{31} \\ 0 & 0 & d_{33} \\ 0 & d_{15} & 0 \\ d_{15} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} E_1 \\ E_2 \\ E_3 \end{pmatrix} \quad (3)$$

Another frequently used piezoelectric constant is g which gives the electric field produced when a stress is applied ($E = gX$). The g constant is related to the d constant through the relative permittivity ε :

$$g = d/\varepsilon\varepsilon_0, \quad (4)$$

where ε_0 is the permittivity of free space. A measure of the effectiveness of the electromechanical energy conversion is the electromechanical coupling factor k , which measures the fraction of electrical energy converted to mechanical energy when an electric field is applied or vice versa when a material is stressed (5):

$$k^2 = (\text{stored mechanical energy}/\text{input electrical energy}), \quad (5)$$

or

$$= (\text{stored electrical energy}/\text{input mechanical energy}). \quad (6)$$

Let us calculate k^2 in Eq.(5), when an electric field E is applied to a piezoelectric material. Because the input electrical energy is $(1/2)\varepsilon_0\varepsilon E^2$ per unit volume and the stored mechanical energy per unit volume under zero external stress is given by $(1/2)x^2/s = (1/2)(dE)^2/s$, k^2 can be calculated as

$$\begin{aligned} k^2 &= [(1/2)(dE)^2/s]/[(1/2)\varepsilon_0\varepsilon E^2] \\ &= d^2/\varepsilon_0\varepsilon s. \end{aligned} \quad (7)$$

Note that k is always less than one. Typical values of k are 0.10 for quartz, 0.4 for BaTiO₃ ceramic, 0.5–0.7 for PZT ceramic and 0.1–0.3 for PVDF polymer. Another important material parameter is the mechanical quality factor Q_m that determines frequency characteristics. Q_m is given by an inverse value of mechanical loss:

$$Q_m = 1/\tan \delta. \quad (8)$$

A low Q_m is preferred for sensor applications in general to cover a wide frequency range. Conversely, a high Q_m is most desired for ultrasonic actuators (e.g., ultrasonic motors) to suppress heat generation through the loss.

History of Piezoelectricity

As already stated, Pierre and Jacques Curie discovered piezoelectricity in quartz in 1880. The discovery of ferroelectricity accelerated the creation of useful piezoelectric materials. Rochelle salt was the first ferroelectric discovered in 1921. Until 1940 only two types of ferroelectrics were known, Rochelle salt and potassium dihydrogen phosphate and its isomorph. In 1940 to 1943, unusual dielectric properties such as an abnormally high dielectric constant were found in barium titanate BaTiO₃ independently by Wainer and Salmon, Ogawa, and Wul and Golman. After the discovery, compositional modifications of BaTiO₃ led to improvement in temperature stability and high-voltage output. Piezoelectric transducers based on BaTiO₃ ceramics became well established in a number of devices.

In the 1950s, Jaffe and co-workers established that the lead zirconate–lead titanate system (called the PZT system) induces strong piezoelectric effects. The maximum piezoelectric response was found for PZT compositions near the morphotropic phase boundary between the rhombohedral and tetragonal phases. Since then, the PZT system containing various additives has become the dominant piezoelectric ceramic for a variety of applications. The development of PZT-based ternary solid solutions was a major success of the piezoelectric industry for these applications.

In 1969, Kawai et al. discovered that certain polymers, notably polyvinylidene difluoride (PVDF), are piezoelectric when stretched during fabrication. Such piezoelectric polymers are also useful for some transducer applications.

Table 3. Material Parameters of Representative Piezoelectric Materials

Parameter	Quartz	BaTiO ₃	PZT4	PZT 5H	(Pb,Sm)/TiO ₃	PVDF-TrFE
d_{33} (pC/N)	2.3	190	289	593	65	33
g_{33} (10 ⁻³ Vm/N)	57.8	12.6	26.1	19.7	42	380
k_t	0.09	0.38	0.51	0.50	0.50	0.30
k_p		0.33	0.58	0.65	0.03	
$\varepsilon_3 T/\varepsilon_0$	5	1700	1300	3400	175	6
Q_m	>10 ⁵		500	65	900	3-10
T_C (°C)		120	328	193	355	

In 1978, Newnham et al. improved composite piezoelectric materials by combining a piezoelectric ceramic and a passive polymer whose properties can be tailored to the requirements of various piezoelectric devices.

Another class of ceramic material has recently become important: relaxor-type electrostrictors such as lead magnesium niobate (PMN), typically doped with 10% lead titanate (PT), which have potential applications in the piezoelectric actuator field. A recent breakthrough in the growth of high-quality, large, single-crystal relaxor piezoelectric compositions has created interest in these materials for applications ranging from high strain actuators to high-frequency transducers for medical ultrasound devices due to their superior electromechanical characteristics. More recently, thin films of piezoelectric materials such as zinc oxide (ZnO) and PZT have been extensively investigated and developed for use in microelectromechanical (MEMS) devices.

PIEZOELECTRIC MATERIALS

This section summarizes the current status of piezoelectric materials: single-crystal materials, piezoceramics, piezopolymers, piezocomposites, and piezofilms. Table 3 shows the material parameters of some representative piezoelectric materials described here (6,7).

Single Crystals

Piezoelectric ceramics are widely used at present for a large number of applications. However, single-crystal materials retain their utility; they are essential for applications such as frequency stabilized oscillators and surface acoustic devices. The most popular single-crystal piezoelectric materials are quartz, lithium niobate (LiNbO₃), and lithium tantalate (LiTaO₃). The single crystals are anisotropic in general, and have different properties depending on the cut of the materials and the direction of the bulk or surface wave propagation.

Quartz is a well-known piezoelectric material. α -Quartz belongs to the triclinic crystal system with point group 32 and has a phase transition at 537°C to the β -type that is not piezoelectric. Quartz has a cut with a zero temperature coefficient at the resonant frequency change. For instance, quartz oscillators using the thickness shear mode of the AT-cut are extensively used as clock sources in computers and as frequency stabilized oscillators in TVs, and VCRs. On the other hand, an ST-cut quartz substrate that has X propagation has a zero temperature coefficient for surface

acoustic waves and so is used for SAW devices that have highly stabilized frequencies. Another distinguishing characteristic of quartz is its extremely high mechanical quality factor $Q_m > 10^5$.

Lithium niobate and lithium tantalate belong to an isomorphous crystal system and are composed of oxygen octahedra. The Curie temperatures of LiNbO₃ and LiTaO₃ are 1210 and 660°C, respectively. The crystal symmetry of the ferroelectric phase of these single crystals is 3m, and the polarization direction is along the c axis. These materials have high electromechanical coupling coefficients for surface acoustic waves. In addition, large single crystals can easily be obtained from their melts by using the conventional Czochralski technique. Thus, both materials are very important in SAW device applications.

Perovskite Ceramics

Most of the piezoelectric ceramics have the perovskite structure ABO₃, shown Fig. 1. This ideal structure consists of a simple cubic unit cell that has a large cation A at the corner, a smaller cation B in the body center, and oxygens O in the centers of the faces. The structure is a network of corner-linked oxygen octahedra surrounding B cations. The piezoelectric properties of perovskite-structured materials can be easily tailored for applications by incorporating various cations in the perovskite structure.

Barium Titanate. Barium titanate (BaTiO₃) is one of the most thoroughly studied and most widely used piezoelectric materials. Figure 2 shows the temperature dependence

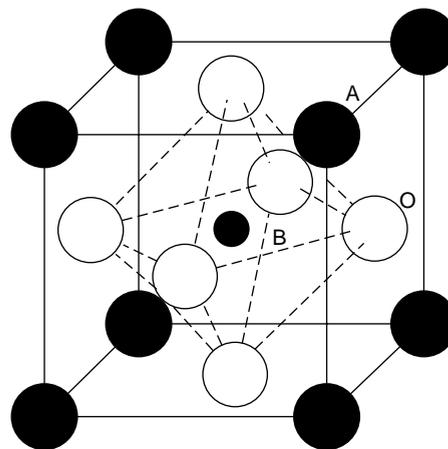


Figure 1. Perovskite structure ABO₃.

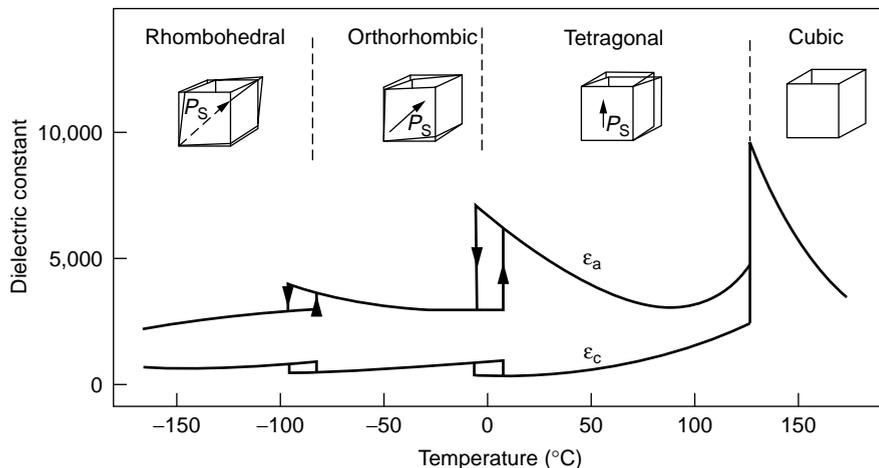


Figure 2. Dielectric constants of BaTiO_3 as a function of temperature.

of dielectric constants in BaTiO_3 , that demonstrate the phase transitions in BaTiO_3 single crystals. Three anomalies can be observed. The discontinuity at the Curie point (130°C) is due to a transition from a ferroelectric to a paraelectric phase. The other two discontinuities are accompanied by transitions from one ferroelectric phase to another. Above the Curie point, the crystal structure is cubic and has no spontaneous dipole moments. At the Curie point, the crystal becomes polar, and the structure changes from a cubic to a tetragonal phase. The dipole moment and the spontaneous polarization are parallel to the tetragonal axis. Just below the Curie temperature, the vector of the spontaneous polarization points in the $[001]$ direction (tetragonal phase), below 5°C it reorients in the $[011]$ (orthorhombic phase), and below -90°C in the $[111]$ (rhombohedral phase). The dielectric and piezoelectric properties of ferroelectric ceramic BaTiO_3 can be affected by its own stoichiometry, microstructure, and by dopants entering into the A or B site solid solution. Modified ceramic BaTiO_3 that contains dopants such as Pb or Ca ions have been used as commercial piezoelectric materials.

Lead Zirconate–Lead Titanate. Piezoelectric $\text{Pb}(\text{Ti}, \text{Zr})\text{O}_3$ solid-solution (PZT) ceramics are widely used because of their superior piezoelectric properties. The phase diagram of the PZT system ($\text{Pb}(\text{Zr}_x\text{Ti}_{1-x})\text{O}_3$) is shown in Fig. 3. The crystalline symmetry of this solid solution is determined by the Zr content. Lead titanate also has a tetragonal ferroelectric phase of the perovskite structure. As the Zr content x increases, the tetragonal distortion decreases, and when $x > 0.52$, the structure changes from the tetragonal $4mm$ phase to another ferroelectric phase of rhombohedral $3m$ symmetry. Figure 4 shows the dependence of several d constants on the composition near the morphotropic phase boundary between the tetragonal and rhombohedral phases. The d constants have their highest values near the morphotropic phase boundary. This enhancement in the piezoelectric effect is attributed to the increased ease of reorientation of the polarization in an electric field. Doping the PZT material with donors or acceptors changes the properties dramatically. Donor doping with ions such as Nb^{5+} or Ta^{5+} provides soft PZTs like PZT-5, because of the facility of domain motion due to the charge compensation

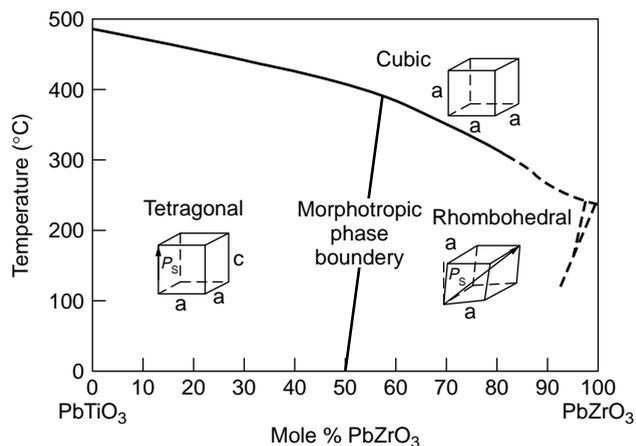


Figure 3. Phase diagram of the PZT system.

of the Pb vacancy which is generated during sintering. On the other hand, acceptor doping with Fe^{3+} or Sc^{3+} leads to hard PZTs such as PZT-8 because oxygen vacancies pin the domain wall motion.

Lead Titanate. PbTiO_3 has a tetragonal structure at room temperature and has large tetragonality $c/a = 1.063$.

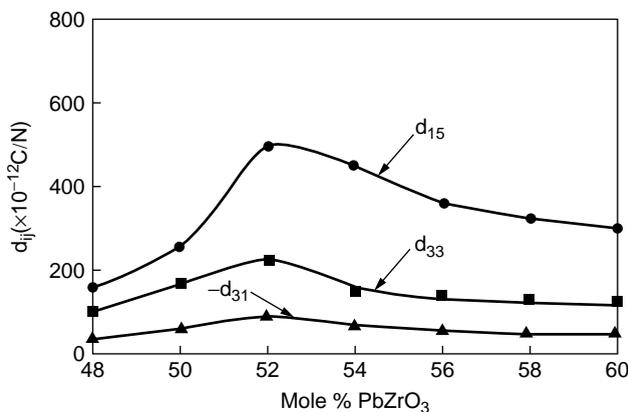


Figure 4. Piezoelectric d strain coefficients versus composition for the PZT system.

The Curie temperature is 490°C. Densely sintered PbTiO_3 ceramics cannot be obtained easily because they break up into powders when cooled through the Curie temperature. This is partly due to the large spontaneous strain that occurs at the transition. Lead titanate ceramics modified by small amounts of additives exhibit high piezoelectric anisotropy. Either $(\text{Pb}, \text{Sm})\text{TiO}_3$ (8) or $(\text{Pb}, \text{Ca})\text{TiO}_3$ (9) has extremely low planar coupling, that is, a large k_t/k_p ratio. Here, k_t and k_p are thickness-extensional and planar electromechanical coupling factors, respectively. $(\text{Pb}, \text{Nd})(\text{Ti}, \text{Mn}, \text{In})\text{O}_3$ ceramics that have a zero temperature coefficient of surface acoustic wave delay have been developed as superior substrate materials for SAW devices (10).

Relaxor Ferroelectrics. Relaxor ferroelectrics differ from normal ferroelectrics; they have broad phase transitions from the paraelectric to the ferroelectric state, strong frequency dependence of the dielectric constant (i.e., dielectric relaxation), and weak remanent polarization at temperatures close to the dielectric maximum. Lead-based relaxor materials have complex disordered perovskite structures of the general formula $\text{Pb}(\text{B}_1, \text{B}_2)\text{O}_3$ ($\text{B}_1 = \text{Mg}^{2+}, \text{Zn}^{2+}, \text{Sc}^{3+}$; $\text{B}_2 = \text{Nb}^{5+}, \text{Ta}^{5+}, \text{W}^{6+}$). The B-site cations are distributed randomly in the crystal. The characteristic of a relaxor is a broad and frequency dispersive dielectric maximum. In addition, relaxor-type materials such as the lead magnesium niobate $\text{Pb}(\text{Mg}_{1/3}\text{Nb}_{2/3})\text{O}_3$ –lead titanate PbTiO_3 solid solution (PMN–PT) exhibit electrostrictive phenomena that are suitable for actuator applications. Figure 5 shows an electric-field-induced strain curve that was observed for 0.9 PMN–0.1 PT and reported by Cross et al. in 1980 (11). Note that a strain of 0.1% can be induced by an electric field as small as 1 kV/mm and that hysteresis is negligibly small for this electrostriction.

Because electrostriction is the secondary electromechanical coupling observed in cubic structures, in principle,

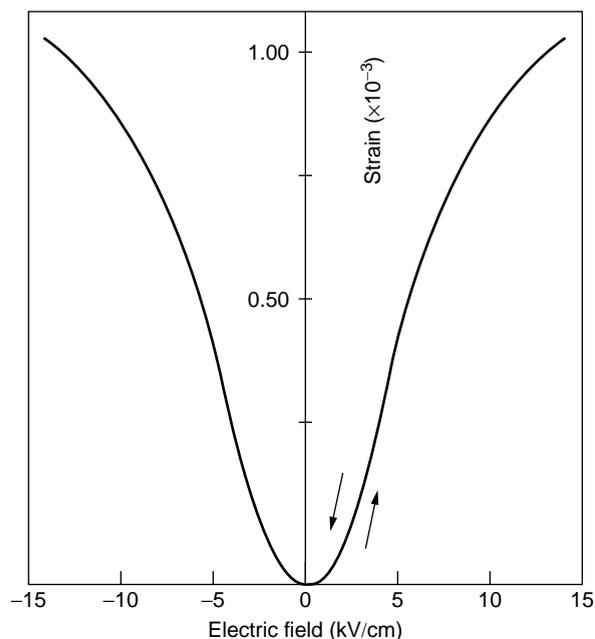


Figure 5. Field-induced electrostrictive strain in 0.9PMN–0.1PT.

the charge is not induced under applied stress. The converse electrostrictive effect, which can be used for sensor applications, means that the permittivity (the first derivative of polarization with respect to an electric field) is changed by stress.

In relaxor ferroelectrics, the piezoelectric effect can be induced under a bias field, that is, the electromechanical coupling factor k_t varies as the applied dc bias field changes. As the dc bias field increases, the coupling increases and saturates. These materials can be used for ultrasonic transducers that are tunable by a bias field (12).

The recent development of single-crystal piezoelectrics started in 1981, when Kuwata et al. first reported an enormously large electromechanical coupling factor $k_{33} = 92$ –95% and a piezoelectric constant $d_{33} = 1500$ pC/N in solid-solution single crystals between relaxor and normal ferroelectrics, $\text{Pb}(\text{Zn}_{1/3}\text{Nb}_{2/3})\text{O}_3$ – PbTiO_3 (13). After about 10 years, Yamashita et al. (Toshiba) and Shrout et al. (Penn State) independently reconfirmed that these values are true, and much more improved data were obtained in these few years, aimed at medical acoustic applications (14,15). Important data have been accumulated for $\text{Pb}(\text{Mg}_{1/3}\text{Nb}_{2/3})\text{O}_3$ (PMN), $\text{Pb}(\text{Zn}_{1/3}\text{Nb}_{2/3})\text{O}_3$ (PZN), and binary systems of these materials combined with PbTiO_3 (PMN–PT and PZN–PT) for actuator applications. Strains as large as 1.7% can be induced practically for a morphotropic phase boundary composition of PZN–PT solid-solution single crystals. Figure 6 shows the field-induced strain curve for [001] oriented 0.92PZN–0.08PT (15). It is notable that the highest values are observed for a rhombohedral composition only when the single crystal is poled along the perovskite [001] axis, not along the [111] spontaneous polarization axis.

Polymers

Polyvinylidene difluoride, (PVDF or PVF_2) is a piezoelectric when stretched during fabrication. Thin sheets of the cast polymer are drawn and stretched in the plane of the

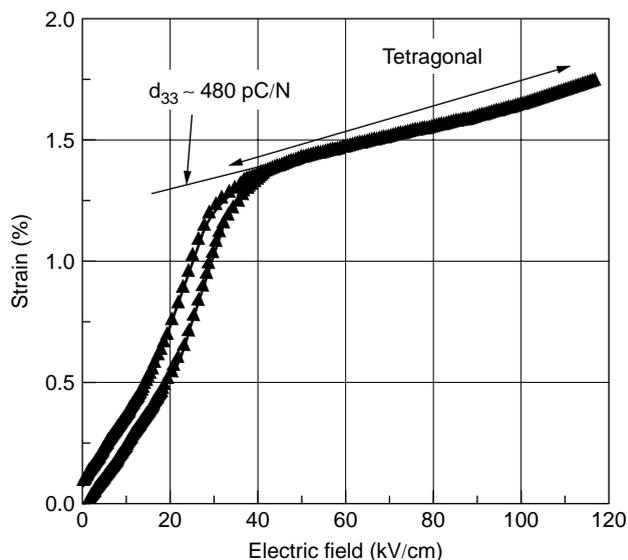


Figure 6. Field induced strain curve for [001] oriented 0.92PZN–0.08PT.

sheet in at least one direction and frequently also in the perpendicular direction to convert the material into its microscopically polar phase. Crystallization from a melt forms the nonpolar α phase, which can be converted into another polar β phase by uniaxial or biaxial drawing; these dipoles are then reoriented by electric poling. Large sheets can be manufactured and thermally formed into complex shapes. Copolymerization of vinylidene difluoride with trifluoroethylene (TrFE) results in a random copolymer (PVDF-TrFE) that has a stable, polar β phase. This polymer does not need to be stretched; it can be poled directly as formed. A thickness-mode coupling coefficient of 0.30 has been reported. Such piezoelectric polymers are used for directional microphones and ultrasonic hydrophones.

Composites

Piezocomposites comprised of piezoelectric ceramics and polymers are promising materials because of excellent tailored properties. The geometry of two-phase composites can be classified according to the connectivity of each phase (0, 1, 2, or 3 dimensionality) into 10 structures; 0-0, 0-1, 0-2, 0-3, 1-1, 1-2, 1-3, 2-2, 2-3, and 3-3 (15). A 1-3 piezocomposite, or PZT-rod/polymer-matrix composite is a most promising candidate. The advantages of this composite are high coupling factors, low acoustic impedance (square root of the product of its density and elastic stiffness), a good match to water and human tissue, mechanical flexibility, a broad bandwidth in combination with a low mechanical quality factor, and the possibility of making undiced arrays by structuring only the electrodes. The thickness-mode electromechanical coupling of the composite can exceed the k_t (0.40–0.50) of the constituent ceramic and almost approaches the value of the rod-mode electromechanical coupling k_{33} (0.70–0.80) of that ceramic (16). The acoustic match to tissue or water (1.5 Mrayls) of typical piezoceramics (20–30 Mrayls) is significantly improved by forming a composite structure, that is, by replacing a heavy, stiff ceramic by a light, soft polymer. Piezoelectric composite materials are especially useful for underwater sonar and medical diagnostic ultrasonic transducers.

Thin Films

Both zinc oxide (ZnO) and aluminum nitride (AlN) are simple binary compounds that have Wurtzite type structures, which can be sputter-deposited in a c -axis oriented thin film on a variety of substrates. ZnO has reasonable piezoelectric coupling, and its thin films are widely used in bulk acoustic and surface acoustic wave devices. The fabrication of highly c -axis oriented ZnO films has been extensively studied and developed. The performance of ZnO devices is, however, limited due to their small piezoelectric coupling (20–30%). PZT thin films are expected to exhibit higher piezoelectric properties. At present, the growth of PZT thin film is being carried out for use in microtransducers and microactuators. A series of theoretical calculations on perovskite type ferroelectric crystals suggests that large d and k values of magnitudes similar to those of PZN-PT can also be expected in PZT. Crystal orientation dependence of piezoelectric properties was phenomenologically

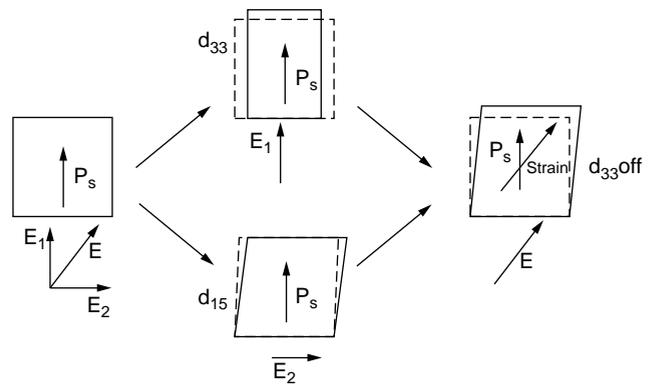


Figure 7. The principle of enhancement in electromechanical couplings in a perovskite piezoelectric.

calculated for compositions around the morphotropic phase boundary of PZT (17). The maximum longitudinal piezoelectric constant d_{33} (4–5 times the enhancement) and the electromechanical coupling factor k_{33} (more than 90%) in the rhombohedral composition were found at angles of 57 and 51° respectively, canted from the spontaneous polarization direction [111], which correspond roughly to the perovskite [100] axis.

Figure 7 shows the principle of the enhancement in electromechanical couplings. Because the shear coupling d_{15} is the highest in perovskite piezoelectric crystals, the applied field should be canted from the spontaneous polarization direction to obtain the maximum strain. Epitaxially grown, [001] oriented thin/thick films using a rhombohedral PZT composition reportedly enhance the effective piezoelectric constant by four to five times.

APPLICATIONS OF PIEZOELECTRICITY

Piezoelectric materials can provide coupling between electrical and mechanical energy and thus have been extensively used in a variety of electromechanical devices. The direct piezoelectric effect is most obviously used to generate charge or high voltage in applications such as the spark ignition of gas in space heaters, cooking stoves, and cigarette lighters. Using the converse effect, small mechanical displacements and vibrations can be produced in actuators by applying an electric field. Acoustic and ultrasonic vibrations can be generated by an alternating field tuned at the mechanical resonant frequency of a piezoelectric device and can be detected by amplifying the field generated by vibration incident on the material, which is usually used for ultrasonic transducers. Another important application of piezoelectricity is frequency control. The application of piezoelectric materials ranges over many fields, including ultrasonic transducers, actuators, and ultrasonic motors; electronic components such as resonators, wave filters, delay lines; SAW devices and transformers and high-voltage applications; and gas igniters, and ultrasonic cleaning and machining. Piezoelectric-based sensors, for instance, accelerometers, automobile knock sensors, vibration sensors, strain gages, and flow meters have been developed because pressure and vibration can be directly sensed as electric

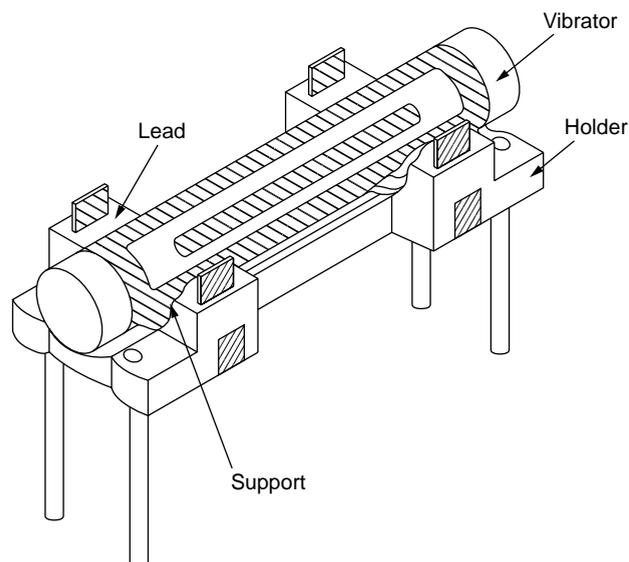


Figure 8. Cylindrical gyroscope commercialized by Tokin (Japan).

signals through the piezoelectric effect. Examples of these applications are given in the following sections.

Pressure Sensor/Accelerometer/Gyroscope

The gas igniter is one of the basic applications of piezoelectric ceramics. Very high voltage generated in a piezoelectric ceramic under applied mechanical stress can cause sparking and ignite a gas.

Piezoelectric ceramics can be employed as stress sensors and acceleration sensors, because of their “direct piezoelectric effect.” Kistler (Switzerland) is manufacturing a 3-D stress sensor. By combining an appropriate number of quartz crystal plates (extensional and shear types), the multilayer device can detect three-dimensional stresses (18).

Figure 8 shows a cylindrical gyroscope commercialized by Tokin (Japan) (19). The cylinder has six divided electrodes; one pair is used to excite the fundamental bending vibration mode, and the other two pairs are used to detect acceleration. When rotational acceleration is applied around the axis of this gyro, the voltage generated on the electrodes is modulated by the Coriolis force. By subtracting the signals between the two sensor electrode pairs, a voltage directly proportional to the acceleration can be obtained. This type of gyroscope has been widely installed in handheld video cameras to monitor the inevitable hand vibration during operation and to compensate for it electronically on a display by using the sensed signal.

Ultrasonic Transducer

One of the most important applications of piezoelectric materials is based on the ultrasonic echo field (20,21). Ultrasonic transducers convert electrical energy into a mechanical form when generating an acoustic pulse and convert mechanical energy into an electrical signal when detecting its echo. Nowadays, piezoelectric transducers are

being used in medical ultrasound for clinical applications that range from diagnosis to therapy and surgery. They are also used for underwater detection, such as sonars and fish finders, and nondestructive testing.

Ultrasonic transducers often operate in a pulse-echo mode. The transducer converts electrical input into an acoustic wave output. The transmitted waves propagate into a body, and echoes are generated that travel back to be received by the same transducer. These echoes vary in intensity according to the type of tissue or body structure, and thereby create images. An ultrasonic image represents the mechanical properties of the tissue, such as density and elasticity. We can recognize anatomical structures in an ultrasonic image because the organ boundaries and fluid-to-tissue interfaces are easily discerned. Ultrasonic imaging can also be done in real time. This means that we can follow rapidly moving structures such as the heart without motional distortion. In addition, ultrasound is one of the safest diagnostic imaging techniques. It does not use ionizing radiation like X rays and thus is routinely used for fetal and obstetrical imaging. Useful areas for ultrasonic imaging include cardiac structures, the vascular system, the fetus, and abdominal organs such as the liver and kidney. In brief, it is possible to see inside the human body by using a beam of ultrasound without breaking the skin.

There are various types of transducers used in ultrasonic imaging. Mechanical sector transducers consist of single, relatively large resonators that provide images by mechanical scanning such as wobbling. Multiple element array transducers permit the imaging systems to access discrete elements individually and enable electronic focusing in the scanning plane at various adjustable penetration depths by using phase delays. The two basic types of array transducers are linear and phased (or sector). Linear array transducers are used for radiological and obstetrical examinations, and phased array transducers are useful for cardiological applications where positioning between the ribs is necessary.

Figure 9 shows the geometry of the basic ultrasonic transducer. The transducer is composed mainly of

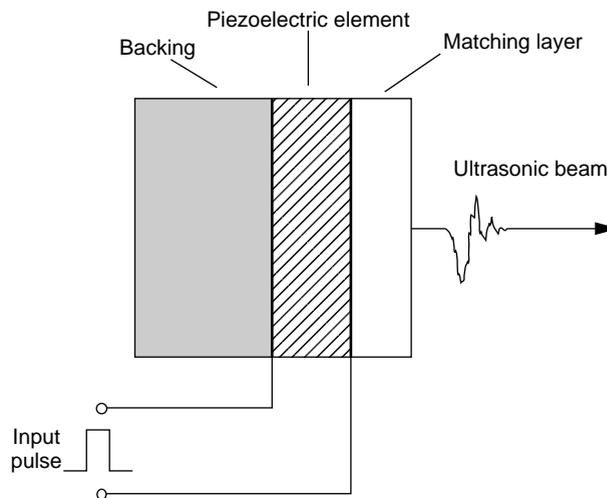


Figure 9. Geometry of the fundamental transducer for acoustic imaging.

matching, piezoelectric material, and backing layers (22). One or more matching layers are used to increase sound transmissions into tissues. The backing is attached to the transducer rear to damp the acoustic returnwave and to reduce the pulse duration. Piezoelectric materials are used to generate and detect ultrasound. In general, broadband transducers should be used for medical ultrasonic imaging. The broad bandwidth response corresponds to a short pulse length that results in better axial resolution. Three factors are important in designing broad bandwidth transducers. The first is acoustic impedance matching, that is, effectively coupling the acoustic energy to the body. The second is high electromechanical coupling coefficient of the transducer. The third is electrical impedance matching, that is, effectively coupling electrical energy from the driving electronics to the transducer across the frequency range of interest. The operator of pulse-echo transducers is based on the thickness mode resonance of the piezoelectric thin plate. The thickness mode coupling coefficient k_t is related to the efficiency of converting electric energy into acoustics and vice versa. Further, a low planar mode coupling coefficient k_p is beneficial for limiting energies from being expended in a nonproductive lateral mode. A large dielectric constant is necessary to enable a good electrical impedance match to the system, especially in tiny piezoelectric sizes.

Table 4 compares the properties of ultrasonic transducer materials (7,23). Ferroelectric ceramics, such as lead zirconate titanate and modified lead titanate, are almost universally used as ultrasonic transducers. The success of ceramics is due to their very high electromechanical coupling coefficients. In particular, soft PZT ceramics such as PZT-5A and 5H type compositions are most widely used because of their exceedingly high coupling properties and because they can be relatively easily tailored, for instance, in the wide dielectric constant range. On the other hand, modified lead titanates such as samarium-doped materials have high piezoelectric anisotropy: the planar coupling factor k_p is much less than the thickness coupling factor k_t . Because the absence of lateral coupling leads to reduced interference from spurious lateral resonances in longitudinal oscillators, this is very useful in high-frequency array transducer applications. One disadvantage of PZT and other lead-based ceramics is their large acoustic impedance (approximately $30 \text{ kgm}^{-2}\text{s}^{-1}$ (Mrayls) compared to body tissue (1.5 Mrayls). Single or multiple matching layers of intermediate impedances need to be used in PZT to improve acoustic matching.

On the other hand, piezoelectric polymers, such as polyvinylidene difluoride-trifluoroethylene, have much lower

acoustic impedance (4–5 Mrayls) than ceramics and thus match soft tissues better. However, piezopolymers are less sensitive than ceramics, and they have relatively low dielectric constants that require large drive voltage and give poor noise performance due to mismatching of electrical impedance.

Piezoelectric ceramic/polymer composites are alternatives to ceramics and polymers. Piezocomposites that have 2–2 or 1–3 connectivity are commonly used in ultrasonic medical applications. They combine the low acoustic impedance advantage of polymers and the high sensitivity and low electrical impedance advantages of ceramics.

The design frequency of a transducer depends on the penetration depth required by the application. Resolution is improved as frequency increases. Although a high-frequency transducer can produce a high-resolution image, higher frequency acoustic energy is more readily attenuated by the body. A lower frequency transducer is used as a compromise when imaging deeper structures. Most medical ultrasound imaging systems operate in the frequency range from 2–10 MHz and can resolve objects approximately 0.2–1 mm in size. At 3.5 MHz, imaging to a depth of 10–20 cm is possible, and at 50 MHz, increased losses limit the depth to less than 1 cm. Higher frequency transducers (10–50 MHz) are used for endoscopic imaging and for catheter-based intravascular imaging. Ultrasound microscopy is being done at frequencies higher than 100 MHz. The operating frequency of the transducer is directly related to the thickness and velocity of sound in the piezoelectric materials employed. As the frequency increases, resonator thickness decreases. For a 3.5 MHz transducer, the PZT ceramic must be roughly 0.4 mm thick. Conventional ceramic transducers, such as PZT, are limited to frequencies below 80 MHz because of the difficulty of fabricating thinner devices (24). Piezoelectric thin-film transducers such as ZnO have to be used for microscopic applications (at frequencies higher than 100 MHz, corresponding to a thickness of less than $20 \mu\text{m}$) (25).

Resonator and Filter

When a piezoelectric body vibrates at its resonant frequency, it absorbs considerably more energy than at other frequencies, resulting in a fall of the impedance. This phenomenon enables using piezoelectric materials as wave filters. A filter is required to pass a certain selected frequency band or to stop a given band. The bandwidth of a filter fabricated from a piezoelectric material is determined by the square of the coupling coefficient k . Quartz crystals that have very low k values of about 0.1 can pass very narrow

Table 4. Comparison of the Properties of Ultrasonic Transducer Materials

	PZT Ceramic	PVDF Polymer	PZT-Polymer Composite	ZnO Film
k_t	0.45–0.55	0.20–0.30	0.60–0.75	0.20–0.30
Z (Mrayls)	20–30	1.5–4	4–20	35
$\epsilon_{33} T / \epsilon_0$	200–5000	10	50–2500	10
$\tan \delta$ (%)	<1	1.5–5	<1	<1
Q_m	10–1000	5–10	2–50	10
ρ (g/cm ³)	5.5–8	1–2	2–5	3–6

frequency bands of approximately 1% of the center resonance frequency. On the other hand, PZT ceramics whose planar coupling coefficient is about 0.5 can easily pass a band of 10% of the center resonance frequency. The sharpness of the passband depends on the mechanical quality factor Q_m of the material. Quartz also has a very high Q_m of about 10^6 which results in a sharp cutoff of the passband and well-defined frequency of the oscillator.

A simple resonator is a thin disk electrode on its plane faces and vibrating radially for applications in filters whose center frequency ranges from 200 kHz to 1 MHz and whose bandwidth is several percent of the center frequency. The disk diameter must be about 5.6 mm for a frequency of 455 kHz. However, if the required frequency is higher than 10 MHz, other modes of vibration such as the thickness extensional mode are exploited, because of its smaller size disk. Trapped-energy type filters made from PZT ceramics have been widely used in the intermediate frequency range, for example, 10.7 MHz for FM radio receivers and transmitters. By employing the trapped-energy phenomenon, the overtone frequencies are suppressed. The plate is partly covered by electrodes of a specific area and thickness. The fundamental frequency of the thickness mode beneath the electrode is less than that of the unelectroded portion because of the extra inertia of the electrode mass. The longer wave characteristic of the electrode region cannot propagate in the unelectroded region. Higher frequency overtones can propagate into the unelectroded region. This is called the trapped-energy principle. Figure 10 shows a schematic drawing of a trapped-energy filter. In this structure, the top electrode is split so that coupling between the two parts is efficient only at resonance. More stable filters suitable for telecommunication systems have been made from single crystals such as quartz or LiTaO_3 .

Piezoelectric Transformer

The transfer of vibrational energy from one set of electrodes to another on a piezoelectric ceramic body can be used to transform voltage. The device is called a piezoelectric transformer. Recently, office automation equipment that has a liquid crystal display such as notebook type personal computers and car navigation systems has been successfully commercialized. This equipment that uses liquid crystal display requires a very thin transformer without electromagnetic noise to start the glow of

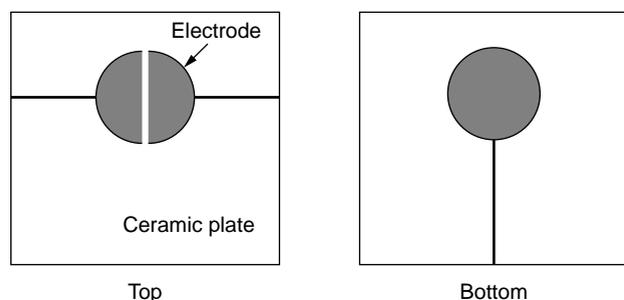


Figure 10. Trapped-energy filter.

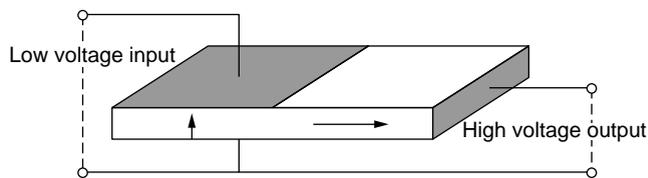


Figure 11. Piezoelectric transformer.

a fluorescent back-lamp. This application has recently accelerated the development of piezoelectric transformers. Figure 11 shows the basic structure, where two differently poled parts coexist in one piezoelectric plate. The plate has electrodes on half of its major faces and on an edge. The plate is then poled in its thickness direction at one end and parallel to the long axis over most of its length. A low-voltage ac supply is applied to the large-area electrodes at a frequency that excites a length extensional mode resonance. Then, a high-voltage output can be taken from the small electrode and from one of the larger electrodes. Following the proposal by Rosen mentioned before, piezoelectric transformers of several different structures have been reported (26). Multilayer type transformers are proposed to increase the voltage step-up ratio (27). The input part has a multilayer structure and has internal electrodes, and the output electrodes are formed at the side surface of the half of the rectangular plate. This transformer uses the piezoelectric transverse mode for the input and output parts.

SAW Device

A surface acoustic wave (SAW), also called a Rayleigh wave, is composed of a coupling between longitudinal and shear waves in which the SAW energy is confined near the surface. An associated electrostatic wave exists for a SAW on a piezoelectric substrate that allows electroacoustic coupling via a transducer. The advantages of SAW technology are that a wave can be electroacoustically accessed and trapped at the substrate surface and its velocity is approximately 10^4 times slower than an electromagnetic wave. The SAW wavelength is of the same order of magnitude as line dimensions that can be photolithographically produced, and the lengths for both short and long delays are achievable on reasonably size substrates (28,29).

There is a very broad range of commercial applications, including front-end and IF (intermediate frequency) filters, CATV (community antenna television) and VCR (video cassette recorder) components, synthesizers, analyzers, and navigators. In SAW transducers, finger electrodes provide the ability to sample or tap the wave, and the electrode gap gives the relative delay. A SAW filter is composed of a minimum of two transducers. A schematic of a simple SAW bidirectional filter is shown in Fig. 12. A bidirectional transducer radiates energy equally from each side of the transducer. Energy not received is absorbed to eliminate spurious reflection.

Various materials are currently being used for SAW devices. The most popular single-crystal SAW materials are lithium niobate and lithium tantalate. The materials have different properties depending on their cuts and the direction of propagation. The fundamental parameters are the

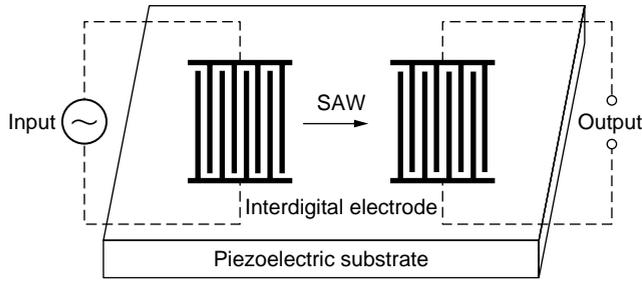


Figure 12. Typical SAW bidirectional filter that consists of two interdigital transducers.

SAW velocity, the temperature coefficient of delay (TCD), the electromechanical coupling factor, and the propagative loss. Surface acoustic waves can be generated and detected by spatially periodic, interdigital electrodes on the plane surface of a piezoelectric plate. A periodic electric field is produced when an RF source is connected to the electrode, thus permitting piezoelectric coupling to a traveling surface wave. If an RF source of a frequency f is applied to an electrode whose periodicity is p , energy conversion from an electrical to mechanical form will be maximum when

$$f = f_0 = V_s/p, \quad (9)$$

where V_s is the SAW velocity and f_0 is the center frequency of the device. SAW velocity is an important parameter that determines the center frequency. Another important parameter for many applications is the temperature sensitivity. For example, the temperature stability of the center frequency of a SAW bandpass filter is a direct function of the temperature coefficient for the velocity and delay time of the material used. The first-order temperature coefficient of delay time is given by

$$(1/t)(dt/dT) = (1/L)(dL/dT) - (1/V_s)(dV_s/dT), \quad (10)$$

where $t = L/V_s$ is the delay time and L is the SAW propagative length. The surface wave coupling factor k_s^2 is defined in terms of the change in SAW velocity that occurs when the wave passes across a surface coated by a thin massless conductor, so that the piezoelectric field associated with the wave is effectively shorted-circuited. The coupling factor

k_s^2 is expressed by

$$k_s^2 = 2(V_f - V_m)/V_f, \quad (11)$$

where V_f is the free surface wave velocity and V_m is the velocity on the metallized surface. In SAW applications, the value of k_s^2 relates to the maximum bandwidth obtainable and the amount of signal loss between input and output that determines the fractional bandwidth versus minimum insertion loss for a given material and a filter. Propagative loss, one of the major factors that determines the insertion loss of a device, is caused by wave scattering by crystalline defects and surface irregularities. Materials that have high electromechanical coupling factors combined with small temperature coefficients of delay time are likely to be required. The free surface velocity V_0 of the material is a function of the cut angle and the propagative direction. The TCD is an indication of the frequency shift expected from a transducer due to a temperature change and is also a function of the cut angle and the propagative direction. The substrate is chosen on the basis of the device's design specifications for operating temperature, fractional bandwidth, and insertion loss.

Table 5 shows some important material parameters of representative SAW materials. Piezoelectric single crystals such as 128°Y-X (128° -rotated Y cut and X propagation) LiNbO_3 and $\text{X-}112^\circ\text{Y}$ (X cut and 112° -rotated Y propagation) LiTaO_3 have been extensively employed as SAW substrates for VIF filters. ZnO thin films c-axis oriented and deposited on a fused quartz, glass, or sapphire substrate have also been commercialized for SAW devices.

Actuators

Currently, another important application of piezoelectric materials exists in the actuator field (30). Using the converse piezoelectric effect, a small displacement can be produced by applying an electric field to a piezoelectric material. Vibrations can be generated by applying an alternating electric field. There is a demand in advanced precision engineering for a variety of types of actuators that can adjust position precisely (micropositioning devices), suppress noise vibrations (dampers), and drive objects dynamically (ultrasonic motors). These devices are used in areas, including optics, astronomy, fluid control, and precision machinery. Piezoelectric strain and electrostriction induced by an electric field are used for actuator applications.

Table 5. Material Parameters for Representative SAW Materials

	Material	Cut-Propagation Direction	k^2 (%)	TCD		
				(ppm/C)	V_0 (m/s)	ϵ_r
Single crystal	Quartz	ST-X	0.16	0	3158	4.5
	LiNbO_3	128°Y-X	5.5	-74	3960	35
	LiTaO_3	$\text{X}112^\circ\text{-Y}$	0.75	-18	3290	42
	$\text{Li}_2\text{B}_4\text{O}_7$	(110)<001>	0.8	0	3467	9.5
Ceramic	$\text{PZT-In}(\text{Li}_{3/5}\text{W}_{2/5})\text{O}_3$		1.0	10	2270	690
	$(\text{Pb,Nd})(\text{Ti,Mn,In})\text{O}_3$		2.6	<1	2554	225
Thin film	ZnO/glass		0.64	-15	3150	8.5
	ZnO/sapphire		1.0	-30	5000	8.5

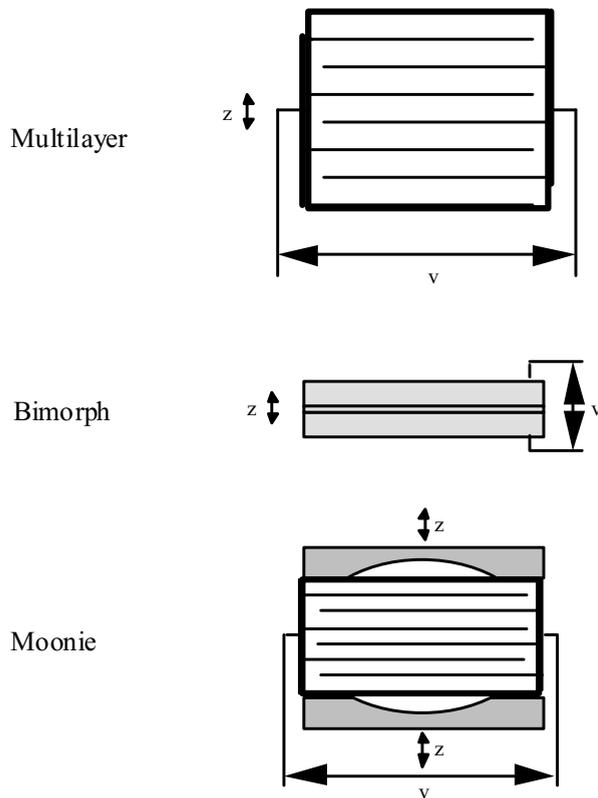


Figure 13. Structures of ceramic actuators.

Figure 13 shows the design classification of ceramic actuators. Simple devices composed of a disk or a multilayer type use the strain induced in a ceramic by the applied electric field directly. Complex devices do not use the induced strain directly but use the amplified displacement through a special magnification mechanism such as a unimorph, bimorph or moonie. The most popularly used multilayer and bimorph types have the following characteristics: The multilayer type does not have a large displacement ($10\ \mu\text{m}$) but has advantages in generation force (1 kN), response speed ($10\ \mu\text{s}$), lifetime (10^{11} cycles), and electromechanical coupling factor k_{33} (0.70). The bimorph type has a large displacement ($300\ \mu\text{m}$) but has disadvantages in generation force (1 N), response speed (1 ms), lifetime (10^8 cycles), and electromechanical coupling factor k_{eff} (0.10). For instance, in a 0.65 PMN–0.35 PT multilayer actuator made of 99 layers of $100\text{-}\mu\text{m}$ thick sheets ($2 \times 3 \times 10\ \text{mm}^3$), an $8.7\text{-}\mu\text{m}$ displacement is generated by a 100 V voltage, accompanied by a slight hysteresis. The transmitted response of the induced displacement after the application of a rectangular voltage is as quick as $10\ \mu\text{s}$. The multilayer has a field-induced strain of 0.1% along the length.

Unimorph and bimorph devices are defined by the number of piezoelectric ceramic plates: only one ceramic plate is bonded onto an elastic shim, or two ceramic plates are bonded together. The bimorph causes bending deformation because each piezoelectric plate bonded together produces extension or contraction in an electric field. In general, there are two types of piezoelectric bimorphs: the antiparallel polarization type and the parallel polarization type,

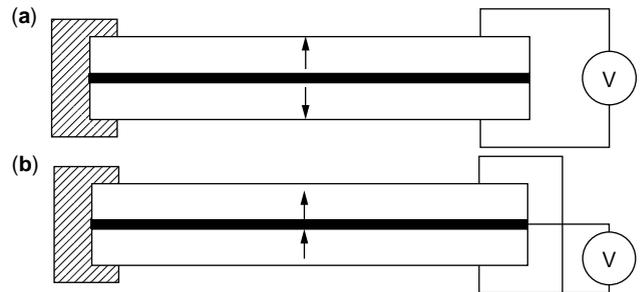


Figure 14. Two types of piezoelectric bimorphs: (a) antiparallel polarization type and (b) parallel polarization type.

as shown in Fig. 14. Two poled piezoelectric plates $t/2$ thick and L long are bonded so that their polarization directions are opposite or parallel to each other. In the cantilever bimorph configuration where one end is clamped, the tip displacement δ_z under an applied voltage V is

$$\delta_z = (3/2)d_{31}(L^2/t^2)V, \quad (\text{antiparallel type}) \quad (12)$$

$$\delta_z = 3d_{31}(L^2/t^2)V. \quad (\text{parallel type}) \quad (13)$$

The resonance frequency f_r for both types is given by

$$f_r = 0.16 t/L^2(\rho s_{11}^E)^{-1/2}, \quad (14)$$

where ρ is the density and s_{11}^E is the elastic compliance. A metallic sheet (called a shim) is occasionally sandwiched between the two piezoelectric plates to increase reliability; the structure can be maintained even if the ceramics fracture. Using the bimorph structure, a large magnification of displacement is easily obtainable. However, the disadvantages include low response speed (1 kHz) and low generative force.

A composite actuator structure called a “moonie” has been developed to amplify the small displacements induced in piezoelectric ceramics. The moonie consists of a thin multilayer element and two metal plates that have narrow moon-shaped cavities bonded together. This device has characteristics intermediate between the conventional multilayer and bimorph actuators; it has an order of magnitude larger displacement ($100\ \mu\text{m}$) than the multilayer, much larger generative force (100 N), and quicker response ($100\ \mu\text{s}$) than the bimorph.

Some examples of applications of piezoelectric and electrostrictive actuators are described here. The piezoelectric impact dot-matrix printer is the first mass-produced device that uses multilayer ceramic actuators (Fig. 15). The advantages of a piezoelectric printer head compared to conventional magnetic type are low energy consumption, low heat generation, and fast printing speed. Longitudinal multilayer actuators do not have a large displacement, and thus a suitable displacement magnification mechanism is necessary. The displacement induced in a multilayer actuator pushes up the force point, and its displacement magnification is carried out through hinge levers to generate a large wire stroke. When the displacement in

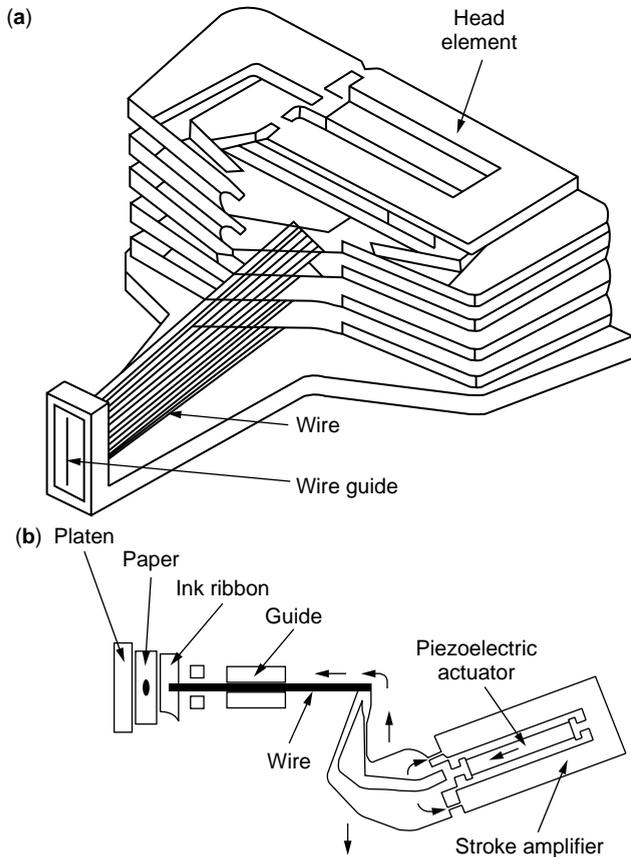


Figure 15. Impact dot-matrix printer head.

the piezoactuator is $8 \mu\text{m}$, a wire stroke of $240 \mu\text{m}$ can be obtained; the magnification rate is 30 times.

Bimorph structures are commonly used for VCR head-tracking actuators because of their large displacements. An autotracking scan system uses a piezoelectric actuator so that the head follows the recording track, even driven in both still and quick modes. As can be anticipated, the bimorph drive is inevitably accompanied by torsional motion. A special mechanism has to be employed to obtain perfectly parallel motion. Piezoelectric pumps for gas or liquid that use the alternating bending motion of the bimorph have been developed for intravenous drip injection in hospitals and for medication dispensing for chemotherapy, chronic pain, and diabetes. Piezoelectric fans for cooling electronic circuits are made from a pair of bimorphs that are driven out of phase so as to blow effectively. Furthermore, piezo-bimorph type camera shutters have been widely commercialized by Minolta.

Lenses and mirrors in optical control systems require micropositioning, and even the shapes of mirrors are adjusted to correct image distortions. For instance, a space-qualified active mirror, called an articulating fold mirror, uses six PMN electrostrictive multilayer actuators to position and tilt a mirror tip precisely to correct the focusing aberration of the Hubble Space Telescope.

Piezoelectric actuators are also useful for vibrational suppression systems in an automobile. An electronic controlled shock absorber was developed by Toyota.

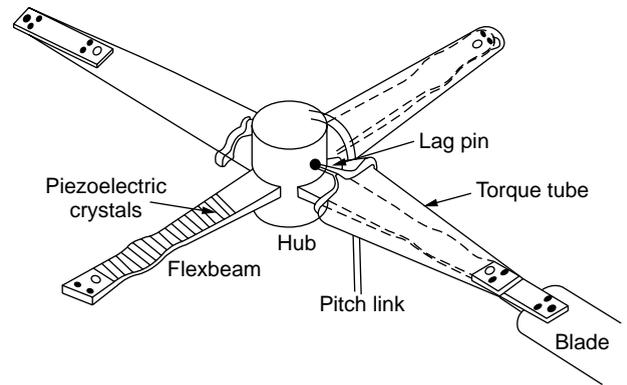


Figure 16. Bearingless rotor flexbeam and attached piezoelectric strips.

Piezoelectric sensors that detecting road roughness are composed of five layers of 0.5-mm thick PZT disks. The actuator is made of 88 layers of 0.5 mm thick disks. Under 500 V , it generates about $50 \mu\text{m}$ displacement, which is magnified 40 times by a piston and plunger pin combination. This stroke pushes the change valve of the damping force down and then opens the bypass oil route, leading to a decrease in flow resistance. This electronically controlled shock absorber has both controllability and provides comfort simultaneously.

The U.S. Army is interested in developing a rotor control system for helicopters because a slight change in the blade angle dramatically enhances controllability. Figure 16 shows a bearingless rotor flexbeam that has piezoelectric strips attached. Various types of PZT-sandwiched beam structures have been investigated for such a flexbeam application and for active vibrational control.

Ultrasonic Motors

An ultrasonic motor (USM) is an example of a piezoelectric actuator that uses resonant vibration. Linear motion in ultrasonic motors is obtained by frictional force from elliptical vibration. The motor consists of a high-frequency power supply, a vibrator, and a slider. The vibrator is composed of a piezoelectric driving component and an elastic vibratory part, and the slider is composed of an elastic moving part and a friction coat. The characteristics of ultrasonic motors are low speed and high torque compared to the high speed and low torque of conventional electromagnetic motors (30,31).

Ultrasonic motors are classified into two types: the standing-wave type and the propagating-wave type. The displacement of a standing wave is expressed by

$$x_s(x, t) = A \cos(kx) \cos(\omega t); \quad (15)$$

for a propagative wave displacement is given by

$$\begin{aligned} x_p(x, t) &= A \cos(kx - \omega t) \\ &= A \cos(kx) \cos(\omega t) + A \cos(kx - \pi/2) \cos(\omega t - \pi/2). \end{aligned} \quad (16)$$

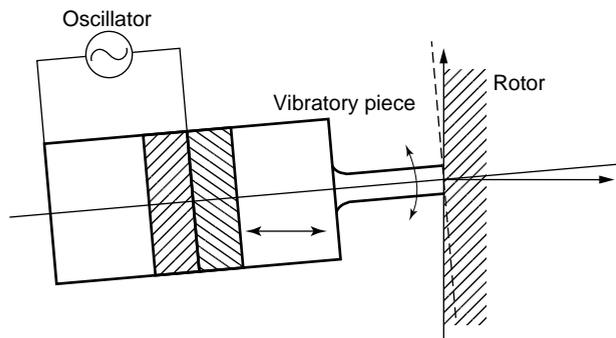


Figure 17. Vibratory-coupler type ultrasonic motor.

A propagating wave can be generated by superimposing two standing waves whose phases differ from each other by 90° in both time and in space. The standing-wave type is sometimes called a vibratory coupler or a "woodpecker" type; a vibratory piece is connected to a piezoelectric driver, and the tip portion generates a flat elliptical movement (Fig. 17). The vibratory piece is attached to a rotor or a slider at a slight cant angle. The standing-wave type has high efficiency, up to 98% of theoretical. However, a problem of this type is lack of control in both clockwise and counterclockwise directions. The principle of the propagative type is shown in Fig. 18. In the propagating-wave type, also called the "surfing-type," a surface particle of the elastic body draws an elliptical locus due to coupling of the longitudinal and transverse waves. This type generally requires two vibrational sources to generate one propagating wave; this leads to low efficiency (not more than 50%), but it is controllable in both rotational directions. An ultrasonic rotary motor is successfully used in an autofocusing camera to produce precise rotational displacements. The advantages of this motor over the conventional electromagnetic motor are silent drive (inaudible), thin motor design, and energy savings.

Tiny conventional electromagnetic motors, smaller than 1cm that have sufficient energy efficiency are rather difficult to produce. Therefore, the ultrasonic motor is gaining widespread attention. Ultrasonic motors whose efficiency is independent of size are superior in the minimotor area. A compact ultrasonic rotary motor as tiny as 3 mm in diameter was developed by Uchino et al. (32). The stator consists of a piezoelectric ring and two concave/convex metal end caps that have windmill-shaped slots bonded together, so as to generate a coupled vibration of the up-down and

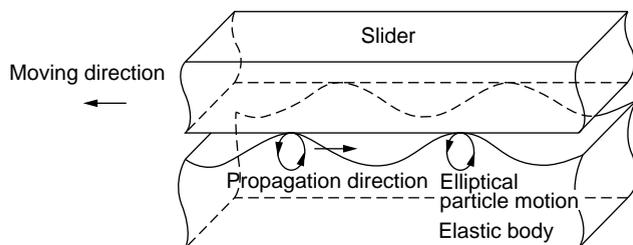


Figure 18. Principle of the propagating-wave type ultrasonic motor.

torsional type. Because the number of components and the fabrication process are minimized, the cost of fabrication will decrease remarkably, and it can be disposable; this is very suitable for medical catheter and endoscopic applications. When driven at 160 kHz, a maximum speed of 600 rpm and maximum torque 1 mN m were obtained.

The following key problems should be systematically studied in developing reliable ultrasonic motors: (1) development of low loss and high vibrational velocity piezoelectric materials; (2) piezoactuator designs that account for heat generation and degradation mechanisms; (3) USM designs, including displacement amplification mechanisms and frictional contact parts.

BIBLIOGRAPHY

1. B. Jaffe, W. Cook, and H. Jaffe, *Piezoelectric Ceramics*. Academic Press, London, 1971.
2. W.G. Cady, *Piezoelectricity*. McGraw-Hill, NY, rev. ed., Dover, 1964.
3. F. Jona and G. Shirane, *Ferroelectric Crystals*. Pergamon Press, London, 1962.
4. M.E. Lines and A.M. Glass, *Principles and Applications of Ferroelectric Materials*. Clarendon Press, Oxford, 1977.
5. IEEE Standard on Piezoelectricity. IEEE, NY, 1978.
6. Landold and Boernstein, *Numerical Data and Functional Relationships in Science and Technology: Crystal and Solid State Physics*, Vol.11. Springer-Verlag, Berlin, 1979.
7. W.A. Smith, *Proc. SPIE* 1733 (1992).
8. H. Takeuchi, S. Jyomura, E. Yamamoto, and Y. Ito, *J. Acoust. Soc. Am.* **74**: 1114 (1982).
9. Y. Yamashita, K. Yokoyama, H. Honda, and T. Takahashi, *Jpn. J. Appl. Phys.* **20** (Suppl. 20-4): 183 (1981).
10. Y. Ito, H. Takeuchi, S. Jyomura, K. Nagatsuma, and S. Ashida, *Appl. Phys. Lett.* **35**: 595 (1979).
11. L.E. Cross, S.J. Jang, R.E. Newnham, S. Nomura, and K. Uchino, *Ferroelectrics* **23**: 187 (1980).
12. H. Takeuchi, H. Masuzawa, C. Nakaya, and Y. Ito, *Proc. IEEE 1990 Ultrasonics Symp.*, 1990, p. 697.
13. J. Kuwata, K. Uchino, and S. Nomura, *Jpn. J. Appl. Phys.* **21**: 1298 (1982).
14. T.R. Shrout, Z.P. Chang, N. Kim, and S. Markgraf, *Ferroelectric Lett.* **12**: 63 (1990).
15. R.E. Newnham, D.P. Skinner, and L.E. Cross, *Mater. Res. Bull.* **13**: 525 (1978).
16. W.A. Smith, *Proc. 1989 IEEE Ultrasonic Symp.*, 1989, p. 755.
17. X.H. Du, J. Zheng, U. Belegundu, and K. Uchino, *J. Appl. Phys. Lett.* **72**: 2421 (1998).
18. Kistler, *Stress Sensor, Production Catalog*. Switzerland,
19. Tokin, *Gyroscope, Production Catalog*. Japan,
20. B.A. Auld, *Acoustic Fields and Waves in Solids*. 2e, Krieger, Melbourne, FL, 1990.
21. G.S. Kino, *Acoustic Waves: Device Imaging and Analog Signal Processing*. Prentice-Hall, Englewood Cliffs, NJ, 1987.
22. C.S. Desilets, J.D. Fraser, and G.S. Kino, *IEEE Trans. Sonics Ultrasonics*, **SU-25**: 115 (1978).
23. T.R. Gururaja, *Am. Ceram. Soc. Bull.* **73**: 50 (1994).
24. F.S. Foster, L.K. Ryan, and D.H. Turnbull, *IEEE Trans. Ultrasonics Ferroelectrics Frequency Control* **38**: 446 (1991).

25. Y.Ito, K. Kushida, K. Sugawara, and H. Takeuchi, *IEEE Trans. Ultrasonics Ferroelectrics Frequency Control* **42**: 316, 1995.
26. C.A. Rosen, *Proc. Electron. Component Symp.*, 1957, p. 205.
27. S. Kawashima, O. Ohnishi, H. Hakamata, S. Tagami, A. Fukuoka, T. Inoue, and S. Hirose, *Proc. IEEE Int. Ultrasonic Symp. '94*, France, Nov. 1994.
28. C. Campbell, *Surface Acoustic Wave Devices and Their Signal Processing Applications*. Academic Press, San Diego, 1989.
29. H. Matthews, *Surface Wave Filters*. Wiley-Interscience, NY, 1977.
30. K. Uchino, *Piezoelectric Actuators and Ultrasonic Motors*. Kluwer Academic, Boston, 1997.
31. S. Ueha and Y. Tomikawa, *Ultrasonic Motors*. Clarendon Press, Oxford, 1993.
32. K. Uchino, *J. Medical Ultrasonics* **24**: 1191 (1997).

CHARACTERIZATION OF PIEZOELECTRIC CERAMIC MATERIALS

T.L. JORDAN
 Z. OUNAIES
 NASA Langley Research Center
 Hampton, VA

INTRODUCTION

This article explores piezoelectric ceramics analysis and characterization. The focus is on polycrystalline ceramics; therefore, single crystals, polymeric materials, and organic/inorganic composites are outside the scope of this review. To grasp the behavior of a piezoelectric polycrystalline ceramic thoroughly, a basic understanding of the ceramic itself should not be overlooked. To this end, we have presented a brief introduction of the history of piezoelectricity and a discussion on processing of the ceramic and the development of the constitutive relationships that define the behavior of a piezoelectric material. We have attempted to cover the most common measurement methods and to introduce parameters of interest. Excellent sources for more in-depth coverage of specific topics can be found in the bibliography. In most cases, we refer to lead zirconate titanate (PZT) to illustrate some of the concepts because it is the most widely used and studied piezoelectric ceramic to date.

PIEZOELECTRIC MATERIALS: HISTORY AND PROCESSING

Smart materials are materials that undergo transformations through physical interactions. An alternate definition is that a smart material is a material that senses a change in its environment and adapts to correct or eliminate such a change by using a feedback system. Piezoelectric materials, shape-memory alloys, electrostrictive materials, magnetostrictive materials, and electrorheological fluids are some examples of currently available smart materials.

Piezoelectricity stems from the Greek word *piezo* for pressure. It follows that a piezoelectric material develops a

potential across its boundaries when subjected to a mechanical stress (or pressure), and vice versa, when an electric field is applied to the material, a mechanical deformation ensues. Therefore, piezoelectric materials fall in the class of smart materials. Ferroelectricity is a subgroup of piezoelectricity, where a spontaneous polarization exists that can be reoriented by applying an ac electric field.

Definition and History

Piezoelectricity is a linear effect that is related to the microscopic structure of a solid. Some ceramic materials become electrically polarized when they are strained; this linear and reversible phenomenon is referred to as *the direct piezoelectric effect*. The direct piezoelectric effect is always accompanied by *the converse piezoelectric effect* where a solid becomes strained when placed in an electric field. The microscopic origin of the piezoelectric effect is the displacement of ionic charges within a crystal structure. In the absence of external strain, the charge distribution within the crystal is symmetrical, and the net electric dipole moment is zero. However when an external stress is applied, the charges are displaced, and the charge distribution is no longer symmetrical. A net polarization develops and results in an internal electric field. A material can be piezoelectric only if the unit cell has no center of inversion.

Piezoelectricity is a property of a group of materials that was discovered in 1880 by Pierre and Jacques Curie during their study of the effects of pressure on the generation of electrical charge by crystals such as quartz, tourmaline, and Rochelle salt. In 1881, the term "piezoelectricity" was first suggested by W. Hankel, and the converse effect was deduced by Lipmann from thermodynamics principles. In the next three decades, collaborations within the European scientific community established the field of piezoelectricity, and by 1910, Voigt's *Lerbuch der Kristallphysic* was published and became a standard reference work detailing the complex electromechanical relationships in piezoelectric crystals (1). However, the complexity of the science of piezoelectricity made it difficult for it to mature to an application until a few years later. Langevin et al. (2) developed a piezoelectric ultrasonic transducer during World War I. Its success opened up opportunities for piezoelectric materials in underwater applications and a host of other applications such as ultrasonic transducers, microphones, and accelerometers. In 1935, Busch and Scherrer discovered piezoelectricity in potassium dihydrogen phosphate (KDP). The KDP family was the first major family of piezoelectrics and ferroelectrics discovered.

During World War II, research in piezoelectric materials expanded to the United State, the Soviet Union, and Japan. Until then, limited performance by these materials inhibited commercialization, but that changed when a major breakthrough came with the discovery of barium titanate and lead zirconate titanate (PZT) in the 1940s and 1950s, respectively. These families of materials exhibited very high dielectric and piezoelectric properties. Furthermore, they offered the possibility of tailoring their behavior to specific responses and applications by using dopants. To date, PZT is one of the most widely used piezoelectric materials. Most commercially available ceramics (such

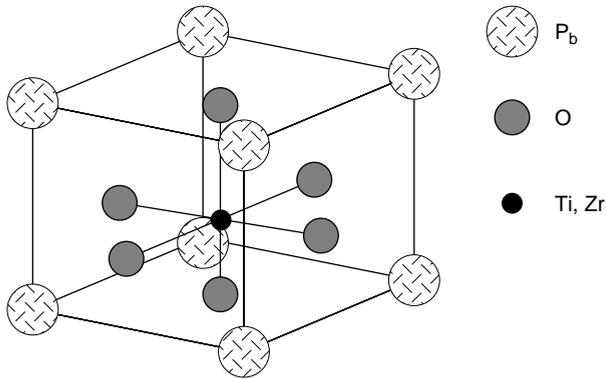


Figure 1. Perovskite structure.

as barium titanate and PZT) are based on the perovskite structure (Fig. 1). The perovskite structure (ABO_3) is the simplest arrangement where the corner-sharing oxygen octahedra are linked together in a regular cubic array; smaller cations (Ti, Zr, Sn, Nb, etc.) occupy the central octahedral B site, and larger cations (Pb, Ba, Sr, Ca, Na, etc.) fill the interstices between octahedra in the larger A site. Compounds such as $BaTiO_3$, $PbTiO_3$, $PbZrO_3$, $NaNbO_3$, and $KNbO_3$ have been studied at length and their high-temperature ferroelectric and antiferroelectric phases have been extensively exploited. This structure also allows multiple substitutions at the A site and B site that result in a number of useful though more complex compounds such as $(Ba,Sr)TiO_3$, $(Pb,Sr)(Zr,Ti)O_3$, $Pb(Fe,Ta)O_3$, and $(KBi)TiO_3$.

Starting around 1965, several Japanese companies focused on developing new processes and applications and opening new commercial markets for piezoelectric devices. The success of the Japanese effort attracted other nations, and today the needs and uses extend from medical applications to the communications field to military applications and the automotive field.

A review of the early history of piezoelectricity is found in the work of Cady (3), and in 1971, Jaffe et al. published the book *Piezoelectric Ceramics* (4) that is still one of the most referenced works on piezoelectricity.

Piezoelectric Ceramic Processing

The fabrication of most bulk piezoelectric ceramics starts with powder preparation. The powder is then pressed to the required shapes and sizes, and the green shapes are in turn processed into mechanically strong and dense ceramics. The more important processes that influence the product characteristics and properties are powder preparation, powder calcining and sintering. The next steps are machining, electroding, and poling, the application of a dc field to orient the dipoles and induce piezoelectricity.

The most common powder preparation is the mixed-oxide. In this process, powder is prepared from the appropriate stoichiometric mixture of the constituent oxides. Lead oxide, titanium oxide, and zirconium oxide are the main compounds for, lead zirconate titanate (PZT). Depending on the application, various dopants are used to tailor the properties of interest. PZT ceramics are rarely used

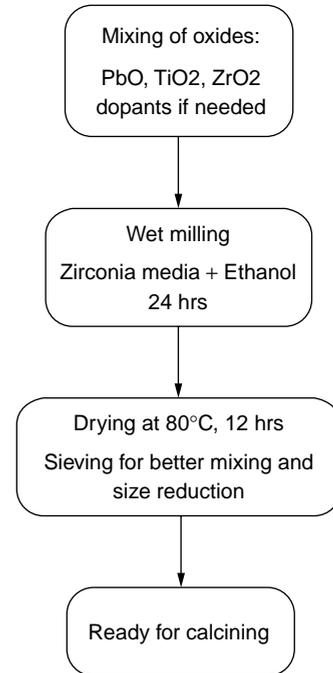


Figure 2. Mixed-oxide route for preparing PZT.

without adding of dopants to modify some of their properties. A-site additives tend to lower the dissipation factor, which affects heat generation, but also lower the piezoelectric coefficients; for this reason they are used mostly in ultrasonics and other high-frequency applications. B-site dopants increase the piezoelectric coefficients but also increase the dielectric constant and loss. B-site doped ceramics used are as actuators in vibrational and noise control, benders, and optical positioning applications.

Figure 2 shows a flowchart of the mixed-oxide route for making PZT ceramics. The powders can be mixed by dry ball milling or wet ball milling; both methods have advantages and disadvantages: wet ball milling is faster than dry milling; however, the disadvantage is the added step of liquid removal. The most common method for making PZT ceramics is wet ball milling; ethanol and stabilized zirconia media are added for wet milling. A vibratory mill may be used rather than a conventional ball mill; Herner (5) showed that this process reduces the risk of contamination by the balls and the jar. Zirconia media are used to reduce the contamination risks further. Calcination is a very crucial step in processing PZT ceramics; it is important for crystallization to be complete because the perovskite phase forms during this step. The goals are to remove any organics, water, or other volatiles left after mixing; to react the oxides to form the desired phase composition before the ceramic is processed into useful devices; and to reduce volume shrinkage and allow for better homogeneity during and after sintering.

After calcining, a binder is added to the powder, and then the mixture is shaped usually by dry pressing in a die for simple shapes, or extrusion, or casting for more complicated bodies. Next, the shapes are sintered—placed in an oven for binder burnout and densification.

The major problem in sintering a PZT ceramic is the volatility of PbO at about 800°C. To minimize this problem, the PZT samples are sintered in the presence of a lead source, such as PbZrO₃, and placed in closed crucibles. Saturation of the sintering atmosphere with PbO minimizes lead loss from the PZT bodies. Sintering can now be carried out at temperatures varying between 1200–1300°C. Despite precautions, usually 2–3% of the initial lead content is lost.

After cutting and machining into desired shapes, electrodes are applied, and a strong dc field is used to orient the domains in the polycrystalline ceramic. Dc poling can be done at room temperature or at higher temperatures, depending on the material and the composition. The poling process only partially aligns the dipoles in a polycrystalline ceramic, and the resulting polarization is lower than that of single crystals.

This processing technique presents many uncertainties; the existence of a wide number of other fabrication techniques is an indication that there is a great need for the production of reliable PZT ceramics whose properties and microstructure are optimal. One problem often encountered is deviation from stoichiometry. This problem is often due to impurities in the raw materials as well as the lead loss during sintering, and invariably results in substantial alterations of the PZT properties. As a result, the elastic properties can vary as much as 5%, the piezoelectric properties 10%, and the dielectric properties 20% within the same batch (6). The piezoelectric and dielectric properties generally suffer also if there is any lack of homogeneity from poor mixing. It is important then that the constituent oxides be intimately mixed. In the method described before, however, the constituents are solid solutions and it has been shown that intimate mixing of solid solutions is difficult if not impossible. More information on the preparation of piezoelectric ceramics can be found in Jaffe et al. (4), and Moulson and Herbert (7). Other processing methods, including hydrothermal processing and coprecipitation methods, are described in (8–10). Noted that there has been a great deal of development in powder processing, shaping, and sintering (11–13) that has resulted in further expanding the application of piezoelectric ceramics.

Ferroelectricity

Some piezoelectric materials are also ferroelectric. A ferroelectric material possesses spontaneous polarization whose direction can be reversed by applying a realizable electric field across some temperature range. Most ferroelectric materials have a Curie temperature T_c below which they are polar and above which they are not. The dielectric permittivity often has a peak at T_c and linearly decreases above it according to the Curie–Weiss law (4,7). The very large permittivity values that are characteristic of ferroelectric materials are greatly exploited in many applications, most widely in the multilayer-capacitor industry.

Applying a large alternating electric field reverses the polarization, and this gives rise to the ferroelectric

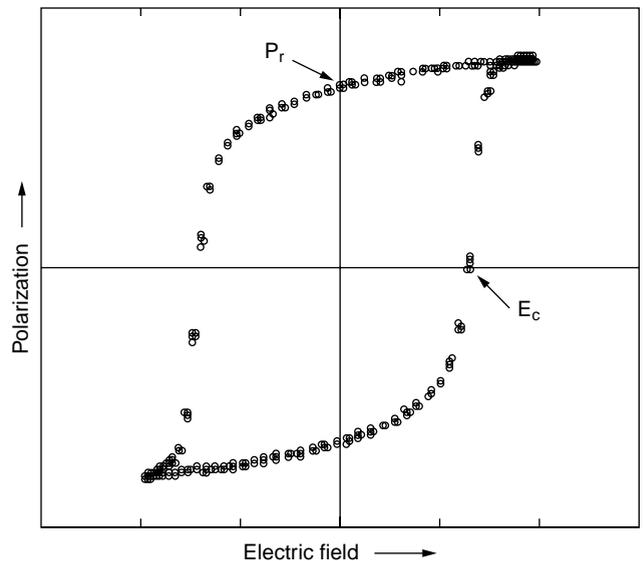


Figure 3. P - E hysteresis loop of a poled piezoelectric ceramic.

hysteretic loop that relates polarization P to applied electric field E . A typical field-polarization loop is illustrated in Fig. 3. For large signals, both the electric displacement D and the polarization P are nonlinear functions of the field E . They are related to each other through the linear equation

$$D_i = P_i + \varepsilon_0 E_i, \quad (1)$$

where ε_0 is the permittivity of free space ($= 8.85 \times 10^{-12}$ C/Vm). The second term in Eq. (1) is negligible for most ferroelectric ceramics, and a D - E loop and P - E loop become interchangeable. Two key characteristics of the P - E loop are the coercive field E_c and the remanent polarization P_r , both defined by analogy to ferromagnetic materials. E_c is the field at which polarization is zero. P_r is the value of the polarization when the electric field is zero. Once the field is switched off, the material's polarization is equal to P_r . Once the P - E loop is saturated, both P_r and E_c can be determined. A loop is said to be saturated once the values of P_r and E_c no longer vary.

Generally, the existence of the P - E loop is considered evidence toward establishing that a material is ferroelectric. A Sawyer–Tower circuit (14), or a modified version of it, is commonly used to obtain a P - E loop. An ac voltage is applied to the electroded sample; the resulting charge stored on the sample is determined by a large reference capacitor placed in series with the sample. An electrometer can be used to detect the voltage across the capacitor; by multiplying this voltage by the value of the reference capacitor, the charge across the sample results. The capacitance of the reference capacitor should be 100 to 1000 times the value of the capacitance of the sample. Note that ferroelectric hysteretic loops are both frequency- and temperature-dependent.

In addition to the P - E loop, polarization switching leads to strain–electric field hysteresis. A typical strain–field

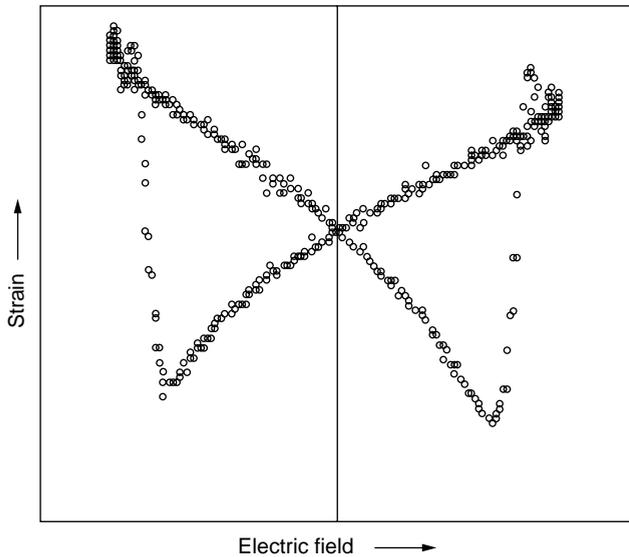


Figure 4. Butterfly loop indicating switching.

response curve is shown in Fig. 4. The shape resembles that of a butterfly, and it is often referred to as the “butterfly loop.” As the electric field is applied, the converse piezoelectric effect dictates that a strain results. As the field is increased, the strain is no longer linear with the field, as domain walls start switching.

For more sources on ferroelectricity, the reader should consult the bibliography (15–19).

PIEZOELECTRIC CONSTITUTIVE RELATIONSHIPS

An understanding of piezoelectricity begins with the structure of the material. To explain it better, let us consider a single crystallite (a small single crystal less than $100\ \mu\text{m}$ in average diameter) from a polycrystalline ceramic. This crystal is made up of negatively and positively charged atoms that occupy specific positions in a repeating unit or cell. The specific symmetry of the unit cell determines the possibility of piezoelectricity in the crystal. All crystals can be divided into 32 classes or point groups (from seven basic crystal systems: triclinic, monoclinic, orthorhombic, tetragonal, rhombohedral, hexagonal, and cubic). Of the 32 classes, 21 do not possess a center of symmetry, and 20 are piezoelectric (although one class lacks a center of symmetry, it is not piezoelectric because of the combination of other symmetry elements). The lack of a center of symmetry means that a net movement of positive and negative ions with respect to each other as a result of stress produces an electric dipole. Because the ceramic is composed of randomly oriented piezoelectric crystallites, it is inactive, that is, the effects of the individual crystals cancel each other and no discernable piezoelectricity is present. Regions of equally oriented polarization vectors are known as domains. “Poling” is a commonly used method to orient the domains by polarizing the ceramic through the application of a static electric field. Electrodes are applied to

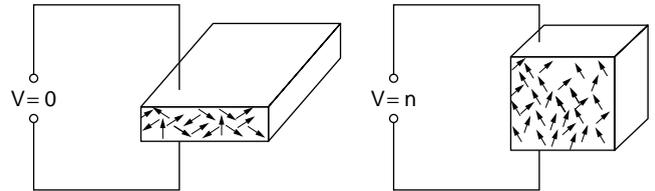


Figure 5. Poling of a piezoelectric, ferroelectric ceramic.

the ceramic, and a sufficiently high electric field is applied such that the domains rotate and switch in the direction of the electric field. Full orientation of all domains never results; however, the polycrystalline ceramic exhibits a large piezoelectric effect. During this process, there is a small expansion of the material along the poling axis and a contraction in both directions perpendicular to it (see Fig. 5). Due to large number of allowable polar directions such as near the morphotropic phase boundary (where the Zr to Ti ratio is 48 to 52), the maximum deviation of the polar axis of a grain from the average polar direction is smaller, and the reduction of polarization is minimized, assuming optimum alignment.

Constitutive Relationships

When writing the constitutive equations for a piezoelectric material, account must be taken of changes of strain and electrical displacement in three orthogonal directions caused by cross-coupling effects due to applied electrical and mechanical stresses. Tensor notation is first adopted, and the reference axes are shown in Fig. 6. The state of strain is described by a second-rank tensor S_{ij} , and the state of stress is also described by a second-rank tensor T_{kl} . The relationships relating the stress tensor to the strain tensor, compliance s_{ijkl} , and stiffness c_{ijkl} , are then fourth-rank tensors. The relationship between the electric field E_j (first-rank tensor) and the electric displacement D_i (also a first-rank tensor) is the permittivity ϵ_{ij} , which is a second-rank tensor. Therefore the piezoelectric

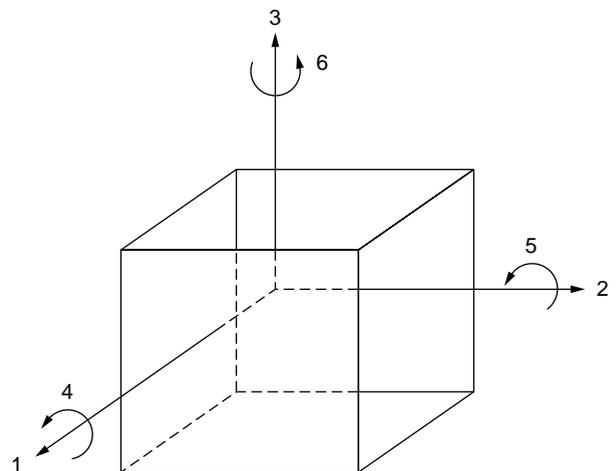


Figure 6. Reference axes.

equations are

$$D_i = \varepsilon_{ij}^T E_j + d_{ijk} T_{jk}, \quad (2)$$

$$S_{ij} = \underline{d}_{ijk} E_k + s_{ijkl}^E T_{kl}, \quad (3)$$

where d_{ijk} , \underline{d}_{ijk} are the piezoelectric constants (third-rank tensor). Superscripts T and E indicate that the dielectric constant ε_{ij} and the elastic constant s_{ijkl} are measured under conditions of constant stress and constant electric field, respectively. In general, a first-rank tensor has three components, a second-rank tensor has nine components, a third-rank tensor has 27 components, and a fourth-rank tensor has 81 components. Not all of the tensor components are independent.

Both of these relationships depend on orientation; they describe a set of equations that relate these properties in different orientations of the material. The crystal symmetry and the choice of reference axes reduce the number of independent components. A convenient way of describing them is by using axis directions, as given by Fig. 6. The convention is to define the poling direction as the 3 axis, the shear planes are indicated by the subscripts 4, 5, and 6 and are perpendicular to directions 1, 2, and 3, respectively. This simplifies the notations introduced before, where a 3-subscript tensor notation ($i, j, k = 1, 2, 3$) is replaced by a 2-subscript matrix notation ($i = 1, 2, 3$ and $j = 1, 2, 3, 4, 5, 6$), and a 2-subscript tensor notation ($i, j = 1, 2, 3$) is replaced by a 1-subscript matrix notation ($i = 1, 2, 3, 4, 5, 6$). A shear strain such as S_4 is a measure of the change of angle between the two initially orthogonal axes in the plane perpendicular to axis 1. The first subscript of the d constant gives the "electrical" direction (field or dielectric displacement), and the second gives the component of mechanical deformation or stress. The planar isotropy of poled ceramics is expressed in their piezoelectric constants by the equalities $d_{32} = d_{31}$ (an electric field parallel to the poling axis 3 interacts in the same way with axial stress along either the 2 axis or the 1 axis) and $d_{24} = d_{15}$ (an electric field parallel to the 2 axis interacts in the same way with a shear in the 2,3 plane as a field along the 1 axis with a shear in the 1,3 plane). Similar relationships hold for the elastic constants because of isotropy in the plane perpendicular to the polar axis.

Property Matrix for a Poled Piezoelectric Ceramic

A piezoelectric ceramic has only one type of piezoelectric matrix, regardless of the symmetry of the constituent crystals. The ceramic is initially isotropic. This isotropy is destroyed in the poling direction. The material is transversely isotropic in the directions perpendicular to the poling direction. The symmetry elements are an axis of rotation of infinite order in the direction of poling and an infinite set of planes parallel to the polar axis as reflection planes. In crystallographic notation, this symmetry is described as ∞mm and is equivalent to the hexagonal polar crystal class, 6 mm.

The elastic, dielectric, and piezoelectric matrices for the cylindrical symmetry of poled PZT are shown in the following equations. Matrices analogous to the piezoelectric also apply to other piezoelectric constants such as g_j (described

in the next section).

$$\begin{array}{cccccc} s_{11} & s_{12} & s_{13} & 0 & 0 & 0 \\ s_{12} & s_{11} & s_{13} & 0 & 0 & 0 \\ s_{13} & s_{13} & s_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & s_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & s_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & 2(s_{11} - s_{12}) \end{array} \quad (4)$$

$$\begin{array}{ccc} \varepsilon_1 & 0 & 0 \\ 0 & \varepsilon_1 & 0 \\ 0 & 0 & \varepsilon_3 \end{array} \quad (5)$$

$$\begin{array}{cccccc} 0 & 0 & 0 & 0 & d_{15} & 0 \\ 0 & 0 & 0 & d_{15} & 0 & 0 \\ d_{31} & d_{31} & d_{33} & 0 & 0 & 0 \end{array} \quad (6)$$

For the symmetry of poled ceramics then, general Equations (1) and (2) are replaced by the these specific equations:

$$D_1 = \varepsilon_1 E_1 + d_{15} T_5, \quad (7)$$

$$D_2 = \varepsilon_2 E_2 + d_{15} T_4, \quad (8)$$

$$D_3 = \varepsilon_3 E_3 + d_{31}(T_1 + T_2) + d_{33} T_3, \quad (9)$$

$$S_1 = s_{11}^E T_1 + s_{12}^E T_2 + s_{13}^E T_3 + d_{31} E_3, \quad (10)$$

$$S_2 = s_{11}^E T_2 + s_{12}^E T_1 + s_{13}^E T_3 + d_{31} E_3, \quad (11)$$

$$S_3 = s_{13}^E (T_1 + T_2) + s_{33}^E T_3 + d_{33} E_3, \quad (12)$$

$$S_4 = s_{44}^E T_4 + d_{15} E_2, \quad (13)$$

$$S_5 = s_{44}^E T_5 + d_{15} E_1, \quad (14)$$

$$S_6 = s_{66}^E T_6. \quad (15)$$

Equations (7)–(9) relate to the direct effect, and Eqs. (10)–(15) relate to the converse effect.

PIEZOELECTRIC PARAMETERS: DEFINITIONS AND CHARACTERIZATION

The parameters that are of interest when considering the electromechanical effects of piezoelectric materials are the piezoelectric charge coefficients (d_{31} and d_{33}), the piezoelectric voltage coefficients (g_{31} and g_{33}), and the piezoelectric coupling factors (k_{31} , k_{33} , k_p , and k_t). The d coefficient is the proportionality constant between electric displacement and stress, or strain and electric field [Eqs. (2) and (3)]. High d coefficients are desirable in materials used as actuators, such as in motional and vibrational applications. The g coefficient is related to the d coefficient by the following expression:

$$d_{mi} = \varepsilon_{nm}^T g_{ni}. \quad (16)$$

where $m, n = 1, 2, 3$ and $i = 1, 2, \dots, 6$. High g coefficients are desirable in materials to be used as sensors to produce voltage in response to mechanical stress.

The piezoelectric coupling factor k is a measurement of the overall strength of the electromechanical effect. It is often defined as the square root of the ratio of electrical

energy output to the total mechanical energy input (in the direct effect) or the mechanical energy available to the total electrical energy (in the converse effect). The value of k is, of course, always less than unity because energy conversion is always incomplete.

Other important properties of PZT are the dielectric properties, namely, the dielectric constant ϵ (F/m) and the dissipation factor or loss tangent $\tan \delta$. The dielectric constant is a measure of the charge stored on an electroded material brought to a given voltage. The dielectric constant of vacuum is $\epsilon_0 = 8.85 \times 10^{-12}$ F/m. The relative dielectric constant K (often referred to as just “the dielectric constant”) is the ratio of ϵ and ϵ_0 and is dimensionless. In an ac field, the dielectric constant has both a real part and an imaginary part; the loss tangent is defined as the ratio of the imaginary part to the real part.

The values of these constants depend on the PZT composition. As an example, the d constants, g constants, and the dielectric constant for compositions near the morphotropic phase boundary show their highest values on the tetragonal side of the transition (4). Then, it is possible to tune the values of these properties for most compositions; one way to achieve this is by adding dopants to the PZT formulation.

Resonant Method and Equivalent Circuit

Resonance Method. Any body has certain characteristic frequencies at which it prefers to resonate. When excited at this resonant frequency f_r , the body will resonate freely at a greater amplitude than at other frequencies. Following this resonant frequency is an antiresonant frequency f_a , where the impedance of the body is at a maximum and the oscillatory amplitude is at a minimum. Piezoelectric ceramics are no different, and measuring these characteristic frequencies provides the means to evaluate the piezoelectric and elastic properties of the ceramic. Different modes of vibration of the ceramic, such as thickness or planar, give insight to the different constants for that mode. A typical resonance plot of impedance versus frequency for a piezoelectric ceramic near a resonance is shown in Fig. 7. Notice the resonant frequency f_r at the point of minimum impedance and the antiresonant frequency f_a at the point of maximum impedance.

At resonance, a piezoelectric element may be modeled by the equivalent circuit shown in Fig. 8. This circuit, commonly referred to as Van Dyke’s model, is recommended by the IEEE Standard on Piezoelectricity (20). An alternate model, proposed by Sherrit et al. (21), eliminates the resistance R_1 and instead represents the remaining components as complex to characterize better the losses of certain piezoelectric elements, especially polymers. All discussion following assumes Van Dyke’s model. Below f_r and above f_a , the ceramic behaves capacitively; however, between these two frequencies, the ceramic behaves inductively. This model is valid only near the resonance. Additionally, the resonance must be sufficiently isolated from other modes to eliminate the effects of any adjacent modes. To ensure that the resonance is isolated, sample geometry must be chosen carefully. Geometries suitable for measuring the different piezoelectric and elastic coefficients are presented in Table 1. Fixturing of the sample

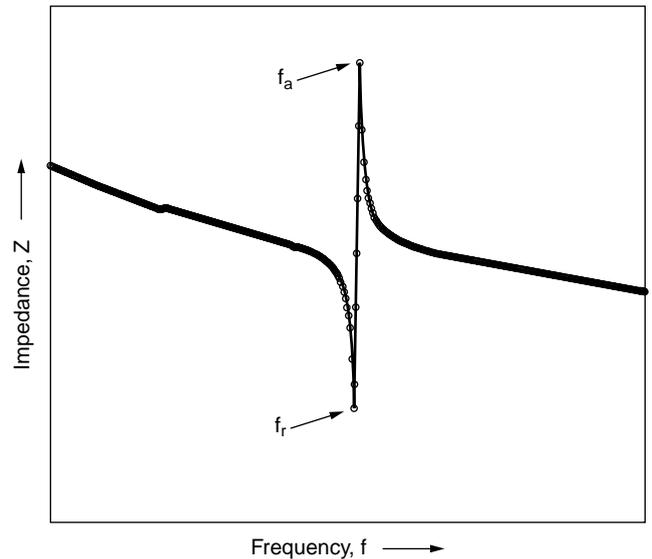


Figure 7. Impedance of a piezoelectric ceramic at resonance.

should not impose any constraints on the vibration of the ceramic. This can be accomplished by using a point holder positioned at a node of vibration. All leads should also be shielded up to the contact point, as much as possible, to avoid any stray capacitances that may arise.

Earlier literature has suggested several circuits for measuring f_r and f_a of a piezoelectric ceramic (22–25). These circuits usually consist of an oscillator for exciting the sample, a voltmeter or other device for measuring current through the circuit, and additional discrete components. To find f_r , the frequency of the oscillator is varied until the maximum current is detected through the circuit. Similarly, for f_a , the frequency of minimum current is determined. Note that there are actually six characteristic frequencies that may be identified for a particular resonance; they include f_m and f_n , the frequencies of maximum and minimum impedance; f_r and f_a , the resonant and antiresonant frequencies; and f_p and f_s , the parallel resonant frequency and series resonant frequency. IEEE Standard 177 (23) identifies these six frequencies and establishes that for many cases, including piezoelectric ceramics, one can assume that $f_m = f_s = f_r$ and $f_a = f_p = f_n$. For lossy materials, such as some piezoelectric thin films, this assumption can introduce appreciable errors, so the six frequencies should be considered separately. The magnitude of the minimum impedance Z_m may be determined by substituting an adjustable resistor into the circuit for the ceramic at the previously identified frequency and adjusting the resistance until the voltmeter reading is the same as for the ceramic. Today, fully integrated

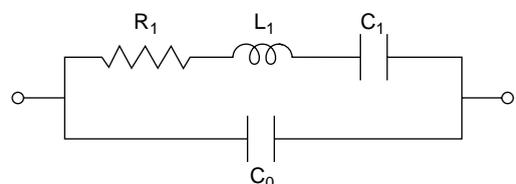
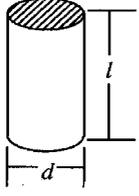
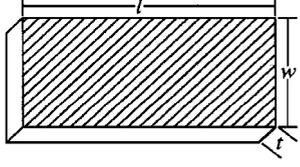
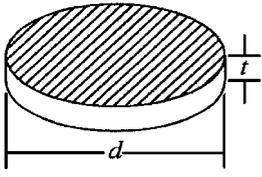


Figure 8. Resonant equivalent circuit of a piezoelectric vibrator.

Table 1. Sample Geometries for Measuring Material Properties

			
Dimensional requirements	Long, slender, length poled rod; $l > 3d$	Thin, flat plate, thickness poled; $l > 3.5 t, w$	Thin flat disc, thickness poled $d > 10 t$
Properties measured	$K_3^T, k_{33}, \tan \delta, s_{33}^D, s_{33}^E, d_{33}, g_{33}, Q_M$	$K_3^T, k_{31}, \tan \delta, s_{11}^D, s_{11}^E, d_{31}, g_{31}, Q_M$	$K_3^T, k_p, \tan \delta, Q_M$

impedance analyzers are commercially available to make this type of measurement, allow the researcher to choose an equivalent circuit model, and report the values of the discrete components of the equivalent circuit along with f_r and f_a . Commercial off-the-shelf software is also available now which can be used in conjunction with an analyzer to evaluate the impedance information and calculate the relative material properties of a piezoelectric device (26). These tools can aid the researcher in evaluating material properties, however, a basic understanding of piezoelectric behavior is an important foundation that should not be overlooked.

Measuring Material Properties. Capacitance measurements are usually carried out at 1 kHz and at low excitatory voltages (mV level). Although research has shown that capacitance and loss vary with excitation voltage and frequency (27,28), the 1-kHz, low-voltage measurement is used to determine material properties. The free relative dielectric constant K^T is defined as the ratio of the permittivity of the material to the permittivity of free space. It is calculated from

$$K^T = \frac{tC}{\epsilon_0 A}, \quad (17)$$

where t is the distance between electrodes in meters, C is the capacitance in farads, ϵ_0 is the permittivity of free space (8.85×10^{-12} F/m), and A is the area of an electrode in meters². The loss tangent, $\tan \delta$, is defined as the ratio of resistance to reactance in the parallel equivalent circuit of Fig. 9b. It is a measure of the dielectric losses in the material and therefore, also a measure of the heat generating capacity of the ceramic when operated under dynamic conditions. This is a direct measurement and is usually formed at the same conditions as the capacitance measurement.

The three most common coupling coefficients are k_p , k_{31} , and k_{33} ; where the p is for planar, and the 31 and 33 subscripts are for length extensional and thickness extensional modes. The coefficients k_{33} and k_{31} are calculated from the frequencies of minimum and maximum impedance and are given by the equations

$$k_{33}^2 = \frac{\frac{\pi}{2}}{1 + \frac{(f_a - f_r)}{f_r}} \tan \left[\frac{\frac{\pi(f_a - f_r)}{2f_r}}{1 + \frac{(f_a - f_r)}{f_r}} \right], \quad (18)$$

$$k_{31}^2 = \frac{\psi}{1 + \psi}, \quad (19)$$

Table 2. Typical Properties of Common Piezoelectric Materials

	α -Quartz	BaTiO ₃	PZT ^a	PZT ^b	LiNbO ₃	PbTiO ₃	Units	
Coupling factors	k_{33}	0.49	0.68	0.72	0.17	0.35		
	k_{31}	0.21	0.34	0.35	0.02	0.052		
	k_p	0.38	0.57	0.61	0.035			
Piezoelectric constants	d_{31}	-79 (-259)	-120 (-394)	-175 (-574)	-0.9 (-2.8)	-7.4 (-24)	10^{-12} m/volt (ft/volt)	
	d_{33}	190 (623)	285 (1970)	375 (1230)	6 (20)	47 (154)	10^{-12} m/volt (ft/volt)	
	g_{31}	-5 (-72)	-10 (-145)	-11 (-160)	-4 (-131)	-4 (-131)	10^{-3} volt.m/N (volt.ft/lbf)	
	g_{33}	11 (160)	25 (365)	25 (365)	23 (76)	26 (85)	10^{-3} volt.m/N (volt.ft/lbf)	
Free dielectric constant	K_3^T	4.6	1300	1760	29	203		
Elastic constants	s_{11}^E	13 (623)	9 (431)	12 (575)	14 (670)	6 (287)	10^{-12} m ² /N (ft ² /lbf)	
	s_{33}^E	10 (480)	10 (480)	15 (718)	19 (910)	5 (240)	10^{-12} m ² /N (ft ² /lbf)	
Dissipation factor	$\tan \delta$	7	0.4	1.8			%	
Density	ρ	2.65 (0.16)	5.7 (0.35)	7.6 (0.47)	7.7 (0.48)	4.64 (0.29)	7.12 (0.44)	10^3 kg/m ³ (lb/ft ³)
Mechanical Q	Q_M	$> 10^6$	500	550	80			
Curie temperature	T_c	130 (266)	330 (626)	360 (680)	1210 (2210)	494 (921)	$^{\circ}\text{C}$ ($^{\circ}\text{F}$)	

^a PZT—hard composition.^b PZT—soft composition.

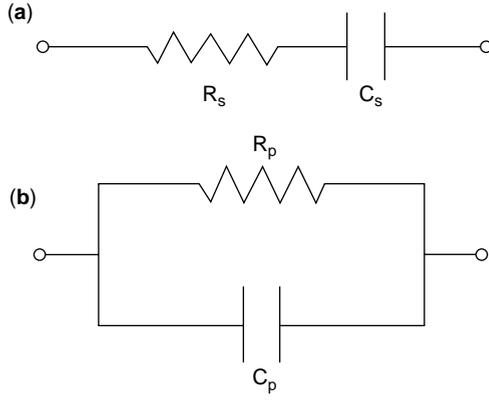


Figure 9. Nonresonant (a) series and (b) parallel equivalent circuit.

where

$$\psi = \frac{\pi}{2} \left[1 + \frac{f_a - f_r}{f_r} \right] \tan \left[\frac{\pi(f_a - f_r)}{2f_r} \right]. \quad (20)$$

The planar coupling coefficient k_p is defined for thin disks and can be approximated by

$$k_p \approx \frac{f_a^2 - f_r^2}{f_r^2}. \quad (21)$$

Elastic compliance is the ratio of a material's change in dimensions (strain) relative to an externally applied load (stress). This is the inverse of Young's modulus. For a piezoelectric material, the compliance depends on whether the strain is parallel or perpendicular to the poling axis and the electrical boundary conditions. Elastic constants are calculated from the following equations:

$$s_{33}^D = \frac{1}{4\rho f_a^2 l^2}, \quad (22)$$

$$s_{33}^E = \frac{s_{33}^D}{1 - k_{33}^2}, \quad (23)$$

$$s_{11}^E = \frac{1}{4\rho f_r^2 w^2}, \quad (24)$$

$$s_{11}^D = s_{11}^E (1 - k_{31}^2), \quad (25)$$

where ρ is the density of the material in kg/m^3 , l is the distance between electrodes, and w is the width of the ceramic. The superscripts D and E stand for constant electric displacement (open circuit) and constant electric field (short circuit), respectively.

The d_{ij} piezoelectric constants, which relate the applied electric field to the strain, can be calculated from the coupling, the elastic coefficients, and the dielectric constant:

$$d_{33} = k_{33} \sqrt{\varepsilon_0 K_3^T s_{33}^E}, \quad (26)$$

$$d_{31} = k_{31} \sqrt{\varepsilon_0 K_3^T s_{11}^E}. \quad (27)$$

The g_{ij} piezoelectric constants are related to the d_{ij} coefficients by the following equations:

$$g_{33} = \frac{d_{33}}{\varepsilon_0 K_3^T}, \quad (28)$$

$$g_{31} = \frac{d_{31}}{\varepsilon_0 K_3^T}. \quad (29)$$

Noted that the piezoelectric coefficients calculated are valid only at frequencies well below resonance and do not account for any nonlinear behavior of the ceramic. They do not depend on the dimensions of the material; however, they vary with the degree of polarization of the ceramic. They also do not provide the sign of the coefficient, which must be determined by direct measurements. The mechanical Q_M , the ratio of reactance to resistance in the series equivalent circuit of Fig. 9a is given by

$$Q_M = \frac{1}{2\pi f_r Z_m C_0} \left(\frac{f_a^2}{f_a^2 - f_r^2} \right). \quad (30)$$

Direct Methods

Direct measurements of the piezoelectric constants are possible and have been used to quantify the direct and converse effects in ceramic samples. Direct methods are also used to investigate the behavior of a ceramic in regard to hysteresis, nonlinearity, frequency response, aging, thermal behavior, and other characteristics that are not resolved by previous methods. These methods typically apply a known input to the ceramic, either an electric field or a force, and record the corresponding output, either a deformation or a charge under various conditions. These methods are in contrast to the bulk material characterization using the electrical resonance techniques described before. Many times, direct measurements are carried out on a ceramic that has been configured as a sensor or actuator. Typical processing may include electroding, laminating, applying preload, mounting, and other assembly procedures to adapt the material effectively for use as a sensor or actuator. These measurements aid the researcher in modeling the behavior of the piezoelectric device and allow efficient integration of the devices into real-world applications.

Displacements of piezoelectric actuators are measured to determine the magnitude and sign of the relationship between the applied electric field and the strain developed, that is, the converse effect. For a PZT wafer, this corresponds to the d_{ij} coefficient; however, for bending type actuators, this relationship does not correlate directly with any of the measured properties for out-of-plane bending using the resonance techniques. Based on Eqs. (10)–(15), it can be seen that when the ceramic is free to expand ($T_i = 0$), then the strain is a function only of the product of the applied field E_i and the d_{ij} coefficient

$$S_j = d_{ij} E_i. \quad (31)$$

Careful attention must be paid to the boundary conditions of the ceramic to ensure that this assumption is valid. In a plot of the strain as a function of applied field, the

slope yields an average value of d_{ij} . Typically, these measurements are made by using a noncontacting displacement transducer (29) to reduce the effects of loading on the actuator. Laser-based and other optical or capacitive displacement measurement techniques are most commonly used (30–32). Displacements may range from submicron levels for single PZT wafers to the centimeter level for bending type actuators. For very small displacements, an optical-lever type measurement system or interferometric techniques (33) have been used to resolve the displacement of the ceramic. Direct application of either foil or optical strain gages has also been used for measuring the actuator strain. These measurements may be either static or dynamic, depending on the measurement system and the intended application of the ceramic. If dynamic measurements are made, excitatory frequencies should be at least an order of magnitude less than any resonant frequency of the device to ensure linear behavior and boundary conditions suitable for the intended measurement.

Another direct method used to measure piezoelectric constants is based on the direct piezoelectric effect (22,34). Here, a known load is either applied to or lifted off a ceramic at rest. The resulting charge, which accumulates on the electrodes, is then measured as a voltage across a capacitor in parallel with the ceramic, or the current from the ceramic can be integrated directly. If E_i is 0 (short circuit), then Eq. (2) reduces to

$$D_i = d_{ij} T_j. \quad (32)$$

Knowing the applied stress and measuring the electric displacement, the appropriate d_{ij} coefficient can be found.

If a piezoelectric ceramic is immersed in a liquid and the pressure of the liquid is varied, then the piezoelectric coefficient d_h can be quantified by measuring the voltage on a large capacitor in parallel with the ceramic. This coefficient represents the response of the ceramic to hydrostatic pressure applied equally to all axes. Convention has dictated that electrodes are perpendicular to the 3 direction for the d_h coefficient. The d_h coefficient is related to the other d

coefficients for a ceramic by the equation,

$$d_h = d_{33} + 2d_{31}. \quad (33)$$

The frequency response of the device may be obtained by varying the frequency of the excitatory voltage to the ceramic while measuring the displacement. Typical resonant frequencies of bulk ceramic material are in the kilohertz to megahertz range depending on the mode of vibration, whereas resonant frequencies of bender types (unimorph or bimorph) may be less than 100 Hz. For maximum strain, a piezoelectric actuator can be excited at its natural frequency; however, this nonlinear behavior must be taken into account if the actuator is to be used across a range of frequencies. Careful attention must also be paid to the instrumentation system's dynamic response in both amplitude and phase distortions, when making dynamic measurements. Measurement systems have their own frequency response characteristics which must be separated from the response of the ceramic under test.

Hysteresis is a phenomenon that is present in all piezoelectric materials. Hysteretic behavior is due to the lossy nature of the ceramic where the current trails the applied voltage by an angle α related to the loss tangent of the material. For actuators, this means that the absolute displacement depends on the excitatory voltage and frequency and also on whether the voltage is increasing or decreasing. To characterize the amount of hysteresis in a ceramic, a sinusoidal voltage is applied to the device, and the displacement is recorded. By plotting the displacement versus driving voltage, as shown in Fig. 10, the hysteretic behavior of the ceramic can be observed. The amount of hysteresis (usually expressed in percent) is defined as the largest difference between the maximum and minimum displacement for any voltage divided by the total displacement. Of note in Fig. 10 is the fact that, as the peak voltage is increased, the amount of hysteresis also increases for any given voltage.

Generally, piezoelectric ceramic actuators exhibit a decrease in their displacement for a given excitatory

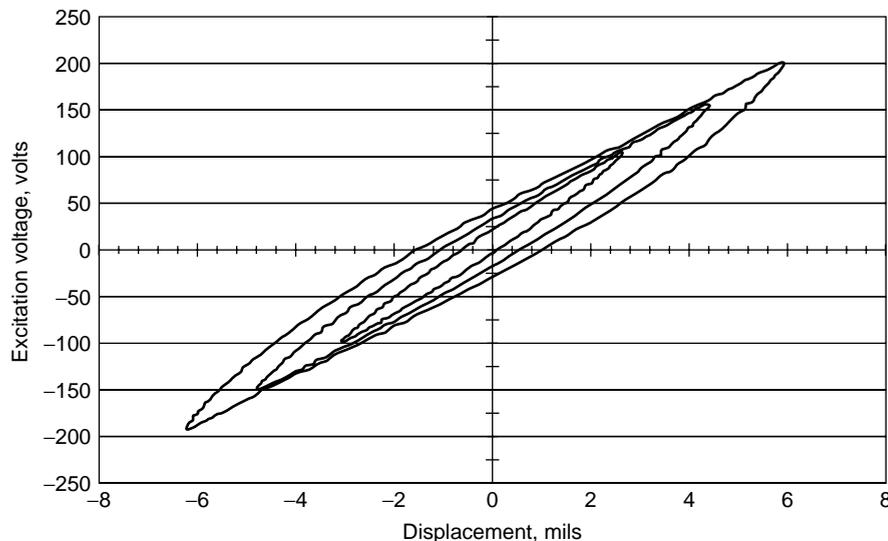


Figure 10. Strain hysteresis of a piezoelectric ceramic unimorph.

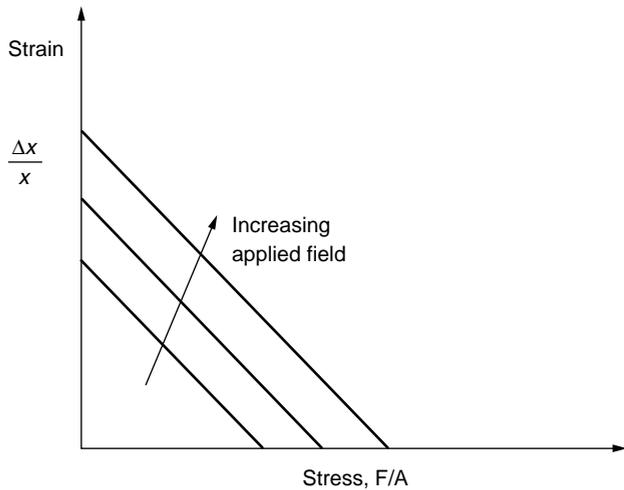


Figure 11. Typical stress/strain relationship for a piezoelectric ceramic.

voltage, as they are loaded. This relationship can be seen in Eq. (3), when $T_k \neq 0$. As the load is increased, the displacement eventually reaches zero, and the actuator provides only a force output. This force is known as the blocked force, and it is the maximum amount of force that the actuator can produce at that voltage. To characterize this relationship, the actuator is loaded with a load less than the blocked force, and the displacement is measured. If the load is varied, then the force/displacement relationship can be determined (Fig. 11). To determine the blocked force, the actuator must be rigidly held so as not to deform, and the force output is measured by using a load cell or other force-measuring device. Because the displacement of some piezoelectric actuators is quite small, this measurement can be difficult. The blocked force F_B can, alternatively be calculated by the equations

$$F_B = \frac{E_3 d_{33} w l}{s_{33}^E}, \quad (34)$$

$$F_B = \frac{E_3 d_{31} w t}{s_{11}^E}, \quad (35)$$

where E is the applied field; and l , w , and t are the length, width, and thickness of the ceramic, respectively. Equation (34) applies to the thickness extensional mode and Eq. (35) applies to the length extensional mode.

Actuators that have greater displacements lend themselves better to blocked force measurement (such as domed prestressed actuators or unimorph/bimorph type actuators). The blocked force may also be determined by extrapolating the force–displacement relationship to zero displacement if a true blocked force measurement is not practical. In most applications, actuators operate somewhere between the free (unloaded) state and the completely constrained state.

It has been previously reported that a constant preload applied to a piezoelectric actuator can actually increase the displacement of the ceramic, compared to an unloaded specimen (34–36). This may result from simply reducing

the compliance or mechanical play in a PZT assembly or may be a real increase in the d coefficient. This effect reaches a maximum and then starts to cause a decrease in the coefficient as the preload is increased up to the blocked force.

Temperature effects on the piezoelectric coefficients of ceramics may also be evaluated. Usually, ceramics must be used well below their Curie temperatures to maintain polarization. The respective Curie temperatures for hard and soft PZTs are of the order of 360°C (680°F) and 330°C (626°F). For operation at lower temperatures, even down to cryogenic levels, the piezoelectric coefficients generally decrease as temperature decreases. This effect can be experimentally quantified through either resonance techniques or direct measurements across the desired temperature range (34).

The power required to drive a piezoelectric ceramic can be calculated from the following equation:

$$P = 2\pi f C \tan \delta V_{\text{rms}}^2, \quad (36)$$

when the ceramic is modeled as in Fig. 9a where f is the driving frequency, C the capacitance, $\tan \delta$ the loss tangent, and V_{rms} the root-mean-square of the excitatory voltage. Typically, it is assumed that both the capacitance and loss tangent of the ceramic are constant when using Eq. (36). Doing so can lead to large errors when estimating the power consumption of a ceramic. To avoid these errors, either the voltage and current supplied to the ceramic should be measured to provide the power consumption directly, or the variation of capacitance and loss of the material as functions of applied field and frequency must be quantified and incorporated into Eq. (36) (28). A number of researchers have investigated the power consumption characteristics of PZT actuators used to excite a host structure (27,37,38) and found a coupling between the mechanical motion of the structure and the electrical characteristics of the piezoelectric actuator. Research by Brennan and McGowan (27) shows that the power consumption of piezoelectric materials used for active vibrational control is independent of the coupling effects of the host structure when the structure is *completely controlled*. From these findings, they conclude that the power requirements of the piezoelectric actuator depend only on its geometry and material properties and the driving voltage and frequency of the control signal. Research (23) has indicated that both capacitance and resistance are nonlinear functions of the peak amplitude and frequency of the excitatory voltage.

In time, piezoelectric effects imparted through poling degrade. Aging of piezoelectric ceramics, like many other materials, is logarithmic with time. In most ceramics, a initial performance levels can be recovered by simply re-poling the sample. Aging levels depend on the composition; the coupling coefficient of a soft PZT composition ages at a rate of -1% per time decade versus -2% for a hard composition. Degradation of piezoelectric behavior also depends on the level of stress to which the ceramic is subjected. High stress levels can lead to switching of the polarization and eventually depoling of the ceramic. High stresses also induce microcracking, which can lead to ceramic breakage and failure.

The methods outlined before can be used either separately or together to investigate the dielectric, piezoelectric, and elastic properties of a ceramic. Resonant techniques, which are the preferred method of measurement in the IEEE standard, are easy to implement, and the associated frequencies can be measured accurately. There is even commercially available hardware and software to assist in these measurements and the evaluation of material properties. However, these methods do not explain any nonlinear behavior that is present in the ceramic. Dependence of material properties on the frequency and amplitude of the applied voltage are among these nonlinear effects. Direct measurements of the piezoelectric constants can quantify the material properties under different operating conditions and provide insight beyond the standard linear behavior predicted by resonance techniques. These methods though, are usually more rigorous in their requirements for material handling and instrumentation.

Modeling of Piezoelectric Ceramics

There are a host of applications for piezoelectric materials, and although they have been studied for more than a century, potential for improvement and innovation still persists. Modeling of piezoelectric ceramics and their properties affords a way to accelerate materials improvement and aid in device design and development. For that reason, we would be remiss not to mention it, albeit briefly. This introduction is in no way meant as a comprehensive review of the vast area of modeling of piezoelectricity; however the references cited provide a good starting place. Care must be taken to differentiate between modeling the piezoelectric material and modeling a "piezoelectric structure;" often, a piezoelectric material is laminated or bonded to a substrate as a unimorph or bimorph.

A number of researchers experimented with commercial packages such as ANSYS[®](39). However, these commercial packages have limitations. Other groups have written their own codes and achieved varying degrees of success (40–43). Finite element schemes that combined piezoelectric and acoustic elements proved useful in characterizing the electromechanical behavior of piezoelectric transducers (44). Most of these schemes are restricted because they assume linearity of the coefficients. Pérez et al. expanded on these models by including nonlinear elements in the equivalent circuit (45). Models of the nonlinear hysteretic behavior of piezoelectric materials are abundant in the literature and can be categorized on the basis of the dimensional scale they probe. Microscopic models stem primarily from energy relationships applied at the atomic or molecular level (46). Macroscopic models (47–49) often use empirical relationships to describe the behavior of the bulk material. Both methods have their advantages and disadvantages; microscopic models require a great number of parameters, often not available, and macroscopic models do not consider the underlying physics. A number of authors proposed a third approach, a mesoscale or semimicroscopic model that combines the advantages of the previous methods, thus allowing a better way to model hysteretic behavior. This is accomplished by starting out from energy principles applied at the microscopic level, then

using a relatively small number of parameters to simulate the behavior of bulk ceramics (50,51).

CONCLUSION

Characterization of the elastic, dielectric and electromechanical properties of piezoelectric ceramics is crucial for several reasons. First, investigations of the material properties provide a link between the manufacturing process and ceramic performance. This enables the developer of the materials to adjust the manufacturing process of the ceramic to produce tailored materials. Second, the engineer can investigate prospective materials for applicability to a specific need. Material parameters obtained through characterization can also be used to develop and validate analytical models of the ceramics. Insights gained through characterization have led to many new devices and uses. For example, investigation of the hydrostatic coefficients of PZT and those of the piezoelectric polymer polyvinylidene fluoride (PVDF) identified the product of d_h and g_h as a figure of merit and led to composite research to combine both materials in a superior device that fits underwater and hydrophone applications better. More than a century after their discovery, piezoelectric ceramics have become commercially viable. Researchers continue diligently to uncover novel ways to characterize the complex electromechanical properties, and as they do so, new processing methods and applications are revealed. Recently, as an example, researchers at MIT successfully grew piezoelectric single crystals (52) that opened opportunities for newer applications. Published articles on composite processing and characterization have also become more abundant. Without question, piezoelectric ceramics have secured a permanent place in the field of material science and engineering.

ACKNOWLEDGMENT

The authors express their sincere appreciation to Dr. Jeffrey A. Hinkley (NASA Langley Research Center) for his review of the manuscript and his helpful comments.

BIBLIOGRAPHY

1. V.W. Voigt, *Lehrbuch der Kristallphysik*. B.G. Teubner, Leipzig, Berlin, 1910.
2. *Œuvres Scientifiques de Paul Langevin*. Centre National de la recherche scientifique, 1950.
3. W.G. Cady, *Piezoelectricity; An Introduction to the Theory and Applications of Electromechanical Phenomena in Crystals*. Dover, NY, 1964.
4. B. Jaffe, W.R. Cook, Jr., and H. Jaffe, *Piezoelectric Ceramics*. Academic Press, London, NY, 1971.
5. S.B. Herner, Thesis, Pennsylvania State University, 1993.
6. Measurement of Properties of Piezoelectric Ceramics, Sensor Technology Limited, BM91-309, Manufacturer Handbook.
7. A.J. Moulson and J.M. Herbert, *Electroceramics: Materials, Properties, Applications*. Chapman and Hall, London, NY, 1992.

8. Q.F. Zhou, H.L.W. Chan, and C.L. Choy, *J. Mater. Process. Technol.* **63**, (1–3): 281–285 (1997).
9. Z. Ounaies, Thesis. Pennsylvania State University, 1995.
10. J.P. Witham, *Hydrothermal Preparation and Fabrication of Lead Zirconate Titanate (PZT) Ceramics*. the Pennsylvania State University, 1993.
11. C. Near, G. Shmidt, K. McNeal, and R. Gentilman, *SPIE 5th Annu. Symp. Smart Struct. Mater. Proc.*, San Diego, 1998, Vol. 3326, pp. 323–331.
12. L.M. Levinson, *Electronic Ceramics: Properties, Devices, and Applications*. General Electric Company, Schenectady, NY, 1988.
13. ACX Controls Expert, www.ACX.com.
14. C.B. Sawyer and C.H. Tower, *Phys. Rev.* **35**: 269–273 (1930).
15. M.E. Lines and A.M. Glass, *Principles and Applications of Ferroelectrics and Related Materials*. Clarendon, Oxford, 1979.
16. Y. Xu, *Ferroelectric Materials and Their Applications*. North-Holland, Amsterdam, 1991.
17. G.A. Smolenskii, V.A. Bokov, V.A. Isupov, N.N. Krainik, R.E. Pasynkov, and A.I. Sokolov, *Ferroelectrics and Related Materials*. Gordon and Breach, NY, 1984.
18. J.C. Burfoot, *Ferroelectrics: An Introduction to the Physical Principles*. Van Nostrand, London, 1967.
19. ANSI/IEEE Std. 180-1986, *IEEE Standard Definitions of Primary Ferroelectric Terms*. The Institute of Electrical and Electronics Engineers, Inc., NY, 1986.
20. *IEEE Standard on Piezoelectricity* (IEEE Standard 176-1987). Institute of Electrical and Electronic Engineers, NY, 1987.
21. S. Sherrit, H.D. Wiederick, B.K. Mukherjee, *IEEE Ultrasonics Symp. Proc.*, Ontario, Canada, 1997, pp. 931–935.
22. W.P. Mason and H. Jaffe, *Proc. IRE* **42**: 921–930 (1954).
23. *IEEE Standard Definitions and Methods of Measurement for Piezoelectric Vibrators* (IEEE Standard 177-1966). Institute of Electrical and Electronic Engineers, NY, 1966.
24. K.W. Kwok, H.L. Chan, and C.L. Choy, *IEEE Trans. Ultrasonics Ferroelectrics Frequency Control* **44**: (4): 733–742 (1997).
25. Procedures for Measuring Properties of Piezoelectric Ceramics, Morgan Matroc Inc., Technical Publication TP-234.
26. Piezoelectric Resonance Analysis Program, TASI Technical Software (1996).
27. M.C. Brennan and A.M. McGowan, *SPIE 4th Annu. Symp. Smart Struct. Mater. Proc.* San Diego, 1997, Vol. 3039, pp. 660–669.
28. T. Jordan, Z. Ounaies, J. Tripp, and P. Tchong, *Mater. Res. Soc. Symp. Proc.* **604**: 203–208 (2000).
29. J.T. Dawley, G. Teowee, B.J.J. Zelinski, and D.R. Uhlmann, Piezoelectric Characterization of Bulk and Thin Film Ferroelectric Materials Using Fiber Optics, MTI Instruments Inc., Application Note.
30. W.Y. Pan, H. Wang, and L.E. Cross, *Jpn. J. Appl. Phys.* **29**: 1570 (1990).
31. Q.M. Zhang, W.Y. Pan, and L.E. Cross, *J. Appl. Phys.* **63**: 2429 (1988).
32. T.L. Jordan, Z. Ounaies, and T.L. Turner, *Mater. Res. Soc. Symp. Proc.* **459**: 231 (1997).
33. H.D. Wiederick, S. Sherrit, R.B. Stimpson, and B.K. Mukherjee, *8th Euro. Meet. Ferroelectricity Proc.*, Nijmegen, The Netherlands, July 4–8, 1995.
34. S. Sherrit, R.B. Stimpson, H.D. Wiederick, B.K. Mukherjee, *SPIE Far East Pac. Rim Symp. Smart Mater.* Bangalore, India, December 1996, Vol. 3321, pp. 74–81.
35. H.H.A. Krueger and D. Berlincourt, *J. Acoust. Soc. Am.* **33**: (10) 1339–1344.
36. The Piezo Book, Burleigh Instruments Inc., Piezo 330 692, Manufacturer Handbook. pages.
37. S.C. Stein, C. Liang, and C.A. Rogers, *J. Acoust. Soc. Am.* **96**: 1598–1604 (1994).
38. C. Liang, F.P. Sun, and C.A. Rogers, *J. Intelligent Mater. Syst. Struct.* **5**: 12–20 (1994).
39. ANSYS/Multiphysics software, ANSYS, Inc., Canonsburg, PA,
40. Y. Kagawa and T. Yamabuchi, *IEEE Trans. Sonics Ultrasonics SU-21*: 275–283 (1974).
41. R. Lerch, *Proc. 9th Conf. Acoust.*, Budapest, Hungary, May 1988.
42. D.F. Ostergaard and T.P. Pawlak, *Proc. IEEE Ultrasonics Symp.*, Williamsburg, VA, 1986, pp. 639–642.
43. R. Lerch, *IEEE Trans. Ultrasonics Ferroelectrics Frequency Control* **37** (2): 233–247 (1990).
44. R. Simkovics, H. Landes, M. Kaltenbacher, and R. Lerch, *IEEE Ultrasonics Symp.*, 1999, pp. 1057–1060.
45. R. Perez, E. Minguella, and J.A. Gorri, *Proc. 11th IEEE Int. Symp. Appl. Ferroelectrics (ISAF)*, 1998, pp. 247–250.
46. M. Omura, H. Adachi, and Y. Ishibashi, *Jpn. J. Appl. Phys.* **30**: 2384–2387 (1991).
47. W.S. Galinaitis and R.C. Rogers, *SPIE Smart Struct. Mater.: Math. Control Smart Mater.*, San Diego, 1997.
48. P.Ge, and M. Jouaneh, *Precision Eng.* **17**: 211–221 (1995).
49. X.D. Zhang and C.A. Rogers, *J. Intelligent Mater. Syst. Struct.* **4**: 307–316 (1993).
50. W. Chen and C.S. Lynch, *J. Intelligent Mater. Syst. Struct.* **9**: 427–431 (1998).
51. R.C. Smith and Z. Ounaies, *J. Intelligent Mater. Syst. Struct.* **11** (1): 62, (2000).
52. A.N. Soukhovjak, H. Wang, G.W. Farrey, and Y.-M. Chiang, *J. Phys. Chem. Solids* **61** (2): 301–304 (2000).

CHEMICAL INDICATING DEVICES

CHRISTOPHER O. ORIAKHI
Hewlett-Packard Company
Corvallis, OR

INTRODUCTION

Most people remember a chemical indicator from their high school chemistry. This kind of indicator is a material that changes color to signify the end point of a titration or to provide a relative indication of the acidity or alkalinity of a chemical substance. The use of indicators extends far beyond this. For example, food, cosmetic, pharmaceutical, and other chemical formulations undergo complex chemical, enzymatic, and microbial interactions when they are exposed to UV light or temperature fluctuations over time. Consequently, product quality may be degraded and may lead to additional safety concerns. The challenges facing the produce industry include successful implementation of Good Manufacturing Practices (GMP), Hazard Analysis and Critical Control Point (HACCP), Total Quality Management (TQM) programs, and other regulations that demand compliance (1).

To address these issues and increase consumer confidence in product quality, safety, and authenticity, many manufacturers incorporate inexpensive monitoring devices into their products during production, packaging, or storage. A large number of consumer-readable indicators are available commercially. Some examples of these types of indicators are tags, labels, seals, and thermometers. Some give a visual color change in response to degradation of product quality, tampering, or to detect a counterfeit. They are used extensively in the chemical, food, and pharmaceutical industries where consumers need assurance of product integrity, quality, and safety during postmanufacture handling.

Generally, chemical indicators may be defined as stimulus responsive materials that can provide useful information about changes in their environment. Organic dyes, hydrogels or “smart polymers,” shape-memory alloys, thermochromic or photochromic inks, and liquid crystals are some examples. They may function by forming structurally altered ionic or molecular complexes with species in their environment through chemical or physical interactions involving proton exchange, chelation, hydrogen bonding, dipole–dipole interactions, or van der Waal forces (1). The resulting characteristic biochemical, chemical, optical, magnetic, thermal, or mechanical changes can be tailored to provide the desired indication response.

This article focuses on inexpensive disposable chemical indicating devices such as pH indicators, temperature indicators, time–temperature indicators (TTI), and tampering and counterfeit indicators. The temperature and TTIs are widely used in the food and pharmaceutical products where date coding on a package may sometimes be inadequate.

CHEMICAL INDICATING DEVICES ARE SMART

Smart materials or devices are defined as materials that produce strong visually perceptible changes in a physical or chemical property in response to small physical or chemical stimuli in the medium. The material properties measured may include pH, concentration, composition, solubility, humidity, pressure, temperature, light intensity, electric and magnetic field, shape, air velocity, heat capacity, thermal conductivity, melting point, or reaction rates (2–6). Chemical indicating devices can respond reversibly or irreversibly to small changes in the physical or chemical properties in their environment in a predictable manner. *They may be regarded as smart materials because of the range of materials properties they encompass.* Typical materials include shape-memory alloys, piezoelectric materials, magnetostrictive substances, electrorheological and magnetorheological fluids, hydrogel polymers, and photo- and thermoresponsive dyes (2–6).

CLASSIFICATION

Indicators can be classified on the basis of the response mechanism, operating principles, or application. Thus there are chemical, biological, biochemical, electrical, magnetic, and mechanical indicators according to the response mechanism. Based on the intended application, indicators can be classified as temperature indicators,

time–temperature indicators, pH indicators, counterfeit indicators, tamper indicators, freeze and thaw indicators, or freshness indicators.

GENERAL OPERATING PRINCIPLES

The response mechanism of most indicators includes one or more of the following: physical, chemical, physicochemical, electrochemical, and biochemical. Physical mechanisms are based on photophysical processes, phase transition, or other critical material properties such as melting, glass transition, crystallization, boiling, swelling, or changes in specific volume. In most cases these transitions are driven by changes in the interactive forces (e.g., hydrophilic–hydrophobic forces) within or around the indicator material. The indicator response mechanism can also be based on chemical, biochemical, and electrochemical reactions. Examples include acid–base, oxidation–reduction, photochemical, polymerization, enzymatic, and microbial reactions. Many of these changes are irreversible, and the onset or termination can be observed visually as a color change, color movement, or mechanical distortion (1–6).

CHOICE OF INDICATORS

Some factors governing the selection of a given indicator device include

- **Cost:** It must be relatively inexpensive. The indicator should not be more expensive than the product it is protecting.
- **Application:** Easy to attach to a variety of containers or packages. Once installed, the device must remain intact and readable during the service life of the package.
- **Response:** The response mechanism must have fast kinetics of the order of seconds to a few hours and must be reproducible. There should be no time delay in response to reactions involving a solid, liquid, or gas. Most applications require the response in the indicator to be irreversible to preserve the needed indication record.
- **Sensitivity:** The indicator must be highly sensitive, accurate, and easily activated. A user-friendly indicator that provides useful information when needed will make both the product manufacturer and the consumer happy.
- **Shelf life:** The indicator must have a shelf life equal to or longer than that of the product it is monitoring.
- **It should be technically difficult to duplicate or counterfeit the indicator’s response.** In this case, the indicator is acting as a “smart” locking mechanism.

pH INDICATORS

The pH indicator is probably the oldest and simplest smart chemical indicating device known. The chemistry of acid–base indicators is well documented (7) and involves proton

Table 1. Selected Indicators for pH Measurement

Common Name	Formula	pK _a ^a	pH Range	Color in Acid	Color in Alkali	λ _{max} [*] (nm)
Bromothymol blue	C ₂₇ H ₂₈ O ₅ Br ₂ S	7.3	6.0–7.6	Yellow	Blue	617
Bromocresol green	C ₂₁ H ₁₄ O ₅ Br ₄ S	4.9	3.8–5.4	Yellow	Blue	618
Bromocresol purple	C ₂₁ H ₁₆ O ₅ Br ₂ S	6.3	5.2–6.8	Yellow	Purple	588
<i>o</i> -C resol red	C ₂₁ H ₁₅ O ₅ S	8.2	0.2–1.8	Red	Yellow	572
Chlorophenol red	C ₁₉ H ₁₂ O ₅ Cl ₂ S	6.25	4.8–6.4	Yellow	Red	572
Methyl red	C ₁₅ H ₁₅ N ₃ O ₂	5.0	4.4–6.3	Red	Yellow	526
Phenol red	C ₁₉ H ₁₄ O ₅ S	8.0	6.4–8.2	Yellow	Red	559
Phenolphthalein	C ₂₀ H ₁₄ O ₄	9.5	8.0–9.8	Colorless	Red–Violet	552
Tetrabromophenol blue	C ₁₉ H ₆ O ₅ Br ₈ S	3.56	3.0–4.6	Yellow	Blue	605
α-Naphtholphthalein	C ₂₈ H ₁₈ O ₄	8.4	7.4–8.8	Colorless	Green–Blue	660
Thymol blue	C ₂₇ H ₃₀ O ₅ S	9.2	1.2–2.8	Red	Yellow	598

^aMeasurements made in solution at zero ionic strength and for a temperature ranging from 15–30°C.

or electron exchange reactions. It will be mentioned only briefly here. By Brønsted and Ostwald's definition, indicators are weak acids or bases. Upon dissociation, they exhibit a structural and color change that is different from the undissociated form. Indicators commonly used in acid–base and redox titrations are conjugated organic dyes that contain one or more light-absorbing groups called chromophores. The electronic structures of these dyes can be changed by redox or proton exchange reactions. This also changes the absorption energy in the visible region of the electromagnetic spectrum and consequently, the color. Therefore, the color change provides a visual indication of the end point of a titration (1,7).

INDICATOR MATERIALS

An indicator for pH measurement may exist as a solution, emulsion, colloidal gel, paper, or electrode. Methods, for preparing them are described in the literature (1,7,8). Table 1 lists examples of common indicators used in solution form or as indicator papers.

There are also non-dye pH sensitive smart polymers that exhibit critical material changes in response to changes in the concentration of hydrogen or other ions. Typically, these are polymer electrolytes derived from homo- or copolymerization of acidic or basic functionalized monomers. Examples include poly(ethylene oxide); poly(dimethylsiloxane); copolymers of *N,N*-dimethylacrylamide, *N-t*-butylacrylamide, and acrylic acid; mixtures of methacrylic acid and methyl methacrylate; mixtures of diethylaminoethyl methacrylate and butyl methacrylate; polystyrenesulfonate; polyacrylamide; and functionalized cellulose copolymers (8,9). In the presence of specific ions, these polymer gels can swell or shrink across a pH range at room temperature, resulting in a visible color change. Indicators based on them are highly selective, sensitive, and inexpensive (8).

TEMPERATURE AND TIME–TEMPERATURE INDICATORS (TTI)

Recently, interest has grown in using critical temperature indicators and time–temperature indicators in intelligent packaging technology. The production, processing,

distribution, storage, and point of use of temperature-sensitive products poses a great challenge to industries. Temperature plays a leading role in the growth of microorganisms in foods and thus in the incidence of food poisoning. As a result, many perishable food and nonfood products are prone to temperature-induced quality loss and degradation. Pharmaceutical products, medical devices, chemical reagents, industrial formulations such as inks, paints and coatings, photographic materials, and other related items are highly sensitive to temperature abuse. Therefore, it is useful to monitor the temperature history from production to the point of consumption of such products.

There has been greater need to monitor and prevent failure of machine and electronic parts due to overheating. Routine quality control and preventive maintenance of automobile and aircraft engines system against overheating may help avoid catastrophic and costly failures.

To ensure product safety, freshness, minimize losses, and retain customer confidence, dedicated temperature indicators or TTI are attached to many commercial products. They show the temperature or record the thermal history of the product and indicate if abuse has occurred, this allows the consumer to make judgments about product quality or safety. The indicator may incorporate a color changing dye, polymeric substance, enzyme, or time–temperature integrating materials that must be thermally activated to function (10).

Critical Temperature Indicators (CTI)

CTIs visually indicate that a material or product has been exposed to an undesirable temperature below or above a reference critical temperature for a time long enough to alter product quality or cause safety concerns (11). Several temperature indicators have been described in the patent literature, although only a few have been commercialized. Their principle of operation depends on temperature changes resulting from freezing, melting transition, liquid crystal formation, polymerization, enzymatic reactions, or electrochemical corrosion (12). Some formulations are derived from a dye and a chemical blend or a thermoresponsive hydrogel polymer that has a specific freezing or melting transition. Some commercial indicators are described here.

High-Temperature Indicating Crayons. These are used mainly for plant maintenance, quality assurance, engine monitoring, diagnosis, and troubleshooting. For example, a crayon would be useful where knowledge of the exact temperature of engine parts is desirable (13). Such crayons are made of materials whose known melting points are calibrated to have a response time of a few seconds and an accuracy of $\pm 1\%$. Two commercial examples include the Telatemp temperature indicating crayon (made by Telatemp Corporation, Fullerton, CA) and Tempilstik temperature indicators (made by Tempil Division, Air Liquide America, South Plainfield, NJ). They are available for temperature ratings from 125–1200°C. In an application, the indicating crayon is used to mark a desired surface of known operating temperatures such as the exhaust manifold, cylinder head, or fuel injection pump. An instantaneous color change upon marking indicates that the rated temperature has been exceeded. Alternatively, if a color change is observed after one or two seconds, the rated temperature of the indicator has been reached. When the material is operating below the rated temperature, a color change may not be observed, or it may take much longer to occur.

Irreversible Temperature Labels. Temperature labels are designed to respond permanently and irreversibly to and record overheating of surfaces. The indicator labels are made by sealing one or more temperature-sensitive materials in a highly stable self-adhesive polymeric or paper strip (14). Temperature ratings are printed on, below, and above each indicator window in degrees centigrade (°C) and Fahrenheit (°F). Several models in custom sizes/shapes and customer-selected temperature ranges are available from Telatemp Corporation. Indicator labels, are installed on a desired surface. When the calibrated rated temperature is exceeded, the indicator window of the label undergoes a noticeable color change (from silver to black in most Telatemp products). The response time is usually less than one second and has a guaranteed tolerance of $\pm 1\%$. Some uses include monitoring overheating of engine components, electronics, transformers, chemicals, foods, and pharmaceutical items.

Defrost Temperature Indicators. These are color-change indicators that are used mostly to monitor chill and frozen temperatures of chemical and food products during storage or distribution (15). They are *go/no go without delay* or *go/no go with some delay time* devices. The *go/no go without delay* indicators provide information that the temperature has risen above a threshold value during a short period. The *go/no go with delay* indicators show that the indicator has been exposed to a predetermined temperature for a specified length of time. Here are some examples:

1. ColdMark indicators (IntroTech, Inc.) are used to show whether a product has been exposed to critical cold temperature conditions as accurately as $\pm 1^\circ\text{C}$. Several designs have been described in the patent literature (16). They consist of a glass bulb filled with a colorless fluid and a tube partially filled

with a colorless and a violet colored fluid separated by a green barrier fluid. The tube is enveloped in a transparent plastic casing provided with a pressure-sensitive adhesive backing for attachment onto the exterior or interior surface of a package (17). For a freeze-monitoring indicator, the response temperature is $32^\circ\text{F}/0^\circ\text{C}$. Upon exposure to temperatures at or below the response temperature for about 30 minutes, the colorless fluid in the bulb freezes and contracts, causing the colored fluid in the tube to flow downward and color the bulb. When warmed above the response temperature, the bulb remains permanently colored. When not in use, the indicator is stored above the response temperature.

2. The TwinMark indicator (IntroTech, Inc.) is a dual purpose indicator that can respond to cold and hot temperatures. It works like the ColdMark and shows irreversible color change in the bulb when the cold side or hot side of the indicator is exposed to a temperature below or above the predetermined value (17).
3. The ColdSNAP indicator (Telatemp Corporation, Fullerton, CA) is a temperature recorder that contains a bimetallic sensing element (14). When attached to a product, a safety tab is pulled to activate the sensor. A clear window in the indicator means that the product storage temperature is safe. When exposed to a damaging critical temperature, the bimetallic sensor snaps into the indicator window and permanently changes it from clear to red. The accuracy of the indicator is $\pm 2^\circ\text{C}$, and customers can select snapping point temperatures ranging from -20 to $+40^\circ\text{C}$.

Liquid Crystal Display (LCD) Labels. LCDs are miniature reversible thermometers engineered on a label. An indicator window and a self-adhesive backing are provided. A liquid crystalline material or polymer is placed behind the indicator window. When installed, the material changes from opaque to transparent at rated temperatures (12,14). It is possible to select the indicator's color change of interest. LCDs can be designed to monitor surface temperatures continuously across various temperature ranges or to indicate selected temperatures. For example, Telatemp markets LCD reversible temperature decals that cover a range of -30°C to 120°C in $2\text{--}5^\circ\text{C}$ increments, visual tan/green/blue color changes permit readings to 1°C .

Time-Temperature Integrator or Indicators (TTI)

A time-temperature integrator/indicator is a continuous monitoring device or tag that measures both the cumulative exposure time and temperature of perishable products from production, distribution, storage, and even point of use. When a TTI is attached to a product and activated, the time-temperature history to which the product has been exposed is recorded and integrated into a single visual result (such as color change). In theory, TTI can be used as an informational, monitoring, and decision-making protocol at any point in the distribution chain of product quality and safety (11,18).

The response mechanism of a TTI is based on irreversible physical, chemical, or biochemical changes such as oxidation and reduction reactions, protein denaturation, enzymatic browning, sublimation of ice, and recrystallization of ice that occurs from the time of activation. The rate of these changes increases as temperature increases. The application and reliability of time–temperature indicators have been studied extensively by several researchers (19–22). Applications of TTI include food produce, pharmaceutical and blood products, cosmetics, adhesives, paints and coatings, photographic and film products, avionics, plastics, and shipment of live plants and animals.

Chemical Kinetic Basis for TTI Application. Major factors responsible for the deterioration of food and related chemical products include microbial growth, physical changes, and biochemical or chemical reactions within the perishable products. Some examples of spoilage reactions include acid–base catalysis, enzymatic reactions, free radical processes, hydrolytic reactions, lipid oxidation, meat pigment oxidation, nonenzymatic browning reactions, and polymerization or cross-linking processes (23). Like any chemical reaction, the rates of degradation or quality loss of perishables are influenced by environmental factors such as temperature, concentration (e.g., composition of an internal gaseous phases), and water activity. Increasing the temperature, for instance, increases the rates of these reactions.

Both the response of time–temperature indicators and changes in food quality (or other perishables) can be modeled by chemical kinetic equations. It is critical to gather accurate and reliable kinetic data for TTI to be successful and reliable. Several studies are available on the subject (23–25). Briefly, studies of several frozen and refrigerated foods indicated that the response of TTIs correlates with storage-related quality changes (26). In another study, the sensory changes in frozen hamburger reportedly correlate with the response of a commercially available TTI (27).

Several investigators have studied the influence of temperature on the rate of quality loss. It is generally accepted that the rate of quality loss behaves as an exponential function of the reciprocal of the absolute temperature. This is the Arrhenius equation shown here:

$$k = Z \exp(-E_A/RT) \quad (1)$$

where k is the reaction rate constant, Z is the temperature independent preexponential factor, E_A is the activation energy that describes the temperature sensitivity of the quality loss reaction, R is the universal gas constant, and T is the absolute temperature in Kelvin (K).

To measure loss of quality or shelf life in perishables, one or more characteristic measurable quality factors, denoted X are selected. The quality factor X can be a chemical, microbiological, or physical parameter. The rate of change of the quality factor with time under isothermal storage conditions is given by the following equation:

$$d[X]/dt = -k[X]^n \quad (2)$$

Table 2. Activation Energies of Selected Reactions That Result in Food Quality Losses^a

Type of Reaction	Activation Energy Range	
	(kJ/mol)	(kcal/mol)
Diffusion controlled	0–60	0–14.3
Acid–base catalysis	80–120	19.1–28.6
Enzymatic reaction	40–60	9.5–14.3
Hydrolysis	50–60	11.9–14.3
Lipid oxidation	80–100	19.1–23.9
Nutrient losses	20–120	4.8–28.6
Maillard reaction	100–180	23.9–43
Protein denaturation	300–500	71.6–119.3
Spore destruction	250–350	59.7–83.5
Vegetable cell destruction	200–600	47.7–143.2
Microbial growth	60–200	14.3–47.7

^aRefs. 20,28.

where t is the reaction time and n is the reaction order. For most food quality losses, a reaction order of zero or one is typical for a simple rate constant. By considering the kinetics of change in food quality at various temperatures, the activation energy E_A for the quality loss reaction can be obtained (19, 24, 25). Table 2 lists typical values of $E_{A(\text{Food})}$ for quality losses. The activation energy of the indicator ($E_{A(\text{TTI})}$) can be obtained from the TTI kinetics.

Correlation of TTI Response With Food Shelf Life

Taoukis and Labuza developed a correlation scheme for predicting the shelf life of a product based on the TTI response [19,20]. This scheme is based on a series of kinetic equations that describe the quality loss and the TTI response (Fig. 1.)

The effective temperature for a variable time–temperature distribution is determined from TTI response kinetics. This value is used in conjunction with food kinetics to estimate the remaining shelf life or quality loss. The scheme shown in Fig. 1 assumes that the effective temperature and the activation energies of both the TTI device and the food are the same. However, this is not always the case. For practical applications, a difference between $E_{A(\text{food})}$ and $E_{A(\text{TTI})}$ of ± 8.4 kJ/mol (2 kcal/mol) generally gives a good prediction [19,20].

Commercially Available TTIs. Several types of TTI devices have been described in the patent literature, but only three have continued to receive significant attention from the industrial and scientific community during the past 15 years (11,18). These include the 3M Monitormark, Lifeline, and VITSAB TTIs. They are available as labels, tags, pallet sticks, or pallets with sheaths. Each is described here.

3M Monitormark TTI. The principles, characteristics, and operation of the 3M Monitormark TTI (3M Identification and Converter Systems Division, St. Paul, MN) are well documented (29,30). It is designed to respond visually only after a critical threshold temperature has been exceeded. The indicator consists of an assemblage of a porous paper wick, a rectangular paper pad saturated with a blue colored fatty acid ester and phthalate mix that has a specific melting point, a polyester barrier layer that separates

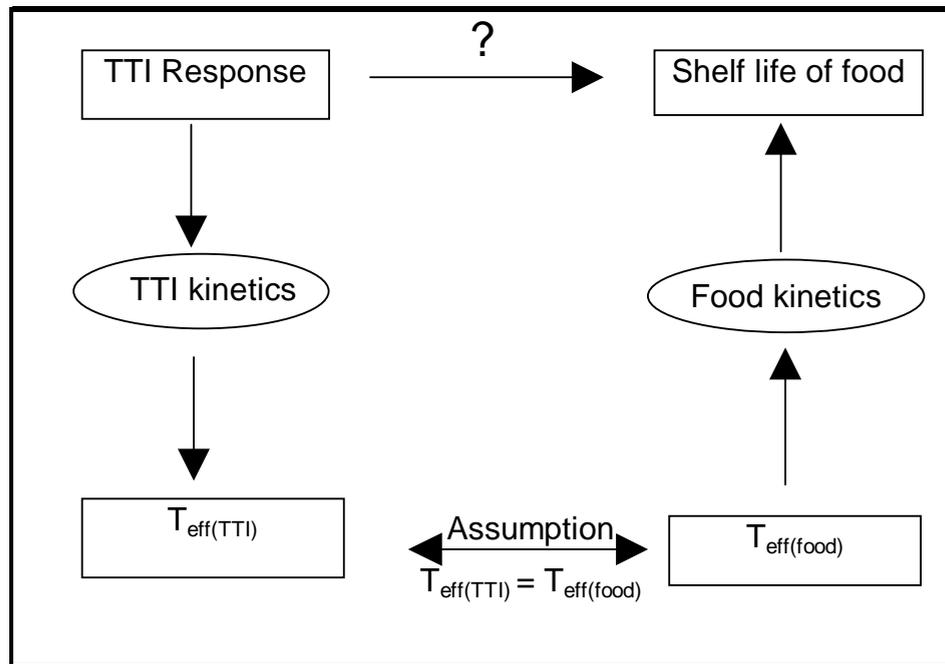


Figure 1. Application of a TTI as a food quality monitor (19).

the wick from the reservoir pad (paper pad), and adhesives. The response temperature is usually the melting point of the specialty chemical. The prepared TTI tag remains inactive until it is activated.

The indicator is activated when a polyester pull-strip is removed and allows contact between the porous paper and the reservoir. When exposed to temperatures higher than the threshold temperature, the chemical in the reservoir melts and migrates into the porous wick. The indicator has five windows that change color progressively as the migrating chemical runs through. Color in the first window implies that the indicator has been exposed to a temperature higher than the preset threshold temperature. The extent of color movement through these windows indicates the cumulative exposure time spent above the threshold temperature. Response cards provided by the manufacturer correlate the time-temperature relationship of each indicator. When all five windows turn blue, the product is considered unacceptable. If stored according to manufacturer's guidelines, the tag shelf life is 2 years from the date of manufacture.

A plot of the square of the run-out distance against time is linear for a Monitormark TTI, and it obeys the Arrhenius equation [20]. The design and operation of this indicator rely on a diffusion mechanism, so that the activation energy of this TTI is limited to 0–60 kJ/mol (0–14.3 kcal/mol). Therefore, it is recommended for use in many enzymatic and diffusion-controlled spoilage reactions.

Lifeline Fresh-Scan and Fresh-Check TTI. Lifeline Technologies (Lifeline Technologies, Morristown, NJ) offers two “full history” time-temperature indicators that monitor the freshness of perishables independent of the temperature threshold. These products are the Fresh-Scan and Fresh-Check indicators that are based on the color change resulting from the solid-state polymerization

of a diacetylenic monomer. The construction, operation, and characteristics of the device are described in detail by Patel and his co-workers (31,32) and by Field and Prusik (33).

The *Fresh-Scan* indicator consists a standard bar code that contains product information and a diacetylenic monomer deposited on a pressure-sensitive band, a portable microcomputer that has a laser scanner to measure changes in the reflectance of the indicator, and a data analysis work station. The indicator band initially shows about 100% reflectance. As the diacetylenic monomer undergoes time-temperature dependent solid-state polymerization during storage, the indicator band darkens and causes a decrease in the measured reflectance. The rate of color change follows Arrhenius behavior, so that higher temperatures enhance the rate of color development. The microcomputer combines this time-temperature characteristic and the product information on the bar code to predict the shelf life of the product.

The *Fresh-Check* indicator is a consumer-readable visual label, which is also based on the color change of an incorporated polymerizable monomer but has a different design and configuration (15). The circular device consists of an inner polymer shell that contains the monomer and an outer nonpolymer shell painted with a reference color. When exposed to some time-temperature storage conditions, the monomer at the center part converts to polymer. This causes progressive color development at a rate that increases as temperature increases. If the polymer center become darker than the outer reference, the consumer is advised to discard the product, regardless of the printed expiration date (15).

These two Lifeline indicators are active as soon as they are manufactured and must be refrigerated at very low temperatures ($< -24^{\circ}\text{C}$) to preserve their high initial reflectance.

Table 3. Activation Energy Values of Some Common Commercial TTIs

Manufacturer	TTI Model	TTI Activation Energy ($E_{A(TTI)}$)		Ref.
		(kJ/mol)	(kcal/mol)	
Visual indicator tag systems, AB-Malmo, Sweden	VITSAB Type 1	57.4	13.7	36
	VITSAB Type 2	94.7	22.6	36
	VITSAB Type 3	126.1	30.1	36
	VITSAB Type 4	194.8	46.5	36
	I-Point 3014 ^a	47.8	11.4	19
	I-Point 4007 ^a	137.0	32.7	19
	I-Point 4014 ^a	101.8	24.3	19
	I-Point 4021 ^a	141.2	33.7	19,20
3M Packagin System Div., St. Paul, MN	All types	41.1	9.8	19,36
Lifeline Technology Inc., Morris Plains, NJ	Fresh-Check A7	155.0	37.0	36
	Fresh-Check A20	81.3	19.4	19
	Fresh-Check A40	81.7	19.5	19
	Fresh-Check B21	88.0	21.0	19
	Fresh-Scan-18	113.1	27	19
	Fresh-Scan-41	85.8	20.5	19,20
	Fresh-Scan-68	82.5	19.7	19,20

^aVitsab was formerly known as I-Point.

Wells and Singh investigated the use of these indicators as a quality change monitor for perishable and semiperishable products such as tomatoes, lettuce, canned fruitcake, and UHT sterilized milk (22). It was shown that the response kinetics of these indicators correlates with quality loss in the products studied. The indicators studied have activation energies of 84 to 105 kJ/mol (20 to 25 kcal/mol). Major limitations of these indicators are that the polymerization reaction is photosensitive and also that they cannot respond to a short time–temperature history.

VITSAB TTI. Numerous studies have been carried out on the time–temperature-dependent enzymatic reaction on which the Vitsab TTI (former known as I-Point) is based [34,35]. It is a “full history” indicator that records the temperature conditions and temporal history of a product independently of threshold temperature. The device uses two color-coded compartments to store the chemicals. The green part houses a mixture of an enzyme solution and a pH indicating dye. The gray part contains a lipase substrate suspension. The line separating the two chambers is a pressure-sensitive barrier. The device is activated by breaking the barrier between the two chambers. This causes the enzyme solution and the substrate to mix to form the indicating solution. As the reaction progresses, the lipase substrate (triglycerides) hydrolyzes into its component fatty acids and causes the pH to drop. A change in pH causes the dye to change color from an initial green to a final yellow at a rate that is temperature-dependent. A variety of VITSAB indicators can be custom designed for specific temperature-sensitive commodities by judiciously selecting the enzyme concentration and the enzyme–substrate combination.

At present there are four different types of indicators that come in one- and three-dot standard configurations, depending on the level of product safety risk to be communicated. The single-dot indicator changes color from

green to yellow above a predetermined temperature. The three-dot indicator conveys the degree of temperature exposure with high reliability.

The activation energies of the major TTI products are summarized in Table 3. By matching the E_A of these TTIs to the E_A values for some common deterioration reactions in food and pharmaceutical products (Table 2), current TTI technology should provide a range of selections that will meet manufacturers’ needs.

ANTICOUNTERFEITING AND TAMPER INDICATOR DEVICES

Counterfeiting, forgery, tampering, and piracy of valuable documents and products has existed from time immemorial. Counterfeiting has become a multibillion dollar business that is prospering more than many of the victim companies. This highly organized trade results in lost revenue to both companies and governments, as well as loss jobs in private and public sectors. The motivation for professional counterfeiters is the ease of making huge profits and the risk of being caught and penalized is very low. It is impossible to enumerate the products or documents that are counterfeited or to quantify their impact on vulnerable consumers. Activities of counterfeiters have increased in developed countries. The incidence of counterfeiting is even greater in developing countries which have become a dumping ground for counterfeit products. A partial list of goods that are commonly faked include antiquities, currency, letters of credit, credit cards, jewelry, chemicals, processed food, pharmaceutical and cosmetic products, product identification marks, videos, CDs, clothing and apparel accessories, electronics, auto and airplane parts, children’s toys, watches, sporting goods, travel documents such as passports, identity cards and driving permits, and textile materials [37–39].

Advances in technology such as computers, lasers, scanners, charge-coupled devices (CCD), color printers, optical holograms, and analytical tools have enhanced the state of the art in anticounterfeiting security and encryption. Ironically, these advances are also being exploited by adversaries to promote high-tech forgery. To combat these activities, research efforts have continued to develop tamper-resistant security mechanisms that will make counterfeiting too expensive to be lucrative. This subject has been reviewed recently (37,38). A summary of some of the available technologies relevant to this article are presented.

Holographic Technology

An optical hologram is a three-dimensional photo created on a photoresist film or plate by the reflection and refraction pattern of laser light incident on an object. Two of the holograms widely used in security applications are the embossed and the Denisyuk hologram. The embossed hologram can be readily mass-produced as a mechanically tough and durable thin film at a very low cost.

The relatively thicker Denisyuk hologram requires a more stringent manufacturing process. In addition to being more expensive to produce, it is not as mechanically robust as the embossed hologram. This hologram is the more easily counterfeited than an optical hologram.

Holographic Dimensions, Inc. offers the Verigram™ security system that consists of two elements. The display hologram and the machine readable Verigram™ hologram afford a completely secure device. Although the display component may be counterfeited, it is claimed that the Verigram™ element is impossible to forge (40).

One of the leading manufacturers of anticounterfeiting holograms is American Bank Note Holographics, (ABNH) Inc. Their products are used for various commercial and security applications. They include holographic stripes, holomagnetic stripes, holographic threads, transparent holographic laminates, hot stamp foils, hologuards, and frangible vinyl and pressure-sensitive and tamper-evident labels. ABNH can manufacture holograms that have a magnetic stripe, visible and invisible bar code, or micro-printing for added security, product tracking, audits, and authentication. The list of producers and vendors of commercial and security holograms is endless. Only two are mentioned here as examples.

Microtaggant Technology

Taggants are microscopic color-coded particles derived from several layers of a highly cross-linked melamine polymer and are used extensively as markers (41). Layers of magnetic and fluorescent materials are added to the taggant particles during formulation to simplify detection and decoding. The technology (originally developed for explosives) finds applications in antiterrorism, authentication, piracy, counterfeiting, and product identification (41). Taggant particles can be coded with information such as the product manufacturer, production date, and batch number as well as the distributor. Taggants for use in antiterrorism are thoroughly mixed with explosives. When a bomb containing tagged explosive is detonated, remnant taggants

may be collected from debris by using a UV light and a magnet. The coded color sequence information can be retrieved by using an optical microscope. This reliably helps in tracing the origin of the explosive.

Tagged particles can also be added to printing ink for security printing. This provides a way to combat counterfeit, forgery, and product diversion and assist in product/property identification. Microtaggants can be added to many products either directly during formulation or by thermal transfer, films, laminates, and spraying. Some of the companies that market microtaggant identification particles are Microtrace Inc., Minneapolis, MN; MICOT Corporation (St. Paul, MN); SW Blasting (BÜLACH, Switzerland); and Plast Labor (Bulle, Switzerland). Patronage comes from government agencies, law enforcement, and industry.

Smart Ink Technologies

Smart inks or coatings change color predictably in response to alteration of their environment. Color change may or may not be reversible, and they are applied in areas such as security printing, lenses, and other optical devices. The two main classes of smart inks include those derived from photochromic and thermochromic materials.

Photochromic inks are formulated from light-sensitive dyes, pigments, polymers, and other colorless chemical compounds. These inks change color when exposed to ultraviolet light. Examples include extrusion lamination inks, invisible ink security markers, tagged inks, and inks based on copper-free metallics.

Thermochromic inks incorporate temperature-sensitive polymers, dyes, or pigments. When exposed to a predefined temperature, the ink undergoes a color change. Typical color-sensitive components include diacetylenic compounds, hydrogel polymers, and thermochromic liquid crystals.

Security inks incorporate fluorescent dyes, photochromic dyes, thermochromic dyes, or organic pigments in their formulations. These dyes or pigments change color reversibly when exposed to ultraviolet light or heat of predefined intensity or temperature. This forms a basis for their use in security printing, counterfeit or forgery indicators, and tamper-evident and applications. Several companies around the world manufacture and distribute counterfeit detector pens or specialty fluorescent light bulbs for quick authentication of currency bills, gift certificates, and other documents.

Tamper-Indicating Devices

Tamper indicators are security seals or labels designed to detect product or document tampering. Indicator devices available are based on fiber-optic active seal technologies, optical thin films, holograms, and chemical reactions. A major requirement for these materials is that they provide unambiguous evidence of tampering. They must be resistant to environmental conditions such as heat or solvents and must be relatively cheap. Several low-cost tamper indicators are commercially available to meet various customer needs.

Telatemp (Telatemp Corporation, Fullerton, CA) sells TelaSEAL security labels constructed from polyester film, high strength instant-bond adhesive, and other proprietary chemicals (14). For additional security, a repeating geometric pattern is printed onto the TelaSEAL label during construction. The seal can be applied to a variety of surfaces including plastics, metals, wood, and painted surfaces. A disrupted repeating pattern that has a custom message "OPENED" is transferred to the applied surface when the device is peeled from the substrate. This provides permanent and unambiguous evidence of tampering.

Meyercord (Sentinel Division, Carol Stream, IL) supplies several tamper-evident, adhesive-base materials that leave a VOID message on surfaces, thus providing an indication of unauthorized opening of product or tampering. The VOID message can be customized to include the company's name, trademark, logos, or serial number.

3M has developed several authenticating devices based on optical thin film technology for counterfeit detection and tamper indication. A proprietary metallic surface coating is formulated and deposited on a pressure-sensitive material. Specialized imaging technologies are used to generate unique covert graphics and print product information or a serial number on the device. When the device is peeled from the substrate, the covert image is revealed followed by an irreversible color change which provides a sure indication of tampering. These devices are used for asset control, package sealing and identification, parts identification, serial numbering, and consumer information.

Indicating Device Issues And Limitations

Interest in indicating devices continues to grow, and numerous applications have been proposed. Despite significant advances made in chemical indicator technology during the last 25 years, only a few devices have been commercialized compared to the number of patents granted each year. There are several reasons that many manufacturers, governments, and consumers are reluctant to adopt chemical indicating devices. Major considerations on the part of manufacturers include device reliability, cost, and education. Manufacturers fear that consumers will be reluctant to embrace products containing these indicators for reasons of additional cost. The addition of a leak indicator to the headspace of packaged food or pharmaceutical products is likely to be greeted by stiff resistance on the part of the consumer for safety reasons. To gain acceptance, commercial indicators must be relatively inexpensive compared to the products to which they are applied. They must also be technically competent to perform their intended functions.

The response mechanisms of many indicators are poorly understood (42). For example, to produce a reliable time-temperature indicator, the quality index, kinetics, and mechanism of food or chemical product degradation under various environmental conditions must be understood. Some degradation reactions may not always follow simple Arrhenius kinetics. Therefore, predictions based on them may be faulty and thus affect product reliability. Consumer surveys (12,30) on the use of TTIs on food products indicate

that the public reception of these indicators is growing. They also raised awareness of the need to educate consumers about the usefulness, reliability, and application of indicator devices to increase the current level of acceptance.

Even though most of the counterfeit and tamper indicator technologies currently available are reliable, they are quite expensive. This makes many industries reluctant to implement them. Some the indicators are limited in their scope of application and are easily forged or copied, as is the case for many optical holograms (37-40). This has been a major problem for many credit and identification cards. The challenge is to produce indicators that are reliable, technically difficult to copy or fake, and are available at a relatively low price.

Recently, several indicating devices have made inroads into the marketplace by making unsubstantiated claims of what the product can offer. In some cases, instructions for consumers on the application and interpretation of indicator response or color change are inadequate. Consumers can be confused. Information about the safety of the indicator should also be provided. Effort should be made to design indicators that are childproof to avoid ingestion and other accidents. A general specification for the various classes of indicators needs to be established. Such a specification should provide performance criteria that each indicator design must meet for a given application. It should also set standards for acceptable qualification testing. This will ensure that commercial indicating devices meet the needs of industry.

CONCLUSION

The application of chemical indicating devices has come to stay. In many cases, they are living up to expectations in reducing the incidence of counterfeiting, forgery, tampering, or piracy and losses of perishable products. Temperature and time-temperature indicators are now used in routine packaging of some critical food and chemical products. For TTIs, only limited information on the kinetics and response mechanism is available in the literature. New indicators that cover a wide range of activation energies for many spoilage reactions need to be designed. This will allow extending the scope of application of TTIs. More research and development effort could be directed toward the present technical and manufacturing challenges that hinder commercialization of many patented inventions. Public education will accelerate interest and ultimate acceptance of these products. Such acceptance may create sufficient volumes that would lead to lower costs for manufacturing these smart devices.

ACKNOWLEDGMENT

The author thanks Dr. Ted Labuza for providing key references on TTI; Dr. Annapoorna Akella for help with patent literature search; Dr. John Evans, Dr. Jim Harvey, and Dr. David Landman, for helpful comments; and companies who graciously provided relevant information.

BIBLIOGRAPHY

1. G. Boside and A. Harmer, *Chemical and Biochemical Sensing with Optical Fibers and Waveguides*. Artech House, Boston, 1996.
2. Y. Nagasaki and K. Kataoka, *Chemtech* p. 3, March 1997.
3. A.S. Hoffman, *Artif. Organs*. **19**: 458–467 (1995).
4. I.Y. Galaev, *Russ. Chem. Rev.* **64**: 471 (1995).
5. F. Mucklich and H. Janocha, *Z. Metallkd* **87**: 357 (1997).
6. M.J. Snowden, M.J. Murry, and B.Z. Chowdry, *Chem. Ind.* **15**: 531 (1996).
7. I.M. Kolthoff and C. Rosenblum, *Acid-Base Indicators*. Macmillan, 1937.
8. L. Bromberg and G. Levin, *Macromol. Rapid Commun.* **17**: 169 (1996).
9. R. Dagani, *Chem. Eng. News* p. 26, June 9, 1997.
10. C.H. Bryne, *Food Technol.* **6**: 66 (1976).
11. P.S. Taoukis, B. Fu, and T.P. Labuza, *Food Technol.* **45**(10): 70 (1991) and references therein.
12. M.L. Woolfe, in *Chilled Foods, A Comprehensive Guide*, C. Dennis and M. Stringer, eds., Ellis Horwood, NY, 1992, Chap. 5.
13. Anon. *Diesel Prog. Eng. Drives* **8**: 36 (1996).
14. Telatemp Corporation Fullerton, CA, product catalog, 1997.
15. J.D. Selman, *Food Manuf.* **65**(8): 30 (1990), and references therein.
16. Critical Temperature Indicator, US Pat. 4,457,252, 1984, W.J. Manske.
17. Introtech, Inc. St. Paul MN, USA Product catalog, 1992.
18. H.M. Schoen and C.H. Byrne, *Food Technol* **26**(10): 46 (1972).
19. B. Fu and T.P. Labuza, *J. Food Distribution Res.* **23**(1): 9 (1992).
20. P.S. Taoukis and T.P. Labuza, *J. Food Sci.* **54**(4): 783 (1989).
21. P.S. Taoukis and T.P. Labuza, *J. Food Sci.* **54**(4): 789 (1989).
22. J.H. Wells and R.P. Singh, *J. Food Sci.* **53**(1): 148 (1988).
23. T.P. Labuza, *J. Chem. Educ.* **61**(4): 348 (1984).
24. M. Hendrickx, G. Maesmans, S. De Cordt, J. Noronha, A. Van Loey, and P. Tobback, *Crit. Rev. Food Sci. Nutr.* **35**(3): 231 (1995) and references therein.
25. R. Villota and J.G. Hawkes, in *Handbook of Food Engineering*, D.R. Heldman and D.B. Lund, eds., Marcel Dekker, NY, 1992, p. 39.
26. J.H. Wells and R.P. Singh, *J. Food Sci.* **53**(6): 1866 (1988).
27. R. Grisius, J.H. Wells, E.L. Barrett, and R.P. Singh, *J. Food Process. Preserv.* **11**: 309 (1987).
28. B. Fu, P.S. Taoukis, and T.P. Labuza, *Drug Dev. Ind. Pharm.* **18**(8): 829 (1992).
29. Selective Time Interval Indicating Device, US Pat. 3,954,011, 1976, W.J. Manske.
30. J.D. Selman, in *Active Food Packaging*, M.L. Rooney, ed., Blackie Academic & Professional, NY, 1995, Chap. 10 and references therein.
31. Diacetylene Time-Temperature Indicator, US Pat. 4,228,126, 1980, G.N. Patel and K.C. Yee.
32. Time-Temperature History Indicators, US Pat. 3,999,946, 1977, G.N. Patel, A.F. Preziosi, and R.H. Baughman.
33. S.C. Field and T. Prusik, in *The Shelf Life of Foods and Beverages*, G. Charalambous, ed., Elsevier, Amsterdam, 1986, p. 23.
34. Enzymatic Substrate Composition Adsorbed on a Carrier, US Pat. 4,043,871, K. Blixt, S.I.A. Tornmarck, R. Juhlin, K.R. Salenstedt, and M. Tiru.
35. K. Blixt and M. Tiru, *Dev. Biol. Std.* **36**: 237 (1977).
36. Vitsab TTI product catalog, 1997, @ www.vitsab.com.
37. R.L. van Renesse, ed., *Optical Document Security*. Artech House, Boston, 1994.
38. I. Lancaster, *Proc. SPIE* **2659**: 2 (1996).
39. S.P. McGrew, *Proc. Opt. Security Syst. Symp.* Zurich, Oct. 14–16, 1987.
40. K.G. Brown, J. Weil, and M.O. Woontner, *Counterfeit Proof and Machine Readable Holographic Technology Verigram Systems*, @http://www.hgrm.com/counterf.htm.
41. A.M. Rouhi, *Chem. Eng. News*, July 24, 1995, p. 10.
42. J.D. Selman, *Food Manufacture*, **63**(12): 36 (1988) and references therein.

CHITOSAN-BASED GELS

KANG DE YAO
 FENG ZHAO
 FANG LI
 YU JI YIN
 Tianjin University
 Tianjin, China

INTRODUCTION

Chitosan is a natural biopolymer obtained by deacetylation of chitin, which is produced from marine shellfish, such as crabs, shrimp, fungal cell walls, and other biological sources. Chemically it is a linear cationic poly(β -(1-4)-2-amino-2-deoxy-D-glucan) derived from chitin, a poly(β -(1-4)-2-acetamido-2-deoxy-D-glucan) by deacetylation. Chitosan is described in terms of the deacetylation degree and average molecular weight. And next to cellulose, it is the second most plentiful biomass and is already known as a biocompatible and biodegradable material. Many researchers have examined tissue response to various chitosan-based implants. Results indicate that these materials evoke minimal foreign body reactions (1).

Gels consist of 3-D polymer networks swollen in swellers, whereas hydrogels swell in water. Smart (intelligent) gels (or hydrogels) can swell or contract in response to stimulus changes, e.g., temperature, pH, ionic strength, chemicals, and fields.

Chitosan has been used for developing smart gels via network or/and complex formation. Due to their various properties that depend on environmental variables such as pH, ionic strength, and electric field, these materials have potential applications such as biomaterials, separation membranes, and field responsive materials.

SUPRAMOLECULAR INTERACTIONS AND GEL FORMATION

As we know, chitosan-based gels are chemically or physically cross-linked networks. To make them smart, special supramolecular interactions are often introduced into the network to provide reversible responsiveness to environmental stimuli. Therefore, different interpolymer complexes are formed within the chitosan-based networks.

These complexes can form or dissociate in response to changes in their environment. They can be divided into these four major categories based on the dominant type of interaction:

1. Hydrogen bonding complexes
2. Polyelectrolyte complexes
3. Coordination complexes
4. Self-assembly

The presence of amino groups in chitosan ($pK_a = 6.5$) results in pH-sensitive behavior. Research efforts have been aimed at tailoring the properties of chitosan through cross-linking, functionalization, copolymerization, and blending to fit smart requirements.

Hydrogen Bond Complexes

The amine groups of glucosamine residue within chitosan chains can serve as cross-linking sites, for example, reacting with glutaraldehyde to form imine cross-links between linear chitosan chains that lead to gel formation (2).

An interpenetrating polymer network (IPN) is defined as a combination of two cross-linked polymers; at least one of them synthesizes or cross-links in the presence of the other. If one of the polymers is linear (not cross-linked), a semi-IPN results. The IPN technique can be used to improve the properties of chitosan-based gels.

Yao et al. (3) examined glutaraldehyde-cross-linked chitosan-poly(propylene glycol) (PPG, MW 3000 daltons) semi-IPN as a pH-sensitive hydrogel. Chitosan-PPG semi-IPN shows pH-dependent swelling; the highest swelling degree is at pH 4.0 and the lowest at pH 7.0 (Fig. 1). The structural features of chitosan-PPG are reversible (e.g. $-\text{NH}_3^+$ to $-\text{NH}_2$) when the gels are transferred from a pH 1.0 to a pH 7.8 buffer (4). Because it is a hydrophobic and water-insoluble polymer, the incorporated PPG is expected to decrease the equilibrium swelling degree in a low pH medium. In addition, linear PPG in the network can enhance the flexibility of semi-IPN.

Polyethylene glycol (PEG) is a water-soluble polymer that has low toxicity and immunogenicity. Hydrophilic PEG can form a diffuse layer containing water molecules

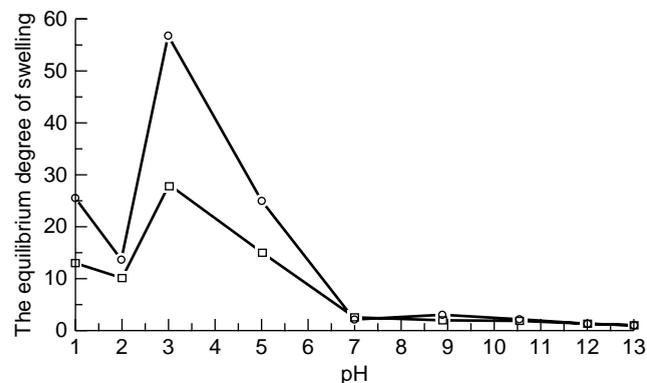


Figure 1. The equilibrium degree of swelling vs. pH for the semi-IPN, synthesized from different amounts of crosslinking agent. Molar ratios of $-\text{CHO} : -\text{NH}_2$ are 2.21(○) and 4.38(□), respectively.

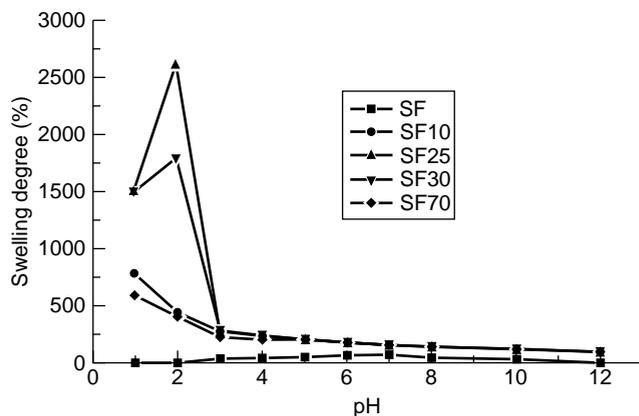


Figure 2. The pH dependence of the degree of swelling of chitosan/silk fibroin semi-IPN. Glutaraldehyde content: 1.5% mol/mol, 30°C. SF10–SF70 denote that the silk fibroin contents in semi-IPN are 10–70% w/w, respectively.

which has protein resistance when PEG is adsorbed on a surface. PEG has been used to improve the biocompatibility of polymers. Glutaraldehyde-cross-linked chitosan-PEG semi-IPN hydrogels were prepared by incorporating PEG (MW 6000) into chitosan. Here, PEG in the network enhances blood compatibility. In addition, because its ether linkage $[-\text{CH}_2\text{CH}_2\text{O}-]_n$ provides rotational freedom, PEG also enhances the flexibility of the network. The swelling degree of semi-IPN drops sharply at high pH (>6). This may be due to the formation of hydrogen bonds between the hydrogens of the amino groups of chitosan and the oxygens of the polyether in alkaline regions, whereas the hydrogen bonds dissociate at acidic pH (5).

Poly(ethylene oxide) (PEO) of average molecular weight ranging from 10,000 to 1 million daltons was used to prepare glyoxal-cross-linked chitosan-PEO semi-IPN; the semi-IPN swells almost ten times more in the low pH environment of gastric fluid (pH 1.2) than in the intestine (pH 7.2) after 6 hours (6).

Silk fibroin, a main component of natural silk, is a fibrous protein whose major amino acids consist of glycine, alanine, and serine residues. Cross-linking chitosan with glutaraldehyde in the presence of silk fibroin creates a chitosan/silk fibroin semi-IPN. The equilibrium swelling degree versus pH of the semi-IPN is shown in Fig. 2. The degree of swelling of silk fibroin is small in the whole pH range, whereas the semi-IPN swells obviously in low pH ($\text{pH} < 5.0$) buffer solutions, and especially at pH 3.3, just like cross-linked chitosan. This may be attributed to the dissociation of hydrogen bonds between chitosan and silk fibroin within the semi-IPN in the strong acidic medium. Moreover, chitosan can bind metal ions, for example, Al^{3+} , resulting in hydrogen bond dissociation. So chitosan/silk fibroin semi-IPN shows ion sensitivity. Figure 3 displays semi-IPN swelling and shrinkage with a change in AlCl_3 concentration. However, relaxation time is longer for ion sensitivity than for pH responsiveness (7).

Gelatin is a denatured form of collagen, composed of glycine, proline, hydroxyproline, arginine, and other amino acids. The amphiphilic protein (isoelectric point = 4.96) can provide amino groups for co-cross-linking with chitosan via

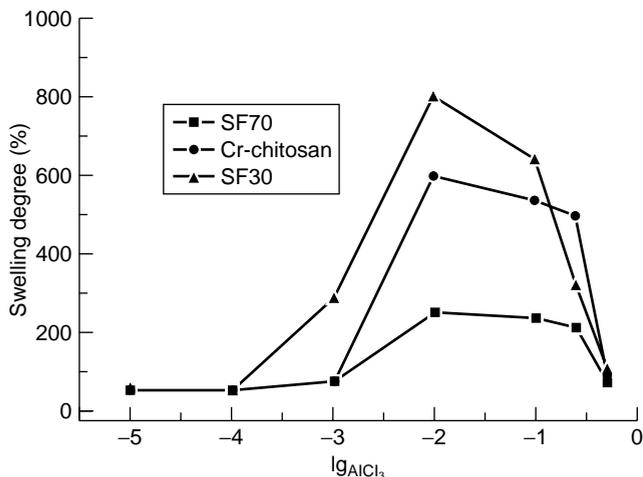


Figure 3. The $[Al^{3+}]$ dependence of the degree of swelling of chitosan/silk fibroin semi-IPN. Glutaraldehyde content: 1.5% mol/mol, 30°C. C_{AlCl_3} denotes the concentration of aqueous $AlCl_3$ solution. SF30 and SF70 denote silk fibroin contents in semi-IPN of 30 and 70% w/w, respectively.

glutaraldehyde to prepare a chitosan/gelatin hybrid polymer network (HPN). The pH-sensitive swelling behavior of the HPN gel is displayed in Fig. 4. The data show that the degree of swelling declines sharply at pH 7.0. This can be explained by the fact that the hydrogen bonds within the chitosan/gelatin HPN dissociate in an acidic medium (8).

Polyelectrolyte Complexes

Polyelectrolyte complexes (PEC) are generally obtained either by the reaction of polycations and polyanions or by polymerizing monomers that have suitable functional groups onto polymeric templates of known structures. PECs have numerous applications such as membranes, medical prosthetic materials, environmental sensors, and chemical detectors. Chitosan, a cationic polysaccharide, has been complexed with anionic polymers.

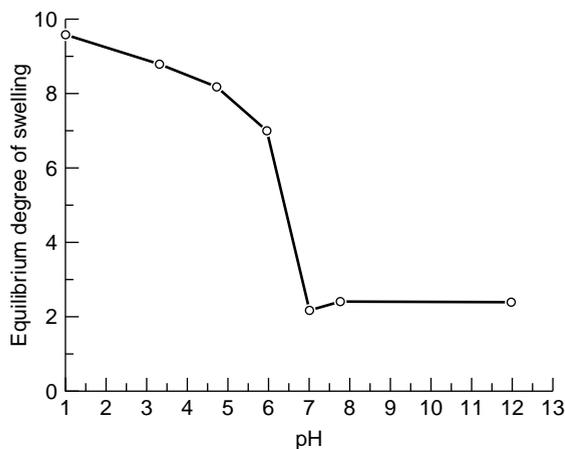


Figure 4. Swelling behavior of chitosan/gelatin hybrid network specimen with a $-CHO : -NH_2$ molar ratio of 10 in solutions of different pH with ionic strength $I = 0.1$ at 37°C.

Alginate comprises a linear chain of (1,4)-linked β -D-mannuronate and α -L-glucuronate residue arranged blockwise. A gel-like chitosan/alginate PEC was formed with coexisting polyions of opposite charges ($-NH_3^+$ and $-COO^-$). Polysaccharides (chitosan and alginate) have bulky pyranose rings and a highly stereoregular configuration in their rigid linear backbone chains. Their conformation changes with pH and leads to pH dependence of the packing density of PEC membranes. Therefore, the drug release behavior of drug-loaded microcapsules depends on pH as well (9).

Pectin is an acidic polysaccharide that has a repeating unit of α (1,4)-D-galacturonic acid, and its methyl ester is interrupted in places by (1,2)-L-rhamnose units. A polyelectrolyte complex is formed from anionic pectin and cationic chitosan. The PEC swells obviously at $pH < 3$ and $pH > 8$ and does not swell in the range of $3 < pH < 8$. Moreover, its degree of swelling in an acidic medium is by far larger than that in an alkaline medium. The swelling of the PEC correlates with its composition and is also affected by the degree of deacetylation and the methoxy level of pectin (10).

Complex formation via hydrogen bonding occurs between two polymers that contain hydrogen bonding donor groups and acceptor groups, whereas a charged polymer forms polyelectrolyte complexes with other polymers that have inverse charged groups. Bovine atelocollagen and a fully deacetylated high molecular weight chitosan (acetyl content $\approx 2.5\%$) form complexes; whether this is by pure electrostatic interaction or by hydrogen bonding depends on the conditions. In the first case, the interaction takes place between $-NH_3^+$ and $-COO^-$ groups carried by chitosan and collagen, respectively, but the maximum proportion of chitosan in the complex is relatively low ($\approx 10\%$) (11). In the presence of a great excess of chitosan, it is also possible to form a hydrogen bonding complex, in addition to the first complex where chitosan chains seem to induce the destabilization of the triple helix organization and denature collagen (12). Moreover, an increase in the content of chitosan in the complex improves the stability toward collagenase (13).

A chitosan-poly(acrylic acid) polyelectrolyte complex can be obtained by radical polymerization of acrylic acid onto chitosan as a template. The template polymerization provides an ordered structure to the PEC compared to that obtained by reacting chitosan with poly(acrylic acid) (14).

Coordination Complex

Chitosan has one amino group in addition to hydroxyl groups in the repeating unit. So chitosan can strongly coordinate with transition-metal ions to form coordination complexes, which may be used as absorbents for metal ions and separation membranes. The order of the stability constants of these chitosan-transition-metal complexes is $Mn < Fe < C < Ni < Cu < Zn$, which is known as the Irving-Williams order. The difference in the stability of complexes may cause the variance in the degree of swelling of a complex membrane whose metal/glucosamine ratio ranges from 1:8 to 1:64 (15).

Self-Assembly

Block-copolymer micelles and hydrophobic water-soluble polymeric amphipaths can form self-aggregates that

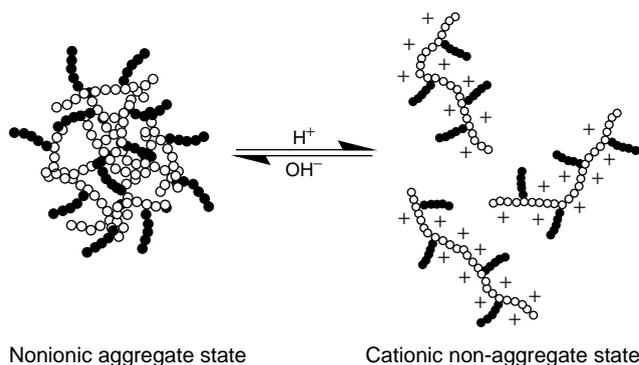


Figure 5. Schematic for aggregation and dissociation behavior of PEG- δ -chitosan in aqueous solution in response to the pH of the solution.

consist of inner cores of hydrophobic segments and outer shells of hydrophilic segments in aqueous media. Chitosan was hydrophobically modified by deoxycholic acid, a main component of bile acid, to yield self-aggregates in aqueous media. The modified chitosan forms self-aggregates in phosphate-buffered saline (PBS pH 7.2) whose mean diameter is less than 180 nm, depending on the degree of substitution of hydrophobic groups. Critical aggregation concentration (*cac*) is the threshold concentration of a self-aggregate formed by intramolecular and/or intermolecular association. In water, the *cac* values ($1.32\text{--}4.47 \times 10^2$ mg/mL) of deoxycholic acid-modified chitosan are lower than the *cac* value (1.0 mg/mL) of deoxycholic acid. Because self-aggregates have positive charges on their outer shells, they can be used as delivery carriers for DNA, a relatively charged polyelectrolyte (16).

PEG is a highly hydrophilic polymer. Grafting PEG onto chitosan results in a water-soluble PEG- δ -chitosan, which is prepared via the coupling reaction of 6-O-triphenylmethylchitosan with MeO-PEG acid, and then deprotecting its triphenylmethyl groups. PEG- δ -chitosan can form intermolecular aggregates in neutral and alkaline conditions and is dissociated in acidic media because of the electrostatic repulsion of protonated amino groups of chitosan moieties (Fig. 5). Therefore, this PEG- δ -chitosan aggregate is expected to be used as a pH dependent material such as a carrier for drug delivery systems (17).

Glycol chitosan (GC, polymerization degree ~ 800) was modified by acylation with palmitic acid *N*-hydroxy-succinimide. The polymer obtained (GCP) can form vesicles by mixing GCP/cholesterol in a 2:1 wt ratio under sonication; the GCP assembles into unilamellar polymeric vesicles in the presence of cholesterol. Anticancer drugs, for example, bleomycin (BLM), can be encapsulated efficiently within the vesicles. The stable GCP-based BLM vesicles were tested *in vivo* (18).

APPLICATIONS

Controlled Release Matrixes

The aim of controlled release of drugs is either to modulate tissue drug levels or to spatially and/or temporarily place a drug in some region of the body to maintain efficacy

and depress side reactions. Therefore, a matrix used for controlled release should offer modulating ability for drug delivery in addition to maintaining drug stability within the matrix. Chitosan has structural characteristics similar to those of glycosaminoglycans; its smart material has been explored for use in controlled release matrixes. The use of chitosan in the development of oral sustained release stemmed from the intragastric "floating tables" of chitosan. Chitosan has gel-forming properties in a low pH range in addition to its antacid, antiulcer characteristics. Thus, it has promising potential for use in oral sustained delivery systems. Gastrointestinal, respiratory, ophthalmologic, cervical, and vaginal mucins are hydrophilic saline gels that are thickened by natural anionic glycoproteins. In searching for drug delivery systems for such substrates, suitable cationic polymers are needed. Chitosan is known to be bioadhesive to mucosal surfaces, such as in the eye, nose, and vagina.

Pluronic[®] (BASF Corp), a well-known biomaterial is a triblock copolymer that contains poly(propylene oxide) (PPO) and poly(ethylene oxide) (PEO) segments in the sequence PEO-PPO-PEO. Here, hydrophobic PPO segments aggregate leading to distinct gelation at body temperature. The PPO aggregation forms micelles that solubilize lipophilic drugs in aqueous media and allow their slow release. Hoffman et al. licensed chitosan grafted with Pluronic[®]. The graft copolymers form clear gels when the temperature is raised from 4 to 37°C at pH 7.4. The copolymer is used as a matrix for receiving proteins via nasal administration. When the proteins are released from the gel *in vitro* through ion exchange and/or diffusion, they are "totally active" (19,20).

The cationic character of the chitosan backbone allows oppositely charged drugs such as insulin, interleukin 1L-2 receptor, IL-4 receptor, and tumor necrosis factor (TNF) receptor to form complexes of Pluronic[®]- δ -chitosan (21).

Proteins and enzymes represent a growing and promising field of therapeutics and are currently administered by injection. As a result of the development of biotechnology, many peptides and proteins have been explored as a new generation of drugs. Maintaining their bioactivity in the drug delivery process is one of the key problems. Oral peptide drug delivery is the easiest method of administration; however, proteins are quickly denatured and degraded in the stomach. Moreover, the intestinal absorption of drugs is generally very poor. Chitosan can enhance the intestinal transport of peptide drugs by increasing the paracellular permeability of the intestinal epithelium. Bioactive peptides, for example, hirudin (a polypeptide whose MW is 10,800 and is a potent antithrombin agent) can be delivered orally by microencapsulation with alginate-PEG/chitosan systems. Actually, oral intake of hirudin via the chitosan alginate matrix can avoid thrombin formation and alter lipid levels of patients beneficially (22).

Chitosan-based materials can be designed as stomach-specific drug delivery systems. Due to the limitations of the gastric emptying time, it is desirable to have a rapid swelling system to release a drug in the stomach. In enzyme-free simulated gastric fluid (pH 1.2), the release rate of metronidazole from freeze-dried chitosan-PEO semi-IPN is significantly higher than that from an air-dried hydrogel. The faster rate is attributed to the higher

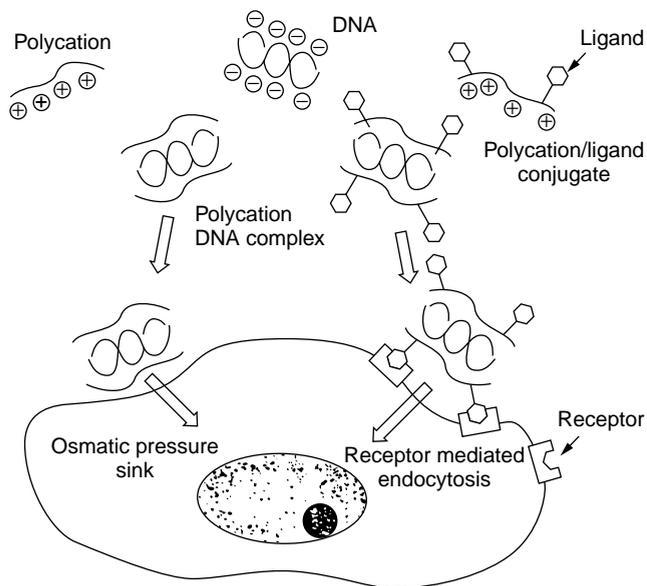


Figure 7. Scheme of gene delivery system via receptor-mediated endocytosis.

Chitosan has been used to form pervaporation membranes for separating water/alcohol mixtures and shows good performance in dehydrating alcohol solutions. To enhance fluxes with more free volume, chitosan membranes cross-linked with hydrophilic sulfosuccinic acid (SSA) were developed where two carboxylic acid groups and one sulfonic acid group offer ionic cross-linking with amine side groups of the chitosan molecules. Because the SSA cross-linked membrane bears more binding sites for the target compound to be separated than the cross-linked chitosan membrane, the former membrane is better than the latter membranes reported earlier (29).

Due to high permeability and good mechanical properties comparable to those of commercial cellulose acetate membranes, the membranes have potential application in pervaporation separation of aqueous organic mixtures (30). The pervaporation properties of isopropanol–water mixture via a chitosan–silk fibroin complex membrane are shown in Fig. 8. The data imply that the flux through the complex membrane expresses ion (Al^{3+}) sensitivity, so the pervaporative flux of the isopropanol–water mixture can be modulated; the high selectivity was maintained by optimizing the AlCl_3 concentrations in the feed (31).

A benzoylchitosan membrane was designed for separating a benzene/cyclohexane mixture, which is very important in the petrochemical industry. The benzene permeation selectivity of the membrane is attributed to the smaller molecular size and higher affinity of the benzene molecules compared to those of cyclohexane (32).

Anionic surfactants for example, sodium lauryl ether sulfate (SLES), can form complexes with the chitosan chain through interaction of their opposite ionic charges. The complexes can survive as a skin layer (ca. 15 μm) on a polyethersulfone ultrafiltration membrane. Here, the ionic property of the surfactant–chitosan complex membrane preferentially promotes the permeation of polar methanol

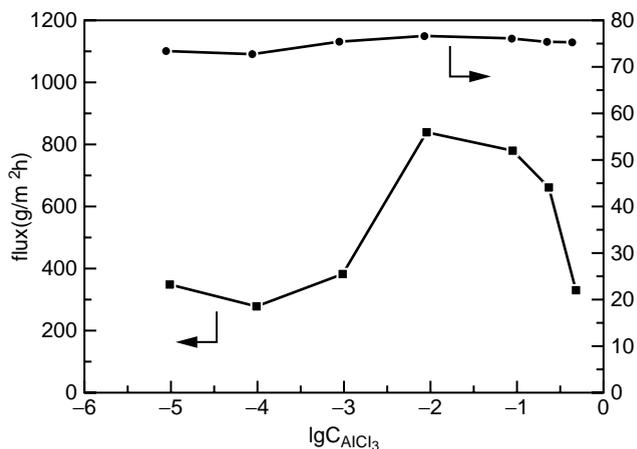


Figure 8. Separation properties of an isopropanol–water mixture through a chitosan–silk fibroin complex membrane by pervaporation [silk fibroin content in membrane: 30% (w/w), isopropanol in feed: 85% (w/w), C_{AlCl_3} denotes the AlCl_3 concentration in the water part of isopropanol–water mixture].

through the membrane, compared with the less polar and relative bulky methyl-*t*-butylether (MTBE) (33).

A formed-in-place (FIP) membrane is dynamically created by depositing polymers on the surface or at the entrance of the pores of macroporous substrates. Chitosan FIP ultrafiltration membranes can be formed on a macroporous titanium dioxide substrate. The chitosan gel membrane formed on the substrate is cross-linked enough by a supramolecular interaction, for example, hydrogen bonding. The estimated mean pore size for the membrane near neutral conditions (pH 6.0 and 8.2) is about 17 nm, and for the membrane at pH 3.6, it is 55 nm. The contraction and swelling of the chitosan membrane are reversible. Therefore, it is possible to control the pore size of the membrane by simply adjusting the pH of the system according to the separation requirement (34).

Smart polymers are the basis for a new protein isolation technique—affinity precipitation. The precipitation applies a ligand coupled to a water-soluble polymer known as an affinity macroligand, which forms a complex with the target. The phase separation of the complex, triggered by small changes in the environment, for example, pH, temperature, ionic strength, or addition of reagents, makes the polymer backbone insoluble; afterward, the target protein can be recovered via elution or dissolution. Chitosan itself has been successfully used as a macroligand for affinity precipitation of isolated wheat germ agglutinin from its extract and glycosides from cellulose preparation by changes in the pH of the media (35).

A partially sulfonated poly(ether sulfone) microporous hollow fiber membrane was coated with chitosan by electrostatic attraction. After cross-linking by reacting the fiber with ethylene glycol diglycidyl ether (EGDGE), the hydroxyl and amino glucose units of chitosan are then modified to bind recombinant protein A (rPrA) as an affinity ligand at 4.77–6.43 mg rPrA/mL fiber. The immobilized rPrA hollow fiber membrane serves as a support for affinity separation of immunoglobulin (IgG) (36).

Immobilization Supports

Immobilization is an effective measure for using enzymes or microbial cells as recoverable, stable, and specific industrial biocatalysts. Immobilization must be carried out under mild conditions.

An ideal support for enzyme immobilization should be chosen to achieve essential properties such as chemical stability, hydrophilicity, rigidity, mechanical stability, larger surface area, and microbial attack resistance. Several methods are used to modify the supports to improve stability and mechanical strength, and to modify different functional groups that may be superior for enzyme immobilization. One of the ways to improve these properties is grafting of monomers onto the matrix. The graft polymeric support, for example, chitosan-poly(glycidyl methacrylate) (PGMA), is used for glucosidase to immobilize urease (37,38).

Immobilized urease can be used in biomedical applications for the removing urea from blood in artificial kidneys, blood detoxification, or in the dialysate regeneration system of artificial kidneys. In the food industry, it may be used to remove traces of urea from beverages and foods and in analytical applications as a urea sensor.

In addition to urease, other enzymes, for example, glutamate dehydrogenase, penicillin acylase, β -galactosidase, and glucosidase, have been immobilized via chitosan gels. Immobilization improves the stability of the enzymes.

Extracellular Matrixes For Tissue Engineering

Various synthetic and naturally derived hydrogels have recently been used as artificial extracellular matrixes (ECMs) for cell immobilization, cell transplantation, and tissue engineering. Native ECMs are complex chemically and physically cross-linked networks of proteins and glycosaminoglycans (GAGs). Artificial ECMs replace many functions of the native ECM, such as organizing cells into a 3-D architecture, providing mechanical integrity to new tissue, and providing a hydrated space for the diffusion of nutrients and metabolites to and from cells. Chitosan is similar to GAG. Therefore, it is promising for application as a biomaterial in addition to use as a controlled delivery matrix.

Chitosan is a basic polysaccharide, so it is possible to evaluate the percentage of amino functions, which remain charged at the pH of cell cultivation (7.2–7.4). These cationic charges have a definite influence on cell attachment by their possible interaction with negative charges located at the cell surface. Chitosan materials give the best results in cell attachment and cell proliferation for chondrocytes and keratinocytes of young rabbits compared with its polyelectrolyte complex with glycosaminoglycans, such as chondroitin 4 and/or 6 sulfate and hyaluronic acid (39).

Field Responsive Materials

Chitosan-based gels consist of a positive charged network and a fluid (e.g., water) that fills the interstitial space of the network. The gels exhibit a variety of unique field responsive behaviors, such as electromechanical phenomena.

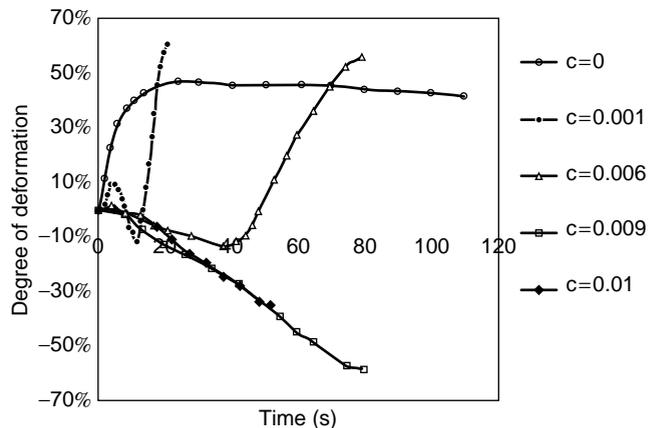


Figure 9. The EMC behaviors of chitosan/PEG crosslinked with different concentrations of ECH.

Electromechanochemical (EMC) behavior deals with contraction of polymers in an electric field. The effects of an electric field on polyelectrolyte hydrogels relate to the protonation of its alkaline amino groups and redistribution of mobile counterions when chitosan/PEG composite fibers is cross-linked with epichlorohydrin (ECH) and glutaraldehyde (GA), respectively. The EMC behavior of fibers in a 0.1% aqueous HCl solution in a 25-V dc electric field is shown in Fig. 9.

The bending direction of the fiber specimen inverts at a critical concentration of the cross-linking agents. When the ECH concentration is more than 9.0×10^3 M or the GA concentration is larger than 5.64×10^4 M, the fiber specimens bend toward the cathode. If the ECH or GA concentration is less than the critical values, they bend toward the anode. The reason may be attributed to variation in the mobile ions within the network (40).

Thin films of chitosan and chitosan doped with rare-earth metal ions can be used as wave-guiding materials. They are transparent across the wavelength range of 300–3000 nm and exhibit low optical loss (less than 0.5 db/cm²) (41,42).

Chitosan/acetic anhydride and acrylate/chitosan hydrogels have an excellent laser-damage threshold (LDT) up to 35 times higher than commercial poly(methyl methacrylate) (PMMA) bulk materials, and their LDT increases as water content increases. As we know, absorbed laser energy can lead to rapid local heating of a laser “hot spot.” A hydrogel can be considered a composite of statistically distributed microchannels and/or fluctuating pores created by the movements of polymer segments within the network in the presence of water. When a hydrogel is irradiated, the energy generated by laser light can be absorbed and dispersed by the water and the polymer frame. These hydrogels have potential applications as new materials for high-power laser-damage usage (43).

BIBLIOGRAPHY

1. J.-K. F. Sub and H.W.K. Matthew, Application of chitosan-based polysaccharide biomaterials in cartilage tissue engineering: A review. *Biomaterials* **21**; 2589–2598 (2000).

2. G.A.F. Robert and K.E. Taylor, The formation of gels by reaction of chitosan with glutaraldehyde. *Makromol. Chem.* **190**, 955–960 (1989).
3. K.D. Yao, T. Peng, M.F.A. Goosen, J.M. Min, and Y.Y. He, pH-sensitivity of hydrogels based on complex forming chitosan-polyether interpenetrating polymer network. *J. Appl. Polym. Sci.* **48**(2), 345–353 (1993).
4. T. Peng, K.D. Yao, C. Yuan, and M.F.A. Goosen, Structural changes of pH-sensitive chitosan/polyether hydrogels in different pH solution. *J. Polym. Sci., Part A.* **32**(3), 591–596 (1994).
5. M.S. Beena, T. Chandy, and C.P. Sharma, Heparin immobilized chitosan-polyethylene glycol interpenetrating network: Antithrombogenicity art. *Cells, Blood Substitutes Immobil. Biotechnol.* **23**(2), 175–192 (1995).
6. M.M. Amiji and V.R. Patel, Chitosan-poly(ethylene oxide) semi-IPN as a pH-sensitive drug delivery system. *Polym. Prepr.* **35**(1), 403–404 (1994).
7. X. Chen, W.J. Li, W. Zhong, Y.H. Lu, and T.Y. Yu, pH sensitivity and ion sensitivity of hydrogels based on complex-forming chitosan/silk fibroin interpenetrating polymer network. *J. Appl. Polym. Sci.* **65**(11), 2257–2262 (1997).
8. K.D. Yao, Y.J. Yin, M.X. Xu, and Y.F. Wang, Investigation of pH-sensitive drug delivery system of chitosan/gelatin hybrid polymer network. *Polym. Int.* **38**(1), 77–82 (1995).
9. K.Y. Lee, W.H. Park, and W.S. Ha, Polyelectrolyte complexes of sodium alginate with chitosan or its derivatives for microcapsules. *J. Appl. Polym. Sci.* **63**(4), 425–432 (1997).
10. K.D. Yao, H.L. Tu, F. Chang, J.W. Zhang, and J. Liu, pH-sensitivity of the swelling of a chitosan-pectin polyelectrolyte complex. *Angew. Makromol. Chem.* **245**, 63–72 (1997).
11. M.N. Taravel and A. Domard, Relation between the physicochemical characteristics of collagen and its interactions with chitosan: I. *Biomaterials* **14**(12), 930–939 (1993).
12. M.N. Taravel and A. Domard, Collagen and its interaction with chitosan: II. Influence of the physicochemical characteristics of collagen. *Biomaterials* **16**(11), 865–876 (1995).
13. M.N. Taravel and A. Domard, Collagen and its interactions with chitosan: III. Some biological and mechanical properties. *Biomaterials* **17**(4), 451–455 (1996).
14. P. Cerrai, G.D. Cuerra, and M. Tricoli, Polyelectrolyte complexes obtained by radical polymerization in the presence of chitosan. *Macromol. Chem. Phys.* **197**(11), 3567–3579 (1996).
15. N. Kobota, Permeability properties of chitosan-transition metal complex membrane. *J. Appl. Polym. Sci.* **64**, 819–822 (1997).
16. K.Y. Lee, W.H. Jo, L.C. Kwon, Y.H. Kim, and S.Y. Jeong, Structure determination and interior polarity of self-aggregates prepared from deoxycholic acid modified chitosan in water. *Macromolecules* **31**(2), 378–383 (1998).
17. T. Ouchi, H. Nishizawa, and Y. Ohya, Aggregation phenomenon of PEG-grafted chitosan in aqueous solution. *Polymer*. **39**(21), 5171–5175 (1998).
18. A.G. Schatzlein, J. Sludden, L. Tetley, E. Mosha, and I.F. Uchegbu, Chitosan based polymeric vesicles as anticancer drug carriers. *Proc.—Int. Controlled Release Bioact. Mater.* **25**, 435–436 (1998).
19. A.S. Hoffman, G.H. Chen, X.D. Wu, and Z.L. Ding, Graft copolymers of PEO-PPO-PEO triblock polyethers on bioadhesive polymer backbones: Synthesis and properties. *Polym. Prepr.* **38**(1), 524–525 (1997).
20. R. Dagani, Intelligent gels. *Chem. Eng. News* **75**(23), 26–37 (1997).
21. A.S. Hoffman, P.S. Stayton, Z. Ding, X. Bulmus, Y. Hayashi, T. Furusona, H. Saito, C. Leng, G. Chen, X. Wu, J.E. Matsuvra, and W.R. Gombotz, Graft copolymers of stimuli-responsive polymers on biomolecule backbone: Synthesis and biomedical applications. *Polym. Prepr.* **38**(2), 532–533 (1997).
22. T. Chandy, D.L. Mooradian, and G.H.R. Rao, Chitosan/polyethylene glycol-alginate microcapsules for oral delivery of hirudin. *J. Appl. Polym. Sci.* **20**(11), 2143–2153 (1998).
23. V.R. Patel and M.M. Amigi, Preparation and characterization of freeze-dried chitosan-poly(ethylene oxide) hydrogels for site-specific antibiotic delivery in the stomach. *Pharm. Res.* **13**(4), 588–593 (1996).
24. P.R. Hari, T. Chandy, and C.P. Sharma, Chitosan/calcium-alginate beads for oral delivery of insulin. *J. Appl. Polym. Sci.* **59**(11), 1795–1801 (1996).
25. P. Calvo, C. Ramona-Lopez, J.L. Vila-Jato, and M.J. Alonso, Novel hydrophilic chitosan-polyethylene oxide nanoparticles as protein carriers. *J. Appl. Polym. Sci.* **63**(1), 125–132 (1997).
26. P. Calvo, C. Ramona-Lopez, J.L. Vila-Jato, and M.J. Alonso, Chitosan and chitosan/ethylene oxide-propylene oxide block copolymer nanoparticles as novel carriers for proteins and vaccines. *Pharm. Res.* **14**(10), 1431–1436 (1997).
27. A.B. Schmuck and M.E. Krajicek, Mucoadhesive polymers as platforms for peroral peptide delivery and absorption: Synthesis and evaluation of different chitosan-EDTA conjugates. *J. Controlled Release* **50**(1–3), 215–223 (1998).
28. J.I. Morata and T. Ouchi, Chitosan derivative having a function carrying DNA. *High Polym., Jpn.* **46**(5), 339 (1997).
29. J.G. Jegsl and K.H. Lee, Chitosan membranes crosslinked with sulfosuccinic acid for the pervaporation of water/alcohol mixtures. *J. Appl. Polym. Sci.* **71**(4), 671–675 (1999).
30. E.E. Skorikova, R.I. Kalyvzhnaya, G.A. Vikhoreva, L.S. Galbraikh, S.L. Kotova, E.F. Ageev, and A.B. Zezin, Complexes of chitosan and poly(acrylic acid). *Macromol. Polym. Sci. USSR, Ser. A.* **38**(1), 61–65 (1996).
31. E.F. Ageev, S.L. Kotova, E.E. Skorikova, and A.B. Zezin, Pervaporation membranes based on polyelectrolyte complexes of chitosan and poly(acrylic acid). *Polym. Sci. USSR (Engl. Transl.), Ser. A* **38**(2), 323–329 (1996).
32. T. Uragami, K. Tsukamoto, K. Invi, and T. Miyata, Pervaporation characteristics of a benzoylchitosan membrane for benzene-cyclohexane mixtures. *Macromol. Chem. Phys.* **199**(1), 49–54 (1998).
33. S.Y. Nam and Y.M. Lee, Pervaporation separation of methanol/methyl-t-butyl ether through chitosan composite membrane modified with surfactants. *J. Membr. Sci.* **157**(1), 63–71 (1999).
34. W. Wang, and H.G. Spencer, Formation and characterization of chitosan formed-in-place ultrafiltration membranes. *J. Appl. Polym. Sci.* **67**(3), 513–519 (1998).
35. I.Y. Galaev, M.N. Gupta, and B. Mattiasson, Use smart polymers for bioseparations. *CHEMTECH* **26**(12), 19–25 (1996).
36. E. Klein, E. Eichholz, F. Theimer, and D. Yeagor, Chitosan modified sulfonated poly(ether sulfone) as a support for affinity separations. *J. Membr. Sci.* **95**(2), 199–204 (1994).
37. M. Chellapandian, and M.R.V. Kridhnsn, Chitosan-poly(glycidyl methacrylate) copolymer for immobilization of urease. *Process Biochem.* **33**(6), 595–600 (1998).
38. A. Gallifuoco, L.D. Ercole, F. Alfani, M. Cantarella, G. Spagna, and P.G. Pifferi, On the use of chitosan-immobilized

- β -glucosidase in wine-making: Kinetics and enzyme inhibition. *Process Biochem.* **33**(2), 163–168 (1998).
39. A. Denuziere, D. Ferrier, O. Damour, and A. Domarel, Chitosan-chondroitin sulfate and chitosan-hyaluronate poly-electrolyte complexes biological properties. *Biomaterials* **19**(14), 1275–1285 (1998).
40. S. Sun, Y.W. Wong, K.D. Yao, and A.F.T. Mak, A study on mechano-electro-chemical behavior of chitosan (polyethylene glycol) composite fibers. *J. Appl. Polym. Sci.* **6**(4), 542–551 (2000).
41. H. Jing, W. Su, S. Caracci, T.G. Bunning, T. Cooper, and W.W. Adams, Optical waveguiding and morphology of chitosan thin films. *J. Appl. Polym. Sci.* **61**(7), 1163–1171 (1996).
42. H. Jing, J. Liang, J. Grant, W. Su, T.G. Bunning, T. Cooper, and W.W. Adams, Characterization of chitosan and rare-earth-metal-ion doped chitosan films. *Macromol. Chem. Phys.* **198**(4), 1561–1578 (1997).
43. H. Jing, W. Su, M. Brant, M.E. De Rosa, and T.G. Bunning, Chitosan-based hydrogels: A new polymer-based system with excellent laser-damage threshold properties. *J. Polym. Sci., Part B* **37**(8), 769–778 (1999).

COATINGS

CHAO-NAN XU

National Institute of Advanced
Industrial Science and Technology (AIST)
Tosu, Saga, Japan

INTRODUCTION

Many materials emit light during the application of a mechanical energy. This phenomenon is usually referred to as mechanoluminescence (ML) or triboluminescence (1). The more historical term is “triboluminescence.” It stands for tribo-induced luminescence, and this was the term used for more than a century to refer to light emission induced by any type of mechanical energy (2). The term “mechano-luminescence” was not proposed until 1978 (3). The prefix “mechano” is correlated to the general mechanical way used for exciting luminescence, including concepts such as deformation, piezo, tribo, stress, cutting, grinding, rubbing, and fracto. In recent years mechanoluminescence (ML) has become the preferred nomenclature (4). Although the transfer of mechanical stress into light radiation is very complex, successes in experimental applications suggest possible uses of the ML phenomena in stress sensors, mechanical displays, and various smart systems.

In general, ML can be divided into fractoluminescence (destructive ML) and deformation luminescence (nondestructive ML); these correspond to the luminescence induced by fracture and mechanical deformation of solid, respectively. Roughly 50% of solid materials gives fractoluminescence by fracture (5): the well-known materials include sugar (6), molecular crystals (7,8), alkali halides (9,10), quartz (11), silica glass (12–14), phosphors (15), piezoelectric complex (16), metals (17), various minerals (18,19), and biomaterials (20). Recently, the fractoluminescence of rare-earth complexes was investigated

in order to build smart damage sensors capable of simple real-time detection of the magnitude and location of structural damage within materials (21). Deformation luminescence can be induced by mechanical deformation without fracture, and this is of interest in nondestructive evaluation. Deformation luminescence can be further divided into plasticoluminescence and elasticoluminescence. The former is produced during plastic deformation of solids, where fracture is not required, and the later is produced during the elastic deformation of solids where neither plastic deformation nor fracture is required. Nondestructive ML due to plastic deformation has been observed in several materials such as colored alkali halides (22,23), II–VI semiconductors (24), and rubbers (25). However, ML in the elastic region has been observed only for the irradiated alkali halides (4,26), and some piezoelectric materials (27). So far nondestructive luminescence intensities of materials have been reported to be too weak and difficult to repeat, and this has deferred any practical application of the phenomenon. For application of ML in developing new materials, repetitive ML must occur with undiminished intensity.

Devices for ML Measurement

Both mechanical and optical devices are being used to measure ML. The objective is to apply the measured mechanical energy to the ML sample, and then to detect the light induced by the mechanical energy. The various techniques already investigated include compression, bending, stretching, loading, piston impact, needle impact, cleaving and cutting, laser, shaking, air-blast, scratching, grinding and milling, and tribo- and rubbing (4). Figure 1 gives popular measurement devices for nondestructive ML; these devices measure compression, tension, bending, and shearing. Figure 1(a) shows a schematic diagram of an ML measurement device capable of measuring ML strain-stress relations simultaneously. Stress is applied on each sample by a materials test machine. The ML intensity is measured by a photon-counting system that consists of a photo multiplier tube (R464S, Hamamatsu Photonics) and a photon counter (C5410-51, Hamamatsu Photonics) controlled by a computer. The ML emission light is guided to the photo multiplier through a quartz glass fiber. The ML spectrum is obtained with a photon multichannel analyzer system (PMA 100, Hamamatsu Photonics). The ML images are recorded with an image intensified charge coupled device (ICCD) controlled by a computer system (C6394, Hamamatsu Photonics Corp.). Simultaneously, the stress and strain of the sample are measured by an in-situ sensor. In addition to compressive test, the materials test machine shown in Fig. 1(a) is able to apply tensile and bending stresses by exchanging the sample holder.

Figure 1(b) shows a schematic diagram of an ML measurement device for applying friction (shear stress); the same equipment as shown in Fig. 1(a) is used to measure the ML intensity and spectrum. The mechanical friction is applied by a friction rod under a load. The friction rod material, as well as the load, can be changed for different levels of mechanical stress applied to the test material. As

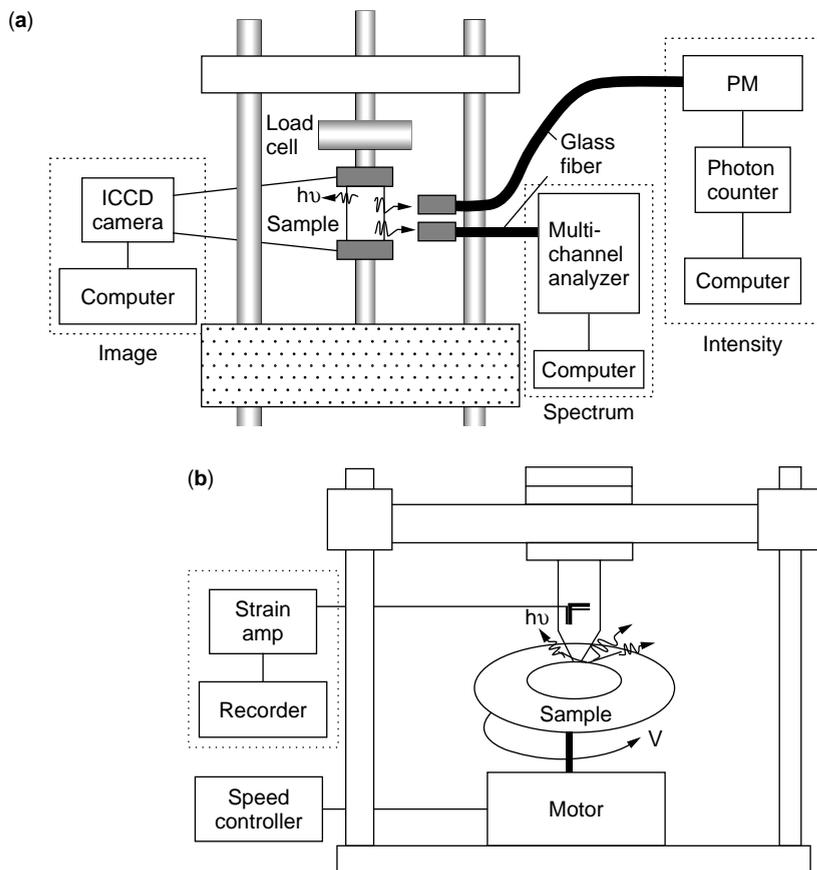


Figure 1. Schematic diagram of ML measurement devices (a) using a materials test machine for the compressive test and (b) using a friction test machine for the friction (shear stress) test.

the test sample is rotated at a controlled speed, as shown in Fig. 1(b), the friction rod draws a concentric circle on the test material. The ML emission light induced by friction is guided to the PM through a quartz glass fiber that is 3-mm in diameter; the distance between the glass fiber and the friction tip is set at 40 mm. The mechanical impact is applied by using a free-falling ball through a steel guide pipe. The impact velocity is adjusted by the height of the falling ball, and the impact energy can also be adjusted by both the weight of the ball and the falling height.

NONDESTRUCTIVE ML FROM ALKALINE ALUMINATES DOPED WITH RARE-EARTH IONS

As previously mentioned, development of materials with strong nondestructive ML intensity is an important goal in exploring applications of ML. Recently, systematic materials research has resulted in producing a variety of materials that emit an intensive and repeatable ML during elastic deformation without destruction: among these are ZnS: Mn, $\text{MAl}_2\text{O}_4\cdot\text{Re}$ ($M = \text{alkaline metals}$, $\text{Re} = \text{rare-earth metals}$), and $\text{SrMgAl}_{10}\text{O}_{17}\cdot\text{Eu}$ (28–32). So far the most promising ML materials are the rare-earth ions doped alkaline aluminates and the transition metal ions doped zinc sulfide. Remarkable upgrading in ML intensity has been achieved in the SrAl_2O_4 doped with europium by controlling the lattice defects in the material.

Preparation of Fine Particles of ML Materials and Their Smart Coatings

The luminescence powders are normally produced by a solid reaction process using a flux. In the solid reaction process, the starting materials of ultrafine powder of SrCO_3 , Al_2O_3 , and $\text{Eu}(\text{NO}_3)_3\cdot 2\text{H}_2\text{O}$, H_3BO_3 (flux) are thoroughly mixed. The mixture is calcined at 900°C for 1 h and then sintered at 1300°C for 4 h in a reducing atmosphere ($\text{H}_2 + \text{N}_2$). However, this process has the limitation that small particles cannot be obtained because of the growth of grains that occurs during the calcinations at high temperatures. To address this problem, a modified sol-gel method has been developed for preparing fine powders of SAO-E (33). In this modified sol-gel process, the starting materials of $\text{Sr}(\text{NO}_3)_2$, $\text{Al}(\text{O}-i\text{-C}_3\text{H}_7)_3$, and $\text{Eu}(\text{NO}_3)_3\cdot 2\text{H}_2\text{O}$ are dissolved in H_2O and thoroughly mixed with $\text{NH}_3\cdot\text{H}_2\text{O}$. The sol solution is then dispersed by $\text{HCON}(\text{CH}_3)_2$, followed by drying, calcining, and finally sintering in a reducing atmosphere at 1300°C for 4 h. In comparison with SAO-E powders synthesized by the solid reaction process, finer particles are obtained by the modified sol-gel process, their mean size being about $1\ \mu\text{m}$; the particles obtained by the solid reaction process are about $15\ \mu\text{m}$. The finer particles exhibit high ML, and as a result smart coating can be applied in uniform layers on the surfaces of the target objects by mixing it with binders and polymers. For example, standard coating techniques such as spin coating and spray

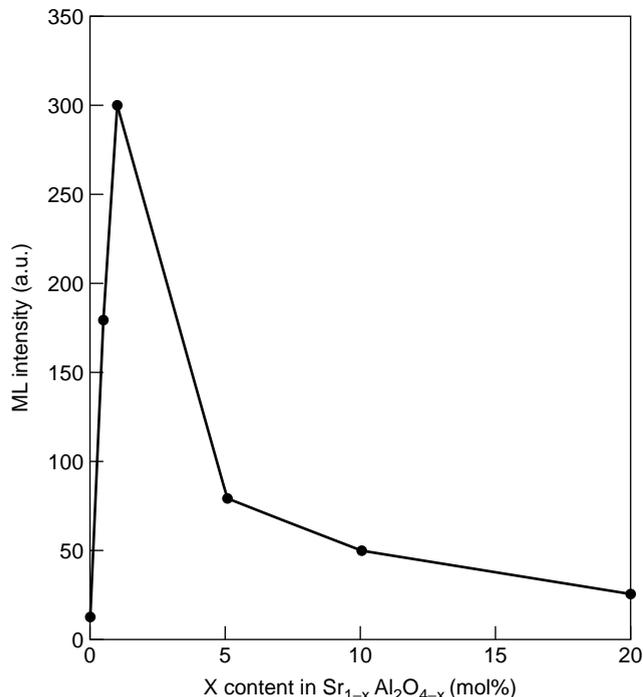


Figure 2. Dependence of ML intensity on defect concentration of Sr.

coating could then be used to create uniform layers (films) of SAO-E/epoxy on the surfaces of plastics, rubbers, glass, ceramics, and metals. The thicknesses of the resultant coating could be controlled from micrometer to millimeter.

Smart coatings made by mixing SAO-E fine powder with an optical epoxy can transfer mechanical energy into the photo energy of light emission, which in turn is capable of sensing the dynamic stress of substrate materials. To obtain the ML intensity and stress distribution, the composite samples of the SAO-E/epoxy have been included in the ML measurement together with the coated samples and the ceramics of SAO-E.

ML Response of SAO-E to Various Stresses

In upgrading the ML intensity of SAO-E for the smart coating application, the composition, the PH value, and the calcination conditions must be controlled. Significant

improvements in ML intensity can be made by optimizing the concentration of defects in the SAO-E system. Figure 2 shows the relationship of ML intensity to the defect concentration of Sr. The ML intensity is seen to strongly correspond to the defect concentration of Sr. The highest ML intensity is obtained by $Sr_{0.975}Al_2O_{3.985}:Eu_{0.01}$ with a lattice defect of 1.5 at% Sr vacancy. Such a nonstoichiometry system is found to produce one order of magnitude higher ML than a stoichiometry system. This is four orders of magnitude higher in intensity than that of the reported strong ML material of a quartz crystal. The luminescence of the defect-controlled SAO-E material gives a high enough ML to monitor the stress of the object it coats.

The influences of the stress and strain rates on the emitted light intensity are measured using the same strain rate but at different peak stresses, and then the same peak stress but at different strain rates. Figure 3(a) shows a time history of a luminescent object recorded during the application of a compressive stress with a constant strain rate for a SAO-E/epoxy composite with a dimension of $55 \times 29 \times 25 \text{ mm}^3$. As can be seen, over time a linearly increased load system results in linearly increased ML intensity. That is, the ML intensity emitted is linearly proportional to the magnitude of the applied stress. Figure 3(b) shows the ML intensity diminished during repetitive cycles of loading. The ML intensity decreased with the repetitive cycles of stress, reaching a stable level at about 20% of its initial strength. The ML intensity of SAO-E recovered completely after UV irradiation (365 nm) using a handy lamp. Such recoverability distinguished SAO-E from other nondestructive ML materials reported, for example, alkali halides which need γ -ray irradiation or very high energy irradiation to recover their intensivity (34). The linear relationship between ML intensivity and stress has been further demonstrated by the tests in which each test had the same strain rate but a different peak stress, as shown in Fig. 4. The ML intensity increased linearly with the increase of strain rate. Such a linear relation was also reported for γ -ray irradiated single crystals of alkali halides during the elastic deformation (35). It is evident that the ML of SAO-E is an elasticoluminescence. In a comparison test, the defect controlled SAO-E was found to give the most intense elasticoluminescence among the materials examined to date (36).

SAO-E was found also to exhibit ML during plastic deformation and fracture. In the region of plastic

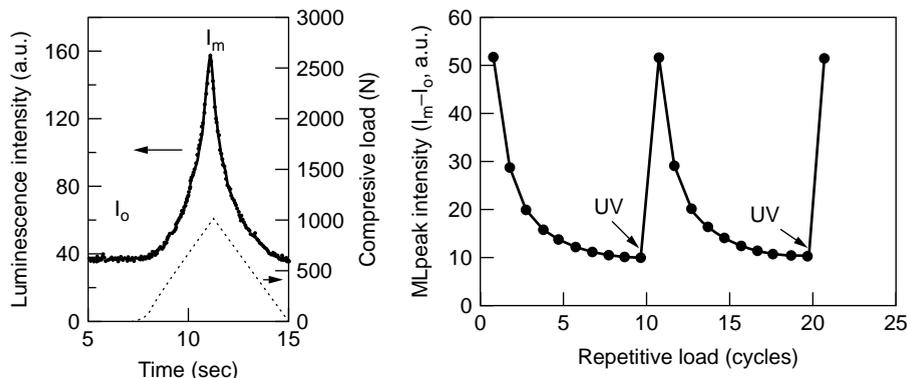


Figure 3. (a) Typical time history of the luminescence intensity recorded during a compressive test for a SAO-E/epoxy composite with a dimension of $55 \times 29 \times 25 \text{ mm}^3$, (b) Diminishment in ML peak intensity during application of repetitive cycles of loading and its recoverability with UV irradiation of 365 nm.

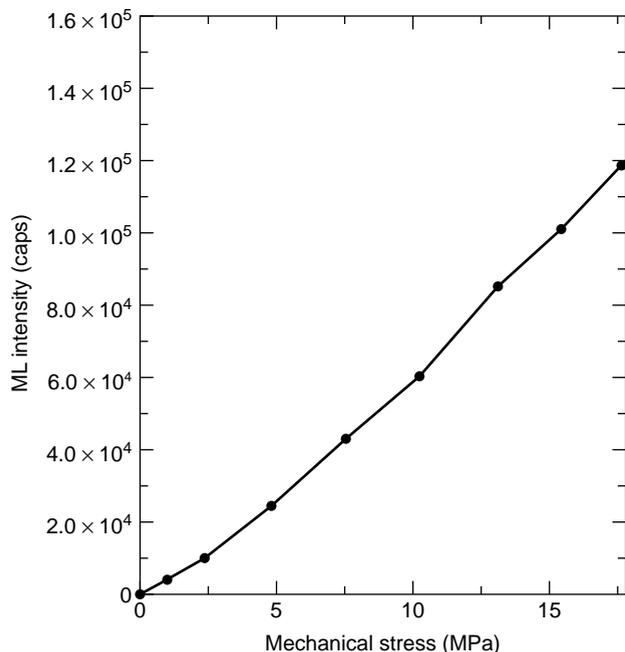


Figure 4. Influence of applied stress on ML intensity (strain rate is 3.0×10^{-4} 1/s).

deformation, ML became intensified because of the stress that was concentrated in this region. Upon further load increases the ML intensity exhibited a sharp rise as the SAO-E material began to crack, revealing the maximum value in ML intensity at fracture. Similar results were obtained for ceramic samples of SAO-E. The results for alkali halide crystals confirmed the presence of intense ML in plastic deformation and that it sharply increases during fracture of the crystals (4,5). Clearly, the linear relation between strain and stress is an important factor in the application of ML stress sensors.

Figure 5 shows the relationship between the strain rate and ML intensity. The relationship is almost linear, as a higher strain rate produces greater ML intensity.

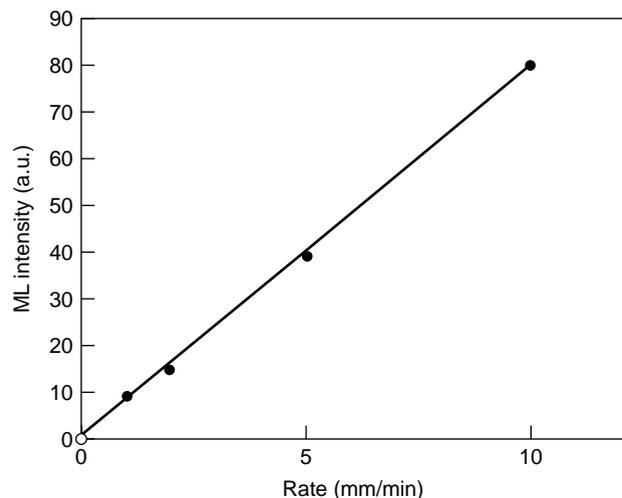


Figure 5. Influence of the strain rate on the ML intensity.

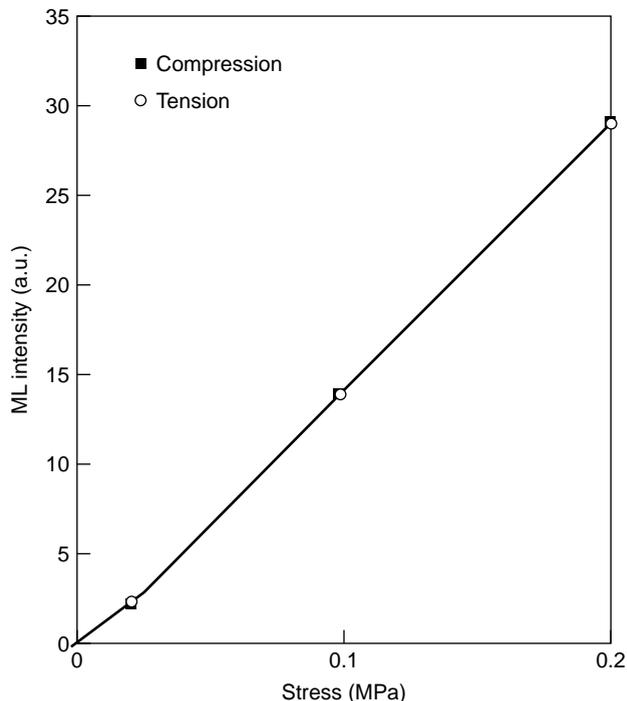


Figure 6. Dependence of ML intensity on tensile stress for a SAO-E/epoxy composite sample with a dimension of $100 \times 25 \times 5$ mm³. The ML intensity induced by compressive stress is also plotted for comparison.

Expressed mathematically, the ML intensity in terms of the stress and strain rate is

$$I - I_0 = K\sigma(t)\dot{\epsilon}(t), \quad (1)$$

where σ is the stress and $\dot{\epsilon}$ is the strain rate. Equation (1) indicates that the linear intensity of ML is consistent with the concept of an elastic (linear) region.

The dependence of ML on tensile stress is shown in Fig. 6 for a SAO-E/epoxy composite sample with measurements of $100 \times 25 \times 5$ mm³. For comparison, the ML intensity induced by compressive stress is also plotted. Note that the SAO-E exhibits the same ML intensity whether or not the stress is compressive or tensile. Figure 7 shows the dependence of ML on torsion measured by a shearing test machine for a SAO-E composite sample with dimensions of $\phi 10$ mm \times 110 mm. Note that the ML intensity increases linearly with the increase of torsion. Clearly, the nondestructive ML of SAO-E can detect shear stress and torsion changes without any volume changes. This is very different from the thermography technique that can detect stress based on the thermo-elastic effect when stress is accompanied by a volume change.

As previously mentioned, SAO-E ceramics and composites exhibit distinct ML behavior. Smart coating with SAO-E/epoxy can be applied to objects to sense changes of stress. Figure 8 shows the dependence of ML intensity on the stress for a plastic coated with an SAO-E/epoxy layer. The results are similar to those of composites and ceramics. The ML intensity almost linearly increases with the

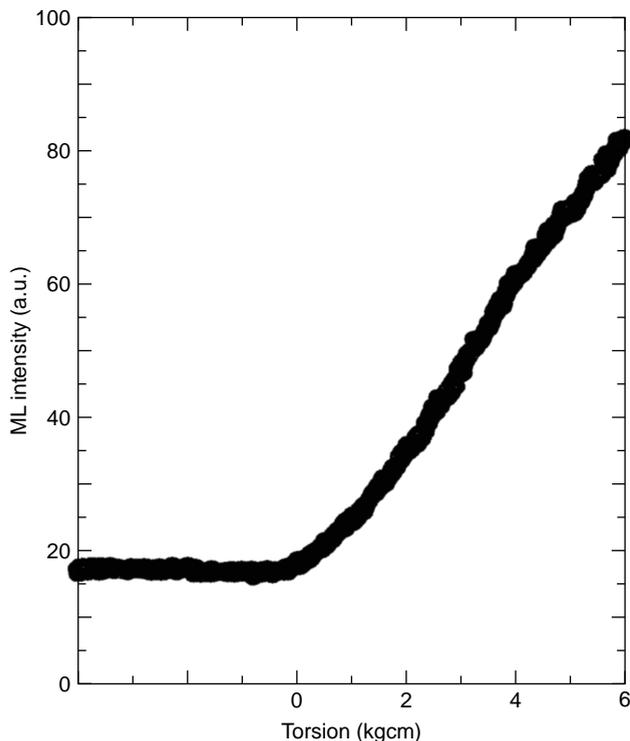


Figure 7. Dependence of ML intensity on torsion measured by a shear test machine for a SAO-E composite sample with dimensions of $\phi 10 \text{ mm} \times 110 \text{ mm}$.

increase of stress when the strain rate is kept constant. Additionally, the ML intensity increases with the increase of the thickness in the ML layer. A thick layer is generally not suitable for stress detection because it produces strain or stress in the ML layer that differs from that of the object beneath it. As these results indicate, a uniform layer is necessary for an accurate display of the stress distribution of the object.

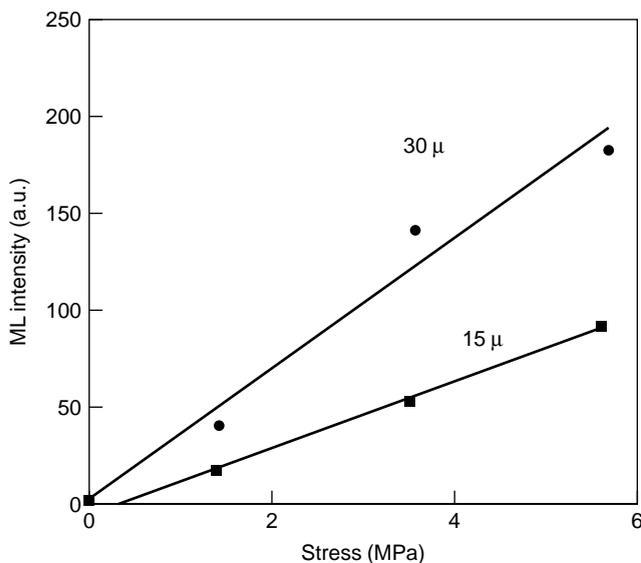


Figure 8. Dependence of ML intensity on the stress for a plastic block coated with SAO-E/epoxy layers with different thickness.

Contrary to the behavior of the destructive ML (37,38), the width of the ML emission band at the peak wavelength is independent of the stress level during elastic deformation. However, the ML intensity at peak emission increases linearly with the increase of stress level. Furthermore, the integrated intensity under the band also increases linearly with stress. As noted earlier, the ML of SAO-E can be induced by compressive, tensile, or shear stress. Smart coating using ML materials provides a simple way to detect these stresses dynamically and remotely.

ML Mechanism of SAO-E

The ML spectrum is measured by a photon multichannel analyzer. The broadband emission peaks at a wavelength of about 520 nm, which is the same as the photoluminescence (PL) spectrum measured by a fluorescence spectrometer. As shown in Fig. 9, the PL and ML spectra from SAO-E are characterized by emissions that peak near 520 nm. No other emission bands are found in the ML spectrum at 300 to 700 nm. This implies that ML is emitted from the same emission center of Eu ions as PL; both are produced by the transition of Eu^{2+} ions between $4f^7$ and $4f^65d^1$ (39,40). Emissions due to N_2 discharge have not been observed, which generally occur in destructive ML (fractoluminescence) (41). These results confirm that the ML of SAO-E described here is produced by a nondestructive deformation of SAO-E. Moreover, the recovery of ML intensity by UV irradiation suggests that the traps in SAO-E sample can be filled by UV irradiation. Measurements of the Hall effect of UV-activated SAO-E reveal traps of holes filled by UV, and this is consistent with other reports (39,44). The depths of these hole traps can be evaluated by the Hoogenstraaten method (42). This technique calls first for thermoluminescence glow curves to be measured at different rates of heating (β) to obtain the glow peak temperature (T_m) for each heating rate β , and then for the depth of trap (E_t) to be

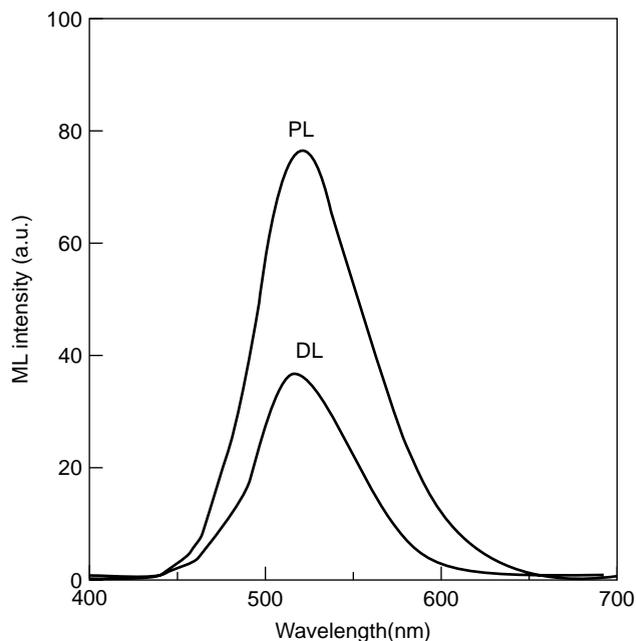


Figure 9. ML and photoluminescence (PL) spectra of SAO-E.

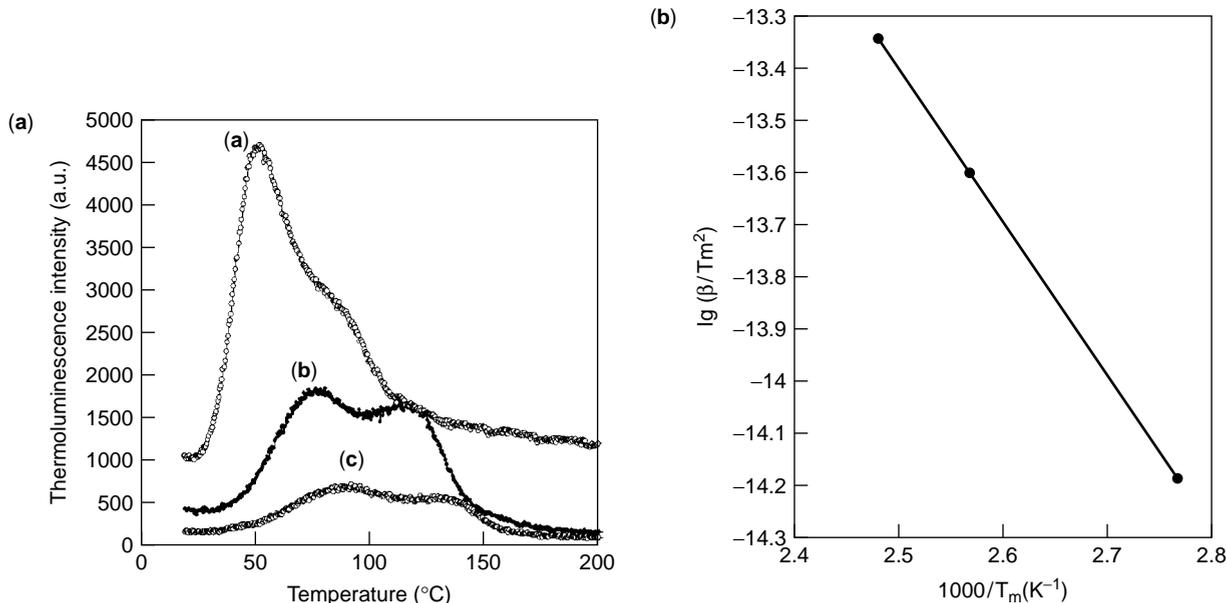


Figure 10. (a) Glow curves of thermoluminescence from SAO-E at different heating rate of 0.091, 0.187, and 0.259 K/s for A, B, and C, respectively; (b) resultant Hoogenstraate plot, where β is a heating rate and T_m is a glow peak temperature.

calculated from the slope of the Hoogenstraaten plots using the equation

$$E_t = \frac{-k \log_e (\beta / T_m^2)}{1/T_m}, \quad (2)$$

where k is the Boltzmann constant. Figure 10(a) and (b) shows the glow curves of SAO-E and the resultant Hoogenstraaten plots, respectively. Two glow peaks are found for SAO-E as shown in Fig. 10 (a), implying that there are different kinds of traps in the material. The depth of the trap associated with lower T_m is about 0.2 ± 0.1 eV, which is much higher than the thermal energy of 350 K (0.03 eV). Consequently these hole traps can not be thermally activated at room temperature. On the other hand, it is evident that traps at levels near 0.1 eV can be activated by deformation to release electrons to the conduction band in the case of KCl (41). The ML kinetic model for SAO-E proposed in Fig. 11 takes these results into account. During deformation the strain energy excites the filled traps (T^+) to release holes to the valence band (process 1). The holes then excite Eu^+ to produce Eu^{2+*} (process 2), and return to ground state by emitting a green light of about 520 nm (process 3).

The hole traps in the model are attributable to the lattice defects of Sr^{2+} (44), which can greatly affect the ML intensity as was indicated in Fig. 12. For this reason the SAO-E is classified as a defect-controlled ML material whose intensity is substantially altered by the existence of the trap (defect). The well-known alkali halides also belong to this outgoing. However, ZnS:Mn is a piezoelectric-induced material as will be described in the next section.

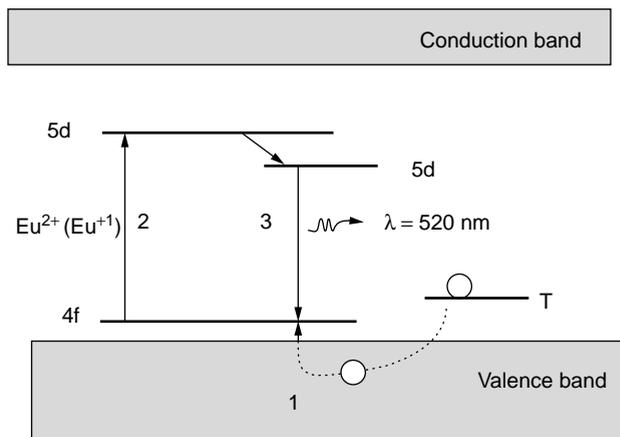


Figure 11. ML kinetic model for SAO-E.

REPEATABLE ML OF TRANSITION METAL IONS DOPED ZINC SULFIDE

As was previously noted, the nondestructive ML of SAO-E showed diminished intensity during the application of a repetitive stress cycle, similar to that of alkali halides (45). ZnS-doped Mn has been found to give undiminished ML intensity during the elastic deformation, so it is the representative material to achieve the most repetitive ML.

Preparation of Highly Oriented ZnS:Mn Films

Thin films can be prepared from a ZnS pellet on various substrate materials, including quartz glass, stainless, carbon, and ceramics (Al_2O_3 , SiC, Si_3N_4), by physical vapor depositions of the ion plating method (46). The substrates

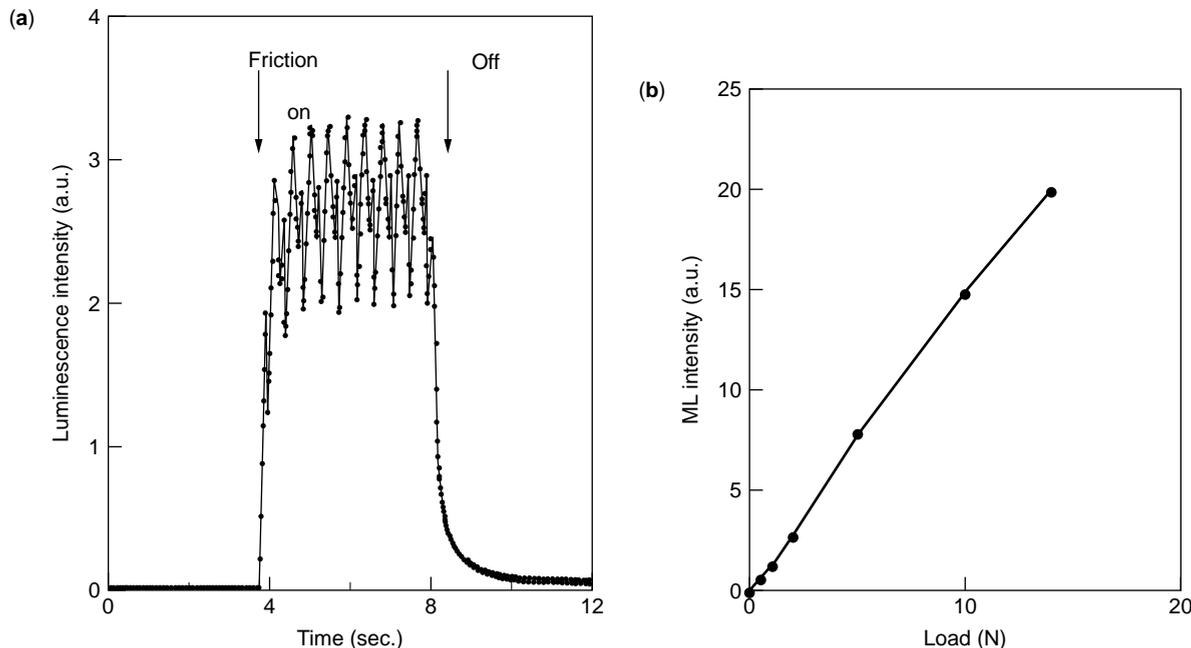


Figure 12. (a) ML response of ZnS:Mn to a friction (shear stress); (b) dependence of ML intensity on the friction load.

are kept at 160°C, and the deposition is carried out in a vacuum of 0.2 Pa in Ar atmosphere. The source pellet is prepared from $\text{Zn}_{0.985}\text{Mn}_{0.015}\text{S}$ powder by a cold isotropic press method followed by sintering in a vacuum-sealed quartz glass tube at 1000°C for 10 h. Such a pretreatment is applied to fabricate highly crystallized pellets whose deposition rate is very stable and easily controlled. A vacuum-sealed technique is used because ZnS begins to sublime at temperatures above 700°C. The as-grown films are annealed at 500 to 1000°C for 1 h in a vacuum-sealed quartz glass tube. Their chemical composition is determined by fluorescent X-ray spectrometric analysis. The Mn amount in the film showed the same as that in the source material, namely 1.5 at%. The XRD pattern exhibited a strong diffraction peak at 28.49° in the 2θ range of 10 to 90°, which was attributed to the (111) plane of the ZnS film was highly oriented.

ML Characteristics of ZnS:Mn Films

The luminescence intensity depends on the microstructure of film. Post-heat treatment in vacuum is effective for obtaining high ML in that it increases the crystallinity of ZnS:Mn and decreases the defects and residual stress in the film. Moreover, the connection between the film and substrate is strengthened by annealing, even for film thicker than 1 μm , and thus prevents deprivation due to moisture or mechanical attack. The ML intensity is enhanced by two magnitudes in order after annealing at 700°C, and the mechanical strength of the ZnS:Mn film is also remarkably strengthened (46). Surface observation shows that the film consists of ZnS grains with a mean size of several nano-meters.

Figure 12(a) shows the ML response of ZnS:Mn to friction (shear stress) measured by the device shown in

Fig. 1(b). The ML intensity increases steeply when the friction on the ZnS:Mn film is turned on, and the frequency of the oscillating change equals the rotating speed of the test sample. This oscillating change is also found in the friction force by a strain gauge attached to the friction rod. The oscillation may be caused by nonuniformity of the test material, as is similarly the case in bulk ceramics (47). The response curve is reproducible, and this indicates that the film is combined strongly onto the substrate, as is confirmed by the SEM image and adhesive strength test. This reproducibility distinguishes the ML of ZnS:Mn from the other destructive ML. It has been found that ZnS:Mn shows a repetitive ML response not only to friction but also to other types of stresses such as compressive stress (30). These results are substantially different from the other elasticoluminescent materials reported so far like gamma colored alkali halide crystals and the SAO-E, where the intensity decreased in a great deal during the application of repetitive stress. Apparently the reproducibility is essentially important for self-diagnosis materials and applications in various novel smart systems including stress sensors.

The ML intensity of ZnS:Mn increases with increasing the mechanical stress. Figure 12(b) shows that the intensity increases linearly with the increasing applied load. Correspondingly, the mechanical friction can be monitored without any mechanical contacts. The linear relation between ML and stress of the ZnS:Mn has also been found in the case of compressive stress.

Figure 13 shows the ML response to a mechanical impact. Note that the ML response transits of ZnS:Mn are similar to those of piezoelectric voltage responses, as reported previously (48). The energy conversion efficiency for converting mechanical energy to photon energy, roughly estimated from the experimental data, is on the order of

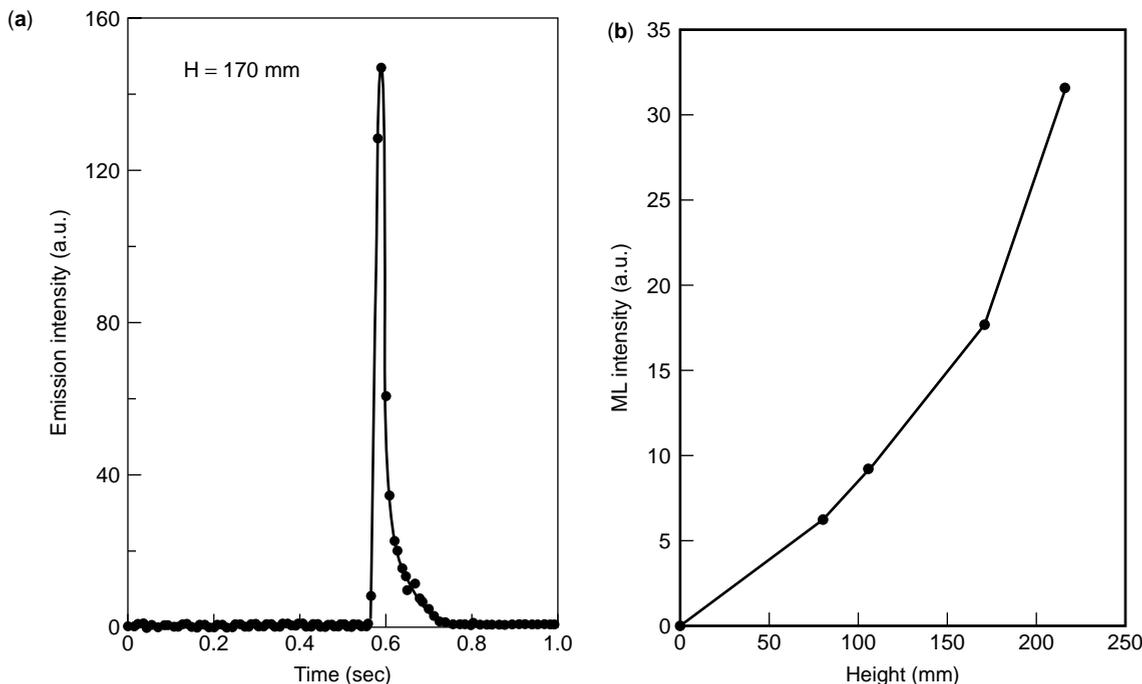


Figure 13. (a) ML response to a mechanical impact; (b) dependence of ML on the falling height.

10^{-2} and 10^{-6} for a slowly applied stress and impact cases, respectively; this is also on the same order as was reported previously for the piezoelectrics. The ML intensity increases linearly with the increase of the falling height of the free ball; that is, ML intensity increases with the increasing impact energy. Correspondingly, the mechanical impact can be remotely monitored using an ML film. More important, the ML intensity emitted by such a high-oriented films is much higher than that of the bulk material by more than one magnitude. The significant improvement in the ML intensity is attributed to the high orientation of the created film and the nano-sized grains of which it is composed.

Mechanism of ZnS:Mn

ZnS is both a piezoelectric and electroluminescent material. Figure 14 gives the spectrum of the ML for ZnS:Mn thin films, along with the photoluminescence (PL) and electroluminescence (EL). The ML exhibits a maximum emission band at 585 nm, which is consistent with the spectra for PL and EL of ZnS:Mn. No additional emissions due to the discharge of N_2 are found in the ML spectrum of ZnS:Mn. This indicates that the ML is introduced by stress and the emission arises from the emitting center of Mn^{2+} ions, due to the transition of ${}^4T_1 \rightarrow {}^6A_1$. The stress-activated mechanism is supported by other experiments; for example, when covering the ZnS:Mn film with a transparent film of AlN, similar emission is seen. Figure 15 shows the ML and PL intensities for ZnS:Mn films deposited on various substrates. It is seen that the ML intensities for conductor substrates like stainless and carbon are much lower than those for dielectric substrates such as quartz, alumina, silicon nitride, and silicon carbide. When

a shear stress is applied on the ZnS:Mn film, a piezoelectric voltage based on the piezoelectric coefficient of d_{14} will be generated between the opposite sides of the thin film surfaces. If the film is deposited on a conducting substrate, then electrical leakage may occur. The presence of such leakage is indicated by low ML intensity for ZnS:Mn on steel and carbon substrates as shown in Fig. 15.

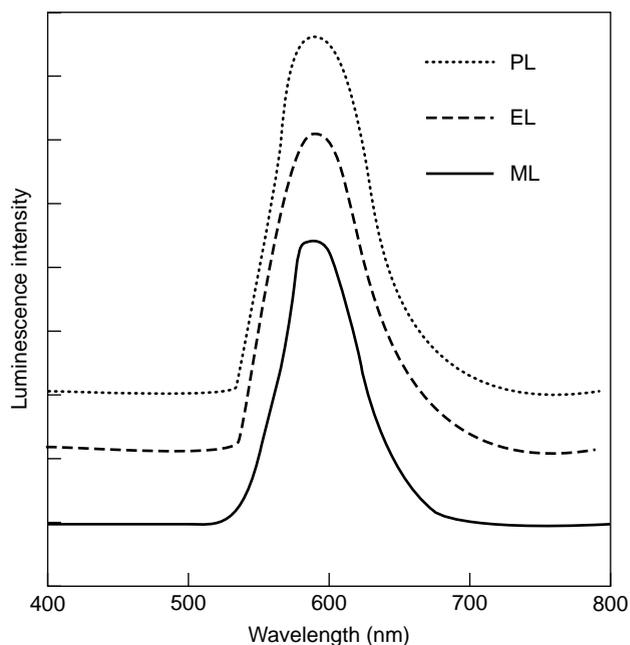


Figure 14. Spectrum of the ML for ZnS:Mn thin films, along with the photoluminescence (PL) and electroluminescence (EL) spectra.

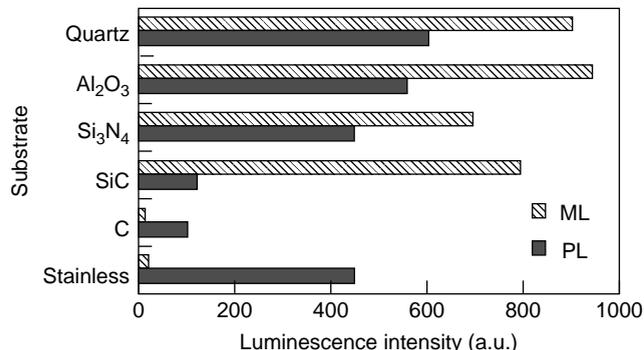


Figure 15. ML and PL intensities for ZnS:Mn films coated on various substrates.

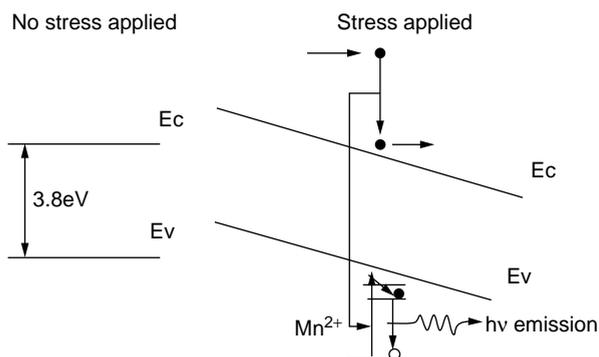


Figure 16. ML model of ZnS:Mn film, where the ML of ZnS:Mn is proposed to be based on the superposition of piezoelectric effect and electroluminescence.

Taking these results into account, the ML phenomenon of ZnS:Mn can be interpreted by a piezoelectric-induced electroluminescence model as shown in Fig. 16. In the figure the ML of ZnS:Mn is considered to be the superposition of piezoelectric effect and electroluminescence, and this can also be considered as the inverse effect of photostriction (49). The proposed ML model can well explain the distinguished behavior of ML. It is evident that the high orientation and crystallinity of the ZnS:Mn film give higher piezoelectric performance. The higher piezoelectric voltage produced on the opposite sides of the nano-sized grains of ZnS leads to a higher intensity of electroluminescence. Meanwhile, the EL efficiency is also improved with the increase in crystallinity (50–52). Therefore the total effect is higher ML intensity. The repeatable ML with an undiminished intensity of ZnS:Mn is attributed to the reproducibility of the piezoelectric effect.

APPLICATION OF SMART COATING WITH ML MATERIALS FOR NOVEL STRESS DISPLAY

Since ML intensity increases linearly with the increase of stress and strain rate in the elastic region, the ML layer is

believed to be able to display a stress distribution of any object that it covers.

Stress distribution is measured in solids to improve their reliability and extend their applications. The distribution of stress in a solid is conventionally evaluated using several techniques (53,54). Electric resistance strain gauges and piezoelectric sensors are typical techniques using electrical signals. The limitation to these methods becomes evident in analyzing the distribution on the micro scale because of their size. The sensors must maintain electrical contact to the target objects, so it is difficult to measure the stress distribution of a dynamic moving part such as cutting tool or gas turbine. Remote detection has permitted the use of optical signals in experimental stress analyses utilizing photoelastic and photoplastic effects, X-ray diffraction, optical fiber networks embedded in composites, and thermography based on thermoelastic analysis, for example. However, until recently there were no simple techniques for the direct visualization of the stress distribution in real time. Current studies to solve this problem have focused on building self-diagnosis systems using smart coatings of piezoelectric (55–57) and the ML materials (30,58,59).

Visualizing Stress Distribution

To view the stress distribution of an object, a smart coating of ML material is applied on the surface of the object. The ML images are recorded during the application of stress on the ML-layer/object. Figure 17 shows an example of a plastic disk coated with a SAO-E/epoxy film (50 μm). The intense green light emitted into the atmosphere from the two ends of the stressed sample can be clearly observed by the naked eye. Also various stress images can be simulated using the finite element method (60). It has been found that the strain energy distribution is in agreement with the ML image, as compared in Fig. 17(b) and (c). This is supported by the argument in the previous sections. In addition, since the strain rate is a constant in the measurement, this ML image is consistent with the stress distribution. Figure 18 illustrates the stressed sample and the line distribution of ML intensity and stress along CC' axis. The ML intensity distribution along $CO C'$, which was obtained experimentally by the ICCD camera, is given by a solid line, and the stress distribution of the sample, which was simulated theoretically based on elastics (61), is shown by a dotted line. As compared in Fig. 18(b), the simulated compressive stress along $CO C'$ increases exponentially with the increasing of r/R . This is consistent with the line profile of the ML image, indicating that the ML image reflects well the stress distribution (stress image) under the experimental conditions. Therefore, smart coating with ML materials can directly display the stress distribution of the object beneath the layer.

Monitoring in Dynamic Stress and Impact

Dynamic ML images have been successfully recorded during the application of different stresses, including bending, tension, compression, and impact (62). The ML images

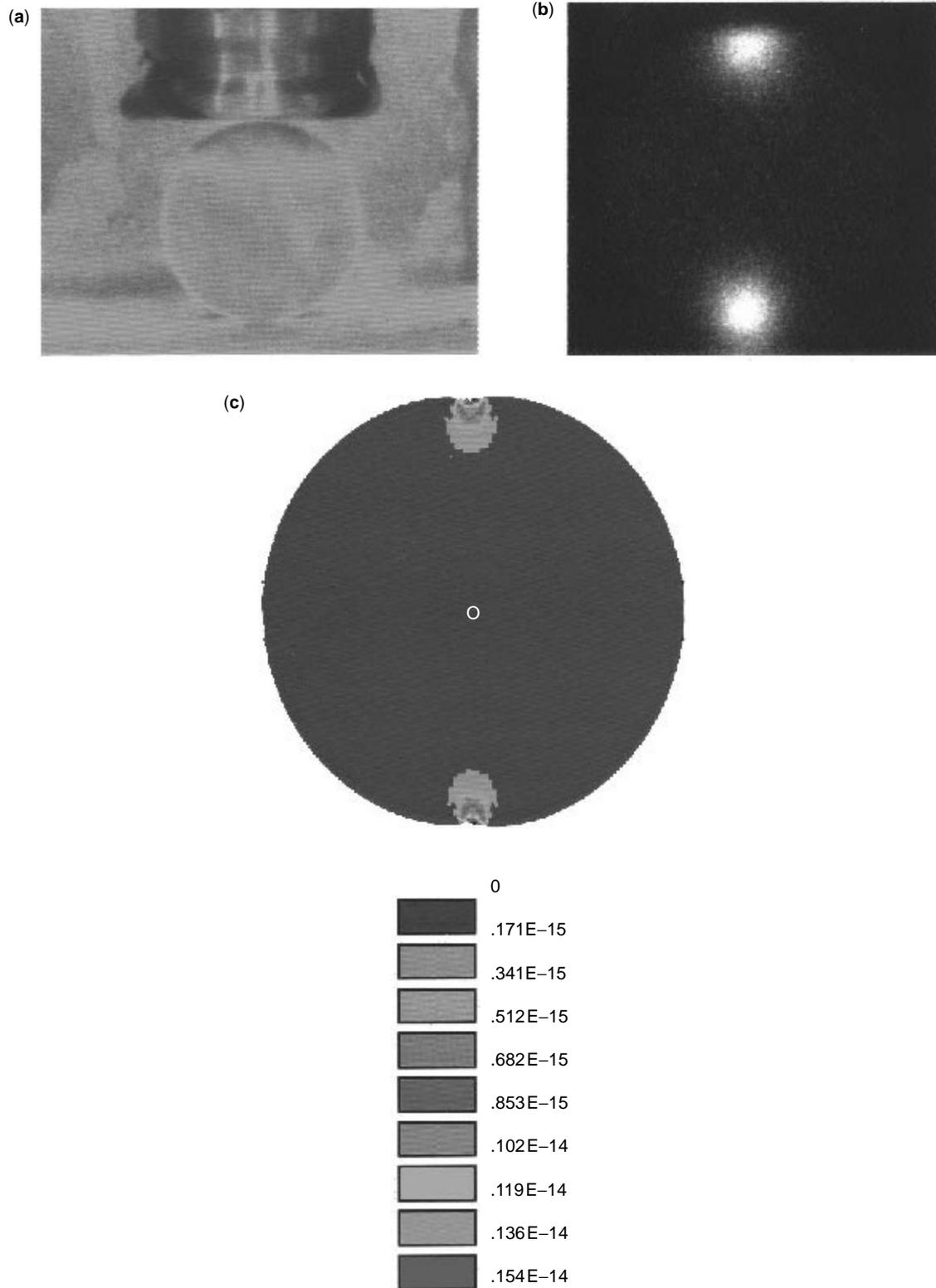


Figure 17. Stress distribution images for plastic disc ($\phi 25 \text{ mm} \times 10 \text{ t mm}$) coated with SAO-E/epoxy layer under compressive test: Stressed sample (a); ML image at a load of 500 N (b); simulated image of strain energy distribution using the finite element method (c).

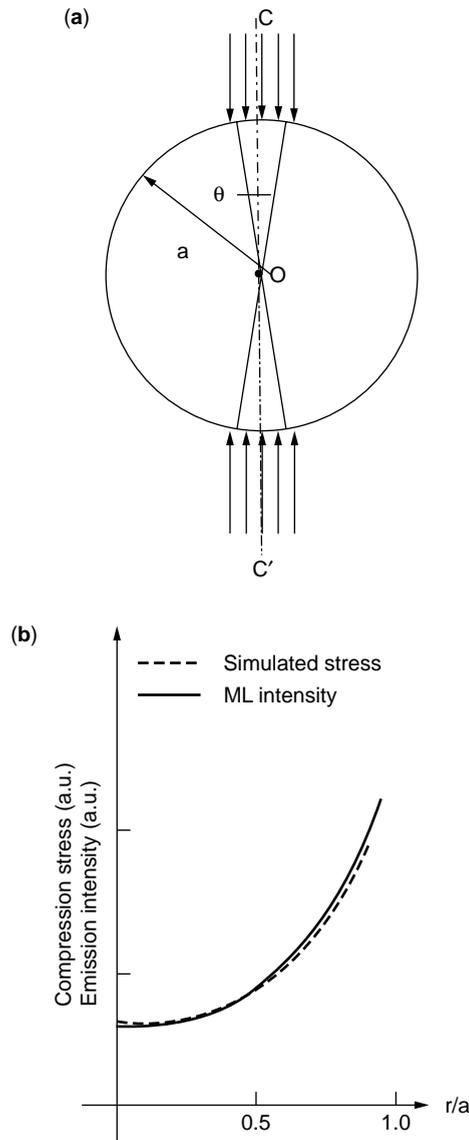


Figure 18. Stressed sample (a) and the line distribution profiles of ML intensity and stress along CC' axis (b).

dynamically changed with the loading, and were found to be in good agreement with the stress concentration results obtained by computer simulation and other experimental stress analyses. This imaging method gives the dynamic stress distribution in real time. It is distinguished from thermography, which requires repetitive cycles of stresses and thus is limited in evaluating stress during the fatigue process. Moreover, the present ML images are strong enough to be seen by the naked eye in a darkened room.

Photographs in Fig. 19(a) and (b) show the dynamic visualization of impact and friction, respectively, for a quartz substrate coated with the (111)-plane-oriented ZnS:Mn film. After applying mechanical impact caused by a free-falling ball, the yellow emission shown in Fig. 19(a) was recorded. Mechanical friction caused the strong ML

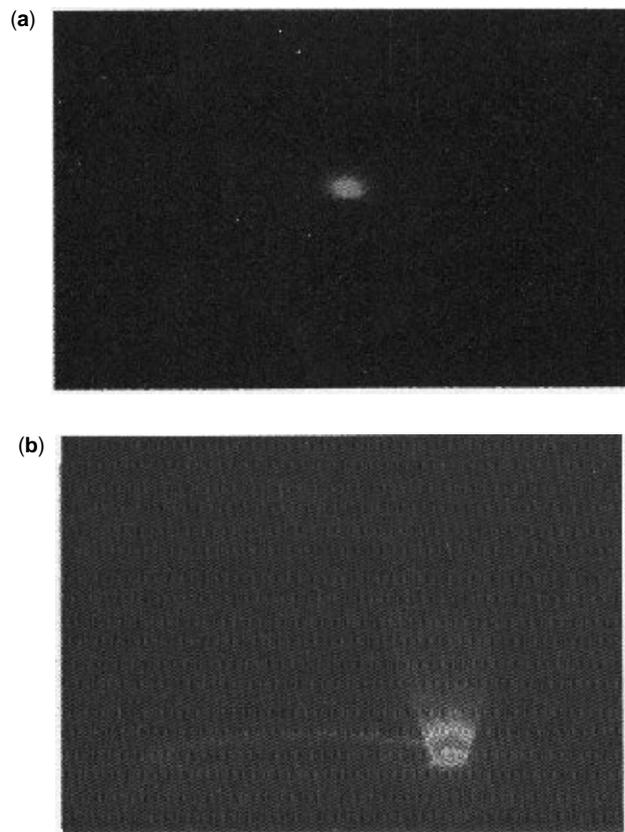


Figure 19. Photographs of the dynamic visualization of impact (a) and friction (b) for a quartz substrate coated with the (111)-plane-oriented ZnS:Mn film.

recorded in real time. The ML emission from the ZnS:Mn film was strong enough to be clearly seen by the naked eye.

The same method can be applied in an aqueous environment. Real-time ML images of stress distributions were obtained in water, ethanol, acetone, and 0.1 M HCl, for example, although the ML intensity values were dependent on the environment due to the different refraction and adsorption values. These results show the practicability of the present method in environment uses.

In particular, the ML smart coating technique holds much promise for observing the stress distribution with high spatial resolution using ML materials with nanoscale particles and the optical microscopy with high resolution. Although more experiments are needed, in the near future stress distributions in a scale smaller than micrometers should become observable as image techniques are combined with microscopy. Already the smart coating technique has enabled us to view stress distributions on both macro and micro scales.

The application of smart coating with an ML layer to analyze dynamic stress not only provides a new method for nondestructive evaluation of materials. In addition to the mechanical display, ML smart coating has opened a window on developing new smart systems and optomechanical devices.

BIBLIOGRAPHY

1. B.P. Chandra and A.S. Rathore. *Cryst. Res. Technol.* **30**(7): 885–896 (1995).
2. G. Wiedemann and F. Schmidt. *Ann. Phys. Chem.* **54**(4): 604–625 (1895).
3. B.P. Chandra and K.K. Shrivastava. *J. Phys. Chem. Solids* **39**: 939–940 (1978).
4. D.R. Vij. *Luminescence of Solids*. Plenum, New York, 1998 (B.P. Chandra), pp. 361–390.
5. A.J. Walton. *Adv. Phys.* **26**(6): 887–948 (1977).
6. J.I. Zink, G.E. Hardy, and J.E. Sutton. *J. Phys. Chem.* **80**(3): 248–249 (1976).
7. B.P. Chandra, M.S. Khan, and M.H. Ansari. *Cryst. Res. Tech.* **33**(2): 291–302 (1998).
8. B.P. Chandra and J.I. Zink. *J. Chem. Phys.* **73**: 5933–5941 (1980).
9. G.N. Chapman and A.T. Walton. *J. Phys. C: Solid State Phys.* **16**: 5542–5543 (1983).
10. K. Meryer, D. Obrilcat, and M. Rossberg. *Kristall U. Tech.* **5**: 5–49, 181–205 (1970).
11. G.N. Chapman and A.T. Walton. *J. Appl. Phys.* **54**(10): 5961–5965 (1983).
12. J.T. Dickinson, S.C. Langford, L.C. Jensen, G.L. Mcvay, J.F. Kelson, and C.G. Pantano. *J. Vac. Sci. Technol.* **A6**(3): 1084–1089 (1988).
13. J.I. Zink, W. Beese, J.W. Schindler, and A.T. Smiel. *Appl. Phys. Lett.* **40**: 110–112 (1982).
14. Y. Kawaguchi. *Phys. Rev.* **B52**: 9224–9228 (1995).
15. T. Ishihara, K. Tanaka, K. Hirao, and N. Soga. *Jpn. J. Appl. Phys.* **36B**(6): L781–783 (1997).
16. N. Takada, S. Hieda, J. Sugiyama, R. Katoh, and N. Minami. *Synthetic Met.* **111**: 587–590 (2000).
17. K.B. Abramova, I.P. Shcherbakov, A.I. Rusakov, and A.A. Semenov. *Phys. Solid State.* **41**(5): 761–762 (1999).
18. Y. Enomoto and H. Hashimoto. *Nature* **346**(6285): 641–643 (1990).
19. B.T. Brady and G.A. Rowell. *Nature* **321**(6069): 488–492 (1986).
20. V.E. Orel. *Medical Hypotheses* **40**: 267–268 (1993).
21. I. Sage, R. Badcock, L. Humberstone, N. Geddes, M. Kemp, and G. Bourhill. *Smart Mater. Struct.* **8**(4): 504–510 (1999).
22. B.P. Chandra. *Radiat. Eff. Defect.* **S138**(1–2): 119–137 (1996).
23. C.T. Butler. *Phys. Rev.* **141**: 750–758 (1966).
24. G. Alzetta, I. Chudacek, and R. Scarmozzino. *Phys. Stat. Sol.* **1a**: 775–785 (1970).
25. I. Grabec. *Non-destruct. Test.* **8**: 258–260 (1975).
26. N.A. Atari. *Phys. Lett.* **90A**(1–2): 93–96 (1982).
27. G.T. Reynolds. *J. Lumin.* **75**(4): 295–299 (1997).
28. C.N. Xu, T. Watanabe, M. Akiyama, and X.G. Zheng. *Mater. Res. Bull.* **34**(12): 1491–1500 (1999).
29. C.N. Xu, T. Watanabe, M. Akiyama, P. Sun, and X.G. Zheng. *Proc. 6th Int. Symp. Ceramic Materials and Composites for Engines*, Technoplaza, Tokyo, pp. 937–941 (1997).
30. C.N. Xu, T. Watanabe, M. Akiyama, and X.G. Zheng. *Appl. Phys. Lett.* **74**(9): 1236–1238 (1999).
31. C.N. Xu, T. Watanabe, M. Akiyama, and X.G. Zheng. *Appl. Phys. Lett.* **74**(17): 2414–2416 (1999).
32. C.N. Xu, H. Matusi, T. Watanabe, M. Akiyama, Y. Liu, and X.G. Zheng. *Proc. 4th Int. Conf. Intelligent Materials*, edited by T. Takagi, M. Aizawa, T. Okano, and N. Shinya. Intelligent Materials Forum, Tokyo, 1998, pp.146–148.
33. C.N. Xu, Y. Liu, and H. Tateyama. Japanese Patent 2001–017891 (2001).
34. B.P. Chandra and N.P. Bisen. *Crystal. Phys. Stat. Sol.* **132**(A): K101–K104 (1992).
35. B.P. Chandra, M. Elyas, and B. Majumdar. *Crystal. Sol. State Commu.* **42**(10): 753–757 (1982).
36. C.N. Xu, T. Watanabe, M. Akiyama, and K. Nonaka. US Patent 09/468, 899 (1999).
37. J.I. Zink. *Crystal. J. Am. Chem. Soc.* **96**(21): 6775–6777 (1974).
38. L.M. Sweeting. *ACS Symp. Series* **415**: 245–260 (1989).
39. T. Matuszawa, Y. Aoki, N. Takeuchi, and Y. Murayama. *J. Electrochem. Soc.* **143**: 2670 (1996).
40. M. Ohta, M. Maruyama, T. Hayakawa, and T. Nishijo. *J. Ceram. Soc. Jpn.* **108**(3): 284–289 (2000).
41. K. Tanaka, K. Fujita, T. Taniguchi, K. Hirao, and T. Ishihara. *J. Appl. Phys.* **88**(7): 4069–4074 (2000).
42. W. Hoogenstraaten. *Philips Res. Rept.* **13**: 515–693 (1958).
43. Y. Hayashiuchi, T. Hagihara, and T. Okada. *Phys. Lett.* **A147**: 245 (1990).
44. V. Abbruscato. *J. Electrochem. Soc.* **118**(6): 930–933 (1971).
45. B.P. Chandra, M. Ramrakhiani, P. Sahu, and A.M. Ratstogi. *Indian Acad. Sci.* **54**(2): 287–303 (2000).
46. C.N. Xu, X.G. Zheng, T. Watanabe, M. Akiyama, and I. Usui. *Thin Solid Film*, **352**: 273–277 (1999).
47. K. Nakayama. *Wear.* **194**: 185–189 (1996).
48. C.N. Xu, M. Akiyama, and T. Watanabe. *IEEE Trans. UFFC.* **45**(4): 1065–1071 (1998).
49. K. Uchino. *Mater. Res. Innov.* **1**(3): 163–168 (1997).
50. O. Kanehisa, M. Shiiki, T. Suzuki, and K. Uenae. *J. Alloys and Compounds.* **223**: 154–159 (1995).
51. T. Mitsui, N. Yamamoto, J. Yoshino, T. Tadokoro, S. Ohta, K. Tamashita, and K. Inoue. *J. Appl. Surf. Sci.* **100/101**: 625–672 (1996).
52. T. Kajiya and H. Mori. *Technology Reports of Kyushu Univ.* **68**(4): 305–311 (1995).
53. J.N. Tsuji, M. Nishida, and K. Kawada. *Photoelastic Experimental Measurements* (in Japanese). Nikkan Kougyo, Tokyo, 1965.
54. K. Kawada, S. Taira, and Y. Tada. *Experimental Stress Analysis Manual*. OHM, Tokyo, 1972.
55. C.N. Xu, M. Akiyama P. Sun, and T. Watanabe. *Appl. Phys. Lett.* **70**(13): 1639–1640 (1997).
56. M. Akiyama, C.N. Xu, K. Nonaka, T. Watanabe, and I. Usui. *J. Mater. Sci. Lett.* **17**(24): 2093–2095 (1998).
57. Y. Liu, C.N. Xu, and T. Watanabe. *J. Mater. Sci.* **34**(17): 4129–4132 (1999).
58. O. Agyman, C.N. Xu, T. Watanabe, Y. Liu, M. Akiyama, M. Taira, X.G. Zheng, and M. Suzuki. *Trans. Mater. Res. Soc. Jpn.* **24**(4): 603–606 (1999).
59. C.N. Xu, Y. Liu, M. Akiyama, K. Nonaka, and X.G. Zheng. *SPIE Proc.* **4220**: 344–349 (2000).
60. T. Miyoshi. *Introduction to the Finite Element Method*. Baifukan, Tokyo, 1994.
61. T. Timoshenko and J.N. Goodier. *Theory of Elasticity*, 3rd ed. McGraw-Hill, New York, 1970.
62. C.N. Xu, X.G. Zheng, M. Akiyama, K. Nonaka, and T. Watanabe. *Appl. Phys. Lett.* **76**(2): 179–181 (2000).

COLOSSAL MAGNETORESISTIVE MATERIALS

A. MAIGNAN

Laboratoire CRISMAT, ISMRA
CAEN Cedex, FRANCE

INTRODUCTION

Magnetoresistance (MR) is defined as the relative change in the electrical resistivity of a material upon the application of a magnetic field and is generally given by $\%MR = 100 \times [(\rho(H) - \rho(0))/\rho(0)]$, where $\rho(H)$ and $\rho(0)$ are the resistivities at a given temperature in and in the absence of a magnetic field, respectively. MR is positive for most nonmagnetic metals, and its magnitude is limited to a few percent, whereas MR can be negative in magnetic materials because the magnetic field tends to reduce the spin disorder. For instance, the $\%MR$ of Co and Fe is $\sim -15\%$.

MR is of considerable technological interest. IBM is using the Permalloy (composition: 80% Ni and 20% Fe) MR of about 3% in a small magnetic field at room temperature for the magnetic storage of information. More recently, larger magnetoresistance also called giant MR (GMR) was observed in thin films of magnetic superlattices (for instance, Fe, Cr) for which metallic layers of a ferromagnet and a nonmagnetic material (or an antiferromagnet) are alternately deposited on a substrate (1,2). By doing so, the MR magnitude is increased by an order of magnitude. Small ferromagnetic particles deposited on a paramagnetic thin film also provide an alternative way to obtain GMR devices (3). For both material classes, small magnetic field applications (a few oersteds) are sufficient to align the magnetizations ferromagnetically and thus to induce a resistivity decrease originating in decreased scattering.

In hole-doped perovskite manganites $\text{Ln}_{1-x}\text{AE}_x\text{MnO}_3$ where $x \sim 0.3$, magnetoresistance values of $\sim -100\%$ in large magnetic fields (several teslas) have been discovered. These effects are called CMR to distinguish them from GMR (4–11). CMR has motivated a large number of experimental studies of these oxides in bulk (ceramics and crystals) and in thin films and also of theoretical work to understand the origin of the phenomenon. In the 1950s, the double-exchange model (DE) was proposed to explain the simultaneous appearance of ferromagnetism and metallicity when $\text{Mn}^{3+}/\text{Mn}^{4+}$ valency is created in $\text{La}_{1-x}\text{Sr}_x\text{MnO}_3$ (12–14). However, after the CMR discovery, several theoretical studies have shown that double exchange alone cannot explain the magnitude of the resistivity drop upon the application of a magnetic field (15). The distorted Jahn–Teller Mn^{3+}O_6 octahedron introduces an interaction between the charge carriers and the crystal lattice so that the bound-state charge and a lattice called a “polaron” has been proposed and experimentally evidenced (15–19). Consequently, the Jahn–Teller distortion, static or dynamic, must be incorporated in any model, built to describe CMR. This time-dependent increasing complexity has been more recently confirmed by the relevancy of the phase-separation scenario for manganese oxides (20–25). Roughly, recent computational studies in which extended coulombic interactions are included have revealed

that, as the $\text{Mn}^{3+}/\text{Mn}^{4+}$ mixed valency is created by varying x in $\text{Ln}_{1-x}\text{AE}_x\text{MnO}_3$, the transition from the antiferromagnetic insulating state for $x = 0$ toward ferromagnetism for hole-doped compositions, where $x \gg 0$, does not occur via intermediate phases (canted phases) but rather through a mixed-phase process (20). An inhomogeneous electronic state should be stabilized, and several experimental studies have confirmed this phase-separation scenario (21–25). Consequently, the large dropping resistivity at the origin of the qualifying “colossal” magnetoresistance is interpreted in a percolation framework: a magnetic field increases the ferromagnetic metallic regions at the expense of the insulating antiferromagnetic areas, so that the macroscopic insulating state becomes metallic beyond the percolation threshold (26).

In this article, several representative examples of perovskite manganites are given to illustrate the richness of their phase diagrams. More particularly, the chemical key factor governing the CMR of hole-doped manganites that contain 30% Mn^{4+} and have the $\text{Ln}_{0.7}\text{AE}_{0.3}\text{MnO}_3$ formula are reviewed. The existence of $\text{Mn}^{3+}/\text{Mn}^{4+}$ charge ordering in the Mn lattice for half-doped manganites ($\text{Mn}^{3+}:\text{Mn}^{4+} = 50:50$, i. e., $x = 0.5$) and also for Mn^{4+} -rich compositions (electron-doped, $x > 0.5$) are discussed. Finally, the possibility of obtaining CMR properties in Mn^{4+} -rich manganites is shown.

CMR IN HOLE-DOPED $\text{Ln}_{0.7}\text{AE}_{0.3}\text{MnO}_3$ PEROVSKITES

Among the first compositions that were investigated, those where $x = 0.3$ in $\text{Ln}_{1-x}\text{AE}_x\text{MnO}_3$, have the best CMR properties (4–10). This is the case for $\text{Ln} = \text{Pr}^{3+}$ and $\text{AE} = \text{Ca}^{2+}/\text{Sr}^{2+}$ that have the formula $\text{Pr}_{0.7}\text{Ca}_{0.3-x}\text{Sr}_x\text{MnO}_3$ (27,28). Some of the latter compositions show resistivity (ρ) drops in a magnetic field ($\mu_0 H = 5\text{T}$) when the ratio $\rho(0)/\rho(5\text{T})$ is from 10^4 to 10^{11} , as shown in Fig. 1 for $\text{Pr}_{0.7}\text{Ca}_{0.25}\text{Sr}_{0.05}\text{MnO}_3$ ($x = 0.05$) and $\text{Pr}_{0.7}\text{Ca}_{0.26}\text{Sr}_{0.04}\text{MnO}_3$ ($x = 0.04$), respectively. By cooling the $x = 0.05$ sample from 300 K to 5 K in the absence of a field, the activated character of ρ observed till 90 K evolves to a metallic character below that temperature; ρ decreases by about four orders of magnitude at 20 K (Fig. 1a). Then, by registering the ρ data using the same process but in a 5-T magnetic field applied at 300 K before cooling, one can clearly see in Fig. 1a the dramatic ρ decrease induced by the field in the temperature vicinity of the insulator–metal (I–M) transition. Five orders of magnitude are obtained from the isothermal $\rho(0)/\rho(5\text{T})$ at 88 K and consequently, the magnetoresistance $\%MR = 100 \times [(\rho(H) - \rho(0))]$, reaches -100% , demonstrating the “colossal” character of this negative magnetoresistance. Furthermore, a composition shift of only 0.01 Ca for Sr (from $x = 0.05$ to $x = 0.04$) prevents the I–M transition (Fig. 1b), and high resistivities are measured at the lowest measurable temperature of ~ 30 K. Again, the field application seems to quench a quasi-T independent metallic state and the $\rho(0)/\rho(5\text{T})$ ratio reaches $\sim 10^{12}$. This behavior has also been confirmed by measuring isothermal field dependent $\rho(H)$ curves ($T = 50$ K, Fig. 2). At 50 K, the curve shows that a ρ drop of 10^7 is reached beyond the critical field of 0.6T.

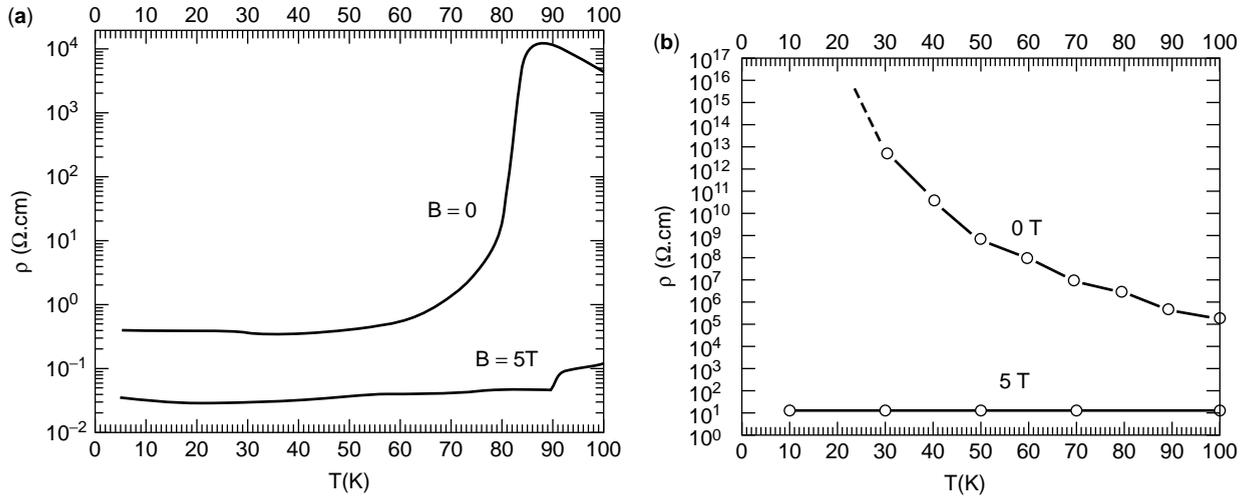


Figure 1. T dependence of the resistivity ρ upon cooling in (5 T) and in the absence (0 T) of a magnetic field for (a) $\text{Pr}_{0.7}\text{Ca}_{0.25}\text{Sr}_{0.05}\text{MnO}_3$ and (b) $\text{Pr}_{0.7}\text{Ca}_{0.26}\text{Sr}_{0.04}\text{MnO}_3$.

These CMR properties are connected with the magnetic properties that show the great interplay between the carriers and the spins. Clear transitions from paramagnetic (PM) to ferromagnetic (FM) are observed from the T -dependent magnetization (M) curves of the $x = 0.05$ and $x = 0.04$ compositions (Fig. 3). The corresponding Curie temperatures (T_C) taken at the inflection point of the transition coincide with the I–M transition temperatures. Thus, for $\text{Pr}_{0.7}\text{Ca}_{0.25}\text{Sr}_{0.05}\text{MnO}_3$, the metallicity appears as the sample becomes ferromagnetic. For the other composition, $\text{Pr}_{0.7}\text{Ca}_{0.26}\text{Sr}_{0.04}\text{MnO}_3$, the $\rho(T)$ curve (Fig. 1b) does not show an I–M transition in the absence of a magnetic field, although this ceramic exhibits a ferromagnetic state (Fig. 3). However, the $M(T)$ curve has been obtained by measuring in an applied field of 1.45 T and a $\rho(T)$ curve registered in the same field shows the I–M transition.

This first set of data allows two important conclusions: high magnetic values are required to obtain CMR effects,

and small chemical changes are responsible for dramatic modifications in physical properties.

ORIGIN OF THE CMR EFFECT: MANGANESE MIXED VALENCY AND DOUBLE EXCHANGE

Manganese oxides $\text{Ln}_{1-x}\text{AE}_x\text{MnO}_3$ crystallize in a perovskite structure (Fig. 4), but their structures differ from that of the ideal cubic perovskite ABO_3 (29,30). The structure can be described as a tridimensional network of MnO_6 octahedra linked by their apexes, so that cages are formed and are filled by the Ln^{3+} and AE^{2+} cations (A site of the perovskite). The distortion of the structure in manganites is a consequence of the small size of the A-site cations

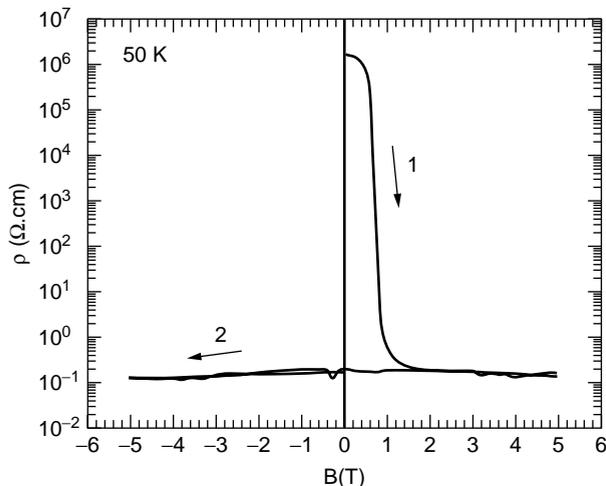


Figure 2. Field-dependent ρ curve for $\text{Pr}_{0.7}\text{Ca}_{0.26}\text{Sr}_{0.04}\text{MnO}_3$.

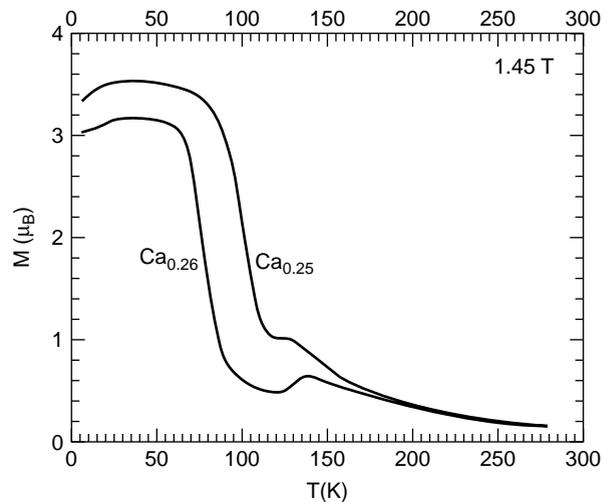


Figure 3. T -dependent magnetization curves found upon warming in 1.45 T after a zero-field-cooling process (ZFC) for $\text{Pr}_{0.7}\text{Ca}_{0.26}\text{Sr}_{0.04}\text{MnO}_3$ ($\text{Ca}_{0.26}$) and $\text{Pr}_{0.7}\text{Ca}_{0.25}\text{Sr}_{0.05}\text{MnO}_3$ ($\text{Ca}_{0.25}$).

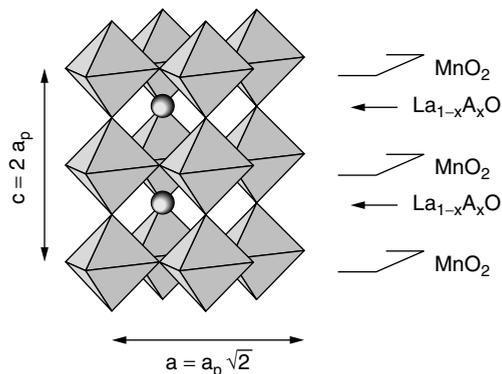


Figure 4. Idealized structure of a $\text{Ln}_{1-x}\text{A}_x\text{MnO}_3$ distorted perovskite.

which gives rise to the tilting of the MnO_6 octahedra. This distortion is quantified by the Goldsmith tolerance factor $t = d_{\text{A-O}}/[\sqrt{2}(d_{\text{Mn-O}})]$, where $d_{\text{A-O}}$ and $d_{\text{Mn-O}}$ are the A cation–oxygen and Mn–oxygen bond lengths, respectively. Usually, for manganites t is ~ 1 or $t < 1$ and, consequently, because the tilting mode depends on t , several kinds of crystallographic space groups can be evidenced as the A-site average cationic size $\langle r_{\text{A}} \rangle$ changes or as the manganese valency (which controls the Mn–O distance) varies.

To understand ferromagnetic metallic properties (31), the electronic configurations of the Mn^{3+} ($3d^4$) and Mn^{4+} ($3d^3$) magnetic species must be considered. Full rotational invariance is broken in the octahedral environment, and this creates the splitting of 3d orbitals in two e_g and three t_{2g} . Due to the strong Hund coupling (J_{H}) for this system, the spins are aligned in the 3d shell (high-spin configuration): three localized electrons populate the t_{2g} orbitals (t_{2g}^3), whereas one electron (e_{g1}) or no electron (e_{g0}) populates the e_g orbital for Mn^{3+} and Mn^{4+} , respectively. Moreover, the e_g filling for Mn^{3+} creates a Jahn–Teller distortion that degenerates the e_g orbitals in two levels, d_{z^2} and $d_{x^2-y^2}$; only the former is occupied (Fig. 5). The e_g electrons of Mn^{3+} are mobile and they use the bridging orbitals of the oxygens to reach an empty e_g orbital of a Mn^{4+} nearest neighbor. This leads to the double-exchange model proposed by Zener (12): the e_g electron delocalization between nearest neighbor manganese ions that have t_{2g} parallel spins (Fig. 6) allows paying the energy J_{H} and gains some kinetic energy for the mobile carriers by minimizing

Figure 5. Electronic configuration of Mn^{4+} ($3d^3$) and Mn^{3+} ($3d^4$) cations.

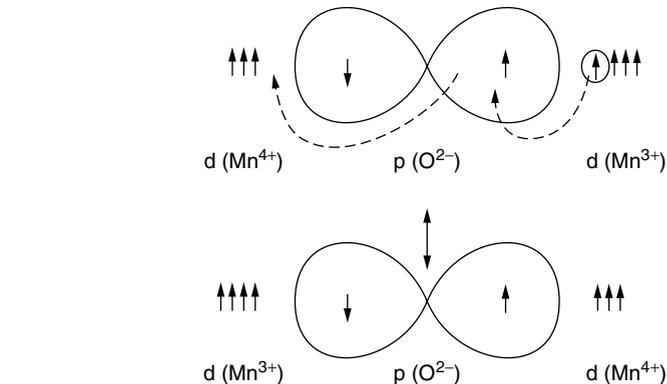


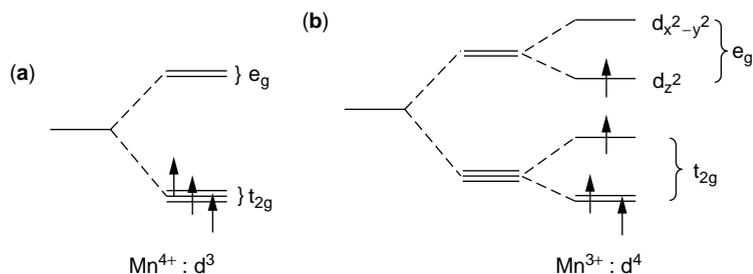
Figure 6. Double-exchange mechanism according to Zener.

the hole–spin scattering. Consequently, the FM droplets around the holes start to overlap as holes (Mn^{4+}) are injected in the Mn^{3+} matrix, and a fully FM metallic state can be reached.

From this model, one can understand that the CMR effect in the T_{C} vicinity results from the field-induced ferromagnetic alignment, which creates delocalization and thus the resistivity decrease. But, several experimental results exist, for instance, coexistence of FM and charge ordering (21–26) that have suggested that more complex ideas are needed to explain CMR properties. At present, the phase-separation scenario (20) seems to be relevant for manganese oxides. Several examples that support this model are described in the following.

CHEMICAL FACTORS GOVERNING CMR PROPERTIES

Two important factors have to be considered to control the magnetism in these systems: the hole concentration and the overlap of the Mn and O orbitals (11,32,33). The first corresponds to the content of Mn^{4+} in the Mn^{3+} matrix and can be tuned by varying the A-site cationic $\text{Ln}^{3+}/\text{AE}^{2+}$ ratio. The best concentration for obtaining the highest T_{C} corresponds to about 30–40% Mn^{4+} ; far below this content, the FM regions do not percolate (FM insulating “FMI” state), and beyond, other complications arise from the closeness to the “half-doped” $\text{Ln}_{0.5}\text{AE}_{0.5}\text{MnO}_3$ compositions that are highly favorable for antiferromagnetism (AFM). This is clearly seen in Fig. 7 where the $\text{Pr}_{1-x}\text{Sr}_x\text{MnO}_3$ phase diagram (34) is given; if one concentrates on the hole region ($x < 0.5$), a clear T_{C} optimum of ~ 280 K is



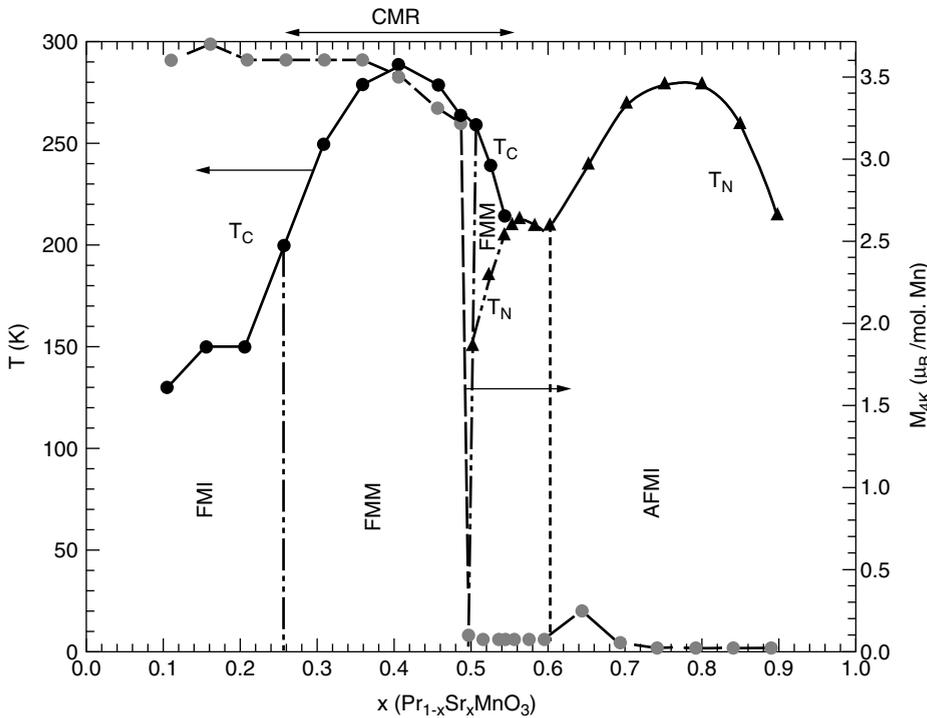


Figure 7. Magnetic and electronic phase diagram of $\text{Pr}_{1-x}\text{Sr}_x\text{MnO}_3$ that shows the great complexity of these systems as the Mn valency varies. The magnetic transition temperatures Néel (T_N) and Curie (T_C) are symbolized by black triangles and circles, respectively. The gray curve (gray circles) corresponds to the magnetization values at 4.2 K in 1.45 T (ZFC). The highest T_C of 280 K is reached for $\text{Pr}_{0.6}\text{Sr}_{0.4}\text{MnO}_3$ ($x = 0.4$).

observed for $x \sim 0.4$. For the same Mn valency, the T_C maximum of $\text{La}_{1-x}\text{Sr}_x\text{MnO}_3$ reaches 370 K (35), and the T_C of $\text{La}_{1-x}\text{Ca}_x\text{MnO}_3$ is 280 K (36).

The overlap of the 3d orbitals of the Mn species and of the oxygen p orbitals are controlled by varying the Mn–O–Mn angle, which can be done by changing $\langle r_A \rangle$. The effect of $\langle r_A \rangle$ on CMR properties was shown simultaneously by several groups (11,32,33). If we return to the $\text{Pr}_{0.7}\text{Ca}_{0.3-x}\text{Sr}_x\text{MnO}_3$ series and more especially to the $\rho(T)$ and $M(T)$ curves where x varies by 0.01 increments from 0.04 to 0.10 (Fig. 8), the following remarks can be made: (1) the resistivity drop at the I–M transition $\rho_{T_{I-M}}/\rho_{5\text{K}}$ increases as the strontium content decreases, from 170 for

$x = 0.10$ up to 3×10^5 for $x = 0.05$ (Fig. 8a); (2) both T_{I-M} and the Curie temperature T_C (Fig. 8b) increase as x increases. These are very important results because they demonstrate that the physical properties are highly sensitive to $\langle r_A \rangle$. The ionic radius of Sr^{2+} is larger than that of Ca^{2+} [1.31 Å versus 1.18 Å (37)], and thus as x increases, $\langle r_A \rangle$ increases; the Mn–O–Mn angle increases as x increases so that the bandwidth (W) increases. Consequently, T_C increases as $\langle r_A \rangle$ increases. The largest T_C of ~ 370 K is thus observed for the larger $\langle r_A \rangle$ such as for $\text{La}_{0.7}\text{Sr}_{0.3}\text{MnO}_3$ (35).

Finally, a third important parameter exists that governs the T_C of these perovskites: the local disorder tends to weaken the DE process. Particularly, the same $\langle r_A \rangle$

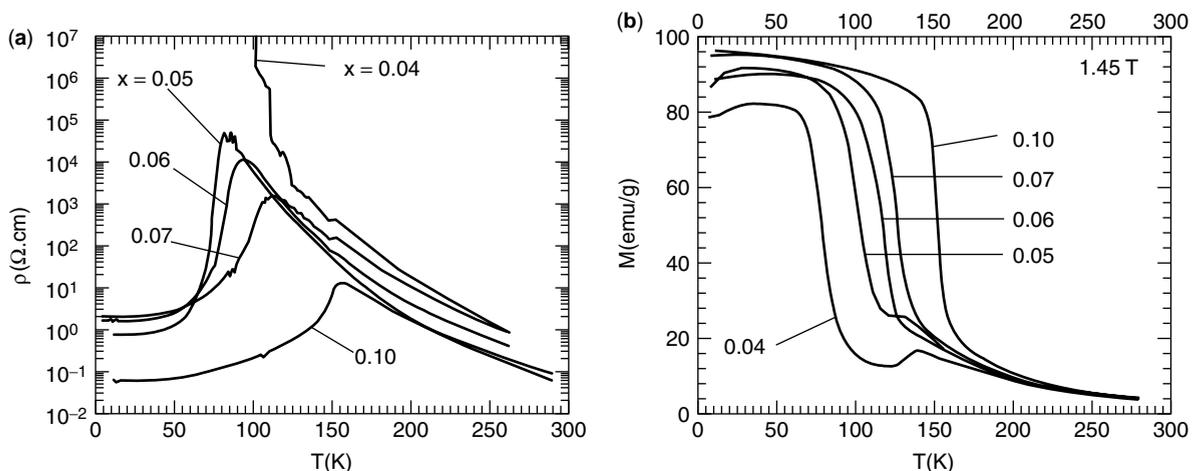


Figure 8. (a) $\rho(T)$ and (b) $M(T)$ curves of $\text{Pr}_{0.7}\text{Ca}_{0.3-x}\text{Sr}_x\text{MnO}_3$. x values are labeled on the graphs.

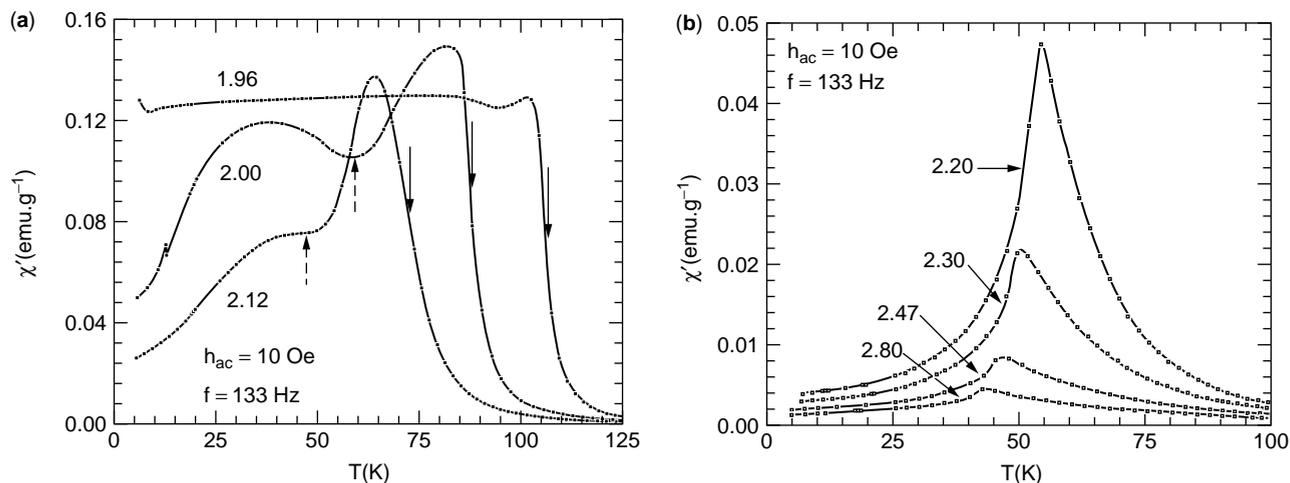


Figure 9. T dependence of the real part of the AC susceptibility (χ') for several $\text{Th}_{0.35}\text{Ae}_{0.65}\text{MnO}_3$ samples characterized by a fixed $\langle r_A \rangle = 1.255 \text{ \AA}$ value but varying A-site mismatch (σ^2). The σ^2 values $\times 10^4$ in nm^2 are indicated on the graph. (a) $\sigma^2 = 1.96, 2.00,$ and 2.12 ; (b) $\sigma^2 = 2.20, 2.30, 2.47,$ and 2.80 .

value can be obtained by using different sets of cations: as shown by Rodriguez-Martinez and Attfield (38), both $\text{La}_{0.7}\text{Ca}_{0.11}\text{Sr}_{0.19}\text{MnO}_3$ and $\text{Sm}_{0.7}\text{Ba}_{0.3}\text{MnO}_3$ are characterized by the same $\langle r_A \rangle = 1.23 \text{ \AA}$ value, but their T_C s differ strongly, 360 K and 60 K for the former and the latter, respectively. This difference was ascribed to the size mismatch of the A site represented by the variance σ^2 , defined by $\sigma^2 = \sum y_i r_i^2 - \langle r_i \rangle^2$, where y_i and r_i are the fractional occupancies and the ionic radii of the i cations. As σ^2 increases, local distortions are generated that reduce T_C . This can be inferred from the results obtained for the highly mismatched $\text{Th}_{0.35}^{4+}\text{AE}_{0.65}^{2+}\text{MnO}_3$ samples (AE = Ba, Sr, Ca) which are characterized by $\langle r_A \rangle = 1.255 \text{ \AA}$, that is, large enough to attain a ferromagnetic state at a sufficiently small σ^2 and for which the mismatch can be varied across a wide range (39). Starting from a FMM sample

where $\sigma^2 = 1.96 \cdot 10^{-4} \text{ nm}^2$, the ferromagnetism can be reduced by increasing the A-site cationic size mismatch, as shown in Fig. 9 from the T -dependent AC-susceptibility [$\chi'(T)$] curves (Fig. 9a,b) and corresponding $\rho(T)$ curves (Fig. 10). Furthermore, for the highest mismatch values, the $\chi'(T)$ curves exhibit a cusp shape characteristic of spin-glass (Fig. 9b), and the $\rho(T)$ curves show insulating behavior (Fig. 10). Clearly, the A-site disorder is an important parameter that strongly affects the FMM state of perovskite manganites and can be controlled to induce a change from FMM samples to spin-glass insulators (SGI), as shown in the electronic and magnetic diagram proposed in Fig. 11.

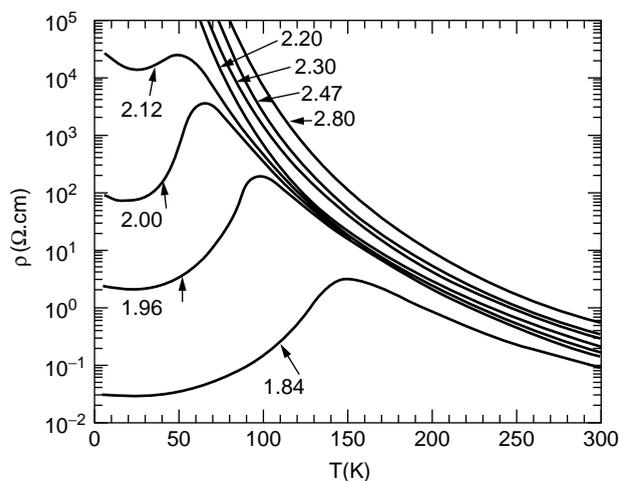


Figure 10. $\rho(T)$ curves of $\text{Th}_{0.35}\text{Ae}_{0.65}\text{MnO}_3$.

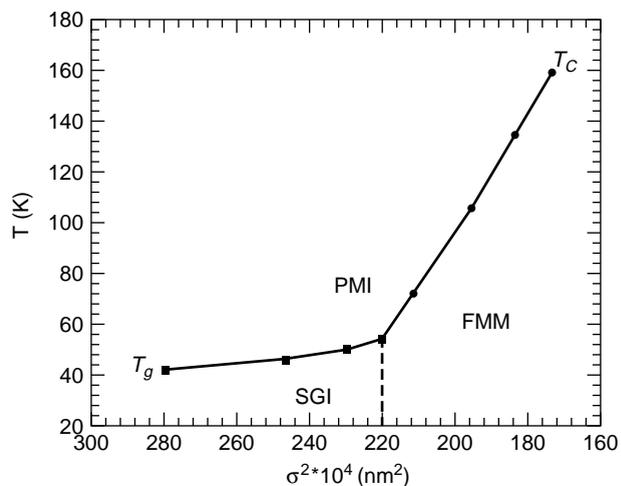


Figure 11. Electronic and magnetic diagram established for the $\text{Th}_{0.35}\text{Ae}_{0.65}\text{MnO}_3$ series. Circles and squares are for the T_C (Curie) and T_g (glass) characteristic temperatures as a function of the mismatch (σ^2). The dotted line is the boundary between the FMM and SGI regions.

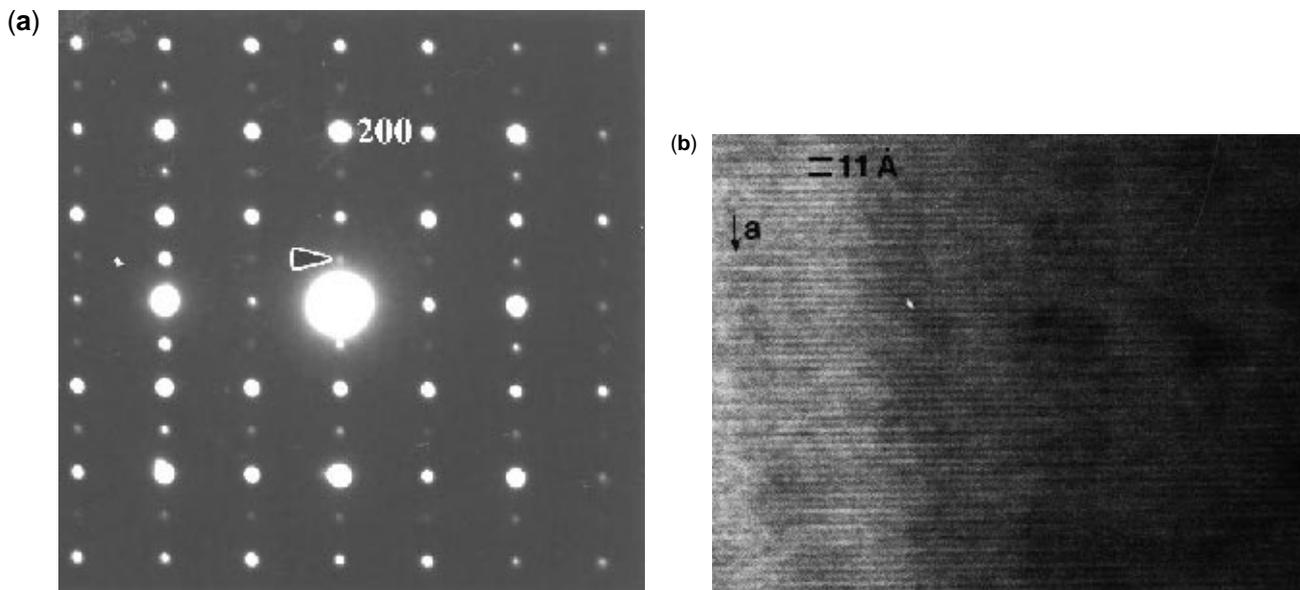


Figure 12. (a) 92-K electron diffraction pattern obtained by a transmission electron microscope for $\text{Sm}_{0.5}\text{Ca}_{0.5}\text{MnO}_3$. The doubling of the a cell parameter observed at 92 K (extra peak indicated by the arrow) is induced by $\text{Mn}^{3+}/\text{Mn}^{4+}$ orbital ordering. (b) Corresponding lattice image that shows the alternation of Mn^{3+} and Mn^{4+} stripes.

CHARGE ORDERING IN PEROVSKITE MANGANITES

Charge carriers doped into antiferromagnetic insulators as in La_2NiO_4 and La_2CuO_4 tend to be arranged in stripes for favorable doping levels. An electronic phase separation is created by stripes of holes interspaced by antiferromagnetic electron-rich regions (40). A different kind of electronic phase separation also occurs in half-doped $\text{Ln}_{0.5}\text{AE}_{0.5}\text{MnO}_3$ manganites, where the e_g carriers $e_{g^0} - e_{g^1}$ are delocalized in the paramagnetic state but localized in alternating Mn^{4+} and Mn^{3+} planes below the characteristic charge-ordering temperature T_{CO} . This model was first proposed by J. B. Goodenough (41) after the pioneering work by Wollan and Koehler on $\text{La}_{0.5}\text{Ca}_{0.5}\text{MnO}_3$ (42). This charge-ordering phenomenon has been proved more recently by the electron diffraction patterns and lattice images collected below T_{CO} by transmission electron microscopy of several $\text{Ln}_{0.5}\text{Ca}_{0.5}\text{MnO}_3$ manganites (43,44). One typical pattern and corresponding lattice image are given in Fig. 12. On the one hand, additional peaks (indicated by an arrow), corresponding to the doubling of the a parameter (where $a \sim 5.5$ Å in the primitive cell) as a $\text{Sm}_{0.5}\text{Ca}_{0.5}\text{MnO}_3$ sample is cooled down below T_{CO} , are observed in the diffraction pattern (Fig. 12a). On the other hand, the alternating bright and dark stripes of Mn^{3+} and Mn^{4+} are visible in the image that lead to an interfringe distance $2a \sim 11$ Å (Fig. 12b). Besides the FMM state driven by the double-exchange mechanism, thus a second phenomenon exists that is driven by long-range coulombic repulsion which tends to separate the Mn^{3+} and Mn^{4+} species. The Jahn–Teller distortion of Mn^{3+} plays a crucial role in the CO process, because the d_{z^2} orbitals of Mn^{3+} are

arranged in 90° zigzag chains in the CO phase (as shown in the drawn projection of Fig. 13): thus CO (T_{CO}) and orbital ordering (T_{OO}) occur simultaneously (41). In Fig. 13, the Mn^{3+} and Mn^{4+} stripe of charges are planes running along the (b, c) planes that alternate in the a direction. The size difference between Mn^{3+}O_6 and Mn^{4+}O_6 octahedra in this checkered pattern is responsible for the doubling of the a parameter, as this CO manganite is cooled below the T_{CO} .

Here again, $\langle r_A \rangle$ is a crucial parameter that governs T_{CO} (and/or T_{OO}): as $\langle r_A \rangle$ decreases, T_{CO} increases, as shown in Fig. 14 by the $\text{Ln}_{0.5}\text{Ca}_{0.5}\text{MnO}_3$ compositions (44). In other words, as the Mn–O–Mn angle decreases, the charge ordering is favored at the expense of the FMM state. For all of the CO $\text{Ln}_{0.5}\text{Ca}_{0.5}\text{MnO}_3$, the spin order in a CE-type AFM

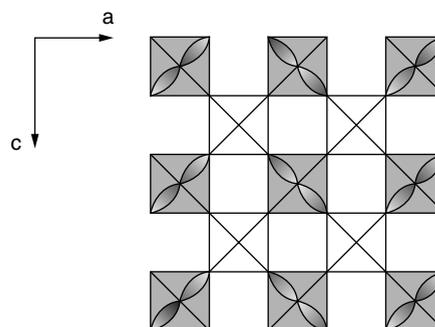


Figure 13. 2-D drawing obtained by projecting the 3-D charge-ordered structure of a $\text{Ln}_{0.5}\text{Ca}_{0.5}\text{MnO}_3$ perovskite. The 90° zigzag chains of the Mn^{3+} d_{z^2} orbitals are clearly visible. The bright octahedra correspond to Mn^{4+}O_6 . The distortion of the Mn^{3+}O_6 octahedron is not shown.

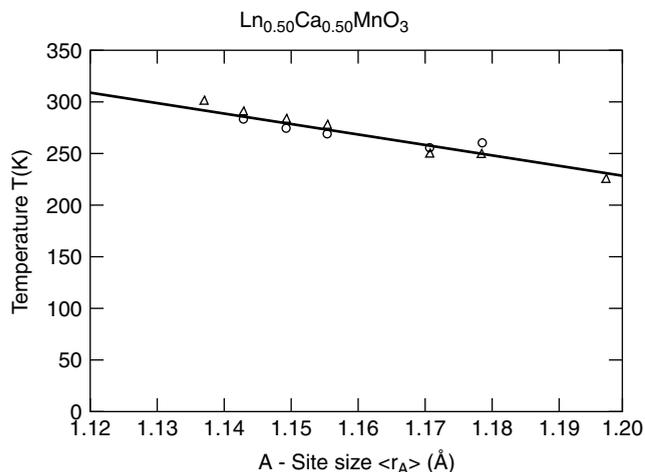


Figure 14. Charge-ordering (CO) temperatures (T_{CO}) as a function of (r_A) determined from the study of several $\text{Ln}_{0.5}\text{Ca}_{0.5}\text{MnO}_3$ charge-ordered manganites by electron microscopy or magnetization.

structure (Fig. 15) below the Néel temperature T_N (42), and the T_N value is always such that $T_N \leq T_{\text{CO}}$. To sum up, the CO process in $\text{Ln}_{0.5}\text{Ca}_{0.5}\text{MnO}_3$ half-doped manganites induces a structural transition [see the doubling of one cell parameter in the CO structure (Fig. 12 and 13)] and an AFM arrangement of the spins, confirming the important interplay between the lattice, the charges, and the spins. The strong electron–phonon coupling in the CO phase has been confirmed in the CO $\text{La}_{0.5}\text{Ca}_{0.5}\text{MnO}_3$ half-doped manganite by an oxygen isotopic effect (18).

At first glance, the robustness of the low temperature CO-AFMI state to a high magnetic field does not seem very attractive for the CMR effect because, for instance, 30T are necessary to melt the CO state of $\text{Pr}_{0.5}\text{Ca}_{0.5}\text{MnO}_3$ (45). However, several possibilities exist for weakening the CO for the benefit of the FMM state.

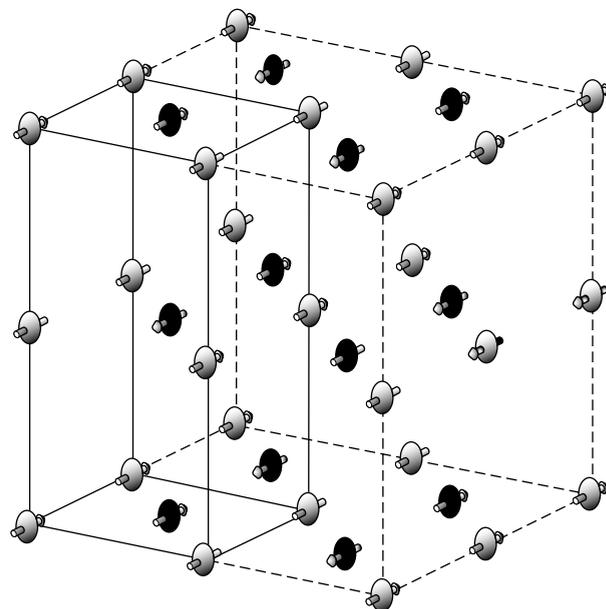


Figure 15. CE-type AFM structure of $\text{Ln}_{0.5}\text{Ca}_{0.5}\text{MnO}_3$ charge-ordered manganites. The bright and black circles are for the Mn^{3+} and Mn^{4+} cations, respectively. The small arrows are for the magnetic moments. The solid line and dotted line are for the nuclear and magnetic cells, respectively.

The first is controlling (r_A); for sufficiently large (r_A) values, as for $\text{Nd}_{0.5}\text{Sr}_{0.5}\text{MnO}_3$ (46), the FMM state exists above T_{CO} , and consequently, these half-doped manganites are such that $T_C > T_{\text{CO}}$. One example of such a realization is the $\text{Pr}_{0.5}\text{Sr}_{0.5-x}\text{Ca}_x\text{MnO}_3$ series (47): starting from the end member $\text{Pr}_{0.5}\text{Ca}_{0.5}\text{MnO}_3$, which is a CO-AFM compound without CMR properties (48), a FMM state can be obtained by increasing (r_A) by substituting Sr for Ca, as shown (Fig. 16a,b) by the $M(T)$ and $\rho(T)$ curves of $\text{Pr}_{0.5}\text{Sr}_{0.5-x}\text{Ca}_x\text{MnO}_3$. The closeness of both CO-AFMI and FMM states creates the metastability of these phases. The

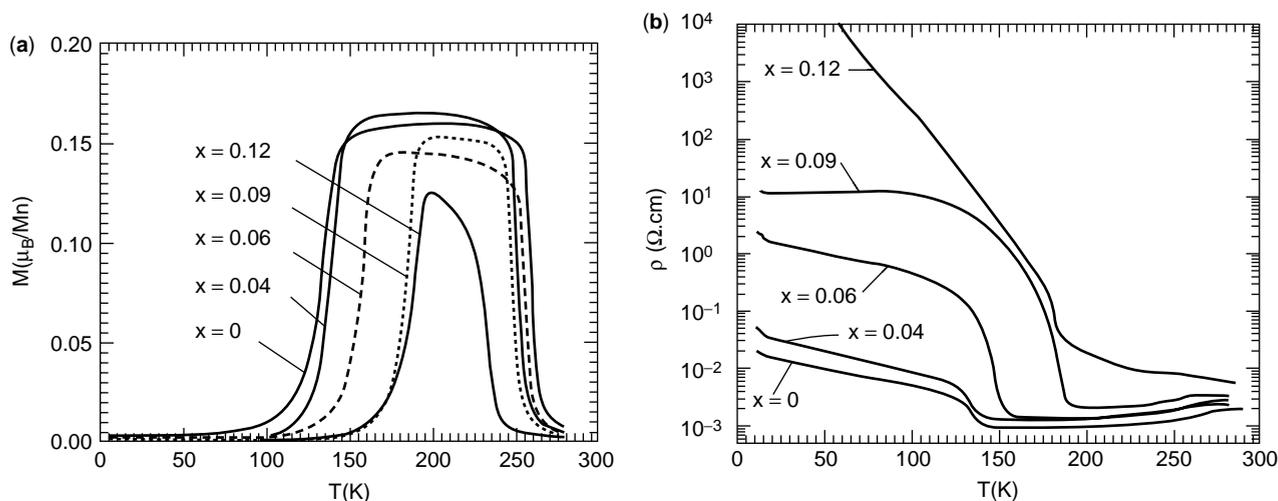


Figure 16. $\text{Pr}_{0.5}\text{Sr}_{0.5-x}\text{Ca}_x\text{MnO}_3$ samples. (a) $M(T)$ curves at $\mu_0 H = 10^{-2}$ T and (b) corresponding $\rho(T)$ curves.

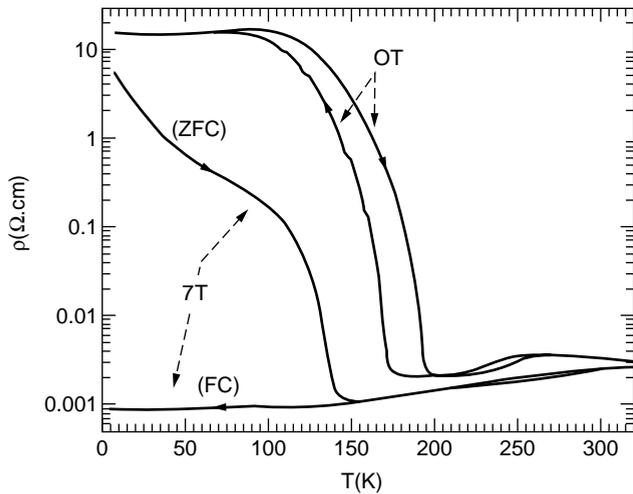


Figure 17. $\rho(T)$ curve of the CO compound $\text{Pr}_{0.5}\text{Sr}_{0.41}\text{Ca}_{0.09}\text{MnO}_3$ ($x = 0.09$) at 0 and 7 T. ZFC and FC are for zero-field-cooling and field-cooling, respectively.

application of a 7-T magnetic field on $\text{Pr}_{0.5}\text{Sr}_{0.41}\text{Ca}_{0.09}\text{MnO}_3$ (47) is sufficient to destabilize the CO state and thus to restore a FMM state as shown in Fig. 17. A resistivity ratio $\rho(0)/\rho(7\text{ T})$ of 10^4 can be reached. Thus, the CO phase can be a “precursor” to CMR effects. The CO CE-type AFM phase can be transformed into the FMM by applying a magnetic field. Direct evidence from a neutron diffraction study as a function of magnetic field has been given for $\text{Nd}_{0.5}\text{Sr}_{0.5}\text{MnO}_3$ (49): at 125 K and under 6 T, the monoclinic CO CE phase collapses into the FMM orthorhombic phase. This field-induced structural transition is facilitated by the coexistence of electronic and magnetic phase segregation at nearly comparable free energies.

A second route to obtaining CMR effects starting from CO compounds is to act on the B site of the perovskite. This impurity effect, it has been shown, is the most efficient when substituting metals such as Cr, Co, and Ni (50). As one can judge from the $\rho(T)$ curve that shows a I–M transition and the $M(T)$ curve that shows ferromagnetic behavior (Fig. 18a and b), 2% of Cr per Mn site in $\text{Pr}_{0.5}\text{Ca}_{0.5}\text{MnO}_3$

is sufficient to induce a FMM. Accordingly, a CMR effect is obtained in the T_{I-M} vicinity. Induced long-range ferromagnetism for $\text{Pr}_{0.5}\text{Ca}_{0.5}\text{Mn}_{0.95}\text{Cr}_{0.05}\text{O}_3$ has been probed by neutron powder diffraction (51). However, the low-temperature electron microscopy study revealed a more complex situation; small charge-ordered monoclinic (AFM) regions of few tens of a nanometer that still remain in the orthorhombic matrix are responsible for the FM observed by neutron diffraction. This coexistence makes all interpretations of the physical properties very complex (26,52). Very recently, some authors proposed that the metallic state observed in the Cr-doped $\text{Nd}_{0.5}\text{Ca}_{0.5}\text{MnO}_3$ phase originates in the percolation of FMM clusters (26). Finally, it should be emphasized that nonmagnetic doping cations—divalent, trivalent, and tetravalent—such as Mg^{2+} , Al^{3+} , Ga^{3+} , Ti^{4+} , Sn^{4+} substituted for Mn, destabilize the CO but do not induce the I–M transition observed for magnetic cations such as Cr^{3+} (50).

It is of prime importance for understanding CMR to realize that the CO tendency is not restricted to the half-doped $\text{Ln}_{0.5}\text{AE}_{0.5}\text{MnO}_3$ compositions, but that it can be observed far from the $\text{Mn}^{3+}:\text{Mn}^{4+} = 50:50$ ratio. Several types of $\text{Mn}^{3+}/\text{Mn}^{4+}$ arrangements below the T_{CO} have been observed by transmission electron microscopy (53–55). Some examples, electron diffraction pattern and corresponding lattice image, are shown in Fig. 19 for $\text{Sm}_{1/4}\text{Ca}_{3/4}\text{MnO}_3$ and $\text{Sm}_{1/3}\text{Ca}_{2/3}\text{MnO}_3$ together with $\text{Sm}_{1/2}\text{Ca}_{1/2}\text{MnO}_3$ [from (55)]. From these contrasts, schematic drawings for the Mn^{3+} and Mn^{4+} planes can be proposed (Fig. 20). They underline the lack of cation intermixing; the extra Mn^{4+} compared to $\text{Sm}_{0.5}\text{Ca}_{0.5}\text{MnO}_3$ forms new Mn^{4+} planes between the Mn^{3+} planes (blocks of 1, 2, and 3 Mn^{4+} planes for Mn valencies of 3.5, 3.66, and 3.75, respectively). For these commensurate $\text{Mn}^{3+}:\text{Mn}^{4+}$ ratios, the doubling of the a cell parameter observed for $\text{Mn}^{3+}:\text{Mn}^{4+} = 50:50$ evolves toward a tripling and a quadrupling for $\text{Sm}_{1/3}\text{Ca}_{2/3}\text{MnO}_3$ and for $\text{Sm}_{1/4}\text{Ca}_{3/4}\text{MnO}_3$, respectively. Obviously, as the charges order, the samples become insulators and antiferromagnetic. The antiferromagnetic structure, CE-type for $\text{Ln}_{0.5}\text{Ca}_{0.5}\text{MnO}_3$ (Fig. 15), changes into a C-type structure (Fig. 21) for $\text{Ln}_{1-x}\text{Ca}_x\text{MnO}_3$ where $0.5 < x$ (CE and C are for AFM structures described

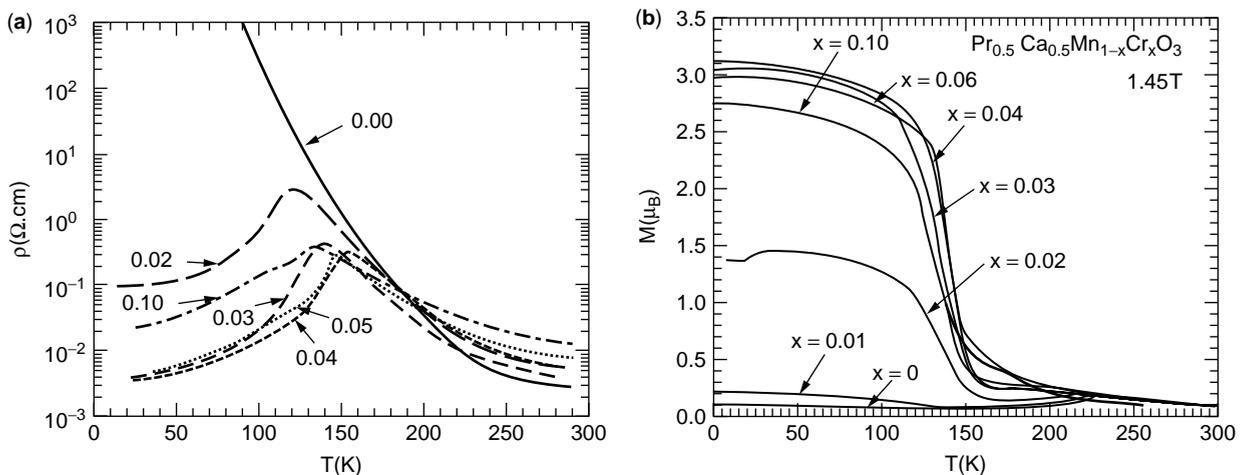


Figure 18. (a) $\rho(T)$ and (b) $M(T)_{1.45\text{T}}$ curves of the $\text{Pr}_{0.5}\text{Ca}_{0.5}\text{Mn}_{1-x}\text{Cr}_x\text{O}_3$ series.

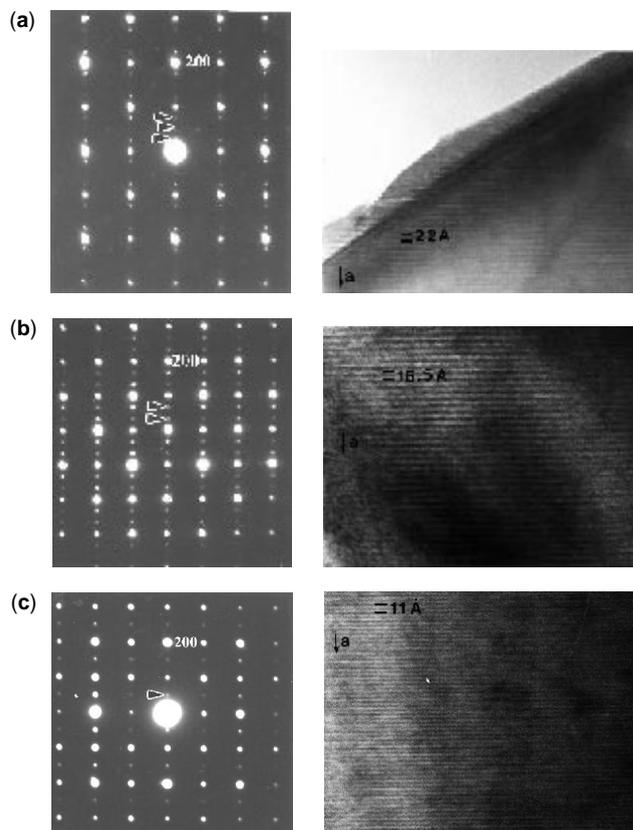


Figure 19. 92-K electron diffraction pattern (left) and corresponding lattice images (right) of $\text{Sm}_{1-x}\text{Ca}_x\text{MnO}_3$ charge-ordered manganites.

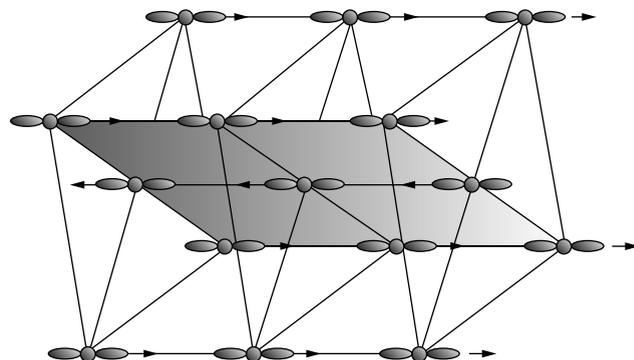


Figure 21. C-type AFM structure. The d_{z^2} orbitals are polarized along the FM chains.

by octants as explained in Ref. 42). At low temperature, incommensurate values of the $\text{Mn}^{3+}:\text{Mn}^{4+}$ ratio lead to extra reflections in the system of intense reflections, observed at room temperature for the $Pnma$ structure, in incommensurate positions on the electron diffraction patterns. In that case, the lattice images show the absence of long-range charge ordering: the alternating $\text{Mn}^{3+}/\text{Mn}^{4+}$ planes cannot order regularly manner in all of the microcrystal. For each lanthanide (Ln), the properties of $\text{Ln}_{1-x}\text{Ca}_x\text{MnO}_3$ manganites, characterized by small $\langle r_A \rangle$ values, can be summarized by a magnetic phase diagram where all of the characteristic transitions are indicated. An example is given in Fig. 22 for the $\text{Sm}_{1-x}\text{Ca}_x\text{MnO}_3$ series (34).

One can see in this graph that a broad composition range exists where charge ordering occurs. However, this

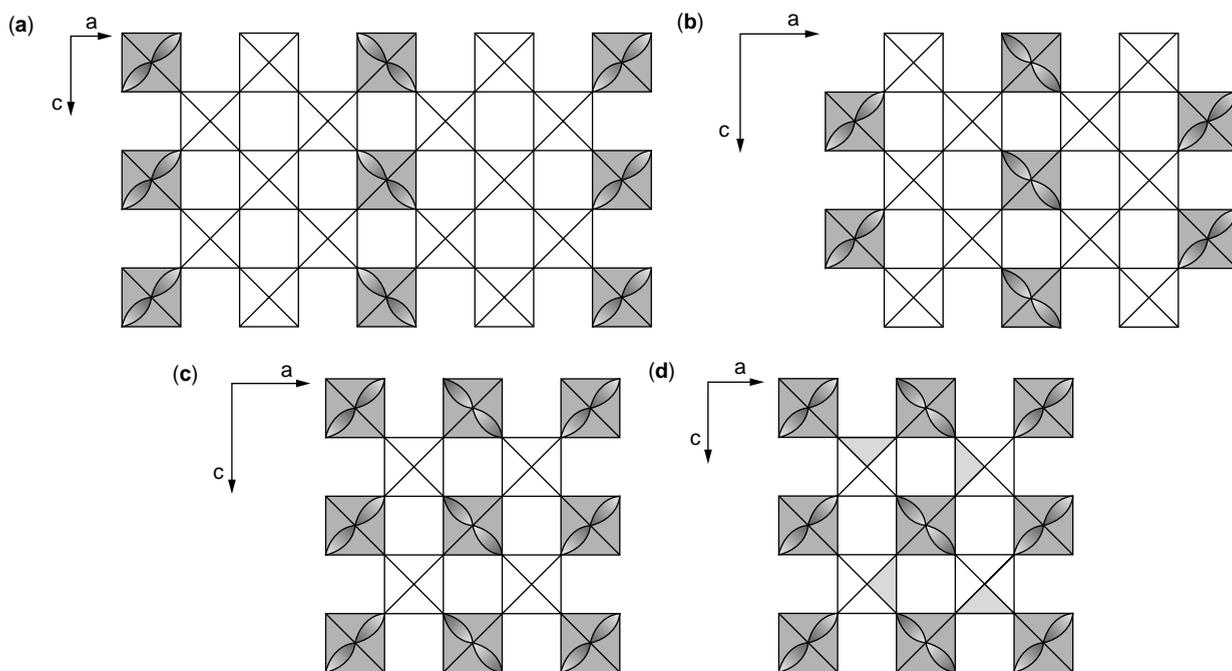


Figure 20. Corresponding $\text{Mn}^{3+}/\text{Mn}^{4+}$ arrangements (a) $x = 1/4$, (b) $x = 1/3$, (c) $x = 1/2$, and (d) the schematic drawing for $\text{Mn}^{3+}/\text{Mn}^{4+} > 1$ is also given.

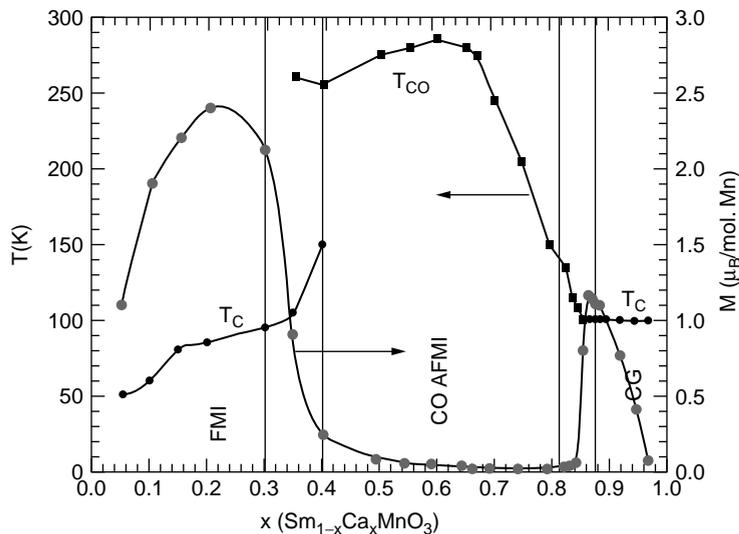


Figure 22. Magnetic and electronic phase diagram of $\text{Sm}_{1-x}\text{Ca}_x\text{MnO}_3$ as a function of x controlling the Mn valency. CG is for cluster-glass.

region covers a part of the Mn^{3+} compositions because CO is detected for $0.35 < x \leq 0.8$ in $\text{Sm}_{1-x}\text{Ca}_x\text{MnO}_3$. In fact, a mixture of two phases is obtained for $0.3 \leq x \leq 0.4$, where CO regions coexist with FM regions. For these hole-doped compositions, the extra reflections characteristic of the CO in the electronic diffraction patterns are in commensurate positions and indicate a doubling of the structural parameter similar to $\text{Sm}_{0.5}\text{Ca}_{0.5}\text{MnO}_3$. This result is not intuitive if one considers that the $\text{Mn}^{3+} : \text{Mn}^{4+}$ ratio is far from the 50:50 value of the half-doped manganites. Most probably, the extra Mn^{3+} species tend to be substituted for Mn^{4+} so that the CO can be viewed as alternating 1:1 planes of Mn^{3+} and mixed $\text{Mn}^{4+}/\text{Mn}^{3+}$ planes (Fig. 20d).

The existence of hole-doped compositions characterized at low temperature by the coexistence of CO and FM regions is very important for CMR. Recent observations by electron microscopy (90 K) of CMR manganites $\text{Pr}_{0.7}\text{Ca}_{0.3-x}\text{Sr}_x\text{MnO}_3$ where $0 \leq x \leq 0.05$, discussed earlier, showed the presence of short-range charge-ordered domains in these FM manganites. Consequently, the application of a magnetic field tends to transform the CO regions in FM regions and thus to generate CMR effects.

Finally, it should be mentioned, that no CO is observed for half-doped manganites of larger (r_A), such as $\text{Pr}_{0.5}\text{Sr}_{0.5}\text{MnO}_3$ (56,57). The low-temperature AFM structure is A type, that is, antiferromagnetically coupled FM planes (Fig. 23). According to the large (r_A) value of this compound, 1.24 Å, the Mn^{3+} octahedron is more flattened than in the CO phases because the $d_{x^2-y^2}$ orbitals are filled rather than the d_{z^2} . The $d_{x^2-y^2}$ orbitals (half-filled and empty for Mn^{3+} and Mn^{4+} , respectively), form MnO_2 FM planes in which the charges can be delocalized. In the out-of-plane direction, the empty d_{z^2} orbitals prevent any charge delocalization. The conductive nature of the FM planes of the A-type AFM structure has been probed by (in-plane and out-of-plane) resistivity measurements of $\text{Nd}_{0.45}\text{Sr}_{0.55}\text{MnO}_3$ (58) which show large anisotropy in transport properties. This anisotropic character can be made more isotropic by field application which restores

3-D FM metallic behavior in $\text{Pr}_{0.5}\text{Sr}_{0.5}\text{MnO}_3$ (56). Such half-doped manganites that have no CO but have an A-type AFM state can also show an I–M transition in a magnetic field and thus are CMR compounds.

OTHER CMR MANGANITES

CMR is also found in compositions very close to the G-type AFMI CaMnO_3 manganite, that is, for Mn valency close to four (59–61). Inspection of the $\text{Sm}_{1-x}\text{Ca}_x\text{MnO}_3$ phase diagram of Fig. 22 reveals the existence of “cluster-glass” (CG) compositions (62). The latter exhibit metallic behavior in the paramagnetic state that, it is believed, are connected with the small content of Jahn–Teller Mn^{3+} species, as contrasted to hole-doped manganites. Surprisingly, these compositions exhibit a G-type AFM structure below $T_N \sim 110$ K (G-type: each Mn moment is antiferromagnetically coupled to its six Mn nearest neighbors, Fig. 24) but also a nonnegligible FM moment ($M_{5K} \sim 1 \mu_B/\text{mole}$ of Mn for $\text{Sm}_{0.1}\text{Ca}_{0.9}\text{MnO}_3$) supposed to be the

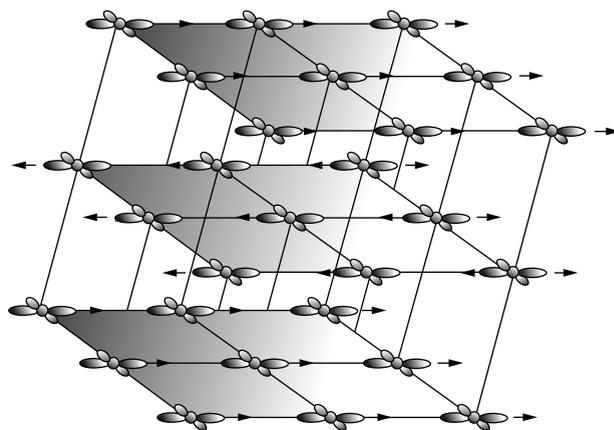


Figure 23. A-type AFM structure. The $d_{x^2-y^2}$ are ordered in the FM planes.

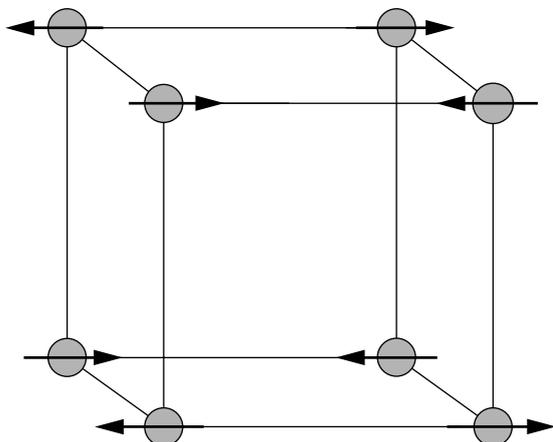


Figure 24. G-type AFM structure of CaMnO_3 .

signature of FM regions that coexist with the G-type AFM matrix that has very close T_N and T_C (63). Percolative pathways between these FM regions, as Mn^{3+} increases from CaMnO_3 , are responsible for the metallic behavior observed below T_N in $\text{Sm}_{0.1}\text{Ca}_{0.9}\text{MnO}_3$. However, at higher Mn^{3+} content, the materials are insulators at low temperature; they exhibit a C-type AFM structure (Fig. 21) with CO as in $\text{Sm}_{1-x}\text{Ca}_x\text{MnO}_3$ where $0.6 \leq x \leq 0.8$. At the boundary between CG and CO compositions, a very narrow composition range exists where CMR effects are observed (61). This is exemplified in Fig. 25 by the comparison of the $\rho(T)$ curves of $\text{Sm}_{0.15}\text{Ca}_{0.85}\text{MnO}_3$, registered upon cooling in and in the absence of a 7-T magnetic field, which show a ρ decrease by several orders of magnitude upon application of a magnetic field.

It should be pointed out that the CMR of these electron-doped compositions (Mn valency is greater than 3.5) is controlled by the valency and also by $\langle r_A \rangle$. First, for the $\text{Ln}_{1-x}\text{AE}_x\text{MnO}_3$ series, characterized by large $\langle r_A \rangle$ such as $\text{Pr}_{1-x}\text{Sr}_x\text{MnO}_3$, the electron-doped compositions lying close to SrMnO_3 crystallize in a hexagonal phase and are thus

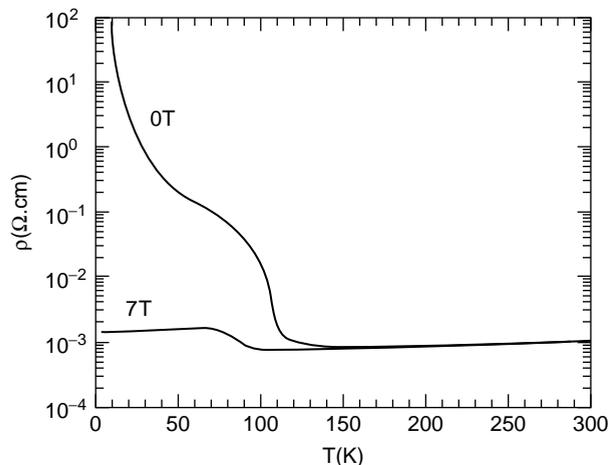


Figure 25. $\rho(T)$ curve of $\text{Sm}_{0.15}\text{Ca}_{0.85}\text{MnO}_3$ upon cooling at 0 and 7 T.

very difficult to prepare as pure cubic perovskite. Second, if cubic perovskites are stabilized by using a different synthetic approach, their magnetic properties show an absence of any FM component, in contrast to $\text{Ln}_{1-x}\text{AE}_x\text{MnO}_3$ manganites of small $\langle r_A \rangle$ (64). Moreover, the AFM state (C- or G-type) is much stronger according to their higher T_N (Fig. 7 and 22). Thus, the $\langle r_A \rangle$ effect on magnetism is opposite to that observed for hole-doped compounds. For the latter, the larger $\langle r_A \rangle$ values favor FM but are unfavorable in electron-doped manganites.

Besides 3-D perovskite manganites, other manganites exist that crystallize in layered structures, as in the Ruddlesden–Popper phases (SrO) $(\text{La}_{1-x}\text{Sr}_x\text{MnO}_3)_n$ (65–67) or pyrochlore as $\text{Tl}_2\text{Mn}_2\text{O}_7$ (68) that exhibit CMR properties. For the former, the $n = 2$ member $\text{La}_{1.8}\text{Sr}_{1.2}\text{Mn}_2\text{O}_7$ is a CMR compound consistent with its FMM state below T_C (67). Interestingly, when the current flows through the $[\text{La}/\text{SrO}]_\infty$ insulating layer, the out-of-plane transport properties are governed by inter-layer tunneling which is of particular interest for its low magnetic field magnetoresistance (69). Their properties (structural, magnetic, electronic) are also very sensitive to the size of the Ln/AE cations and to the Mn valency (70) as in the perovskite but are more complicated by the existence of two structural sites for Ln/AE cations (70)—perovskite between the two $[\text{MnO}_2]_\infty$ layers and NaCl-type in the $[(\text{Ln}/\text{AE})_2\text{O}_2]_\infty$ separating layers. By choosing a $\text{Mn}^{3+} : \text{Mn}^{4+} = 1 : 1$ ratio, CO in the planes has also been observed as in $\text{La}_{0.5}\text{Sr}_{1.5}\text{MnO}_4$ ($n = 1$) (71) and $\text{LaSr}_2\text{Mn}_2\text{O}_7$ ($n = 2$) (72).

The second example is the pyrochlore $\text{Tl}_2\text{Mn}_2\text{O}_7$ manganite (68). This structure has already attracted considerable attention: it is made of Mn–O–Mn angles of only 133° ; there is no mixed valency (it is a pure Mn^{4+} phase), but it is a ferromagnetic (by superexchange) metal below 142 K. Its conductivity is much larger than that of other $\text{A}_2\text{Mn}_2\text{O}_7$ pyrochlores ($A = \text{Y}, \text{Sc}, \text{In}$) and arises from the contribution of Tl ($6s^2$) states to the density of states as has been proposed (73).

CONCLUSION

The goal of this paper was to emphasize the complexity of the underlying phenomena governing the CMR properties of manganites. These compounds have attracted a lot of scientific interest and numerous papers on the topic have been published. It is very difficult and too ambitious to give an exhaustive overview of the state of the art. For more informations and details, the readers can refer to books (74,75) and review articles (76–80). However, a few important notions emerge in the examples given in this article.

The strong interplay among charge, spin, and structure in manganese oxides should always be considered in apprehending the properties of these fascinating materials. Note that this guideline can be extended to other transition-metal oxides such as the high T_C superconducting cuprates, the $\text{La}_2\text{NiO}_{4+\delta}$ nickelates, the charge-ordered ferrites [for instance $(\text{La}/\text{Sr})\text{FeO}_3$], etc.

A second important parameter is the competition between the charge-ordering process that leads to a localized

antiferromagnetic state and a ferromagnetic state that allows charge delocalization. Subtle changes in composition can be used to change their respective proportions and induce dramatic effects on structural, magnetic, and electronic properties. In this context, percolation models become more and more pertinent for describing experimental observations.

Finally, one actual limitation for the application is a magnetic field too high to obtain magnetoresistance effects at room temperature. Several technological approaches are possible for overcoming this problem, such as control of grain-boundary quality and growth of multilayered heterostructures.

BIBLIOGRAPHY

- M.N. Baibich, J.M. Broto, A. Fert, F. Nguyen Van Dan, F. Petroff, P. Etienne, G. Creuzet, A. Friedrich, and J. Chazeles, *Phys. Rev. Lett.* **61**: 2472 (1988).
- G. Binash, P. Grünberg, F. Saurenbach, and W. Zin, *Phys. Rev. B* **39**: 4828 (1989).
- P.M. Levy, *Solid State Phys.* **47**: 367 (1994); P.M. Levy and S. Zhang, *J. Magn. Magn. Mater.* **151**: 315 (1995).
- R.M. Kusters, D.A. Singleton, D.A. Kenn, R. McGreevy, and W. Hayes, *Physica B* **155**: 362 (1989).
- R. Von Helmut, B. Holzapfel, L. Schultz, and K. Samwer, *Phys. Rev. Lett.* **71**: 2331 (1993).
- K. Chahara, T. Ohno, M. Kasai, and Y. Kozono, *Appl. Phys. Lett.* **63**: 1990 (1993).
- M. McCormack, S. Jin, T. Tiefel, R.M. Fleming, J. Philipps, and R. Ramesh, *Appl. Phys. Lett.* **64**: 3045 (1994).
- S. Jin, T.H. Tiefel, M. McCormack, R.A. Fastnacht, R. Ramesh, and L.H. Chen, *Science* **264**: 413 (1994).
- H.L. Ju, C. Kwon, R.L. Greene, and T. Venkatesan, *Appl. Phys. Lett.* **65**: 2108 (1994).
- Y. Tokura, A. Urushibara, Y. Moritomo, T. Arima, A. Asamitsu, G. Kido, and N. Furukawa, *J. Phys. Soc. Jpn.* **63**: 3931 (1994).
- R. Mahesh, R. Mahendiran, A.K. Raychaudhuri, and C.N.R. Rao, *J. Solid State Chem.* **114**: 297 (1995); **120**: 204 (1995).
- C. Zener, *Phys. Rev.* **82**: 403 (1951).
- P.G. de Gennes, *Phys. Rev.* **118**: 141 (1960).
- P.W. Anderson and H. Hasegawa, *Phys. Rev.* **100**: 675 (1955).
- A.J. Millis, P.B. Littlewood, and B.I. Shraiman, *Phys. Rev. Lett.* **74**: 5144 (1995).
- H. Röder, J. Zhang, and A. Bishop, *Phys. Rev. Lett.* **76**: 1356 (1996).
- M.R. Ibarra, P.A. Algarabel, C. Marquina, J. Blasco, and J. Garcia, *Phys. Rev. Lett.* **75**: 3541 (1995).
- G. Zhao, K. Conder, H. Keller, and K.A. Müller, *Nature* **381**: 676 (1996); G. Zhao, K. Gosh, H. Keller, and R.L. Greene, *Phys. Rev. B* **59**: 81 (1999).
- J.M. De Teresa, M.R. Ibarra, P.A. Algarabel, C. Ritter, C. Marquina, J. Blasco, J. Garcia, A. del Moral, and Z. Arnold, *Nature (London)* **386**: 256 (1997).
- A. Moreo, S. Yunoki, and E. Dagotto, *Science* **283**: 2034 (1999) and references therein.
- M. Uehara, S. Mori, C.H. Chen, and S.W. Cheong, *Nature* **399**: 560 (1999).
- M. Fäth, S. Freisem, A.A. Menovsky, Y. Tomioka, J. Aarts, and J. A. Mydosh, *Science* **285**: 1540 (1999).
- M. Hennion, F. Moussa, J. Rodriguez-Carvajal, L. Pinsard, and A. Revcolevschi, *Phys. Rev. B* **56**: R497 (1997).
- G. Allodi, R. De Renzi, F. Licci, and M.W. Pieper, *Phys. Rev. Lett.* **81**: 4736 (1999).
- R. Mahendiran, M.R. Ibarra, A. Maignan, F. Millange, A. Arulraj, R. Mahesh, B. Raveau, and C.N.R. Rao, *Phys. Rev. Lett.* **82**: 2191 (1999).
- T. Kimura, Y. Tomioka, R. Kumai, Y. Okimoto, and Y. Tokura, *Phys. Rev. Lett.* **83**: 3940 (1999).
- A. Maignan, Ch. Simon, V. Caignaert, and B. Raveau, *Solid State Commun.* **96**: 623 (1995).
- B. Raveau, A. Maignan, and C. Simon, *J. Solid State Chem.* **117**: 424 (1995).
- A.M. Glazer, *Acta Crystallogr. B* **28**: 3384 (1972).
- P.M. Woodward, *Acta Crystallogr.* **53**: 32 (1992).
- G.H. Jonker and J.H. Van Santen, *Physica* **16**: 337 (1950).
- V. Caignaert, A. Maignan, and B. Raveau, *Solid State Comm.* **95**: 357 (1995); A. Maignan, C. Simon, V. Caignaert, and B. Raveau, *Z. Phys. B* **99**: 305 (1996).
- P.G. Radaelli, M. Marezio, H.Y. Hwang, and S.W. Cheong, *J. Solid State Chem.* **122**: 444 (1996).
- C. Martin, A. Maignan, M. Hervieu, and B. Raveau, *Phys. Rev. B* **60**: 12191 (1999).
- A. Urushibara, Y. Moritomo, T. Arima, A. Asamitsu, G. Kido, and Y. Tokura, *Phys. Rev. B* **51**: 14103 (1995).
- P. Schiffer, A.P. Ramirez, W. Bao, and S.W. Cheong, *Phys. Rev. Lett.* **75**: 3336 (1995).
- R.D. Shannon, *Acta Crystallogr. A* **32**: 751 (1976).
- L.M. Rodriguez-Martinez and J.P. Attfield, *Phys. Rev. B* **54**: R15622 (1996).
- A. Maignan, C. Martin, G. Van Tendeloo, M. Hervieu, and B. Raveau, *Phys. Rev. B* **60**: 15214 (1999).
- J.M. Tranquada et al., *Phys. Rev. Lett.* **79**: 2133 (1997); *Nature* **375**: 561 (1995).
- J.B. Goodenough, *Phys. Rev. B* **100**: 564 (1955).
- E.O. Wollan and W.C. Koehler, *Phys. Rev.* **100**: 545 (1955).
- S. Mori, C.H. Chen, and S.W. Cheong, *Phys. Rev. Lett.* **81**: 3972 (1998).
- A. Barnabé, M. Hervieu, A. Maignan, C. Martin, and B. Raveau, *J. Appl. Phys.* **84**: 5506 (1998).
- Y. Tomioka, A. Asamitsu, Y. Moritomo, and Y. Tokura, *J. Phys. Soc. Jpn.* **64**: 3626 (1995).
- H. Kuwahara, Y. Tomioka, and Y. Tokura, *Science* **270**: 961 (1995); H. Kawano, H. Kuwahara, and Y. Tokura, *Phys. Rev. Lett.* **78**: 4253 (1997).
- J. Wolfman, C. Simon, M. Hervieu, A. Maignan, and B. Raveau, *J. Solid State Chem.* **123**: 413 (1996); F. Damay, C. Martin, A. Maignan, M. Hervieu, B. Raveau, and Z. Jirak, G. André, and F. Bourée, *Chem. Mater.* **11**: 536 (1999).
- Z. Jirak, S. Krupicka, Z. Simsa, M. Dhoula, and S. Vratslav, *J. Magn. Magn. Mater.* **53**: 153 (1985).
- C. Ritter, R. Mahendiran, M.R. Ibarra, L. Morellon, A. Maignan, B. Raveau, and C.N.R. Rao, *Phys. Rev. B* **61**: 9229 (2000).
- B. Raveau, A. Maignan, and C. Martin, *J. Solid State Chem.* **130**: 162 (1997); A. Maignan, F. Damay, C. Martin, and B. Raveau, *Mater. Res. Bull.* **32**: 965 (1997); A. Maignan, F. Damay, A. Barnabé, C. Martin, M. Hervieu, and B. Raveau, *Philos. Trans. R. Soc. London A* **356**: 1635 (1998).

51. F. Damay, C. Martin, A. Maignan, M. Hervieu, B. Raveau, F. Bourée, and G. André, *Appl. Phys. Lett.* **73**: 3772 (1998).
52. R. Mahendiran, M. Hervieu, A. Maignan, C. Martin, and B. Raveau, *Solid State Comm.* **114**: 429 (2000).
53. W. Bao, J.D. Axe, C.H. Chen, and S.W. Cheong, *Phys. Rev. Lett.* **78**: 543 (1997).
54. Y. Murakami, D. Shindo, H. Chiba, M. Kikuchi, and Y. Syono, *Phys. Rev. B* **55**: 15043 (1997).
55. M. Hervieu, A. Barnabé, C. Martin, A. Maignan, F. Damay, and B. Raveau, *Eur. Phys. J. B* **8**: 31 (1999).
56. H. Kawano, R. Kajimoto, H. Yoshizawa, Y. Tomioka, H. Kuwahara, and Y. Tokura, *Phys. Rev. Lett.* **78**: 4253 (1997).
57. F. Damay, C. Martin, M. Hervieu, A. Maignan, and B. Raveau, *J. Magn. Magn. Mater.* **184**: 71 (1998).
58. H. Kuwahara, T. Okuda, Y. Tomioka, A. Asamitsu, and B. Tokura, *Phys. Rev. Lett.* **82**: 4316 (1999).
59. H. Chiba, M. Kikuchi, K. Kusaba, Y. Muraoka, and Y. Syono, *Solid State Commun.* **99**: 446 (1996).
60. I.O. Troyanchuk, N.V. Samsonenko, H. Szymczak, and A. Nabialek, *J. Solid State Chem.* **131**: 144 (1997).
61. C. Martin, A. Maignan, F. Damay, M. Hervieu, and B. Raveau, *J. Solid State Chem.* **134**: 198 (1997); A. Maignan, C. Martin, F. Damay, and B. Raveau, *Chem. Mater.* **10**: 950 (1998).
62. A. Maignan, C. Martin, F. Damay, B. Raveau, and J. Hejtmanek, *Phys. Rev. B* **58**: 2578 (1998).
63. C. Martin, A. Maignan, M. Hervieu, B. Raveau, Z. Jirak, M.M. Savosta, A. Kurbakov, V. Trounov, G. André, and F. Bourée, *Phys. Rev. B* **62**: 6442 (2000).
64. M. Hervieu, C. Martin, A. Maignan, G. Van Tendeloo, Z. Jirak, J. Hejtmanek, A. Barnabé, D. Thopart, and B. Raveau, *Chem. Mater.*, **12**: 1456 (2000).
65. R.A. Ram, P. Ganguly, and C.N.R. Rao, *J. Solid State Chem.* **70**: 82 (1987).
66. Y. Moritomo, Y. Tomioka, A. Asamitsu, and Y. Tokura, *Phys. Rev. B* **51**: 3297 (1995).
67. Y. Moritomo, A. Asamitsu, H. Kuwahara, and Y. Tokura, *Nature* **380**: 141 (1996).
68. Y. Shimakawa, Y. Kubo, and T. Marako, *Nature (London)* **379**: 53 (1996).
69. T. Kimura et al., *Science* **274**: 1698 (1996).
70. P. Laffez, G. Van Tendeloo, R. Seshadri, M. Hervieu, C. Martin, A. Maignan, and B. Raveau, *J. Appl. Phys.* **80**: 5850 (1996).
71. Y. Murakami, H. Kawada, M. Tanaka, T. Arima, Y. Moritomo, and Y. Tokura, *Phys. Rev. Lett.* **80**: 1932 (1998).
72. J.Q. Li, Y. Mutsui, T. Kimura, and Y. Tokura, *Phys. Rev. B* **57**: R3205 (1998).
73. M.A. Subramanian, B.H. Toby, A.P. Ramirez, et al., *Science* **273**: 81 (1996).
74. C.N.R. Rao and B. Raveau, eds., *Colossal magnetoresistance, charge Ordering and Related Properties of Manganese Oxides*, World Scientific P, Singapore, 1998.
75. Y. Tokura, ed., *Colossal Magnetoresistive Oxides*, Gordon and Breach, 1999.
76. A.P. Ramirez, *J. Phys. : Condens. Matter* **9**: 8171 (1997).
77. C.N.R. Rao, A.K. Cheetham, and R. Mahesh, *Chem. Mater.* **8**: 2421 (1996).
78. M. Imada, A. Fujimori, and Y. Tokura, *Rev. Mod. Phys.* **70**: 1039 (1998).
79. J.M.D. Coey, M. Viret, and S. von Molmar, *Adv. Phys.* **48**: 167 (1999).
80. C.N.R. Rao, A. Arulraj, A.K. Cheetham, and B. Raveau, *J. Phys. Condens. Matter* **12**: 83 (2000).

COMPOSITES, FUTURE CONCEPTS

BRIAN S. THOMPSON

Michigan State University
East Lansing, MI

INTRODUCTION

The composite materials considered here are solid objects with a macrostructure. The constituents of these, solids can be observed with the naked eye. Solid objects are said to be *smart* if they embody additional functionality capabilities beyond their inherent structural attributes. These capabilities might be attributed to an embedded network of interconnected sensors, actuators and computers, for example. Synthetic inhomogeneous materials with these capabilities comprise the basis for a new generation of materials. These materials have the potential to revolutionize many types of products, and usher into existence unforeseen manufactured goods.

Humankind's traditional quest for superior materials may be satisfied in the near future by ideas furnished by Mother Nature. The design and manufacturing methodologies needed for creating new generations of materials will come from a meticulous study of flora and fauna. The future lies with the development of synthetic materials that mimic naturally occurring biological materials.

HISTORICAL PROLOGUE

Materials science has come the full circle. The evolution of this important field began with humankind's use of naturally occurring materials. Materials have had a profound impact on the evolution of world civilizations. Historians have classified periods in this evolution by the materials that were the state-of-the-art during these periods. Thus the vocabulary now contains phrases like the Stone Age, the Bronze Age, and the Iron Age. Each of these eras, illustrated in Fig. 1, is characterized by the material that was the most advanced of its time. An alternative classification is presented in Fig. 2. Here the eras are classified by the type of properties embodied by each material.

Homo habilis chose unrefined naturally occurring materials for weapons and tools during the Paleolithic period, some one million years ago. This decision was responsible for the selection of flint, a fine-grained very hard abrasion-resistant siliceous rock, rather than sandstone or bone to be lashed to a long shaft of wood to create an innovative weapon system for hunting: namely the spear. About 3500 BC, *Homo sapiens sapiens* began to create bronzes by smelting ores. This accomplishment required metallurgical prowess, and it also exploited the ability to generate and control heat. These new nonferrous materials were alloys of copper and tin, and they were instrumental in creating superior classes of weapons, tools, and utensils.

The development of foundry technology enabled higher furnace temperatures to be generated, and a different class of ores to be smelted. This new technology was responsible for the demise of nonferrous alloys as the materials of choice. They were superseded by a new class of

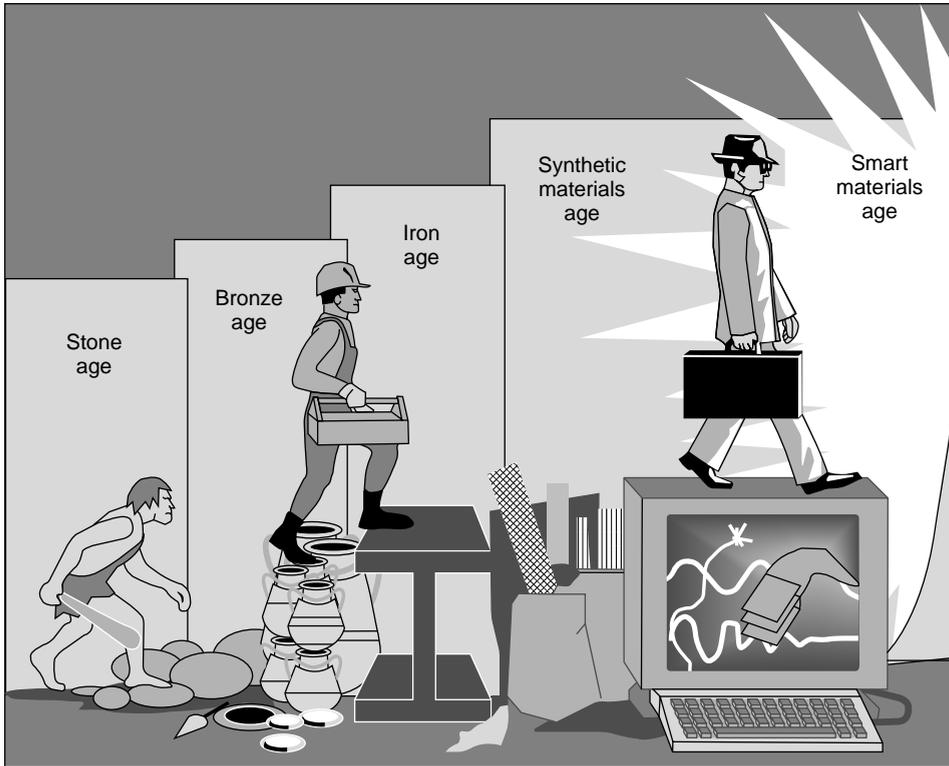


Figure 1. Historical eras of materials science.

state-of-the-art materials that were more versatile and harder than their predecessors. These new ferrous materials were the irons. The irons were exploited in many aspects of life to generate innovative classes of products, ranging from domestic articles to military systems and agricultural implements.

The societal consequences of these new materials were immeasurable. For example, Archimedes, Ctesibius, and Hero exploited the properties of these new materials to create numerous mechanical devices. These included the pulley, the lever and the screw, hydraulic devices, compressed air devices and screw-cutting machines. In addition, plough shares wrought in iron facilitated the development of deeper plowing techniques and the cultivation of inferior quality soils. Consequently this new capability enabled human settlement to occur in these less desirable regions, thereby leading to the growth of different civilizations. Clearly, the cascading consequences of this new generation of ferrous materials were immeasurable.

In 1856 the demise of iron was initiated by Henry Bessemer's design of a converter to create steel by inexpensively refining iron. The consequences of this new class of materials were titanic. Steel fueled the Industrial Revolution, and at that significant technological, socio-economic, and cultural changes. These changes increased confidence in exploiting the properties of new materials, and utilizing scientific data; this has resulted in new generations of industrial machines, new organizational structures for the human component of the enterprises, and improved communications and transportation technologies. A nexus of mass production environments was created and consequently a dramatic increase in the consumption of raw materials.

These cascading phenomena were far reaching. They included an increase in international trade. Employees were required to develop machine-oriented skills while being subjected to the discipline of an industrial environment, and industrialization mandated that the legal system respond with new legislation. Suburban areas flourished, wealth was distributed more uniformly, agriculture was mechanized to feed the industrial communities, and finally humankind received a psychological boost from this new-found success in further exploiting natural resources in a controlled manner. Thus at the dusk of the nineteenth century, there were complex interactions between the naturally occurring materials that fueled the industrialization, technological innovations, and rapid changes in society.

By the twentieth century an embryonic technology involving synthetic materials had emerged. This was a profound departure from the traditional approach of exploiting natural materials with their known defects and limitations. The new generation of materials was the plastics. The early successes included celluloid, xylonite, cellophane, and bakelite. The growth of the plastics field was astronomical. Indeed, by 1979, the volume of plastics manufactured in the United States exceeded steel production for the first time. There were thousands of different plastics in the marketplace to satisfy a diverse range of design requirements for materials with appropriate stiffness, strength, hardness, wear resistance, heat resistance, transparency, and so on. This diversity was achieved by carefully synthesizing the appropriate atomic structure from primarily the chemical elements carbon, hydrogen, oxygen, and nitrogen.

Synthetic plastic materials replaced the traditional materials in a diverse range of industries from transportation

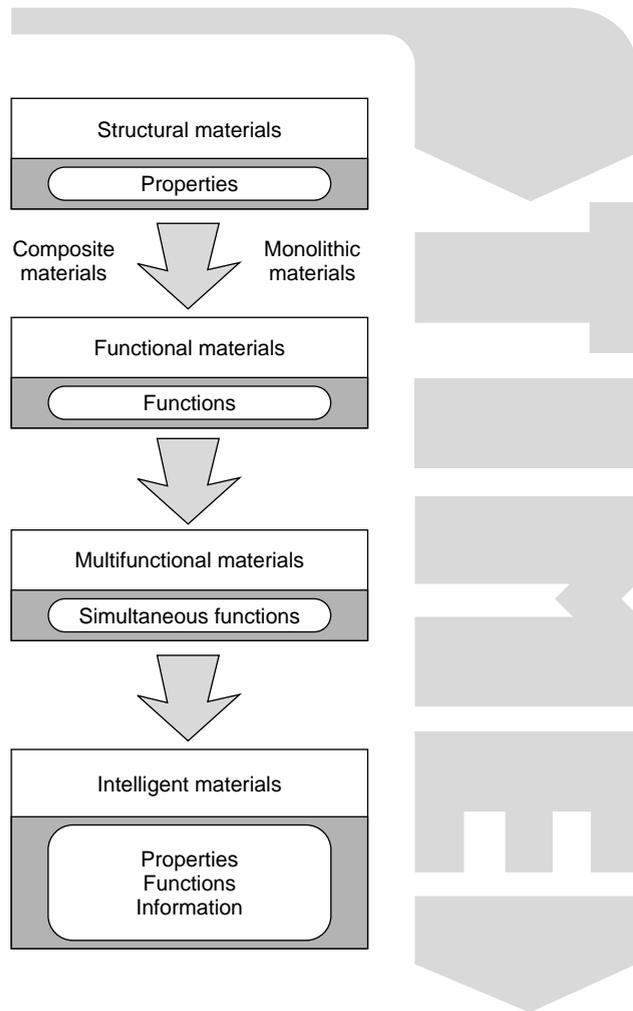


Figure 2. Evolution of the design of materials with macrostructure and microstructure.

to domestic appliances and the medical supply industry. The reason for this universal utilization of these new materials is their extensive range of physical properties. Their physical properties could comply precisely with the design specification of a product. In the past, metallurgists were primarily responsible for the development of new materials. Now these new materials are synthesized by eclectic teams of specialists.

The trend of employing eclectic teams of specialists was responsible for the explosive development during the latter half of the twentieth century of a plethora of material classifications too numerous to discuss here. These include the development of a variety of functional materials, such as gallium arsenide or magnetostriuctive materials where the functional properties are exploited in practice instead of the structural properties as in traditional practice. However, there is one classification worthy of mention because it is important commercially and is central to chronicling the evolution of a branch of materials science.

This class of modern materials is the advanced composites. These engineered materials are synthesized within

two distinct phases comprising a load-bearing material housed in a relatively weak protective matrix. The reinforcement is typically particles, whiskers, or fibers, while the matrix can be polymeric, ceramic, or metallic materials. A characteristic of these composite materials is that the combination of two or more constituent materials creates a material with engineering properties superior to those of the constituents—albeit at the expense of more challenging fabrication technologies.

As an aside, it should be noted that the field of fibrous composite materials technologies is not entirely new. Consider the recording in Exodus, chapter 5 of the Bible, of the Israelites manufacturing bricks from a mixture of clay and straw for the pharaoh in Egypt in 450 BC. Over a thousand years later, the French used a combination of horse hair and plaster to create ornate ceilings in stately homes.

Advanced polymeric composite materials have afforded the engineering community the opportunity to fabricate products with the strongest and stiffest parts per unit weight. Furthermore, by appropriately designing the macrostructure of these materials, the engineer can develop composites with different properties in different directions, or alternatively, different properties in different domains of a structural member. Thus, not only are the geometrical and surface attributes of the part being designed, but in addition the material's macrostructure is being design too. The infusion of these materials into numerous industries has been responsible for the creation of many generations of products in the defense, automotive, biomedical, and sporting goods industries, for example.

Thus at the dusk of the twentieth century the creation of materials with an engineered macrostructure was the state-of-the-art. In the context of advanced fibrous composite materials, materials science had come the full circle. In the beginning, naturally occurring materials were studied by *Homo habilis* to select the appropriate mineral for the tip of the spear or arrow. Now, one million years later, *Homo sapiens sapiens* is studying naturally occurring fibrous composite materials in order to emulate the placement of fibers for the creation of synthetic composite parts. One of the primary motivations for the growth of this field of scientific endeavor is that naturally occurring fibrous composites are embodied in numerous species of flora and fauna. They are one of the basic building blocks of life. They are an intrinsic design of Nature's. Clearly, if these biological systems have evolved to these mature states through the millennia, then they are indeed worthy of study and perhaps emulation in an engineering context. This observation provides the underpinning of the subsequent section: biologically inspired materials.

Historically, then, each era during the evolution of humankind was motivated by the insatiable demand for superior weapons and more innovative products. In turn, this demand propelled the maturing of materials science, because the triumvirate of product design comprises materials, manufacturing and conceptual design. And of course this quest is very evident to this day. Currently the age of synthetic materials has been driven by the same motives. While plastics shall undoubtedly remain an increasingly dominant component of modern lifestyles, the new millennium shall witness the emergence of a superior class

of advanced materials. These smart materials, or intelligent materials, materials with diverse characteristics that mimic flora and fauna, shall be responsible for historians classifying it as *The Age of Smart Materials*, because they will dominate the materials selection process when technologists are seeking state-of-the-art materials.

BIOLOGICALLY-INSPIRED CREATIVITY IN ENGINEERING

Although human genius through various inventions makes instruments corresponding to the same ends. It will never discover an invention more beautiful nor more ready nor more economical than does Nature, because in her inventions nothing is lacking and nothing is superfluous.

Leonardo da Vinci

(Ms RL 19115v; K/P 114r; Royal Library, Windsor)

These profound words penned some 500 years ago by Leonardo, that brilliant painter, sculptor, draftsman, architect and visionary engineer, are the earliest that formally recognize the power of Nature's creativity in the fields of natural science such as botany, zoology, and entomology. The naturally occurring materials and systems associated with these academic disciplines have developed their properties and characteristics over millions of years through processes of Darwinian evolution. They have been required to survive when subjected to dynamically changing environmental conditions. Indeed, these conditions demand that only the most adaptable and fittest survive. Therefore, these biological systems are truly optimal designs engineered by Mother Nature in response to a set of unwritten design specifications. Hence they merit meticulous scrutiny and potential emulation by humankind.

Since Charles Darwin's seminal work in 1859, entitled "On the Origin of Species by Means of Natural Selection," the notion of natural selection has remained the central theme of evolutionary biology. Thus the proposition is that all life forms evolved because organisms with traits that promoted reproduction and survival somehow passed on those traits to future generations. Organisms without those traits simply became extinct. They failed to survive. Indeed, this powerful assertion has motivated theories in several disciplines beyond the central theme of materials science. Hence the proposition of learning from biological systems to advance the field of engineering has credence. Evolutionary psychology, for example, proposes that the human mind is not a vacuous medium, but instead comprises specialized mental protocols that were honed by the solving of problems faced long ago. Sociobiology, on the other hand, employs natural selection and other biological phenomena to explain the social behavior of animals.

This emulation, or mimicking, of biological systems is often called *biomimetics*. The name is derived from the Greek *bios* (life) and *mimesis* (imitation). It can be employed by creative designers to develop solutions to engineering problems through the use of a direct analogy between a naturally occurring system and an engineering system. Thus a meticulous and comprehensive study of a living organism can yield invaluable insight into the subtleties of its refined design attributes developed by a lengthy process of evolution and optimization.

Already during recent centuries Nature's creative prowess has been recognized and exploited by many gifted individuals. Consider, for example, Sir George Cayley's work in 1810 when he was designing a low-drag shape for his fixed wing flying machine. He exploited his knowledge of ichthyology to propose that the geometry of the wing cross section should mimic the streamlined low-drag cross section of the trout. Sir Marc Isambard Brunel proposed the use of caissons to facilitate the underwater construction of civil engineering structures by the serendipitous observation of a shipworm tunneling in timber. George de Mistral developed the hook-and-loop fastener, such as those manufactured by Velcro USA, Inc, by studying how cocklebur plants tenaciously adhered to his trousers after he had walked through woodland thickets. Of course, there are many others too who have used a direct analogy between a biological system and an engineering system in order to create a new artifact for the betterment of humankind.

In the context of the development of advanced materials from studies of naturally occurring systems with fibers, such as the stalks of celery or the skin of a banana, the fibrous composites of engineering practice are an obvious consideration. They emulate the fibrous structures of muscles or plants. For example, the structure of the humble tree comprises flexible cellulose fibers in a rigid lignin matrix. On the other hand, the insect cuticle comprises chitin fibers in a proteinaceous matrix.

On a separate level of comparison between materials engineered by humankind and those created by Mother Nature, consider the variety of materials in both classifications. In 1990, it was estimated that the marketplace contained about 60,000 different plastics. This contrasts sharply with the observation that there are only two groups of substances from which almost all skeletal tissues of animals and plants are formed. These groups are the amino-acid-based proteins and polysaccharides that occur together in different proportions in the vast majority of biological materials. Clearly, the apparently limited chemistry of these naturally occurring materials is compensated for by the tremendous diversity of their microstructure. In addition, the creation and manufacture of these diverse microstructures is accomplished when subjected to a small range of temperature, humidity, and pressure. These limited conditions contrast markedly with the extreme thermodynamic environments employed to produce plastics. Finally, these processes of design and manufacture occur simultaneously. They occur in unison. Again, this contrasts sharply with the protocols implemented in a large percentage of industrial enterprises where design departments and manufacturing departments function autonomously. They do not practice simultaneous engineering.

SMART MATERIALS AND STRUCTURES: CURRENT NONCOMMERCIAL TECHNOLOGIES

Biological systems are extremely complex, and in the context of biomimetics, it is therefore only inevitable that an eclectic team of specialists must be assembled to prosecute the research if progress is to be assured in the

creation of new materials. Teams of trained professionals are mandatory because biomimetics lies at the boundaries of numerous artificial disciplines that have been created by the academic community in order to study the phenomenological problems of life. Thus a team with expertise in materials science, biological sciences, biotechnology, nanotechnology, molecular electronics, artificial intelligence, data-processing, microprocessors, automatic control theory, manufacturing, applied mathematics, chemistry, physics and the like, is required, along with the associated equipment. The best configuration would be a blend of academics, industrialists, and government employees. The creation of this type of infrastructure is often a challenging undertaking. Thus there is a need for large grants and perhaps centers of excellence.

Unfortunately, in many academic institutions the traditional research philosophy is for each professor to seek their own funding. Interdisciplinary collaboration is minimal. Unfortunately, this is an outcome of the academic system where professors are promoted because of their individual accomplishments and natural phenomena are broken down into distinct fields and the artificial interfaces are well defined. This infrastructure hinders potential interactions between individuals in adjacent fields. Thus the conventional model is of a single investigator guiding graduate students and postdoctoral fellows. Therefore there needs to be a major shift in the current research paradigm if progress in this field is to proceed effectively.

However, using the typical single-discipline non-embracing philosophy, humankind has indeed created a multitude of macroscopically smart materials that mimic naturally occurring materials. While these material systems are still quite crude relative to Nature's materials, they do exhibit, at the most advanced level, the behavior of muscles (by using actuators), nerves (by using sensors), and brains (by using control schemes, microprocessors, fuzzy logic, artificial intelligence, etc.) at various levels of sophistication.

The diverse structural materials, functional materials, and multifunctional materials can be combined in building-block style, as shown in Fig. 3. The materials that mimic naturally occurring materials can have, at the most advanced levels sensors, actuators, information pathways, control schemes, and microprocessors. Some people classify these materials as *intelligent materials* because they can respond effectively to dynamically changing environmental conditions.

An intelligent material could comprise a graphite-epoxy fibrous polymeric structural material that houses some embedded fiber optic Mach Zender interferometric strain sensors and conventional surface-mounted electrical resistance strain gauge strain sensors. The signals from these sensors could be processed by a microprocessor that employs an appropriate control algorithm to activate an array of actuators exploiting piezoelectric and shape-memory phenomena to create an optimal performance under variable external stimuli.

It should be noted that there are naturally several levels of complexity in this field. At the lowest level, a material might feature only sensors to emulate the nerves of a

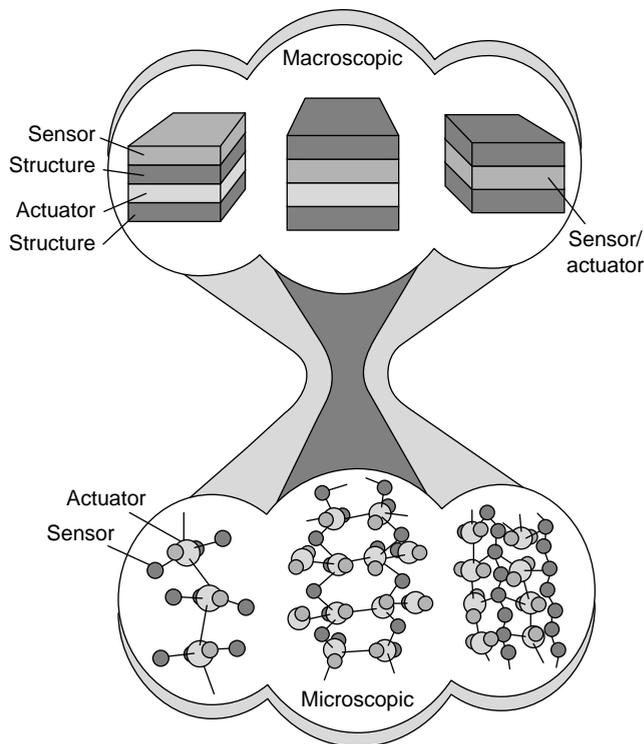


Figure 3. Classification of materials design.

biological material: photonic, piezoelectric, or conventional strain gauges, for example, depending on the design specification. Alternatively, the smart material might only feature actuators that are controlled manually to emulate the muscles of a biological material. These actuators include a variety of functional materials that demonstrate the diverse phenomena associated with piezoelectricity, electro-rheology, magnetorheology, magnetostriction, and electrostriction, for example. They are typically excited by an electrical stimulus that changes the geometrical configuration, stiffness, or energy dissipation characteristics in a controlled manner. At the highest level, the material might include the brains of a biological system. These complex entities can be imperfectly represented by the exploitation and integration of diverse theories and technologies associated with neural networks, rule-based systems, a multitude of control schemes, signal processing, nanotechnology, and microprocessors, for example. Figure 4 illustrates some of the primary components of these materials.

The last decade has witnessed an explosion of articles on smart materials, and the first book dedicated to the field was published in 1992 by the author. While two journals are dedicated to this field—namely the *Journal of Intelligent Material Systems and Structures* and the *Journal of Smart Materials and Structures*—the eclectic nature of this field, coupled with the importance of materials science in the product realization process, mandates that most academic journals now publish regularly articles on some aspect of the field. The World Wide Web is an obvious source of information on these activities, but the

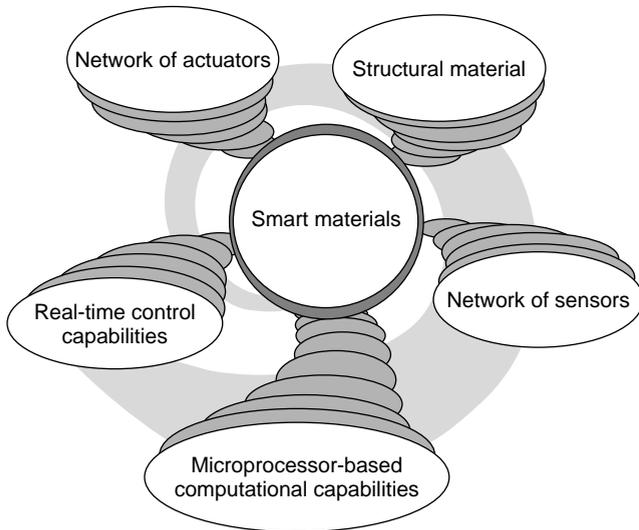


Figure 4. Ingredients of the premier class of smart materials.

professional wishing to enter this rapidly evolving field would be well advised to attend the SPIE's international symposium on smart materials and structures for an overview of current thinking.

SMART MATERIALS AND STRUCTURES: THE FUTURE

Predicting the future of a technical field with any certainty is fraught with errors and subsequent embarrassment. Therefore I will refrain from peering to the horizon with binoculars, preferring to view the future without an optical aid. After all, there are numerous records of truly gifted individuals failing to accomplish this task of predicting the evolution of a scientific disciplines or a technological field. For example, the famous British scientist Lord Kelvin, who was president of the Royal Society in 1895, stated that "Heavier than air flying machines are impossible." Four years later, Charles H. Duell, commissioner of the U.S. Office of Patents, commented that "Everything that can be invented has been invented." More recently, in 1977, Ken Olsen, the founder of Digital Equipment Corporation, said "There is no reason for any individuals to have a computer in their home." Thus the evidence is clear. Nevertheless, future objectives will be explored on the tenet that engineers and scientists should mimic naturally occurring systems as one method of advancing the state-of-the art.

The most sophisticated class of smart materials in the foreseeable future will be that which emulates biological systems. This class of multifunctional materials will possess the capability to select and execute specific functions intelligently in order to respond to changes in the local environment. Furthermore, these materials could have the ability to anticipate challenges based on the ability to recognize, analyze, and discriminate. These capabilities should include self-diagnosis, self-repair, self-multiplication, self-degradation, self-learning, and homeostasis.

Table 1. Nature's Rules for Sustaining Ecosystems

Use waste as a resource
Diversify and cooperate to fully use the habitat
Gather and use energy efficiently
Optimize rather than maximize
Use material sparingly
Don't foul nests
Don't draw down resources
Remain in balance with the biosphere
Run on information
Shop locally

Environmental Design Issues

The maturing nexuses between professionals in the biological and engineering communities have been responsible for a deeper appreciation of the awesome prowess of Nature's creations. This appreciation should be used to establish new operating procedures for the design and manufacture of the next generation of materials. They canon can be distilled from studies of complex ecosystems that have evolved through the millennia. These winning strategies have been adopted by diverse organisms; therefore, they have been employed by both animals and plants comprising diverse organs and parts. This diversity nevertheless does not defer their harmonious functioning together to sustain life and its activities. Indeed, these diverse organisms provide a set of rules that are worthy of analysis and potential implementation in engineering practice by humankind. After all, *Homo sapiens sapiens* is an organism too. They are presented in Table 1.

The information in Table 1 has very broad ramifications in terms of environmental design protocols for the engineering profession, in addition to the creation of new classes of smart materials.

Health Monitoring of Smart Materials

There is a cycle in all ecosystems. Biological systems experience life and death. Thus there is a recycling of the material. Mature trees in a forest absorb nutrients from the soil, but they eventually die and collapse to the forest floor. There they crumble and decompose, providing enrichment to the soil. This enrichment provides nourishment for the next generation of trees and, of course, other forms of vegetation too.

It appears therefore that materials that emulate this natural cycle could be developed. Indeed, there are already a number of biodegradable materials on the market that decompose after serving their useful purpose. In addition this philosophy has been responsible for new engineering protocols concerning the environment. Thus parts are being recycled, others are being re-manufactured, and there is great concern for reducing the consumption of natural resources.

The continuous monitoring of the health of a part or product is one of the many new ideas set for development based on the re-evaluation of practices in the context of naturally occurring systems. The engineering approach



Figure 5. Health monitoring of *Homo sapiens sapiens*.

requires sensors to be embedded in a material, or attached to the surface of a part, in order to provide information on the behavior of the engineering artifact or its structural integrity. It is called “health monitoring” or the cradle-through-grave” approach. This approach of continually monitoring the status of a part during manufacturing, service, and failure mimics the spasmodic health-monitoring activities of *Homo sapiens sapiens*. Noninvasive techniques are often employed to monitor the health of individuals before appropriate medical action is prescribed. These techniques, illustrated in Fig. 5, often begin in the womb and continue throughout life in response to various stimuli and conditions.

The analogous engineering situations employ in situ sensors to monitor the behavior of the part during its manufacture, the service life, and at failure as shown in Fig. 6. Thus the sensing system would be employed initially to monitor the state of cure of a smart part during the fabrication process, in an autoclave perhaps. Subsequently, the same system would monitor continuously the health of the part during the service life. For example, dynamical stresses could be monitored and compared with limits imposed by the design specification. If limits are exceeded,

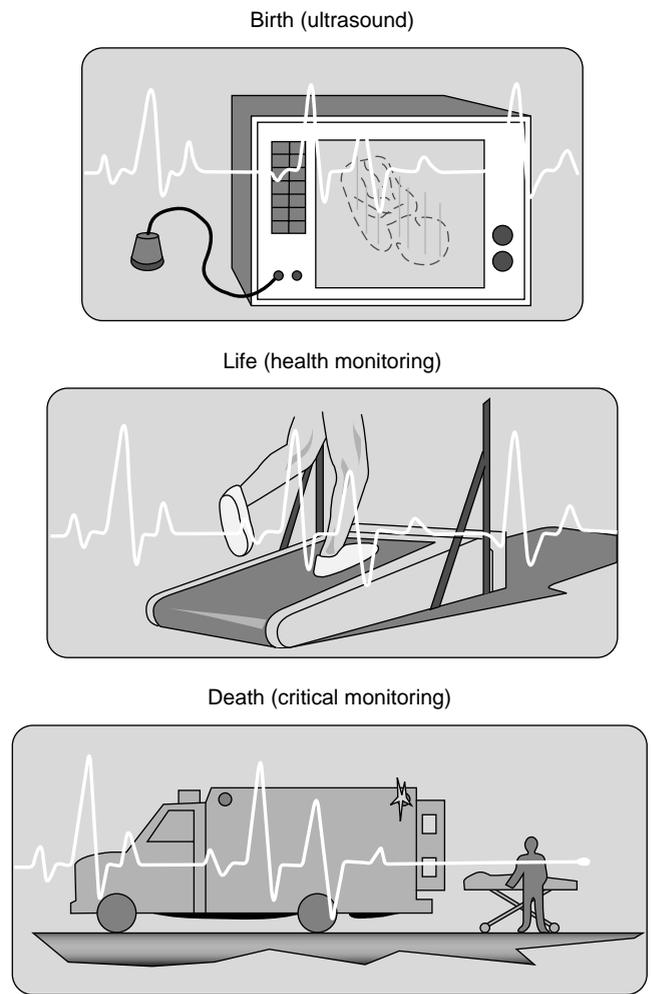


Figure 6. Health monitoring of smart structures.

then the operational envelope of the machine could be automatically changed to restore it to the desired domain.

Critical parts, subjected to complex fatigue environments, could be monitored for structural integrity and impending failure. These would include primary structures of aircraft and large civil engineering structures such as dams, pipelines, and highway bridges. The task of maintaining these structures by maintenance and inspection would be greatly enhanced by this health-monitoring capability.

Intelligence in Biological Materials

The intelligence to be imbued in a synthetic material developed by humankind should emulate the intelligent attributes found in biological systems. These attributes do not require human involvement, and they function autonomously, as evidenced by self-learning, self-degradation, and regeneration. Thus the rusting of iron in a humid environment could be considered to be a simple form of self-degradation. Other functions could include the availability to recognize and subsequently discriminate, redundancy, hierarchical control schemes, and the

election of an appropriate action based on sensory data. Furthermore, a material that has been damaged and is undergoing a process of self-repair would reduce its level of performance in order to survive.

Intelligence that should be inherent in future generations of smart materials can comprise several categories that are derived from studies of biological systems. These are highlighted, with some of the consequences, for the variety of terms listed below.

Crystal Structure. Changes occur in the orientation of crystals, in the atomic configuration, and in the interatomic spacings: These changes are responsible phase transformations, and in polymeric materials, the molecular chain can be re-configured from a folded state to an extended state.

Molecular Structure. Changes occur in the molecular structure caused by molecular chains breaks, re-configured intramolecular bonds, three-dimensional intramolecular spacings, and antigenic or enzymatic reactions.

Macroscopic Structure. Global changes occur in the macroscopic structure after the diffusion and bulk transfer of fluids and ultra-fine powder.

Composition. The interaction between a material and the neighboring environment at the interface results in a change in the material composition at the surface because of the attendant chemical reactions.

Interfaces. Interfacial changes typically pertain to grain boundary phenomena and also reactions at the surface of a material resulting from interactions with the adjacent environment.

Energy. The interaction between a material and the adjacent environment at the interface trigger an energy change or the release of electrons or photons.

Ion Transfer. Ions become transformed ion materials and in addition radicals are transferred along polymer chains.

Charge Transfer. Charge transfer occurs by conductivity in metals and charges are transferred along polymer chains in organic materials.

Electronic Structure. The magnetic properties of materials are changed by changes in the orientation of electronic spin vectors.

This is the main vocabulary being used as attention shifts to the attributes of biological materials and how they can relate to the development of intelligence in smart materials. The following paragraphs highlight some of the primary attributes of biological systems and their interpretation in the context of materials science. Some are illustrated in Fig. 7.

Mitosis. Self-multiplication, self-breeding, or growth in a biological system involves a cell creating similar cells by mitosis to replicate itself. To mimic these processes, smart materials will be self-producing with the additional implicit constraint of terminating the process when a prescribed state has been achieved.

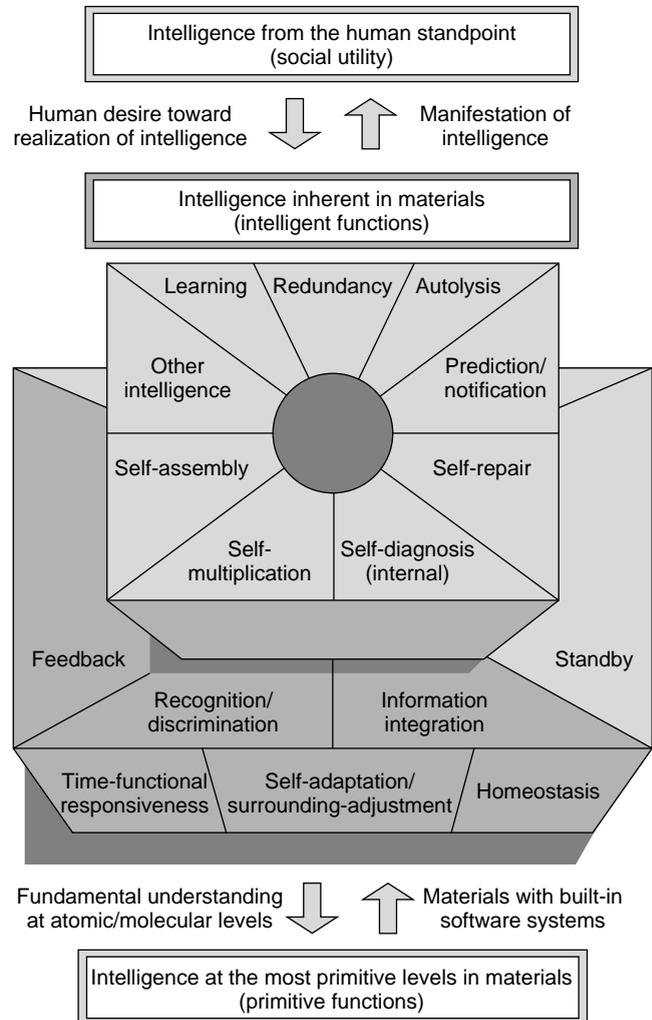


Figure 7. Intelligent functions inherent in biological materials.

This might require changes in their molecular or crystalline structure through the absorption of substances from the neighboring environment or other regions of the material.

Self-repair. This biological activity is closely related to the self-multiplication function or growth. It requires the identification of damage and the extent of the damage before the repair process is initiated and the material restored to its state of normality. The state is manifest in materials that undergo changes in crystalline structure and in the interfacial conditions at the surface of the material or at grain boundaries. Future classes of smart materials could autonomously regain their original shape through phase transformation, even after the material has suffered permanent deformation from surface impact.

Autolysis. Biological systems die and decompose when they sustain severe injuries or damage, or they cease to receive nutrition. Smart materials with these attributes would decompose upon completion of their

useful life and be assimilated within the environment. Smart materials with this attribute would be able to change their molecular structure or their macrostructure.

Redundancy. To enhance survivability, biological systems possess some degree of redundancy in their structures and functions in order to survive. Smart materials with this attribute would feature redundant functions that would be dormant under normal conditions. However, as the material structure changes in response to dynamic environmental conditions, the microstructure would be reconfigured to maintain equilibrium through the activation of new functions. These functions would trigger a change of the molecular structure of the crystalline structure. For example, a stress-induced transformation might be triggered at the tip of a crack in order to reduce the stress concentration in that region and hinder the crack propagation. Alternatively, parts that demand high reliability and are being subjected to very high loads could embody an innovative function that triggers a phase transformation in order to develop higher strength properties.

Learning. The ability to learn is fundamental to many aspects of the behavior of biological systems. Learning would be manifest in smart materials through changes in their physical constants, and changes in the molecular or crystalline structure. The associated knowledge-base is associated with some innate attributes; others are acquired through experience and through interactions with the local environment. These may involve inductive logic where general conclusions are distilled from the observations of an event. Smart materials would benefit from this ability and respond more appropriately to external stimuli through the recollection of previous experiences. Some of the underpinnings would include sensory abilities, data processing, control schemes, and actuators.

Autonomous Diagnosis. Some biological systems contain a self-diagnosis health-monitoring capability that permits the identification of problems, degradations, malfunctions, and judgmental errors. This is facilitated by a comparison between the current conditions and the past. Classes of smart materials would be able to monitor their own welfare so that the effects of damage on the performance of the material could be ascertained and corrective measures implemented.

This class of functions would be achieved by molecular structural changes, changes in the crystalline structure, or changes in the interfacial properties at the surface or the grain boundaries. Typically changes occur in materials that are traditionally used in nonequilibrium states, and their functional properties change when they reach a state of equilibrium. In addition, changes are found in materials that determine autonomously the appropriate time to quickly terminate their functional behavior in response to the ambient environmental conditions, and

in materials that detect degradation before they trigger a stress-induced transformation that reveals the damaged state, perhaps using energy.

Prediction. Some biological systems can predict the immediate future and take appropriate action by exploiting sensory data and by learning from their past experiences. Classes of smart materials designed to emulate these biological systems would need to use a combination of sensors, control algorithms, and knowledge. This is manifest in materials that undergo an energy change, a change in their molecular structure, or a change in the crystalline structure. Examples include the change in the crystalline structure associated with a phase change, the transfer of electrons precipitated by the re-configuring of the crystalline structure and stress induced transformations.

Standby. Smart materials should embody the state of readiness for action displayed by biological systems. This state would probably require several subfunctions such as sensing, diagnosis, prediction, learning, and some class of actuators. This responsiveness could be governed by the role of the material and the ability to analyze the dynamically changing external environmental conditions. It would be manifest by changes in energy, molecular structure, and crystalline structure.

Information Integration. Biological systems generally evolve by storing the integrated experiences of previous generations extending over a great period of time before it is transferred to subsequent generations. Thus the emulation of genes or DNA would provide the basis for the creation of materials with innovative functions that integrate and maintain memory banks.

Recognition. Organisms frequently possess the ability to recognize and discriminate when evaluating information. Smart materials with this ability would not only embody analytical skills but also some measure of fuzzy logic to interpret the data from sensing functions.

Homeostasis. Biological systems are generally subjected to dynamically changing environmental conditions that cause internal systems and structures to change continually. Such systems survive and maintain stable physiological states by coordinated responses that autonomously compensate for these ever-changing external conditions. Innovative functions would need to be established to accomplish these tasks. These functions would require closed-loop feedback systems where sensory information from both the input and the output states need to be compared. Furthermore, the desired conditions would require the careful orchestration of sensors, processors and actuators.

Feedback is manifest in a variety of forms including the transfer of charges, the change in composition, the change in molecular structure, the transfer of radicals and ions, and the change in the crystalline structure. Thus, for example, materials

with a transformation temperature would undergo changes when external conditions subject the material to an elevated temperature.

Adaptation. Biological systems adapt to ever-changing environmental conditions through the evolution of their physiological state. These abilities would have great utility in engineering practice. Examples of this class of materials include films of lubricants that assume solid or fluidic states depending on the local thermal or dynamic conditions. Others include materials that control their optical properties in response to changes in external stimuli like magnetic fields, electrical fields, or heat.

There are probably other biological attributes that are worthy of emulation in the design and manufacture of smart materials that mimic naturally occurring materials, but this list provides a beginning. The road ahead will have many positive and negative undulations and many winding turns. The primary provinces will include theoretical studies, experimental studies, and computational studies, followed by design and manufacture, the creation of primitive functions, assembly and integration of these functions, material characterization, and the development of the supporting technologies.

CONCLUDING COMMENTS

With the passing of time, the technologies associated with materials are becoming ever more sophisticated. It is unreasonable to contemplate the receding of this tide of knowledge and complexity because it is motivated by both commercial and military demands for superior materials. Composite materials, based on ideas emulating from naturally occurring materials, could be considered the most sophisticated class of materials created by humankind. If this is true, then it is only natural to invoke induction and suggest that biological materials are the basis from which new generations of materials will be developed.

COMPOSITES, INTRINSICALLY SMART STRUCTURES

D.D.L. CHUNG
State University of New York at Buffalo
Buffalo, NY

INTRODUCTION

Smart structures are important because of their use in hazard mitigation, structural vibration control, structural health monitoring, transportation engineering, and thermal control. Research on smart structures has emphasized incorporating of various devices in a structure to provide sensing, energy dissipation, actuation, control, or other functions. Research on smart composites has emphasized incorporating of a smart material in a matrix to enhance smartness or durability. Research on smart materials has

emphasized the study of materials (e.g., piezoelectric materials) used for making the devices. However, relatively little attention has been given to the development of structural materials (e.g., concrete and composites) that can inherently provide some of the smart functions, so that the need for embedded or attached devices is reduced or eliminated, thereby lowering cost, enhancing durability, increasing the smart volume, and minimizing mechanical property degradation (which is usually caused by embedded devices).

Smart structures are structures that can sense certain stimuli and respond to the stimuli appropriately, somewhat like a human being. Sensing is the most fundamental aspect of a smart structure. A structural composite, which is itself a sensor, is multifunctional.

This article focuses on structural composites for smart structures. It addresses cement-matrix and polymer-matrix composites. The smart functions addressed include strain sensing (for structural vibration control and traffic monitoring), damage sensing (both mechanical and thermal damage related to structural health monitoring), temperature sensing (for thermal control, hazard mitigation, and structural performance control), thermoelectricity (for thermal control and saving energy), and vibration reduction (for structural vibration control). These functional abilities of the structural composites have been shown in the laboratory. Applications in the field are forthcoming.

CEMENT-MATRIX COMPOSITES FOR SMART STRUCTURES

Cement-matrix composites include concrete (containing coarse and fine aggregates), mortar (containing fine aggregate but no coarse aggregate), and cement paste (containing no aggregate, whether coarse or fine). Other fillers, called admixtures, can be added to the mix to improve the properties of the composite. Admixtures are discontinuous, so that they can be included in the mix. They can be particles such as silica fume (a fine particulate) and latex (a polymer in the form of a dispersion). They can be short fibers such as polymer, steel, glass, or carbon fibers. They can be liquids such as aqueous methylcellulose solutions, water reducing agents, and defoamers. Admixtures to render the composite smart while maintaining or even improving the structural properties are the focus of this section.

Background on Cement-Matrix Composites

Cement-matrix composites for smart structures include those that contain short carbon fibers (for sensing strain, damage, and temperature and for thermal control), short steel fibers (for sensing temperature and for thermal control), and silica fume (for vibration reduction). This section provides background on cement-matrix composites and emphasizes the carbon-fiber cement-matrix composite due to its dominance among intrinsically smart cement-matrix composites.

Carbon-fiber cement-matrix composites are structural materials that are gaining in importance quite rapidly due

to the decrease in carbon-fiber cost (1) and the increasing demand for superior structural and functional properties. These composites contain short carbon fibers, typically 5 mm long; the short fibers can be used as an admixture in concrete (whereas continuous fibers cannot be simply added to the concrete mix), and short fibers are less expensive than continuous fibers. However, due to the weak bond between carbon fiber and the cement matrix, continuous fibers (2–4) are much more effective than short fibers in reinforcing concrete. Surface treatment of carbon fiber [e.g., by heating (5) or by using ozone (6,7), silane (8), SiO₂ particles (9), or hot NaOH solution (10)] is useful for improving the bond between fiber and matrix, thereby improving the properties of the composite. Surface treatment by ozone or silane improves the bond due to the enhanced wettability by water. Admixtures such as latex (6,11), methylcellulose (6), and silica fume (12) also improve the bond.

The effect of carbon fiber addition on the properties of concrete increases with fiber volume fraction (13), unless the fiber volume fraction is so high that the air void content becomes excessively high (14). (The air void content increases with fiber content and air voids tend to have a negative effect on many properties, such as compressive strength.) In addition, the workability of the mix decreases with fiber content (13). Moreover, the cost increases with fiber content. Therefore, a rather low volume fraction of fibers is desirable. A fiber content as low as 0.2 vol.% is effective (15), although fiber contents exceeding 1 vol.% are more common (16–20). The required fiber content increases with the particle size of the aggregate, because flexural strength decreases as particle size increases (21).

The effective use of carbon fibers in concrete requires dispersing of the fibers in the mix. Dispersion is enhanced by using silica fume (a fine particulate) as an admixture (14,22–24). Typical silica fume content is 15% by weight of cement (14). The silica fume is usually used along with a small amount (0.4% by weight of cement) of methylcellulose to help the dispersion of the fibers and the workability of the mix (14). Latex (typically 15–20% by weight of cement) is much less effective than silica fume in helping fiber dispersion, but it enhances the workability, flexural strength, flexural toughness, impact resistance, frost resistance, and acid resistance (14,25,26). The ease of dispersion increases with decreasing fiber length (24).

The structural properties improved by carbon fiber addition are increased tensile and flexible strengths, increased tensile ductility and flexural toughness, enhanced impact resistance, reduced drying shrinkage, and improved freeze–thaw durability (13–15,17–25,27–38). Tensile and flexural strengths decrease as specimen size increases, so that the size effect becomes larger as the fiber length increases (39). Low drying shrinkage is valuable for large structures, for repair (40,41), and in joining bricks in a brick structure (42,43). The functional properties created by carbon fiber addition are strain sensing (7,44–58) (for smart structures), temperature sensing (59–62), damage sensing (44,48,63–65), thermoelectric behavior (60–62), thermal insulation (66–68) (to save energy for buildings), electrical conduction (69–78) (to facilitate cathodic protection of embedded steel and to provide electrical grounding or connection), and radio wave reflection/absorption

(79–84) (for electromagnetic interference or EMI shielding, for lateral guidance in automatic highways, and for television image transmission).

In relation to structural properties, carbon fibers compete with glass, polymer, and steel fibers (18,27–29,32,36–38,85). Carbon fibers (isotropic pitch-based) (1,85) are advantageous in their superior ability to increase the tensile strength of concrete, even though the tensile strength, modulus, and ductility of isotropic pitch-based carbon fibers are low compared to most other fibers. Carbon fibers are also advantageous in relative chemical inertness (86). PAN-based carbon fibers are also used (17,19,22,33), although they are more commonly used as continuous fibers than short fibers. Carbon-coated glass fibers (87,88) and submicron diameter carbon filaments (77–79) are even less commonly used, although the former are attractive for the low cost of glass fibers and the latter are attractive for their high radio wave reflectivity (which results from the skin effect). C-shaped carbon fibers are more effective for strengthening than round carbon fibers (89), but their relatively large diameter makes them less attractive. Carbon fibers can be used in concrete together with steel fibers; the addition of short carbon fibers to steel-fiber-reinforced mortar increases the fracture toughness of the interfacial zone between the steel fiber and the cement matrix (90). Carbon fibers can also be used in concrete together with steel bars (91,92) or together with carbon-fiber-reinforced polymer rods (93).

Carbon fibers are exceptional in most functional properties, compared to the other fiber types. Carbon fibers are electrically conducting, in contrast to glass and polymer fibers, which are not conducting. Steel fibers are conducting, but their typical diameter ($\geq 60 \mu\text{m}$) is much larger than the diameter of a typical carbon fiber ($15 \mu\text{m}$). The combination of electrical conductivity and small diameter makes carbon fibers superior to the other fiber types in the strain sensing and electrical conduction. However, carbon fibers are inferior to steel fibers for thermoelectric composites, due to the high electron concentration in steel and the low hole concentration in carbon.

Although carbon fibers are thermally conducting, the addition of carbon fibers to concrete lowers the thermal conductivity (66), thus allowing applications for thermal insulation. This effect of carbon fiber addition is due to the increase in air void content. The electrical conductivity of carbon fibers is higher than that of the cement matrix by about eight orders of magnitude, whereas the thermal conductivity of carbon fibers is higher than that of the cement matrix by only one or two orders of magnitude. As a result, electrical conductivity is increased upon carbon fiber addition despite the increase in air void content, but thermal conductivity is decreased upon fiber addition.

The use of pressure after casting (94) and extrusion (95, 96) can result in composites that have superior microstructure and properties. Moreover, extrusion improves the shapability (95).

Cement-Matrix Composites for Strain Sensing

The electrical resistance of strain-sensing concrete (without embedded or attached sensors) changes reversibly with

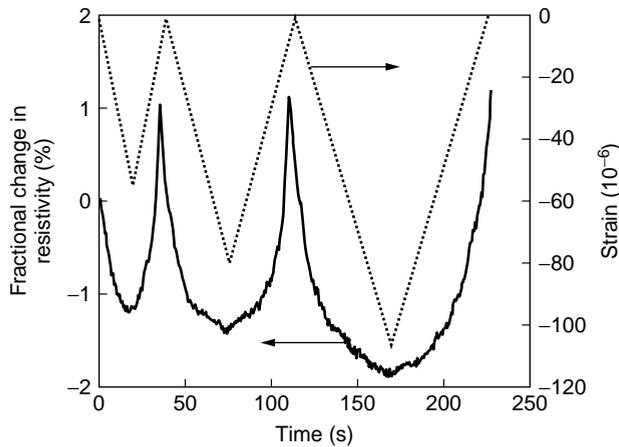


Figure 1. Variation of the fractional change in volume electrical resistivity and of the strain (negative for compressive strain) with time during dynamic compressive loading at increasing stress amplitudes within the elastic regime for a carbon-fiber latex cement paste after 28 days of curing.

strain, such that the gauge factor (fractional change in resistance per unit strain) is up to 700 under compression or tension (7,44–58). The resistance (dc/ac) increases reversibly under tension and decreases reversibly upon compression due to fiber pull-out upon microcrack opening ($<1\ \mu\text{m}$) and the consequent increase in fiber–matrix contact resistivity. The concrete contains as little as 0.2 vol.% short carbon fibers, which are preferably those that have been surface-treated. The fibers do not need to touch one another in the composite. The treatment improves wettability with water. The presence of a large aggregate decreases the gauge factor, but the strain-sensing ability remains sufficient for practical use. Strain-sensing concrete works even when data acquisition is wireless. The applications include structural vibration control and traffic monitoring.

Figure 1 shows the fractional change in resistivity along the stress axis, as well as the strain during repeated compressive loading at an increasing stress amplitude for carbon-fiber latex cement paste after 28 days of curing. The strain varies linearly with the stress up to the highest stress amplitude. The strain returns to zero at the end of each loading cycle. The resistivity decreases upon loading in every cycle (due to fiber push-in) and increases upon unloading in every cycle (due to fiber pull-out). The resistivity has a net increase after the first cycle, due to damage. Little further damage occurs in subsequent cycles; the resistivity after unloading does not increase much after the first cycle. The greater the strain amplitude, the more the resistivity decreases during loading, although the resistivity and strain are not linearly related. The effects in Fig. 1 were also observed in carbon-fiber silica-fume cement paste after 28 days of curing.

Figures 2 and 3 show the fractional changes in the longitudinal and transverse resistivities, respectively, for carbon-fiber silica-fume cement paste after 28 days of curing during repeated uniaxial tensile loading at increasing strain amplitudes. The strain essentially returns to zero at

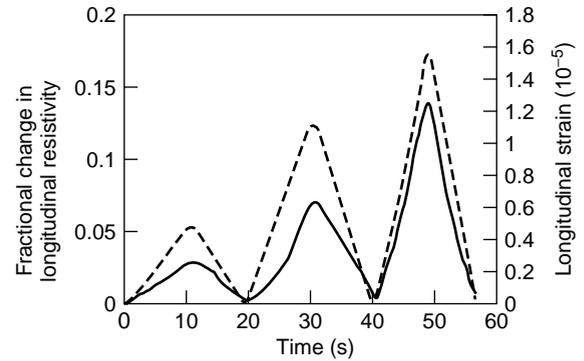


Figure 2. Variation of the fractional change in longitudinal electrical resistivity (solid curve) and of strain with time (dashed curve) during dynamic uniaxial tensile loading at increasing stress amplitudes within the elastic regime for a carbon-fiber silica-fume cement paste.

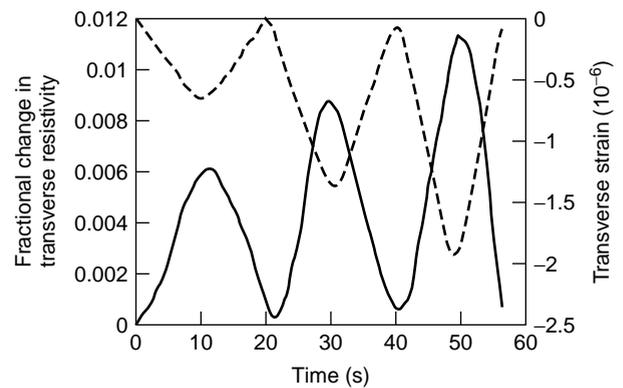


Figure 3. Variation of the fractional change in transverse electrical resistivity (solid curve) and of strain with time (dashed curve) during dynamic uniaxial tensile loading at increasing stress amplitudes within the elastic regime for a carbon-fiber silica-fume cement paste.

the end of each cycle, indicating elastic deformation. The longitudinal strain is positive (i.e., elongation); the transverse strain is negative (i.e., shrinkage due to the Poisson effect). Both longitudinal and transverse resistivities increase reversibly under uniaxial tension. The reversibility of both strain and resistivity is more complete in the longitudinal direction than in the transverse direction. The gauge factor is 89 and -59 for the longitudinal and transverse resistances, respectively.

Figures 4 and 5 show corresponding results for silica-fume cement paste. The strain is essentially totally reversible in both the longitudinal and transverse directions, but the resistivity is only partly reversible in both directions, in contrast to the reversibility of the resistivity when fibers are present (Figs. 2 and 3). As with fibers, both longitudinal and transverse resistivities increase under uniaxial tension. However, the gauge factor is only 7.2 and -7.1 for Figs. 4 and 5, respectively.

Comparison of Figs. 2 and 3 (with fibers) with Figs. 4 and 5 (without fibers) shows that fibers greatly enhance the

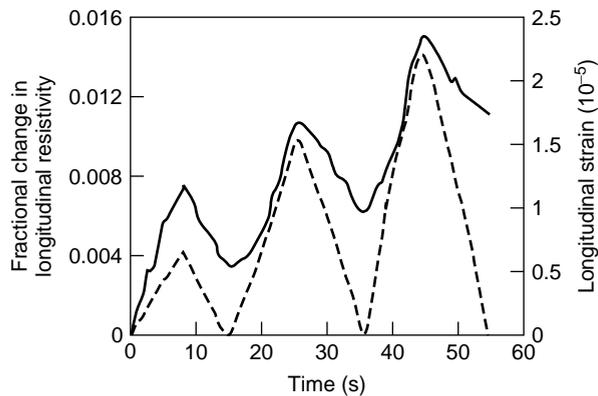


Figure 4. Variation of the fractional change in longitudinal electrical resistivity (solid curve) and of strain with time (dashed curve) during dynamic uniaxial tensile loading at increasing stress amplitudes within the elastic regime for a silica-fume cement paste.

magnitude and reversibility of the resistivity effect. The gauge factors are much smaller when fibers are absent.

The increase in both longitudinal and transverse resistivities under uniaxial tension for cement pastes, whether with or without fibers, is attributed to defect (e.g., microcrack) generation. In the presence of fibers, fiber bridging across microcracks occurs, and slight fiber pull-out occurs under tension, thus enhancing the possibility of microcrack closing and causing more reversibility in the resistivity change. The fibers are much more electrically conductive than the cement matrix. The presence of the fibers introduces interfaces between fibers and matrix. The degradation of the fiber-matrix interface due to fiber pull-out or other mechanisms is an additional type of defect generation which will increase the resistivity of the composite. Therefore, the presence of fibers greatly increases the gauge factor.

The transverse resistivity increases under uniaxial tension, even though the Poisson effect makes the transverse strain negative. This means that the effect of transverse

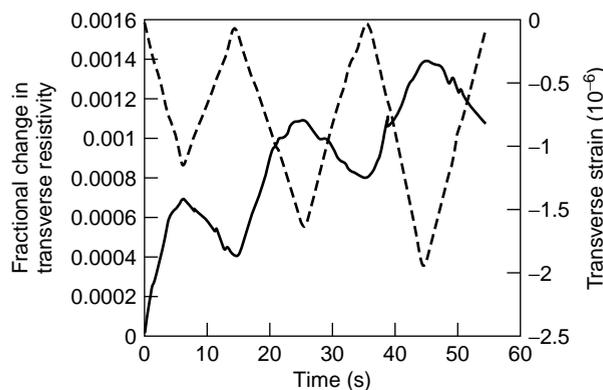


Figure 5. Variation of the fractional change in transverse electrical resistivity (solid curve) and of strain with time (dashed curve) during dynamic uniaxial tensile loading at increasing stress amplitudes within the elastic regime for a silica-fume cement paste.

resistivity increase overshadows the effect of transverse shrinkage. The increase in resistivity is a consequence of uniaxial tension. In contrast, under uniaxial compression, the resistance in the stress direction decreases after 28 days of curing. Hence, the effects of uniaxial tension on transverse resistivity and of uniaxial compression on longitudinal resistivity are different; the gauge factors are negative and positive for these cases, respectively.

The similarity of the resistivity change in the longitudinal and transverse directions under uniaxial tension suggests similarity for other directions as well. This means that the resistance can be measured in any direction to sense the occurrence of tensile loading. Although the gauge factor is comparable in both longitudinal and transverse directions, the fractional change in resistance under uniaxial tension is much higher in the longitudinal direction than in the transverse direction. Thus, the use of the longitudinal resistance for practical self-sensing is preferred.

Cement-Matrix Composites for Damage Sensing

Concrete, with or without admixtures, can sense major and minor damage—even damage during elastic deformation—due to the increase in electrical resistivity that accompanies damage (44,48,63–65). That both strain and damage can be sensed simultaneously through resistance measurement means that the strain/stress condition (during dynamic loading) under which damage occurs can be obtained, thus facilitating damage origin identification. Damage is indicated by an increase in resistance, which is larger and less reversible when the stress amplitude is higher. The increase in resistance can be sudden during loading. It can also be a gradual shift of the baseline resistance.

Figure 6 (64) shows the fractional change in resistivity along the stress axis, as well as the strain during repeated compressive loading at increasing stress amplitude for silica-fume cement paste after 28 days of curing. The strain varies linearly with stress up to the highest stress amplitude. The strain returns to zero at the end of each loading

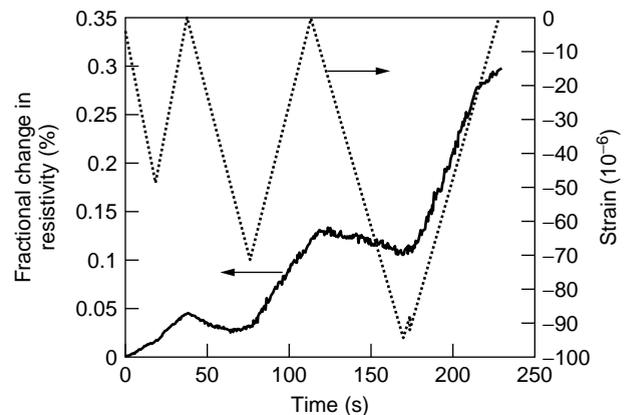


Figure 6. Variation of the fractional change in electrical resistivity and of strain (negative for compressive strain) with time during dynamic compressive loading at increasing stress amplitudes within the elastic regime for a silica-fume cement paste after 28 days of curing.

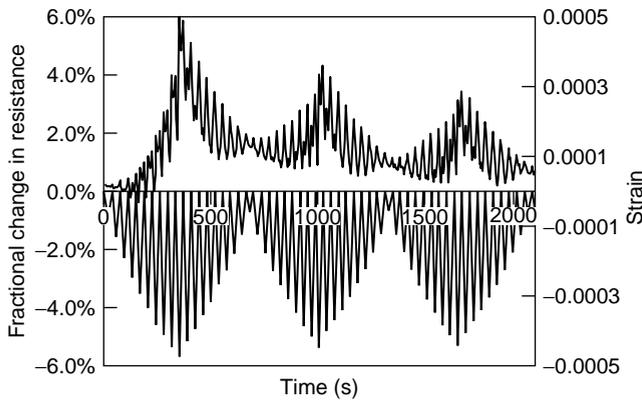


Figure 7. Fractional change in resistance and strain during repeated compressive loading at increasing and decreasing stress amplitudes, the highest of which was 60% of the compressive strength, for carbon-fiber concrete after 28 days of curing.

cycle. During the first loading, the resistivity increases due to damage initiation. During the subsequent unloading, the resistivity continues to increase, probably due to opening of microcracks generated during loading. During the second loading, the resistivity decreases slightly as the stress increases up to the maximum stress of the first cycle (probably due to the closing of microcracks) and then increases as the stress increases beyond this value (probably due to the generation of additional microcracks). During unloading in the second cycle, the resistivity increases significantly (probably due to the opening of microcracks). During the third loading, the resistivity essentially does not change (or decreases very slightly) as the stress increases to the maximum stress of the third cycle (probably due to the balance between microcrack generation and microcrack closing). Subsequent unloading causes the resistivity to increase very significantly (probably due to the opening of microcracks).

Figure 7 shows the fractional change in resistance and strain during repeated compressive loading at increasing and decreasing stress amplitudes for carbon-fiber (0.18 vol.%) concrete (with fine and coarse aggregates) after 28 days of curing. The highest stress amplitude is 60% of the compressive strength. A group of cycles in which the stress amplitude increases cycle by cycle and then decreases cycle by cycle back to the initial low stress amplitude is here referred to as a group. Figure 7 shows the results for three groups. The strain returns to zero at the end of each cycle for any of the stress amplitudes, indicating elastic behavior. The resistance decreases upon loading in each cycle, as in Fig. 1. An extra peak at the maximum stress of a cycle grows as the stress amplitude increases, resulting in two peaks per cycle. The original peak (strain induced) occurs at zero stress, and the extra peak (damage induced) occurs at maximum stress. Hence, during loading from zero stress within a cycle, the resistance drops, then increases sharply, and reaches the maximum resistance of the extra peak at the maximum stress of the cycle. Upon subsequent unloading, the resistance decreases, then increases as unloading continues, and reaches the maximum resistance of the original peak at zero stress. In the part of

this group where the stress amplitude decreases cycle by cycle, the extra peak diminishes and disappears, leaving the original peak as the sole peak. In the part of the second group where the stress amplitude increases cycle by cycle, the original peak (peak at zero stress) is the sole peak, except that the extra peak (peak at the maximum stress) returns in a minor way (more minor than in the first group) as the stress amplitude increases. The extra peak grows as the stress amplitude increases, but, in the part of the second group in which the stress amplitude decreases cycle by cycle, it quickly diminishes and vanishes, as in the first group. Within each group, the amplitude of the variation in resistance increases as the stress amplitude increases and decreases as the stress amplitude subsequently decreases.

The greater the stress amplitude, the larger and the less reversible the damage-induced increase in resistance (the extra peak). If the stress amplitude has been experienced before, the damage-induced increase in resistance (the extra peak) is small, as shown by comparing the result of the second group with that of the first group (Fig. 7), unless the extent of damage is large (Fig. 8 for the highest stress amplitude of >90% of the compressive strength). When the damage is extensive (as shown by a decrease in modulus), a damage-induced increase in resistance occurs in every cycle, even at a decreasing stress amplitude, and it can overshadow the strain-induced decrease in resistance (Fig. 8). Hence, the damage-induced increase in resistance occurs mainly during loading (even within the elastic regime), particularly at a stress above that in prior cycles, unless the stress amplitude is high and/or damage is extensive.

At a high stress amplitude, the damage-induced increase in resistance, cycle by cycle, as the stress amplitude increases, causes the baseline resistance to increase irreversibly (Fig. 8). The baseline resistance in the regime of major damage (that decreases in modulus) provides a measure of the extent of damage (i.e., condition monitoring). This measure works in the loaded or unloaded state. In

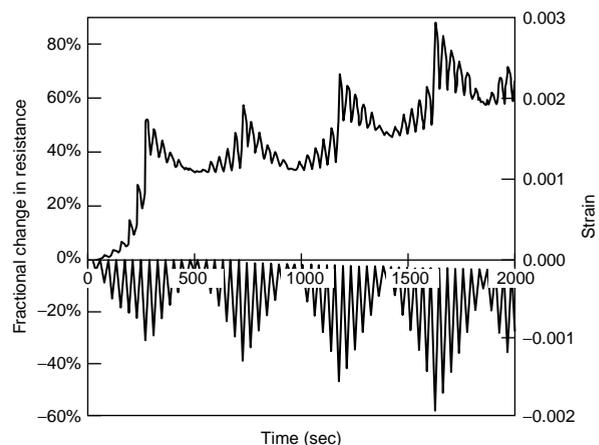


Figure 8. Fractional change in resistance and strain during repeated compressive loading at increasing and decreasing stress amplitudes, the highest of which was >90% of the compressive strength, for a carbon-fiber concrete after 28 days of curing.

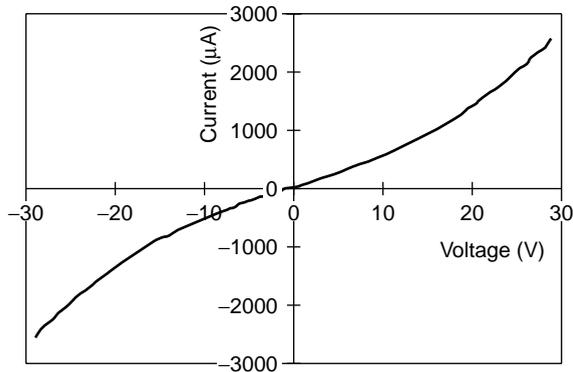


Figure 9. Current-voltage characteristic of a carbon-fiber silica-fume cement paste at 38°C during stepped heating.

contrast, the measure using the damage-induced increase in resistance (Fig. 7) works only during an increase in stress and indicates the occurrence of damage (whether minor or major), as well as the extent of damage.

Cement-Matrix Composites for Temperature Sensing

A thermistor is a thermometric device that consists of a material (typically a semiconductor, but in this case a cement paste) whose electrical resistivity decreases with a rise in temperature. The carbon-fiber concrete described previously for strain sensing is a thermistor because its resistivity decreases reversibly as temperature increases (59); the sensitivity is comparable to that of semiconductor thermistors. (The effect of temperature will need to be compensated for in using the concrete as a strain sensor.)

Figure 9 (59) shows the current-voltage characteristic of carbon-fiber (0.5% by weight of cement) silica-fume (15% by weight of cement) cement paste at 38°C during stepped heating. The characteristic is linear below 5 V and deviates positively from linearity beyond 5 V. The resistivity is obtained from the slope of the linear portion. The voltage at which the characteristic starts to deviate from linearity is called the critical voltage.

Figure 10 shows a plot of resistivity versus temperature for carbon-fiber silica-fume cement paste during heating and cooling. The resistivity decreases upon heating, and the effect is quite reversible upon cooling. That the resistivity increases slightly after a heating/cooling cycle is probably due to thermal degradation of the material. Figure 11 shows the Arrhenius plot of log conductivity (conductivity = 1/resistivity) versus the reciprocal of the absolute temperature. The slope of the plot gives the activation energy, which is 0.390 ± 0.014 and 0.412 ± 0.017 eV during heating and cooling, respectively.

Results similar to those of carbon-fiber silica-fume cement paste were obtained with carbon-fiber (0.5% by weight of cement) latex (20% by weight of cement) cement paste, silica-fume cement paste, latex cement paste, and plain cement paste. However, for all of these four types of cement paste, (1) the resistivity is higher by about an order of magnitude, and (2) the activation energy is lower

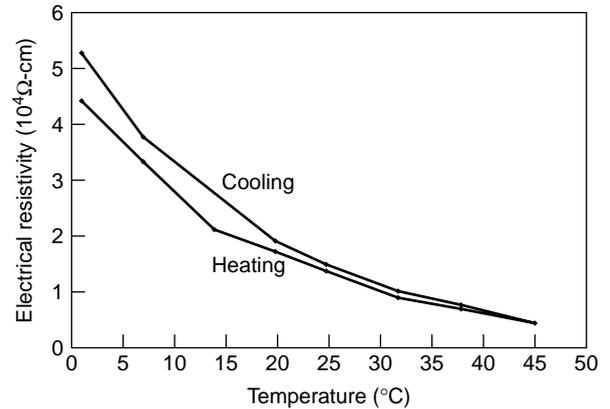


Figure 10. Plot of volume electrical resistivity vs. temperature during heating and cooling for a carbon-fiber silica-fume cement paste.

by about an order of magnitude, as shown in Table 1. The critical voltage is higher when fibers are absent (Table 1).

The Seebeck (60–62,97) effect is a thermoelectric effect which is the basis for using thermocouples for temperature measurement. This effect involves charge carriers that move from a hot point to a cold point within a material, thereby resulting in a voltage difference between the two points. The Seebeck coefficient is the negative of the voltage difference (hot minus cold) per unit temperature difference between the two points (hot minus cold). Negative carriers (electrons) make it more negative, and positive carriers (holes) make it more positive.

The Seebeck effect in carbon-fiber-reinforced cement paste involves electrons from the cement matrix (62) and holes from the fibers (60,61), such that the two contributions are equal at the percolation threshold, a fiber content from 0.5–1.0% by weight of cement (62). The hole contribution increases monotonically as fiber content changes above and below the percolation threshold (62).

Due to the free electrons in a metal, cement that contains metal fibers such as steel fibers is even more negative in thermoelectric power than cement without fiber (97).

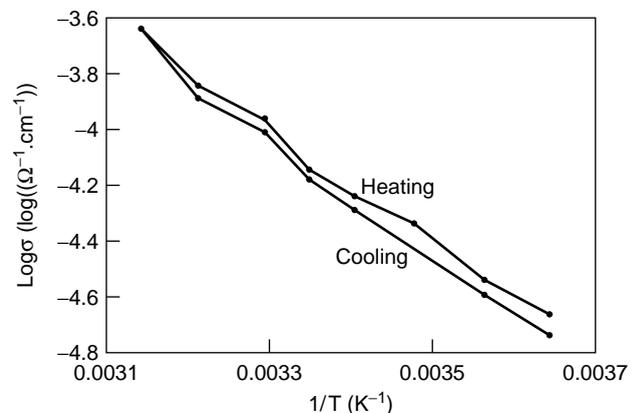


Figure 11. Arrhenius plot of log electrical conductivity vs. reciprocal absolute temperature for a carbon-fiber silica-fume cement paste.

Table 1. Resistivity, Critical Voltage, and Activation Energy of Five Types of Cement Pastes

Formulation	Resistivity at 20°C (Ω cm)	Critical Voltage at 20°C (V)	Activation Energy (eV)	
			Heating	Cooling
Plain	$(4.87 \pm 0.37) \times 10^5$	10.80 ± 0.45	0.040 ± 0.006	0.122 ± 0.006
Silica fume	$(6.12 \pm 0.15) \times 10^5$	11.60 ± 0.37	0.035 ± 0.003	0.084 ± 0.004
Carbon fibers + silica fume	$(1.73 \pm 0.08) \times 10^4$	8.15 ± 0.34	0.390 ± 0.014	0.412 ± 0.017
Latex	$(6.99 \pm 0.12) \times 10^5$	11.80 ± 0.31	0.017 ± 0.001	0.025 ± 0.002
Carbon fibers + latex	$(9.64 \pm 0.08) \times 10^4$	8.76 ± 0.35	0.018 ± 0.001	0.027 ± 0.002

The attainment of a very negative thermoelectric power is attractive because a material whose thermoelectric power is positive and a material whose thermoelectric power is negative are two very dissimilar materials; their junction is a thermocouple junction. The greater the dissimilarity, the more sensitive the thermocouple.

Table 2 and Fig. 12 show the thermoelectric power results. The absolute thermoelectric power is much more negative for all of the steel-fiber cement pastes compared to all of the carbon-fiber cement pastes. An increase in the steel fiber content from 0.5% to 1.0% by weight of cement makes the absolute thermoelectric power more negative, whether or not silica fume (or latex) is present. An increase in the steel fiber content also increases the reversibility and linearity of the change in Seebeck voltage with the temperature difference between the hot and cold ends, as shown by comparing the values of the Seebeck coefficient obtained during heating and cooling in Table 2. The values obtained during heating and cooling are close for the pastes that have the higher steel-fiber content but are not so close for the pastes that have the lower steel-fiber content. In contrast, for pastes that have carbon fibers in place of steel fibers, the change in Seebeck voltage with temperature difference is highly reversible for both carbon-fiber contents of 0.5 and 1.0% by weight of cement, as shown in Table 2 by comparing the values of the Seebeck coefficient obtained during heating and cooling.

Table 2 shows that the volume electrical resistivity is much higher for the steel-fiber cement pastes than the corresponding carbon-fiber cement pastes. This is attributed to the much lower volume fraction of fibers in the former (Table 2). An increase in the steel-or carbon-fiber content from 0.5 to 1.0% by weight of cement decreases the resistivity, though the decrease is more significant for the carbon-fiber cement than for the steel-fiber cement. That the resistivity decrease is not large when the steel fiber content is increased from 0.5 to 1.0% by weight of cement and that the resistivity is still high at a steel-fiber content of 1.0% by weight of cement suggest that a steel-fiber content of 1.0% by weight of cement is below the percolation threshold.

With or without silica fume (or latex), the change of the Seebeck voltage with temperature is more reversible and linear at a steel-fiber content of 1.0% by weight of cement than at a steel fiber content of 0.5% by weight of cement. This is attributed to the larger role of the cement matrix at the lower steel-fiber content and the contribution of the cement matrix to the irreversibility and nonlinearity. Irreversibility and nonlinearity are particularly significant when the cement paste contains no fiber.

From the practical point of view, the steel-fiber silica-fume cement paste that contains 1.0% steel fibers by weight of cement is particularly attractive for use in temperature sensing because the absolute thermoelectric

Table 2. Volume Electrical Resistivity, Seebeck Coefficient (μ V/°C) with Copper as the Reference, and the Absolute Thermoelectric Power (μ V/°C) of Various Cement Pastes with Steel Fibers (S_f) or Carbon Fibers (C_f)

Cement Paste	Volume Fraction Fibers(%)	Resistivity (Ω cm)	Heating		Cooling	
			Seebeck Coefficient	Absolute Thermoelectric Power	Seebeck Coefficient	Absolute Thermoelectric Power
S_f (0.5 ^a)	0.10	$(7.8 \pm 0.5) \times 10^4$	-51.0 ± 4.8	-53.3 ± 4.8	-45.3 ± 4.4	-47.6 ± 4.4
S_f (1.0 ^a)	0.20	$(4.8 \pm 0.4) \times 10^4$	-56.8 ± 5.2	-59.1 ± 5.2	-53.7 ± 4.9	-56.0 ± 4.9
S_f (0.5 ^a) + SF ^b	0.10	$(5.6 \pm 0.5) \times 10^4$	-54.8 ± 3.9	-57.1 ± 3.9	-52.9 ± 4.1	-55.2 ± 4.1
S_f (1.0 ^a) + SF ^b	0.20	$(3.2 \pm 0.3) \times 10^4$	-66.2 ± 4.5	-68.5 ± 4.5	-65.6 ± 4.4	-67.9 ± 4.4
S_f (0.5 ^a) + L ^c	0.085	$(1.4 \pm 0.1) \times 10^5$	-48.1 ± 3.2	-50.4 ± 3.2	-45.4 ± 2.9	-47.7 ± 2.9
S_f (1.0 ^a) + L ^c	0.17	$(1.1 \pm 0.1) \times 10^5$	-55.4 ± 5.0	-57.7 ± 5.0	-54.2 ± 4.5	-56.5 ± 4.5
C_f (0.5 ^a) + SF ^b	0.48	$(1.5 \pm 0.1) \times 10^4$	$+1.45 \pm 0.09$	-0.89 ± 0.09	$+1.45 \pm 0.09$	-0.89 ± 0.09
C_f (1.0 ^a) + SF ^b	0.95	$(8.3 \pm 0.5) \times 10^2$	$+2.82 \pm 0.11$	$+0.48 \pm 0.11$	$+2.82 \pm 0.11$	$+0.48 \pm 0.11$
C_f (0.5 ^a) + L ^c	0.41	$(9.7 \pm 0.6) \times 10^4$	$+1.20 \pm 0.05$	-1.14 ± 0.05	$+1.20 \pm 0.05$	-1.14 ± 0.05
C_f (1.0 ^a) + L ^c	0.82	$(1.8 \pm 0.2) \times 10^3$	$+2.10 \pm 0.08$	-0.24 ± 0.08	$+2.10 \pm 0.08$	-0.24 ± 0.08

^a% by weight of cement.

^bSF: silica fume.

^cL: latex.

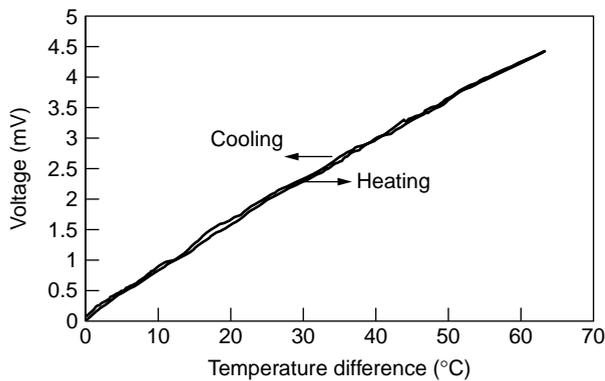


Figure 12. Variation of the Seebeck voltage (with copper as the reference) vs. the temperature difference during heating and cooling for a steel-fiber silica-fume cement paste containing 1.0% steel fibers by weight of cement.

power is the highest ($-68 \mu\text{V}/^\circ\text{C}$) and the variation of the Seebeck voltage with the temperature difference between the hot and cold ends is reversible and linear. The absolute thermoelectric power is as high as that of commercial thermocouple materials.

Joints between concretes that have different values of thermoelectric power, made by multiple pouring, provide concrete thermocouples (98).

Cement-Matrix Composites for Thermal Control

Concretes that can inherently provide heating through Joule heating, provide temperature sensing, or provide temperature stability through a high specific heat (high thermal mass) are highly desirable for thermal control of structures and energy saving in buildings. Concretes of low electrical resistivity (69–78) are useful for Joule heating, concrete thermistors and thermocouples are useful for temperature sensing, and concretes of high specific heat (66–68,99) are useful for heat retention. These concretes involve the use of admixtures such as fibers and silica fume. For example, silica fume introduces interfaces that promote the specific heat (66); short carbon fibers enhance electrical conductivity (74) and produce p-type concrete (62). Plain concrete is n-type (62).

Figure 13 (74) gives the volume electrical resistivity of composites after 7 days of curing. The resistivity decreases a lot as the fiber volume fraction increases, whether or not a second filler (silica fume or sand) is present. When sand is absent, the addition of silica fume decreases the resistivity at all carbon-fiber volume fractions except the highest volume fraction of 4.24%; the decrease is most significant at the lowest fiber volume fraction of 0.53%. When sand is present, the addition of silica fume similarly decreases the resistivity; the decrease is most significant at fiber volume fractions below 1%. When silica fume is absent, the addition of sand decreases the resistivity only when the fiber volume fraction is below about 0.5%; at high fiber volume fractions, the addition of sand even increases the resistivity due to the porosity induced by the sand. Thus, the addition of a second filler (silica fume or sand) that is essentially

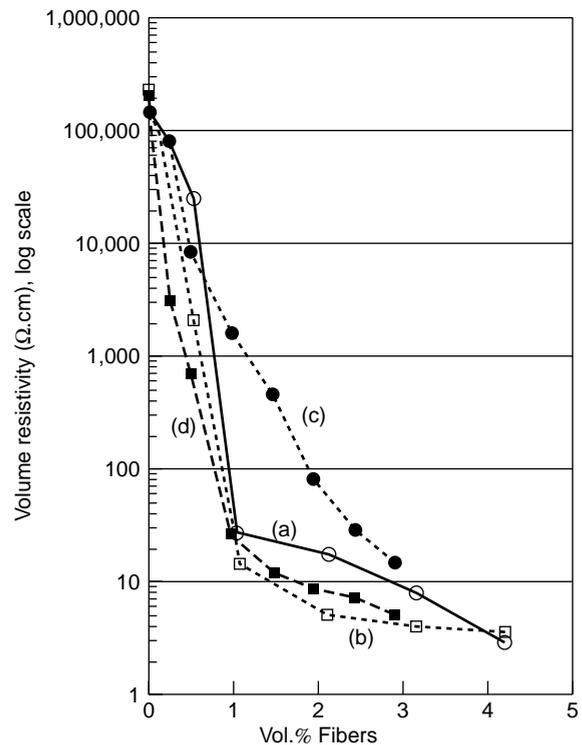


Figure 13. Variation of the volume electrical resistivity of cement-matrix composites with carbon-fiber volume fraction. (a) without sand, with methylcellulose, without silica fume; (b) without sand, with methylcellulose, with silica fume; (c) with sand, with methylcellulose, without silica fume; (d) with sand, with methylcellulose, with silica fume.

nonconducting decreases the resistivity of the composite only at low volume fractions of the carbon fibers, and the maximum fiber volume fraction for decreased resistivity is larger when the particle size of the filler is smaller. The decrease in resistivity is attributed to the improved fiber dispersion that results from the presence of the second filler. Consistent with improved fiber dispersion is increased flexural toughness and strength due to the presence of the second filler.

Table 3 (67,100) shows the specific heats of cement pastes. The specific heat is significantly increased by adding of silica fume. It is further increased by the further addition of methylcellulose and defoamer. It is still further increased by the still further addition of carbon fibers. The effectiveness of fibers in increasing the specific heat increases in the following order: as-received fibers, O_3 -treated fibers, dichromate-treated fibers, and silane-treated fibers. This trend applies whether the silica fume is as-received or silane-treated. For any of the formulations, silane-treated silica fume gives higher specific heat than as-received silica fume. The highest specific heat is exhibited by the cement paste that contains silane-treated silica fume and silane-treated fibers. The specific heat is 12% higher than that of plain cement paste, 5% higher than that of the cement paste containing as-received silica fume and as-received fibers, and 0.5% higher than that of the cement

Table 3. Specific Heats (J/g · K, ± 0.001) of Cement Pastes. The Value for Plain Cement Paste (with Cement and Water only) Is 0.736 J/g · K

Formulation ^a	As-Received Silica Fume	Silane-Treated Silica Fume
A	0.782	0.788
A ⁺	0.793	0.803
A ⁺ F	0.804	0.807
A ⁺ O	0.809	0.813
A ⁺ K	0.812	0.816
A ⁺ S	0.819	0.823

^aA: Cement + water + water reducing agent + silica fume.
^aA⁺: A + methylcellulose + defoamer.
^aA⁺F: A⁺ + as-received fibers.
^aA⁺O: A⁺ + O₃-treated fibers.
^aA⁺K: A⁺ + dichromate-treated fibers.
^aA⁺S: A⁺ + silane-treated fibers.

paste containing as-received silica fume and silane-treated fibers. Hence, silane treatment of fibers is more valuable than treatment of silica fume to increase the specific heat.

Table 4 (67,100) shows the thermal diffusivities of cement pastes. The thermal diffusivity is significantly decreased by adding silica fume. The further addition of methylcellulose and defoamer or the further addition of fibers has relatively little effect on thermal diffusivity. Surface treatment of the fibers by ozone or dichromate slightly increases the thermal diffusivity, whereas surface treatment of the fibers by silane slightly decreases the thermal diffusivity. These trends apply whether the silica fume is as-received or silane-treated. For any of the formulations, silane-treated silica fume gives slightly lower (or essentially the same) thermal diffusivity than as-received silica fume. Silane treatments of silica fume and of fibers are about equally effective in lowering the thermal diffusivity.

Table 5 (67,100) shows the densities of cement pastes. The density is significantly decreased by adding silica fume. It is further decreased slightly by the addition of methylcellulose and defoamer. It is still further decreased by the addition of fibers. The effectiveness of

Table 4. Thermal Diffusivities (mm²/s, ± 0.03) of Cement Pastes. The Value for Plain Cement Paste (with Cement and Water Only) Is 0.36 mm²/s

Formulation ^a	As-Received Silica Fume	Silane-Treated Silica Fume
A	0.26	0.24
A ⁺	0.25	0.22
A ⁺ F	0.27	0.26
A ⁺ O	0.29	0.27
A ⁺ K	0.29	0.27
A ⁺ S	0.25	0.23

^aA: Cement + water + water reducing agent + silica fume.
^aA⁺: A + methylcellulose + defoamer.
^aA⁺F: A⁺ + as-received fibers.
^aA⁺O: A⁺ + O₃-treated fibers.
^aA⁺K: A⁺ + dichromate-treated fibers.
^aA⁺S: A⁺ + silane-treated fibers.

Table 5. Density (g/cm³ ± 0.02) of Cement Pastes. The Value for Plain Cement Paste (with Cement and Water Only) Is 2.01 g/cm³

Formulation ^a	As-Received Silica-Fume	Silane-Treated Silica-Fume
A	1.72	1.73
A ⁺	1.69	1.70
A ⁺ F	1.62	1.64
A ⁺ O	1.64	1.65
A ⁺ K	1.65	1.66
A ⁺ S	1.66	1.68

^aA: Cement + water + water reducing agent + silica fume.
^aA⁺: A + methylcellulose + defoamer.
^aA⁺F: A⁺ + as-received fibers.
^aA⁺O: A⁺ + O₃-treated fibers.
^aA⁺K: A⁺ + dichromate-treated fibers.
^aA⁺S: A⁺ + silane-treated fibers.

fibers in decreasing the density decreases in the following order: as-received fibers, O₃-treated fibers, dichromate-treated fibers, and silane-treated fibers. This trend applies whether the silica fume is as-received or silane-treated. For any of the formulations, silane-treated silica fume gives slightly higher (or essentially the same) specific heat than as-received silica fume. Silane treatment of fibers is more valuable than treatment of silica fume to increase density.

Table 6 (67,100) shows the thermal conductivities of cement pastes. They are significantly decreased by adding silica fume. The further addition of methylcellulose and defoamer or the still further addition of fibers has little effect on density. Surface treatment of fibers by ozone or dichromate slightly increases the thermal conductivity, whereas surface treatment of the fibers by silane has a negligible effect. These trends apply whether the silica fume is as-received or silane-treated. For any of the formulations, silane-treated silica fume gives slightly lower (or essentially the same) thermal conductivity than as-received silica fume. Silane treatments of silica fume and of fibers contribute comparably to reducing the thermal conductivity.

Table 6. Thermal Conductivities (W/mK, ± 0.03) of Cement Pastes. The Value for Plain Cement Paste (with Cement and Water Only) Is 0.53 W/m · K

Formulation ^a	As-Received Silica-Fume	Silane-Treated Silica-Fume
A	0.35	0.33
A ⁺	0.34	0.30
A ⁺ F	0.35	0.34
A ⁺ O	0.38	0.36
A ⁺ K	0.39	0.37
A ⁺ S	0.34	0.32

^aA: Cement + water + water reducing agent + silica fume.
^aA⁺: A + methylcellulose + defoamer.
^aA⁺F: A⁺ + as-received fibers.
^aA⁺O: A⁺ + O₃-treated fibers.
^aA⁺K: A⁺ + dichromate-treated fibers.
^aA⁺S: A⁺ + silane-treated fibers.

Table 7. Thermal Behavior of Cement Pastes and Mortars

	Cement Paste		Mortar	
	Without Silica Fume ^a	With Silica Fume ^a	Without Silica Fume ^a	With Silica Fume ^a
Density (g/cm ³ , ±0.02)	2.01	1.73	2.04	2.20
Specific heat (J/g · K, ±0.001)	0.736	0.788	0.642	0.705
Thermal diffusivity (mm ² /s, ±0.03)	0.36	0.24	0.44	0.35
Thermal conductivity ^b (W/m · K, ±0.03)	0.53	0.33	0.58	0.54

^aSilane-treated.^bProduct of density, specific heat, and thermal diffusivity.

Sand is a much more common component in concrete than silica fume. It differs from silica fume in its relatively large particle size and negligible reactivity with cement. Sand gives effects that are opposite from those of silica fume; sand addition decreases the specific heat and increases the thermal conductivity (99).

Table 7 (99) shows the thermal behavior of cement pastes and mortars. Comparison of the results on cement paste without silica fume and those on mortar without silica fume shows that sand addition decreases the specific heat by 13% and increases the thermal conductivity by 9%. Comparison of the results on cement paste with silica fume and those on mortar with silica fume shows that sand addition decreases the specific heat by 11% and increases the thermal conductivity by 64%. That sand addition has more effect on the thermal conductivity when silica fume is present than when silica fume is absent is due to the low value of the thermal conductivity of cement paste that contains silica fume (Table 7).

Comparison of the results on cement paste without and with silica fume shows that silica fume addition increases the specific heat by 7% and decreases the thermal conductivity by 38%. Comparison of the results on mortar without and with silica fume shows that silica fume addition increases the specific heat by 10% and decreases the thermal conductivity by 6%. Hence, the effects of silica fume addition on mortar and cement paste are in the same direction. That the effect of silica fume on the thermal conductivity is much less for mortar than for cement paste is mainly due to the fact that silica fume addition increases the density of mortar but decreases the density of cement paste (Table 7). That the fractional increase in specific heat due to silica fume addition is higher for mortar than cement paste is attributed to the low value of the specific heat of mortar without silica fume (Table 7).

Comparison of the results on cement paste with silica fume and those on mortar without silica fume shows that sand addition gives a lower specific heat than silica fume addition and a higher thermal conductivity than silica fume addition. Because sand has a much larger particle size than silica fume, sand has much less interfacial area than silica fume, though the interface may be more diffuse for silica fume than for sand. The low interfacial area for sand is believed responsible for the low specific heat and

the higher thermal conductivity because slippage at the interface contributes to the specific heat and the interface acts as a thermal barrier.

Silica fume addition increases the specific heat of cement paste by 7%, whereas sand addition decreases it by 13%. Silica fume addition decreases the thermal conductivity of cement paste by 38%, whereas sand addition increases it by 22%. Hence, silica fume addition and sand addition have opposite effects. The cause is believed to be associated mainly with the low interfacial area for sand and the high interfacial area for silica fume, as explained in the last paragraph. The high reactivity of silica fume compared to sand may contribute to the observed difference between silica fume addition and sand addition, though this contribution is believed to be minor because the reactivity should have tightened up the interface, thus decreasing the specific heat (in contrast to the observed effects). The decrease in specific heat and the increase in thermal conductivity upon sand addition are believed to be due to the higher level of homogeneity within a sand particle than within cement paste.

Cement-Matrix Composites for Vibration Reduction

Vibration reduction requires high damping capacity and high stiffness. Viscoelastic materials such as rubber have high damping capacity but low stiffness. Concretes that have both high damping capacity (two or more orders higher than conventional concrete) (Table 8) (101) and high stiffness (Table 9) (101) can be obtained by adding surface-treated silica fume to concrete. Steel-reinforced concretes that have improved damping capacity and stiffness can be obtained by surface treating the steel (say, by sand blasting) before incorporating it in concrete (Table 10) (102) or by using silica fume in concrete (101). Due to its small

Table 8. Loss Tangent (Tan δ , ±0.01)

Mix	0.2 Hz	0.5 Hz	1.0 Hz	2.0 Hz
Plain ^a	0.016	<10 ⁻⁴	<10 ⁻⁴	<10 ⁻⁴
Sand	<10 ⁻⁴	<10 ⁻⁴	<10 ⁻⁴	<10 ⁻⁴
Sand + silica fume	0.021	0.14	0.01	<10 ⁻⁴

^aNo sand, no silica fume.

Table 9. Storage Modulus (GPa, ± 0.2)

Mix	0.2 Hz	0.5 Hz	1.0 Hz	2.0 Hz
Plain ^a	13.7	14.48	14.02	14.00
Sand	9.43	11.67	10.32	9.56
Sand + silica fume	13.11	14.34	13.17	13.11

^aNo sand, no silica fume.

particle size, silica fume in concrete introduces interfaces that enhance damping. Sand blasting of a steel rebar increases the interfacial area between steel and concrete, thereby enhancing damping. Carbon-fiber addition has relatively small effects on damping capacity and stiffness (103).

POLYMER-MATRIX COMPOSITES FOR SMART STRUCTURES

Polymer-matrix composites for structural applications typically contain continuous fibers such as carbon, polymer, and glass fibers because continuous fibers tend to be more effective as a reinforcement than short fibers. Polymer-matrix composites that contain continuous carbon fibers are used for aerospace, automobile, and civil structures. In contrast, continuous fibers are too expensive for reinforcing concrete. Because carbon fibers are electrically conducting, whereas polymer and glass fibers are not, carbon-fiber composites are predominant among polymer-matrix composites that are intrinsically smart.

Background on Polymer-Matrix Composites

Polymer-matrix composites that contain continuous carbon fibers are important structural materials due to their high tensile strength, high tensile modulus, and low density. They are used for lightweight structures such as satellites, aircraft, automobiles, bicycles, ships, submarines, sporting goods, wheel chairs, armor, and rotating machinery (such as turbine blades and helicopter rotors). Due to the recent emphasis on repair of civil infrastructural systems,

composites are beginning to be used to repair concrete structures and bridges, even though they are much more expensive than concrete. Because the price of carbon fibers has been dropping steadily during the last two decades, the spectrum of applications has been widening tremendously.

The continuous carbon fibers used are primarily based on either polyacrylonitrile (PAN) or mesophase pitch. Mesophase-pitch-based carbon fibers, if heat treated to high temperatures exceeding 2500°C, can be graphitized and attain very high values of tensile modulus and thermal conductivity (in-plane), in addition to improved oxidation resistance. The high thermal conductivity is attractive for thermal management, which is particularly important for electronics (i.e., heat sinks, etc.). However, graphitized fibers tend to have relatively low strength due to the ease of shear between the graphite layers, and they are very expensive. On the other hand, PAN-based fibers cannot be graphitized, though they compete well with mesophase-pitch-based fibers which have not been graphitized, both materials exhibit reasonably high values of both strength and modulus and are not very expensive. These fibers are the most widely used among carbon fibers. The fabrication of both pitch-based and PAN-based carbon fibers involves stabilization (infusibilization) and then carbonization (conversion from hydrocarbon molecules to a carbon network). Graphitization optionally follows carbonization.

Due to the importance of carbon-fiber polymer-matrix composites for structural applications, the mechanical behavior of these materials has been much investigated. Much less work has been done to study the electrical behavior (104–111). On the other hand, due to the fact that carbon fibers are much more conductive than the polymer matrix, the electrical behavior gives much information on the microstructure, such as the degree of fiber alignment, the number of fiber–fiber contacts, the amount of delamination, and the extent of fiber breakage. Such information is useful for scientific understanding of the properties of the composite and is also valuable for giving the composite the ability to sense its strain, damage, and temperature in real time via electrical measurement. In other words, the strain, damage, and temperature affect the electrical

Table 10. Loss Tangent, Storage Modulus, and Loss Modulus of Mortars with and Without Steel Reinforcement

Property	Sample Type ^a	Frequency		
		0.2 Hz	0.5 Hz	1.0 Hz
Loss tangent	A	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
	B	$(2.73 \pm 0.19) \times 10^{-2}$	$(1.56 \pm 0.08) \times 10^{-2}$	$(7.20 \pm 0.37) \times 10^{-3}$
	C	$(3.32 \pm 0.15) \times 10^{-2}$	$(1.98 \pm 0.17) \times 10^{-2}$	$(1.07 \pm 0.09) \times 10^{-2}$
	D	$(3.65 \pm 0.27) \times 10^{-2}$	$(2.50 \pm 0.22) \times 10^{-2}$	$(1.24 \pm 0.16) \times 10^{-2}$
Storage modulus (GPa)	A	20.2 ± 3.5	27.5 ± 4.3	25.8 ± 3.7
	B	44.2 ± 4.8	47.7 ± 5.3	44.4 ± 5.0
	C	36.9 ± 4.3	41.0 ± 3.9	38.4 ± 3.0
	D	46.0 ± 4.0	51.2 ± 6.4	49.3 ± 5.8
Loss modulus (GPa)	A	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$
	B	1.21 ± 0.22	0.74 ± 0.12	0.32 ± 0.05
	C	1.23 ± 0.20	0.81 ± 0.15	0.41 ± 0.07
	D	1.68 ± 0.27	1.28 ± 0.27	0.61 ± 0.51

^aA: No rebar. B: As-received steel rebar. C: Ozone treated steel rebar. D: Sand-blasted steel rebar.

behavior, such as electrical resistance, which thus serves to indicate strain, damage, and temperature. In this way, the composite is self-sensing, that is, intrinsically smart, without the need for attached or embedded sensors (such as optical fibers, acoustic sensors and piezoelectric sensors), which raise the cost, reduce the durability and, in the case of embedded sensors, weaken the structure.

Carbon fibers are electrically conductive, whereas the polymer matrix is electrically insulating [except for the rare situation in which the polymer is electrically conductive (112)]. The continuous fibers in a composite laminate are in the form of layers called laminae. Each lamina comprises many bundles (called tows) of fibers in a polymer matrix. Each tow consists of thousands of fibers. There may or may not be twist in a tow. Each fiber has a diameter that typically ranges from 7–12 μm . The tows within a lamina are usually oriented in the same direction, but tows in different laminae may or may not be in the same direction. A laminate whose tows in all of the laminae are oriented in the same direction is said to be unidirectional. A laminate whose tows in adjacent laminae are oriented at a 90° angle is said to be crossply. In general, an angle of 45° and other angles may also be involved in the various laminae, as desired to attain the mechanical properties required for the laminate in various directions in the plane of the laminate.

Within a lamina whose tows are in the same direction, the electrical conductivity is highest in the fiber direction. In the transverse direction in the plane of the lamina, the conductivity is not zero, even though the polymer matrix is insulating because there are contacts between fibers of adjacent tows (113). In other words, a fraction of the fibers of one tow touch a fraction of the fiber of an adjacent tow here and there along the length of the fibers. These contacts result from the fact that fibers are not perfectly straight or parallel (even though the lamina is said to be unidirectional) and the flow of the polymer matrix (or resin) during composite fabrication can prevent a fiber from being completely covered by the polymer (even though, prior to composite fabrication, each fiber may be completely covered by the polymer, as in the case of a prepreg, a fiber sheet impregnated with polymer). Fiber waviness is known as marcelling. Thus, the transverse conductivity gives information on the number of fiber–fiber contacts in the plane of the lamina.

For similar reasons, the contacts between fibers of adjacent laminae make the conductivity in the through-thickness direction (direction perpendicular to the plane of the laminate) nonzero. Thus, the through-thickness conductivity gives information on the number of fiber–fiber contacts between adjacent laminae.

Matrix cracking between the tows of a lamina decreases the number of fiber–fiber contacts in the plane of the lamina, thus decreasing the transverse conductivity. Similarly, matrix cracking between adjacent laminae [as in delamination (114)] decreases the number of fiber–fiber contacts between adjacent laminae, thus decreasing the through-thickness conductivity. This means that the transverse and through-thickness conductivities can indicate damage as matrix cracking.

Fiber damage (as distinct from fiber fracture) decreases the conductivity of a fiber, thereby decreasing the

longitudinal conductivity (conductivity in the fiber direction). However, due to the brittleness of carbon fibers, the decrease in conductivity due to fiber damage before fiber fracture is rather small (115).

Fiber fracture causes a much larger decrease in the longitudinal conductivity of a lamina than fiber damage. If there is only one fiber, a broken fiber results in an open circuit, i.e., zero conductivity. However, a lamina has a large number of fibers and adjacent fibers can make contact here and there. Therefore, the portions of a broken fiber still contribute to the longitudinal conductivity of the lamina. As a result, the decrease in conductivity due to fiber fracture is less than it would be if a broken fiber did not contribute to conductivity. Nevertheless, the effect of fiber fracture on longitudinal conductivity is significant, so that longitudinal conductivity can indicate damage as fiber fracture (116).

The through-thickness volume resistance of a laminate is the sum of the volume resistance of each of the laminae in the through-thickness direction and the contact resistance of each of the interfaces between adjacent laminae (i.e., the interlaminar interface). For example, a laminate that has eight laminae has eight volume resistances and seven contact resistances, all in the through-thickness direction. Thus, to study the interlaminar interface, it is better to measure the contact resistance between two laminae rather than the through-thickness volume resistance of the entire laminate.

The contact resistance between laminae can be measured by allowing two laminae (strips) to contact at a junction and using the two ends of each strip to make four electrical contacts (117). An end of the top strip and an end of the bottom strip serve as contacts for passing current. The other end of the top strip and the other end of the bottom strip serve as contacts for voltage measurement. The fibers in the two strips can be in the same direction or in different directions. This method is a form of the four-probe method of measuring electrical resistance. The configuration is illustrated in Fig. 14 for crossply and unidirectional laminates. To make sure that the volume resistance within a lamina in the through-thickness direction does not contribute to the measured resistance, the fibers at each end of a lamina strip should be electrically shorted together by using silver paint or other conducting media. The measured resistance is the contact resistance of the junction. This resistance, multiplied by the area of the junction, gives the contact resistivity, which is independent of the area of the junction and depends only on the nature of the interlaminar interface. The unit of contact resistivity is $\Omega\text{ m}^2$, whereas that of the volume resistivity is $\Omega\text{ m}$.

The structure of the interlaminar interface tends to be more prone to change than the structure within a lamina. For example, damage as delamination is much more common than damage as fiber fracture. Moreover, the structure of the interlaminar interface is affected by the interlaminar stress (whether thermal stress or curing stress), which is particularly significant when the laminae are not unidirectional (because the anisotropy within each lamina enhances the interlaminar stress). The structure of the interlaminar interface also depends on the extent of consolidation of the laminae during composite fabrication.

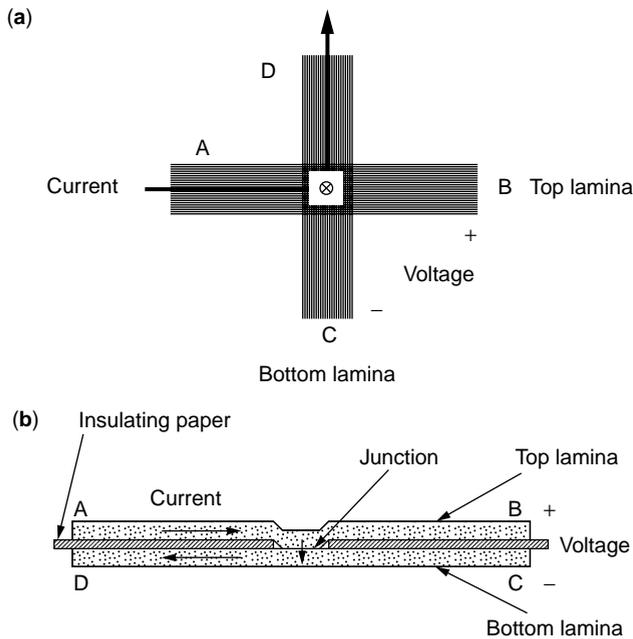


Figure 14. Specimen configuration for measuring of the contact electrical resistivity between laminae. (a) Crossply laminae. (b) Unidirectional laminae.

The contact resistance provides a sensitive probe of the structure of the interlaminar interface.

The volume resistivity in the through-thickness direction can be measured by using the four-probe method, in which each of the two current contacts is a conductor loop (made by silver paint, for example) on each of the two outer surfaces of the laminate in the plane of the laminate, and each of the two voltage contacts is a conductor dot within the loop (114). An alternate method is to have four extra long laminae in the laminate, extend out to serve as electrical leads (118). The two outer leads are for current contacts, and the two inner leads are for voltage contacts. The use of a thin metal wire inserted at an end into the interlaminar space during composite fabrication to serve as an electrical contact is not recommended because the quality of the electrical contact between the metal wire and carbon fibers is hard to control and the wire is intrusive in the composite. The alternate method is less convenient than the method involving loops and dots, but it approaches the ideal four-probe method more closely.

To attain zero conductivity in the through-thickness direction of a laminate, it is necessary to use an insulating layer between two adjacent laminae (119). The insulating layer can be a piece of writing paper. Tissue paper is ineffective in preventing contacts between fibers of adjacent laminae due to its porosity. The attainment of zero conductivity in the through-thickness direction allows the laminate to serve as a capacitor. This means that the structural composite stores energy by serving as a capacitor.

Polymer-Matrix Composites for Strain Sensing

Smart structures that can monitor their own strain are valuable for structural vibration control. Self-monitoring of strain (reversible) has been achieved in carbon-fiber

epoxy-matrix composites without using embedded or attached sensors (118,120–123) because the electrical resistance of the composite in the through-thickness or longitudinal direction changes reversibly with longitudinal strain (gauge factor up to 40) due to a change in the degree of fiber alignment. Tension in the fiber direction of the composite increases the degree of fiber alignment, thereby decreasing the chance for fibers of adjacent laminae to touch one another. As a consequence, the through-thickness resistance increases, and the longitudinal resistance decreases.

Figure 14 (121) shows the change in longitudinal resistance during cyclic longitudinal tension in the elastic regime for a unidirectional continuous-carbon-fiber epoxy-matrix composite that has eight fiber layers (laminae). The stress amplitude is equal to 14% of the breaking stress. The strain returns to zero at the end of each cycle. Because of the small strains involved, the fractional resistance change $\Delta R/R_0$ is essentially equal to the fractional change in resistivity. The longitudinal $\Delta R/R_0$ decreases upon loading and increases upon unloading in every cycle, such that R irreversibly decreases slightly after the first cycle (i.e., $\Delta R/R_0$ does not return to zero at the end of the first cycle). At higher stress amplitudes, the effect is similar, except that both of the reversible and irreversible parts of $\Delta R/R_0$ are larger.

Figure 15 (121) shows the change in the through-thickness resistance during cyclic longitudinal tension in the elastic regime for the same composite. The stress amplitude is equal to 14% of the breaking stress. The through-thickness $\Delta R/R_0$ increases upon loading and decreases

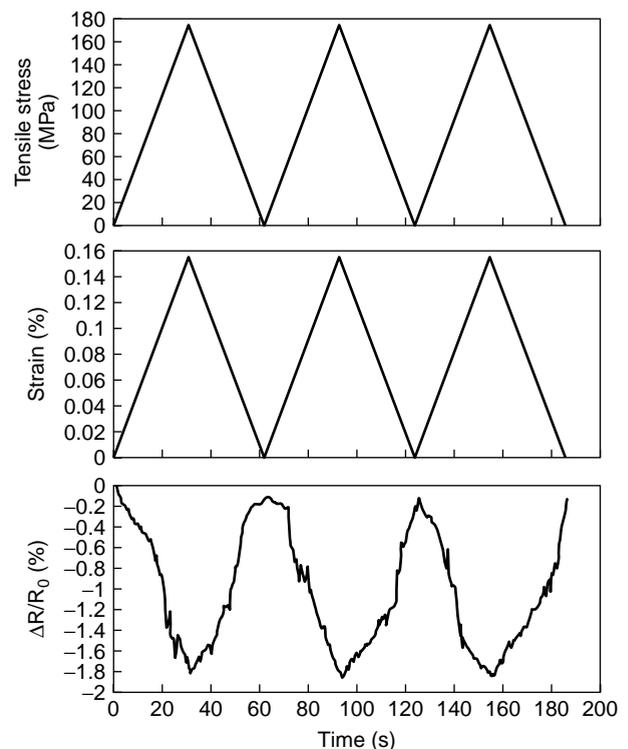


Figure 15. Longitudinal stress and strain and fractional resistance increase ($\Delta R/R_0$) obtained simultaneously during cyclic tension at a stress amplitude equal to 14% of the breaking stress for a continuous-fiber epoxy-matrix composite.

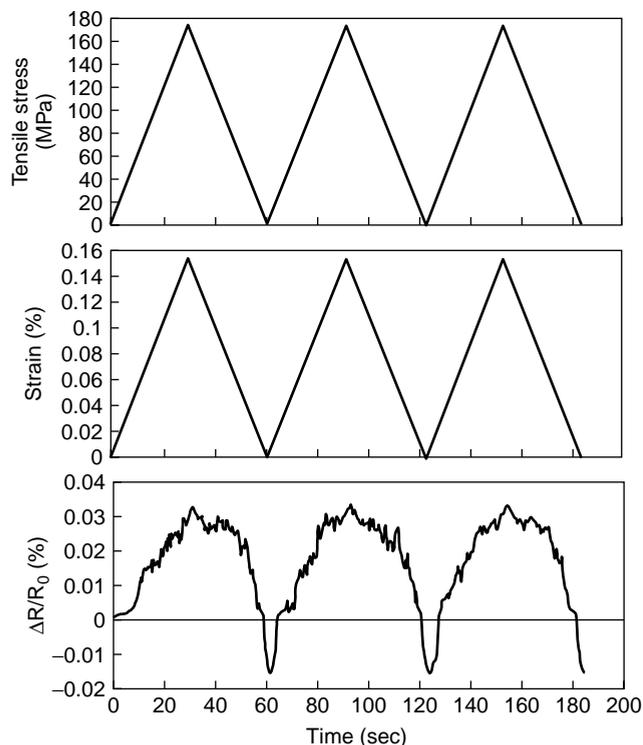


Figure 16. Longitudinal stress and strain and the through-thickness $\Delta R/R_0$ obtained simultaneously during cyclic tension at a stress amplitude equal to 14% of the breaking stress for a continuous-fiber epoxy-matrix composite.

upon unloading in every cycle, such that R irreversibly decreases slightly after the first cycle (i.e., $\Delta R/R_0$ does not return to zero at the end of the first cycle). Upon increasing the stress amplitude, the effect is similar, except that the reversible part of $\Delta R/R_0$ is larger.

The strain sensitivity (gauge factor) is defined as the reversible part of $\Delta R/R_0$ divided by the longitudinal strain amplitude. It is negative (from -18 to -12) for the longitudinal $\Delta R/R_0$ and positive (17 – 24) for the through-thickness $\Delta R/R_0$. The magnitudes are comparable for the longitudinal and through-thickness strain sensitivities. As a result, whether the longitudinal R or the through-thickness R is preferred for strain sensing depends only on the convenience of electrical contact application for the geometry of the particular smart structure.

Figure 16 (121) shows the compressive stress, strain, and longitudinal $\Delta R/R_0$ obtained simultaneously during cyclic compression at stress amplitudes equal to 14% of the breaking stress for a similar composite that has 24 rather than eight fiber layers. The longitudinal $\Delta R/R_0$ increases upon compressive loading and decreases upon unloading in every cycle, such that resistance R irreversibly increases very slightly after the first cycle. The magnitude of the gauge factor is lower in compression (-1.2) than in tension (from -18 to -12).

A dimensional change without any resistivity change would have caused longitudinal R to increase during tensile loading and decrease during compressive loading. In contrast, the longitudinal R decreases upon tensile loading

and increases upon compressive loading. In particular, the magnitude of $\Delta R/R_0$ under tension is 7–11 times that of $\Delta R/R_0$ calculated by assuming that $\Delta R/R_0$ is due only to dimensional change and not due to any change in resistivity. Hence, the contribution of $\Delta R/R_0$ from the dimensional change is negligible compared to that from the change in resistivity.

Though the irreversible behavior is small compared to the reversible behavior, it is such that R (longitudinal or through-thickness) under tension is irreversibly decreased after the first cycle. This behavior is attributed to the irreversible disturbance to the fiber arrangement at the end of the first cycle, such that the fiber arrangement becomes less neat. A less neat fiber arrangement means that there is more chance for the adjacent fiber layers to touch one another.

Polymer-Matrix Composites for Damage Sensing

Self-monitoring of damage (whether due to stress or temperature, under static or dynamic conditions) has been achieved in continuous-carbon-fiber polymer-matrix composites because the electrical resistance of the composite changes with damage (114,120,124–137). Minor damage in the form of slight matrix damage and/or disturbance to the fiber arrangement is indicated by the irreversible decrease of the longitudinal and through-thickness resistance due to an increase in the number of contacts between fibers, as shown after one loading cycle in Figs. 15–17.

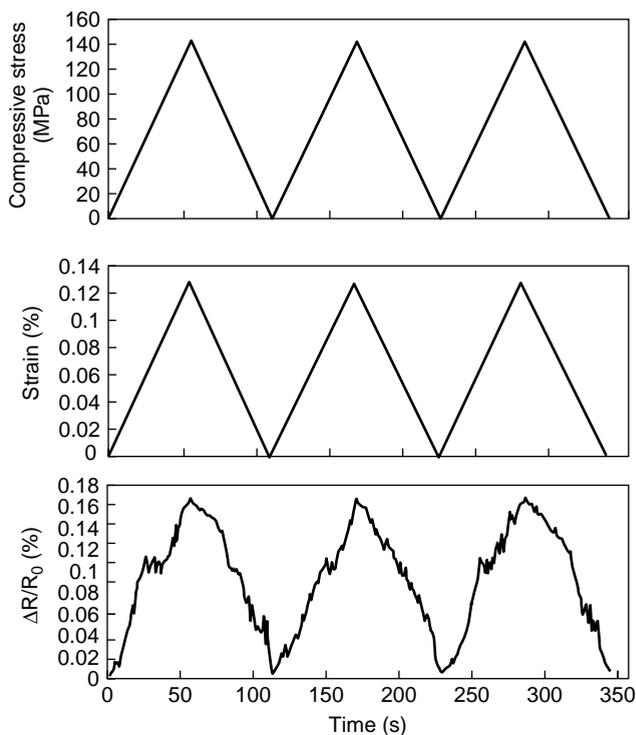


Figure 17. Longitudinal stress, strain, and $\Delta R/R_0$ obtained simultaneously during cyclic compression (longitudinal) at a stress amplitude equal to 14% of the breaking stress for a continuous-fiber epoxy-matrix composite.

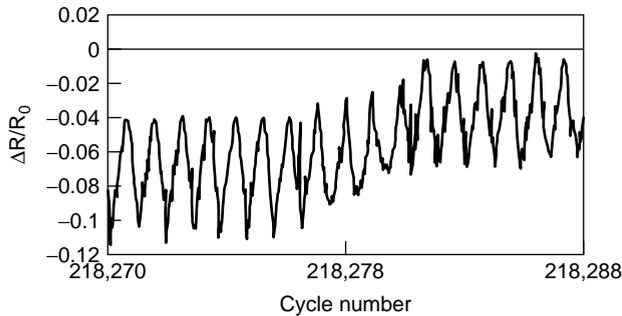


Figure 18. Variation of longitudinal $\Delta R/R_0$ with cycle number during tension–tension fatigue testing for a carbon-fiber epoxy-matrix composite. Each cycle of reversible decrease in resistance is due to strain. The irreversible increase in resistance at around cycle number 218,281 is due to damage as fiber breakage.

More significant damage as delamination or interlaminar interfacial degradation is indicated by the increase in the through-thickness resistance (or more exactly the contact resistivity of the interlaminar interface) due to a decrease in the number of contacts between fibers of different laminae. Major damage as fiber breakage is indicated by the irreversible increase of the longitudinal resistance.

During mechanical fatigue, it was observed that delamination begins at 30% of the fatigue life, whereas it was observed that fiber breakage begins at 50% of the fatigue life. Figure 18 (124) shows an irreversible increase in resistance at about 50% of the fatigue life during tension–tension fatigue testing of a unidirectional, continuous-carbon-fiber epoxy-matrix composite. The resistance and stress are in the fiber direction. The reversible changes in resistance are due to strain, which makes the resistance decrease reversibly in each cycle, as in Fig. 15.

Figure 19 shows the variation in contact resistivity with temperature during thermal cycling. The temperature is repeatedly increased to various levels. A group of cycles in which the temperature amplitude increases cycle by cycle and then decreases cycle by cycle back to the initial low temperature amplitude is called a group. Figure 19a shows the results of the first 10 groups, whereas Fig. 19b shows only the first group. The contact resistivity decreases upon heating in every cycle of every group. At the highest temperature (150°C) of a group, a spike of increase in resistivity occurs, as shown in Fig. 19b. It is attributed to damage at the interlaminar interface. In addition, the baseline resistivity (i.e., the top envelope) gradually and irreversibly shifts downward as cycling progresses, as shown in Fig. 19a. The baseline decrease is probably due to matrix damage within a lamina and the resulting decrease in modulus and hence decrease in residual stress.

Polymer-Matrix Composites for Temperature Sensing

Continuous-carbon-fiber epoxy-matrix composites provide temperature sensing by serving as thermistors (117,138) and thermocouples (139).

The thermistor function stems from the reversible decrease of contact electrical resistivity at the interface between fiber layers (laminae) from temperature change.

Figure 20 shows the variation in contact resistivity ρ_c with temperature during reheating and subsequent cooling, both at 0.15°C/min, for carbon-fiber epoxy-matrix composites cured at zero and 0.33 MPa. The corresponding Arrhenius plots of log contact conductivity (inverse of contact resistivity) versus the inverse of absolute temperature during heating are shown in Fig. 21. The activation energy can be calculated from the slope (negative) of the Arrhenius plot, which is quite linear. The linearity of the Arrhenius plot means that the activation energy does not change throughout the temperature variation. This activation energy is the energy of electrons jumping from one lamina to the other. Electronic excitation across this energy enables conduction in the through-thickness direction.

The activation energies, thicknesses, and room temperature contact resistivities for samples made at different curing pressures and composite configurations are shown in Table 11. All of the activation energies were calculated on the basis of the data at 75–125°C. In this temperature regime, the temperature change was very linear and well controlled. From Table 11, it can be seen that, for the same composite configuration (crossply), the higher the curing pressure, the smaller the composite thickness (because more epoxy is squeezed out), the lower the contact resistivity, and the higher the activation energy. A smaller composite thickness corresponds to a higher fiber volume fraction in the composite. During curing and subsequent cooling, the matrix shrinks, so a longitudinal compressive stress develops in the fibers. The modulus of carbon fibers in the longitudinal direction is much higher than that in the transverse direction. Moreover, the carbon fibers are continuous in the longitudinal direction. Thus, the overall shrinkage in the longitudinal direction tends to be less than that in the transverse direction. Therefore, there will be a residual interlaminar stress in the two crossply layers in a given direction. This stress accentuates the barrier for electrons to jump from one lamina to another. After curing and subsequent cooling, heating will decrease the thermal stress. Both thermal stress and curing stress contribute to residual interlaminar stress. Therefore, the higher the curing pressure, the larger the fiber volume fraction, the greater the residual interlaminar stress, and the higher the activation energy, as shown in Table 11. Besides the residual stress, thermal expansion can also affect contact resistance by changing the contact area. However, calculation shows that the contribution of thermal expansion is less than one-tenth of the observed change in contact resistance with temperature.

The electron jump occurs primarily at points where direct contact occurs between fibers of adjacent laminae. Direct contact is possible due to the flow of epoxy resin during composite fabrication and due to the slight waviness of the fibers, as explained in Wang and Chung (114) in relation to the through-thickness volume resistivity of a carbon-fiber epoxy-matrix composite.

The curing pressure for the sample in the unidirectional composite configuration was higher than that of any of the crossply samples (Table 11). Consequently, the thickness was the lowest. As a result, the fiber volume fraction was the highest. However, the contact resistivity of the unidirectional sample was the second highest rather

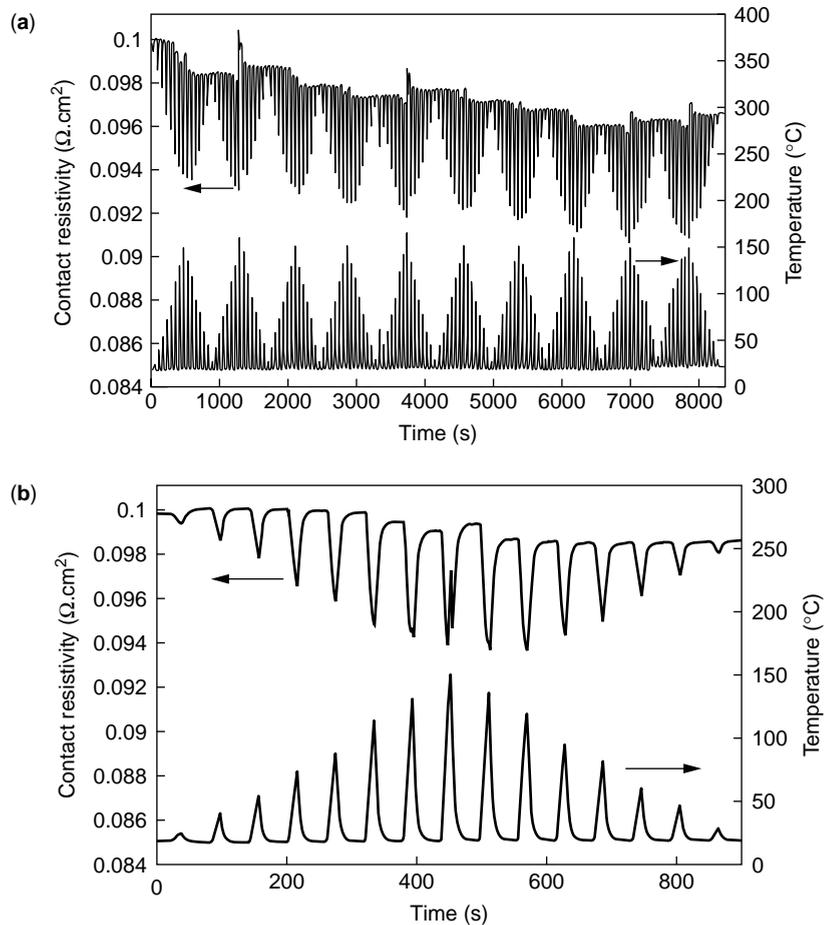


Figure 19. Variation in contact electrical resistivity and of temperature with time during the thermal cycling of a carbon-fiber epoxy-matrix composite. (b) the magnified view of the first 900 s of (a).

than the lowest, and its activation energy was the lowest rather than the highest. The low activation energy is consistent with the fact that there was no mismatch of CTE or curing shrinkage between the two unidirectional laminae and, as a result, no interlaminar stress between the

laminae. This low value supports the notion that interlaminar stress is important in affecting activation energy. The high contact resistivity for the unidirectional case can be explained in the following way. In the crossply samples, the pressure during curing forced the fibers of the

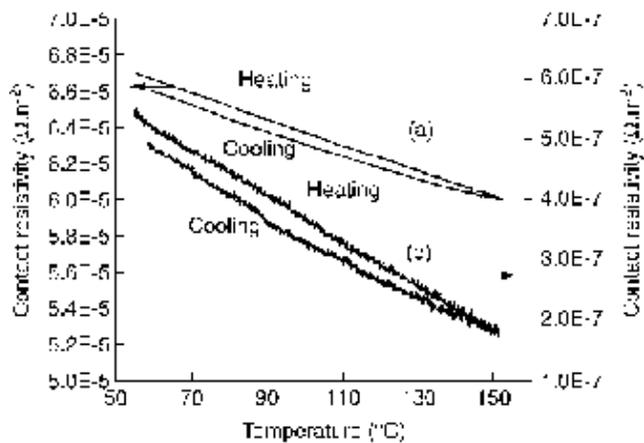


Figure 20. Variation of contact electrical resistivity with temperature during heating and cooling of carbon-fiber epoxy-matrix composites at 0.15°C/min (a) for composite made without any curing pressure and (b) for composite made at a curing pressure of 0.33 MPa.

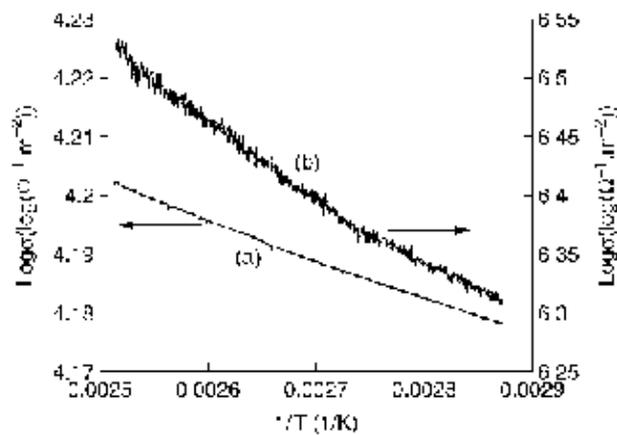


Figure 21. Arrhenius plot of log contact conductivity vs. inverse of absolute temperature during heating of carbon-fiber epoxy-matrix composites at 0.15°C/min (a) for composite made without any curing pressure and (b) for composite made at a curing pressure of 0.33 MPa.

Table 11. Activation Energy for Various Carbon-Fiber Epoxy-Matrix Composites. The Standard Deviations Are Shown in Parentheses

Composite Configuration	Curing Pressure (MPa)	Composite Thickness (mm)	Contact Resistivity ρ_{co} ($\Omega \text{ m}^2$)	Activation Energy (kJ/mol)		
				Heating at 0.15°C/min	Heating at 1°C/min	Cooling at 0.15°C/min
Crossply	0	0.36	7.3×10^{-5}	1.26 (2×10^{-3})	1.24 (3×10^{-3})	1.21 (8×10^{-4})
	0.062	0.32	1.4×10^{-5}	1.26 (4×10^{-3})	1.23 (7×10^{-3})	1.23 (4×10^{-3})
	0.13	0.31	1.8×10^{-5}	1.62 (3×10^{-3})	1.57 (4×10^{-3})	1.55 (2×10^{-3})
	0.19	0.29	5.4×10^{-6}	2.14 (3×10^{-3})	2.15 (3×10^{-3})	2.13 (1×10^{-3})
	0.33	0.26	4.0×10^{-7}	11.4 (4×10^{-2})	12.4 (8×10^{-2})	11.3 (3×10^{-2})
Unidirectional	0.42	0.23	2.9×10^{-5}	1.02 (3×10^{-3})	0.82 (4×10^{-3})	0.78 (2×10^{-3})

two laminae to press onto one another and hence contact tightly. In the unidirectional sample, the fibers of one of the laminae just sank into the other lamina at the junction, so pressure helped relatively little in the contact between fibers of adjacent laminae. Moreover, in the crossply, every fiber at the lamina–lamina interface contacts many fibers of the other lamina, whereas, in the unidirectional situation, every fiber has little chance to contact the fibers of the other lamina. Therefore, the number of contact points between the two laminae is less for the unidirectional sample than for the crossply samples.

The thermocouple function stems from the use of n-type and p-type carbon fibers (as obtained by intercalation) in different laminae. The sensitivity and linearity of the thermocouple are as good as or better than those of commercial thermocouples. By using two laminae that are crossply, a two-dimensional array of thermistors or thermocouple junctions is obtained, thus allowing temperature distribution sensing.

Table 12 shows the Seebeck coefficient and the absolute thermoelectric power of carbon fibers and the thermocouple sensitivity of epoxy-matrix composite junctions. A positive value of the absolute thermoelectric power indicates p-type behavior; a negative value indicates n-type behavior. Pristine P-25 is slightly n-type; pristine

T-300 is strongly n-type. A junction that comprises pristine P-25 and pristine T-300 has a positive thermocouple sensitivity that is close to the difference in the Seebeck coefficients (or the absolute thermoelectric powers) of T-300 and P-25, whether the junction is unidirectional or crossply. Pristine P-100 and pristine P-120 are both slightly n-type. Intercalation with sodium makes P-100 and P-120 strongly n-type. Intercalation with bromine makes P-100 and P-120 strongly p-type. A junction comprising bromine-intercalated P-100 and sodium-intercalated P-100 has a positive thermocouple sensitivity that is close to the sum of the magnitudes of the absolute thermoelectric powers of the bromine-intercalated P-100 and the sodium-intercalated P-100. Similarly, a junction comprising bromine-intercalated P-120 and sodium-intercalated P-120 has a positive thermocouple sensitivity that is close to the sum of the magnitudes of the absolute thermoelectric powers of the bromine-intercalated P-120 and the sodium-intercalated P-120. Figure 22 shows the linear relationship of the measured voltage with the temperature difference between hot and cold points for the junction comprising bromine-intercalated P-100 and sodium-intercalated P-100.

A junction comprising n-type and p-type partners has a thermocouple sensitivity that is close to the sum of the

Table 12. Seebeck Coefficient ($\mu\text{V}/^\circ\text{C}$) and Absolute Thermoelectric Power ($\mu\text{V}/^\circ\text{C}$) of Carbon Fibers and Thermocouple Sensitivity ($\mu\text{V}/^\circ\text{C}$) of Epoxy-Matrix Composite Junctions. All Junctions are Unidirectional Unless Specified as Crossply. The Temperature Range is 20–110°C

	Seebeck Coefficient with Copper as the Reference ($\mu\text{V}/^\circ\text{C}$)	Absolute Thermoelectric Power ($\mu\text{V}/^\circ\text{C}$)	Thermocouple Sensitivity ($\mu\text{V}/^\circ\text{C}$)
P-25 ^a	+0.8	−1.5	
T-300 ^a	−5.0	−7.3	
P-25 ^a + T-300 ^a			+5.5
P-25 ^a + T-300 ^a (crossply)			+5.4
P-100 ^a	−1.7	−4.0	
P-120 ^a	−3.2	−5.5	
P-100 (Na)	−48	−50	
P-100 (Br ₂)	+43	+41	
P-100 (Br ₂) + P-100 (Na)			+82
P-120 (Na)	−42	−44	
P-120 (Br ₂)	+38	+36	
P-120 (Br ₂) + P-120 (Na)			+74

^aPristine (i.e., not intercalated).

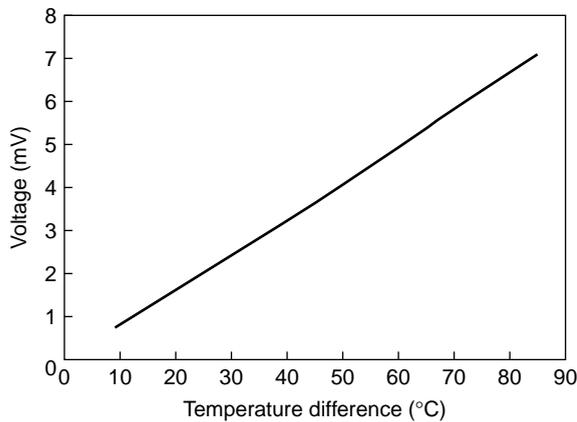


Figure 22. Variation of the measured voltage with the temperature difference between hot and cold points of the epoxy-matrix composite junction comprised of bromine-intercalated P-100 and sodium-intercalated P-100 carbon fibers.

magnitudes of the absolute thermoelectric powers of the two partners because the electrons in the n-type partner, as well as the holes in the p-type partner, move away from the hot point toward the corresponding cold point. As a result, the overall effect on the voltage difference between the two cold ends is additive.

By using junctions comprising strongly n-type and strongly p-type partners, a thermocouple sensitivity as high as $+82 \mu\text{V}/^\circ\text{C}$ was attained. Semiconductors are known to exhibit much higher values of the Seebeck coefficient than metals, but the need to have thermocouples in the form of long wires makes metals the main materials for thermocouples. Intercalated carbon fibers exhibit much higher values of the Seebeck coefficient than metals. Yet, unlike semiconductors, their fiber form and fiber composite form make them convenient for practical use as thermocouples.

The thermocouple sensitivity of carbon-fiber epoxy-matrix composite junctions is independent of the extent of curing and is the same for unidirectional and crossply junctions. This is consistent with the fact that the thermocouple effect hinges on the difference in the bulk properties of the two partners and is not an interfacial phenomenon. This behavior means that the interlaminar interfaces in a fibrous composite serve as thermocouple junctions in the same way, irrespective of the layup configuration of the dissimilar fibers in the laminate. Because a structural composite typically has fibers in multiple directions, this behavior facilitates using a structural composite as a thermocouple array.

It is important to note that the thermocouple junctions do not require any bonding agent other than the epoxy, which serves as the matrix of the composite and is not an electrical contact medium (because it is not conductive). Despite the presence of the epoxy matrix in the junction area, direct contact occurs between a fraction of the fibers of a lamina and a fraction of the fibers of the other lamina, thus resulting in a conduction path in the direction perpendicular to the junction.

Polymer-Matrix Composites for Vibration Reduction

Polymer-matrix composites that contain continuous carbon fibers and exhibit both high damping capacity and high stiffness have been attained by using interlayers in the form of discontinuous $0.1 \mu\text{m}$ diameter carbon filaments (140). The damping enhancement is due to the large interfacial area between filaments and matrix. Viscoelastic constrained interlayers, though common, degrade the strength and modulus and become less effective as the temperature increases (141). Surface treatment of the filaments helps, at least in a thermoplastic matrix (141). Other techniques for damping enhancement are matrix modification (142,143), fiber coating (143), and filament winding angle optimization (144).

CONCLUSION

Intrinsically smart structural composites for strain sensing, damage sensing, temperature sensing, thermal control, and vibration reduction are attractive for smart structures. They include cement-matrix and polymer-matrix composites, particularly cement-matrix composites that contain short carbon fibers and polymer-matrix composites that contain continuous carbon fibers. The electrical conductivity of the fibers enables the dc electrical resistivity of the composites to change in response to strain, damage, or temperature, thereby allowing sensing. In addition, the conduction enables the Seebeck effect, which is particularly large in cement-matrix composites that contain short steel fibers and in polymer-matrix composites that contain intercalated, continuous, carbon fibers. By using the interfaces in composites to enhance damping, cement-matrix and polymer-matrix composites that have both enhanced damping capacity and increased stiffness are obtained. By using composite interfaces, cement-matrix composites that have increased specific heat for thermal control are also obtained.

BIBLIOGRAPHY

1. J.W. Newman, *Int. SAMPE Symp. Exhib.* **32**, 938–944 (1987).
2. S. Furukawa, Y. Tsuji, and S. Otani, *Proc. Jpn. Congr. Mater. Res.* **30**, 149–152 (1987).
3. K. Saito, N. Kawamura, and Y. Kogo, *Adv. Mater. The Big Payoff, Natl. SAMPE Tech. Conf.*, Covina, CA, 1989, Vol. 21, pp. 796–802.
4. S. Wen and D.D.L. Chung, *Cem. Concr. Res.* **29**(3), 445–449 (1999).
5. T. Sugama, L.E. Kukacka, N. Carciello, and D. Stathopoulos, *Cem. Concr. Res.* **19**(3), 355–365 (1989).
6. X. Fu, W. Lu, and D.D.L. Chung, *Cem. Concr. Res.* **26**(7), 1007–1012 (1996).
7. X. Fu, W. Lu, and D.D.L. Chung, *Carbon* **36**(9), 1337–1345 (1998).
8. Y. Xu and D.D.L. Chung, *Cem. Concr. Res.* **29**(5), 773–776 (1999).
9. T. Yamada, K. Yamada, R. Hayashi, and T. Herai, *Int. SAMPE Symp. Exhib.* **36**(pt. 1), 362–371 (1991).

10. T. Sugama, L.E. Kukacka, N. Carciello, and B. Galen, *Cem. Concr. Res.* **18**(2), 290–300 (1988).
11. B.K. Larson, L.T. Drzal, and P. Sorousian, *Composites* **21**(3), 205–215 (1990).
12. A. Katz, V.C. Li, and A. Kazmer, *J. Mater. Civ. Eng.* **7**(2), 125–128 (1995).
13. S.B. Park and B.I. Lee, *Cem. Concr. Compos.* **15**(3), 153–163 (1993).
14. P. Chen, X. Fu, and D.D.L. Chung, *ACI Mater. J.* **94**(2), 147–155 (1997).
15. P. Chen and D.D.L. Chung, *Composites* **24**(1), 33–52 (1993).
16. A.M. Brandt and L. Kucharska, *Materi. New Millennium, Proc. Mater. Eng. Conf.* (1996), Vol. 1, pp. 271–280.
17. H.A. Toutanji, T. El-Korchi, R.N. Katz, and G.L. Leatherman, *Cem. Concr. Res.* **23**(3), 618–626 (1993).
18. N. Banthia and J. Sheng, *Cem. Concr. Compos.* **18**(4), 251–269 (1996).
19. H.A. Toutanji, T. El-Korchi, and R.N. Katz, *Cem. Concr. Compos.* **16**(1), 15–21 (1994).
20. S. Akihama, T. Suenaga, and T. Banno, *Int. J. Cem. Compos. Lightweight Concr.* **6**(3), 159–168 (1984).
21. M. Kamakura, K. Shirakawa, K. Nakagawa, K. Ohta, and S. Kashihara, *Sumitomo Kinzoku* (1983).
22. A. Katz and A. Bentur, *Cem. Concr. Res.* **24**(2), 214–220 (1994).
23. Y. Ohama and M. Amano, *Proc. Jpn. Congr. Mater. Res.* **27**, 187–191 (1983).
24. Y. Ohama, M. Amano, and M. Endo, *Concr. Int.: Des. Constr.* **7**(3), 58–62 (1985).
25. K. Zayat and Z. Bayasi, *ACI Mater. J.* **93**(2), 178–181 (1996).
26. P. Soroushian, F. Aouadi, and M. Nagi, *ACI Mater. J.* **88**(1), 11–18 (1991).
27. B. Mobasher and C.Y. Li, *ACI Mater. J.* **93**(3), 284–292 (1996).
28. N. Banthia, A. Moncef, K. Chokri, and J. Sheng, *Can. J. Civ. Eng.* **21**(6), 999–1011 (1994).
29. B. Mobasher and C.Y. Li, *Infrastructure: New Materials and Methods of Repair*, Proc. Mater. Eng. Conf., No. 804, pp. 551–558. ASCE, New York, 1994.
30. P. Soroushian, M. Nagi, and J. Hsu, *ACI Mater. J.* **89**(3), 267–276 (1992).
31. P. Soroushian, *Constr. Specifier* **43**(12), 102–108 (1990).
32. A.K. Lal, *Batiment Int./Build. Res. Pract.* **18**(3), 153–161 (1990).
33. S.B. Park, B.I. Lee, and Y.S. Lim, *Cem. Concr. Res.* **21**(4), 589–600 (1991).
34. S.B. Park and B.I. Lee, *High Temp.—High Press.* **22**(6), 663–670 (1990).
35. P. Soroushian, M. Nagi, and A. Okwuegbu, *ACI Mater. J.* **89**(5), 491–494 (1992).
36. M. Pigeon, M. Azzabi, and R. Pleau, *Cem. Concr. Res.* **26**(8), 1163–1170 (1996).
37. N. Banthia, K. Chokri, Y. Ohama, and S. Mindess, *Adv. Cem. Based Mater.* **1**(3), 131–141 (1994).
38. N. Banthia, C. Yan, and K. Sakai, *Cem. Concr. Compos.* **20**(5), 393–404 (1998).
39. T. Urano, K. Murakami, Y. Mitsui, and H. Sakai, *Composites, Part A* **27**(3), 183–187 (1996).
40. A. Ali and R. Ambalavanan, *Indian Concr. J.* **72**(12), 669–675.
41. P. Chen, X. Fu, and D.D.L. Chung, *Cem. Concr. Res.* **25**(3), 491–496 (1995).
42. M. Zhu and D.D.L. Chung, *Cem. Concr. Res.* **27**(12), 1829–1839 (1997).
43. M. Zhu, R.C. Wetherhold, and D.D.L. Chung, *Cem. Concr. Res.* **27**(3), 437–451 (1997).
44. P. Chen and D.D.L. Chung, *Smart Mater. Struct.* **2**, 22–30 (1993).
45. P. Chen and D.D.L. Chung, *Composites, Part B* **27B**, 11–23 (1996).
46. P. Chen and D.D.L. Chung, *J. Am. Ceram. Soc.* **78**(3), 816–818 (1995).
47. D.D.L. Chung, *Smart Mater. Struct.* **4**, 59–61 (1995).
48. P. Chen and D.D.L. Chung, *ACI Mater. J.* **93**(4), 341–350 (1996).
49. X. Fu and D.D.L. Chung, *Cem. Concr. Res.* **26**(1), 15–20 (1996).
50. X. Fu, E. Ma, D.D.L. Chung, and W.A. Anderson, *Cem. Concr. Res.* **27**(6), 845–852 (1997).
51. X. Fu and D.D.L. Chung, *Cem. Concr. Res.* **27**(9), 1313–1318 (1997).
52. X. Fu, W. Lu, and D.D.L. Chung, *Cem. Concr. Res.* **28**(2), 183–187 (1998).
53. S. Wen and D.D.L. Chung, *Cem. Concr. Res.* **30**(8), 1289–1294 (2000).
54. Z. Shi and D.D.L. Chung, *Cem. Concr. Res.* **29**(3), 435–439 (1999).
55. Q. Mao, B. Zhao, D. Sheng, and Z. Li, *J. Wuhan Univ. Tech., Mater. Sci. Ed.* **11**(3), 41–45 (1996).
56. Q. Mao, B. Zhao, D. Shen, and Z. Li, *Fuhe Cailiao Xuebao* **13**(4), 8–11 (1996).
57. M. Sun, Q. Mao, and Z. Li, *J. Wuhan Univ. Tech., Mater. Sci. Ed.* **13**(4), 58–61 (1998).
58. B. Zhao, Z. Li, and D. Wu, *J. Wuhan Univ. Tech., Mater. Sci. Ed.* **10**(4), 52–56 (1995).
59. S. Wen and D.D.L. Chung, *Cem. Concr. Res.* **29**(6), 961–965 (1999).
60. M. Sun, Z. Li, Q. Mao, and D. Shen, *Cem. Concr. Res.* **28**(4), 549–554 (1998).
61. M. Sun, Z. Li, Q. Mao, and D. Shen, *Cem. Concr. Res.* **28**(12), 1707–1712 (1998).
62. S. Wen and D.D.L. Chung, *Cem. Concr. Res.* **29**(12), 1989–1993 (1999).
63. D. Bontea, D.D.L. Chung, and G.C. Lee, *Cem. Concr. Res.* **30**(4), 651–659 (2000).
64. S. Wen and D.D.L. Chung, *Cem. Concr. Res.* **30**(12), 1979–1982 (2000).
65. J. Lee and G. Batson, in *Materials for the New Millennium*, Proc. 4th Mater. Eng. Conf., Vol. 2, pp. 887–896. ASCE, New York, 1996.
66. X. Fu and D.D.L. Chung, *ACI Mater. J.* **96**(4), 455–461 (1999).
67. Y. Xu and D.D.L. Chung, *Cem. Concr. Res.* **29**(7), 1117–1121 (1999).
68. Y. Shinozaki, *Natl. SAMPE Tech. Conf.* **22**, 986–997 (1990).
69. X. Fu and D.D.L. Chung, *Cem. Concr. Res.* **25**(4), 689–694 (1995).
70. J. Hou and D.D.L. Chung, *Cem. Concr. Res.* **27**(5), 649–656 (1997).
71. G.G. Clemena, *Materi. Perform.* **27**(3), 19–25 (1988).
72. R.J. Brousseau and G.B. Pye, *ACI Mater. J.* **94**(4), 306–310 (1997).

73. P. Chen and D.D.L. Chung, *Smart Mater. Struct.* **2**, 181–188 (1993).
74. P. Chen and D.D.L. Chung, *J. Electron. Mater.* **24**(1), 47–51 (1995).
75. X. Wang, Y. Wang, and Z. Jin, *Fuhe Cailiao Xuebao* **15**(3), 75–80 (1998).
76. N. Banthia, S. Djeridane, and M. Pigeon, *Cem. Concr. Res.* **22**(5), 804–814 (1992).
77. P. Xie, P. Gu, and J.J. Beaudoin, *J. Mater. Sci.* **31**(15), 4093–4097 (1996).
78. Z. Shui, J. Li, F. Huang, and D. Yang, *J. Wuhan Univ. Tech., Mater. Sci. Ed.* **10**(4), 37–41 (1995).
79. X. Fu and D.D.L. Chung, *Cem. Concr. Res.* **28**(6), 795–801 (1998).
80. X. Fu and D.D.L. Chung, *Carbon* **36**(4), 459–462 (1998).
81. X. Fu and D.D.L. Chung, *Cem. Concr. Res.* **26**(10), 1467–1472 (1996).
82. X. Fu and D.D.L. Chung, *Cem. Concr. Res.* **27**(2), 314 (1997).
83. T. Fujiwara and H. Ujie, *Tohoku Kogyo Daigaku Kiyo, 1: Ritokogakuhen* **7**, 179–188 (1987).
84. Y. Shimizu, A. Nishikata, N. Maruyama, and A. Sugiyama, *Terebijon Gakkaishi* **40**(8), 780–785 (1986).
85. P. Chen and D.D.L. Chung, *ACI Mater. J.* **93**(2), 129–133 (1996).
86. T. Uomoto and F. Katsuki, *Doboku Gakkai Rombun-Hokokushu* **490**(pt. 5–23), 167–174 (1994–1995).
87. C.M. Huang, D. Zhu, C.X. Dong, W.M. Kriven, R. Loh, and J. Huang, *Ceram. Eng. Sci. Proc.* **17**(4), 258–265 (1996).
88. C.M. Huang, D. Zhu, X. Cong, W.M. Kriven, R.R. Loh, and J. Huang, *J. Am. Ceram. Soc.* **80**(9), 2326–2332 (1997).
89. T.-J. Kim and C.-K. Park, *Cem. Concr. Res.* **28**(7), 955–960 (1998).
90. S. Igarashi and M. Kawamura, *Doboku Gakkai Rombun-Hokokushu* **502**(pt. 5–25), 83–92 (1994).
91. M.Z. Bayasi and J. Zeng, *ACI Struct. J.* **94**(4), 442–446 (1997).
92. G. Campione, S. Mindess, and G. Zingone, *ACI Mater. J.* **96**(1), 27–34 (1999).
93. T. Yamada, K. Yamada, and K. Kubomura, *J. Compos. Mater.* **29**(2), 179–194 (1995).
94. S. Delvasto, A.E. Naaman, and J.L. Throne, *Int. J. Cem. Compos. Lightweight Concr.* **8**(3), 181–190 (1986).
95. C. Park, *Nippon Seramikkusu Kyokai Gakujutsu Ronbunshi* **106**(1231), 268–271 (1998).
96. Y. Shao, S. Marikunte, and S.P. Shah, *Concr. Int.* **17**(4), 48–52 (1995).
97. S. Wen and D.D.L. Chung, *Cem. Concr. Res.* **30**(4), 661–664 (2000).
98. S. Wen and D.D.L. Chung, *Cem. Concr. Res.* **31**(3), 507–510 (2001).
99. Y. Xu and D.D.L. Chung, *Cem. Concr. Res.* **30**(7), 1175–1178 (2000).
100. Y. Xu and D.D.L. Chung, *ACI Mater. J.* **97**(3), 333–342 (2000).
101. Y. Wang and D.D.L. Chung, *Cem. Concr. Res.* **28**(10), 1353–1356 (1998).
102. S. Wen and D.D.L. Chung, *Cem. Concr. Res.* **30**(2), 327–330 (2000).
103. Y. Xu and D.D.L. Chung, *Cem. Con. Res.* **29**(7), 1107–1109 (1999).
104. W.F.A. Davies, *J. Phys., D* **7**, 120–130 (1974).
105. W.J. Gadjia, Report RADC-TR-78-158, A059029. 1978.
106. P. Li, W. Strieder, and T. Joy, *J. Comp. Mater.* **16**, 53–64 (1982).
107. T. Choi, P. Ajmera, and W. Strieder, *J. Comp. Mater.* **14**, 130–141 (1980).
108. T. Joy and W. Strieder, *J. Comp. Mater.* **13**, 72–78 (1979).
109. K.W. Tse and C.A. Moyer, *Mater. Sci. Eng.* **49**, 41–46 (1981).
110. V. Volpe, *J. Comp. Mater.* **14**, 189–198 (1980).
111. V.G. Shevchenko, A.T. Ponomarenko, and N.S. Enikolopyan, *Int. J. Appl. Electromag. Mater.* **5**(4), 267–277 (1994).
112. X.B. Chen and D. Billaud, *Ext. Abstr. Program—Bienn. Conf. Carbon* **20**, 274–275 (1991).
113. X. Wang and D.D.L. Chung, *Compos. Interfaces* **5**(3), 191–199 (1998).
114. X. Wang and D.D.L. Chung, *Polym. Compos.* **18**(6), 692–700 (1997).
115. X. Wang and D.D.L. Chung, *Carbon* **35**(5), 706–709 (1997).
116. X. Wang and D.D.L. Chung, *J. Mater. Res.* **14**(11), 4224–4229 (1999).
117. S. Wang and D.D.L. Chung, *Compos. Interfaces* **6**(6), 497–506 (1999).
118. S. Wang and D.D.L. Chung *Polym. Composites* **22**(1), 42–46 (2001).
119. X. Luo and D.D.L. Chung, *Compos. Sci. Technol.* **61**, 885–888 (2001).
120. N. Muto, H. Yanagida, T. Nakatsuji, M. Sugita, Y. Ohtsuka, and Y. Arai, *Smart Mater. Struct.* **1**, 324–329 (1992).
121. X. Wang, X. Fu, and D.D.L. Chung, *J. Mater. Res.* **14**(3), 790–802 (1999).
122. X. Wang and D.D.L. Chung, *Composites: Part B* **29B**(1), 63–73 (1998).
123. P.E. Irving and C. Thiogarajan, *Smart Mater. Struct.* **7**, 456–466 (1998).
124. X. Wang, S. Wang, and D.D.L. Chung, *J. Mater. Sci.* **34**(11), 2703–2714 (1999).
125. S. Wang and D.D.L. Chung, *Polym. Compos.* (in press).
126. N. Muto, H. Yanagida, M. Miyayama, T. Nakatsuji, M. Sugita, and Y. Ohtsuka, *J. Ceram. Soc. Jpn.* **100**(4), 585–588 (1992).
127. N. Muto, H. Yanagida, T. Nakatsuji, M. Sugita, Y. Ohtsuka, Y. Arai, and C. Saito, *Adv. Compos. Mater.* **4**(4), 297–308 (1995).
128. R. Prabhakaran, *Exp. Tech.* **14**(1), 16–20 (1990).
129. M. Sugita, H. Yanagida, and N. Muto, *Smart Mater. Struct.* **4**(1A), A52–A57 (1995).
130. A.S. Kaddour, F.A.R. Al-Salehi, S.T.S. Al-Hassani, and M.J. Hinton, *Compos. Sci. Tech.* **51**, 377–385 (1994).
131. O. Ceysson, M. Salvia, and L. Vincent, *Ser. Mater.* **34**(8), 1273–1280 (1996).
132. K. Schulte and C. Baron, *Compos. Sci. Technol.* **36**, 63–76 (1989).
133. K. Schulte, *J. Phys. IV, Colloq. (Orsay, Fr.)* **3**(C7), 1629–1636 (1993).
134. J.C. Abry, S. Bochard, A. Chateauminois, M. Salvia, and G. Giraud, *Compos. Sci. Technol.* **59**(6), 925–935 (1999).
135. A. Tedoroki, H. Kobayashi, and K. Matuura, *JSME Int. J. Ser. A* **38**(4), 524–530 (1995).
136. S. Hayes, D. Brooks, T. Liu, S. Vickers, and G.F. Fernando, *Proc. SPIE, Smart Struct. Mater.* **2718**, 376–384 (1996).
137. S. Wang and D.D.L. Chung, *Polym. & Polym. Compos.* **9**(2), 135–140 (2001).

138. S. Wang and D.D.L. Chung, *Composites: Part B* **30**(6), 591–601 (1999).
139. S. Wang and D.D.L. Chung, *Compos. Interfaces* **6**(6), 519–530 (1999).
140. S.W. Hudnut and D.D.L. Chung, *Carbon* **33**(11), 1627–1631 (1995).
141. M. Segiet and D.D.L. Chung, *Compos. Interfaces* **7**(4), 257–276 (2000).
142. B. Benchechou, M. Coni, H.V.C. Howarth, and R.G. White, *Composites: Part B* **29B**, 809–817 (1998).
143. I.C. Finegan and R.F. Gibson, *Compos. Struct.* **44**(2), 89–98 (1999).
144. H.L. Wettergren, *J. Compos. Mater.* **32**(7), 652–663 (1998).

COMPOSITES, SURVEY

K.H. SEARLES

Oregon Graduate Institute of
Science and Technology
Beaverton, OR

INTRODUCTION

The definition of a composite material, loosely translated, could point to any material having more than a single phase such that the multiple phases act in a synergistic manner to provide a behavior or set of behaviors more desirable than exhibited singularly. By this definition, a significant amount of creative space is available for scientists and engineers alike to develop exciting new materials, structures, processes, and applications to meet the demands dictated by the evolution of new technologies. As this evolution progresses, the traditional distinction made between the five classes of materials (i.e., metals, ceramics, polymers, elastomers, and glasses) tends to fade. It is here where many new possibilities arise to intelligently process and “smear” these classes together to produce new, smarter composite materials and engineering structures that “think” and respond to stimuli. The paradigm for this is by way of the tailorable structure-performance relationships for composites based on interactions between constituent materials, which may encompass any of the five material classes.

ENGINEERING MATERIALS

The number of materials available to scientists and engineers for translating ideas into usable products and markets is vast, if not almost overwhelming. Current estimations give a range somewhere on the order of between 40,000 and 80,000 engineering materials at our disposal (1). As new materials are developed having novel and exploitable characteristics (properties and responses), the number of choices further expands. The appropriate choice(s) of material(s) for design can be just as overwhelming as the number of materials available. How does one appropriately choose from the vast menu? Through proper rationalizations and design considerations, the search and selection processes can be drastically minimized. Many of

the important product and market design considerations are as follows: safety, strength, weight, cost, reliability, life cycle, processing, wear resistance, corrosion resistance, friction, lubrication, stiffness, flexibility, volume, shape, styling, surface finish (appearance), control, maintenance, utility, and liability.

An adequate methodology of materials selection for design cannot be addressed without considering the function, material, process, and shape. The function, ultimately governed by material behavior, is confined by shape. Restrictions on the ability to achieve desired shapes are placed by limitations in material processing. Considerations in shape and geometry include both the micro- and macroscales. Internal two-dimensional and three-dimensional structures such as interleaves, cells, and honeycombs collectively determine, to an extent, the external shape. In most materials, presence or absence of these structures also influences material behavior (response), ultimately leading back to the desired function or set of functions. As illustrated by Fig. 1, the interactions between function, material, process, and shape are bidirectional. At the center of these interactions, an iterative design process utilizes the processing, materials, and design optimization sciences in forming an integrated system to arrive at a final product and market.

Materials Evolution

Prior to 5000 BC, the engineering materials that defined the peak of technology were natural composites, polymers, ceramics, and glasses. Little metal was consumed and only in the form of naturally occurring gold and silver. Between 5000 BC and AD 500 the increased consumption of copper, bronze, and iron served to mark respective technological ages. However, wood, skins, natural fibers, rubber, flint, and pottery were still the primary sources for materials usage. A perspective given by Ashby (1986) illustrates the relative importance and evolution of advanced materials for mechanical and civil engineering purposes (1).

It is evident from Fig. 2 that engineering steels did not gain prominence until 1850, and from then to the mid-1900s, steels and their alloys had been the preeminent material sources for design. Nonmetal materials were rarely thought of for use as structural materials, perhaps with the exceptions being refractories and Portland cement. Even so, their consumption compared to that of total material consumption was small. Within the last 20 to 30 years though, newly developed materials belonging to the other classes (polymers and elastomers, ceramics, composites) have surfaced due to their increasing technological importance. The growth rate of many of these materials is on par with (if not exceeding) the growth rate enjoyed by steel during the height of the Industrial Revolution.

Demand for New Materials

Several factors must be taken into account regarding the choice of material substitution when designing a part. A considerable driving force behind this is weight savings. However, part functionality must be critically reviewed if weight reduction is obtained through the substitution of a lightweight material for a conventional one. For example,

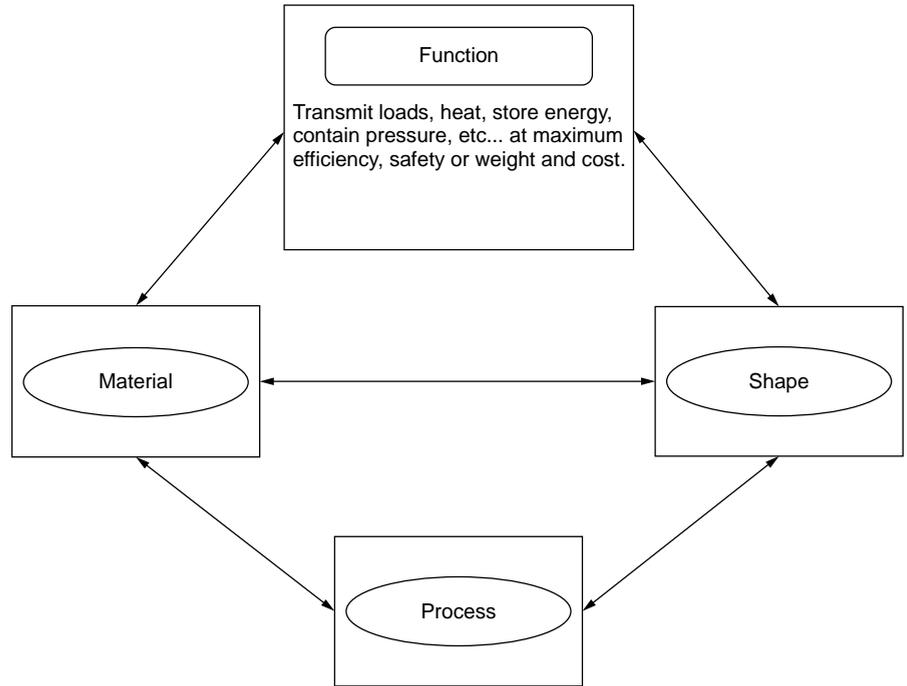


Figure 1. Bidirectional interactions between function, material, shape and process (1).

in a study of materials for ground transportation discussed by Compton and Gjostein (2,3), substitution of a hypothetical aluminum part of the same volume for a cast iron part would result in a 63% weight savings. If equal load sharing were maintained, then weight savings dropped from 63% to 56%. A similar comparison made between mild steel and high-strength steel revealed a weight savings of 18% where structural strength was the main concern, though mild steel is really no lighter than high strength steel by volume.

For high-strength steel replacements using aluminum or fiber-reinforced plastics (FRPs), savings in weight is much smaller when equivalent tensile, compressive, and bending stiffnesses are necessary. Additional to weight savings, another technological aspect of functionality as equally important pertains to the demand of elevated temperature environments. In particular to the commercial aircraft, military aircraft, and aerospace sectors, “skin” temperatures from friction with air have risen 20 fold,

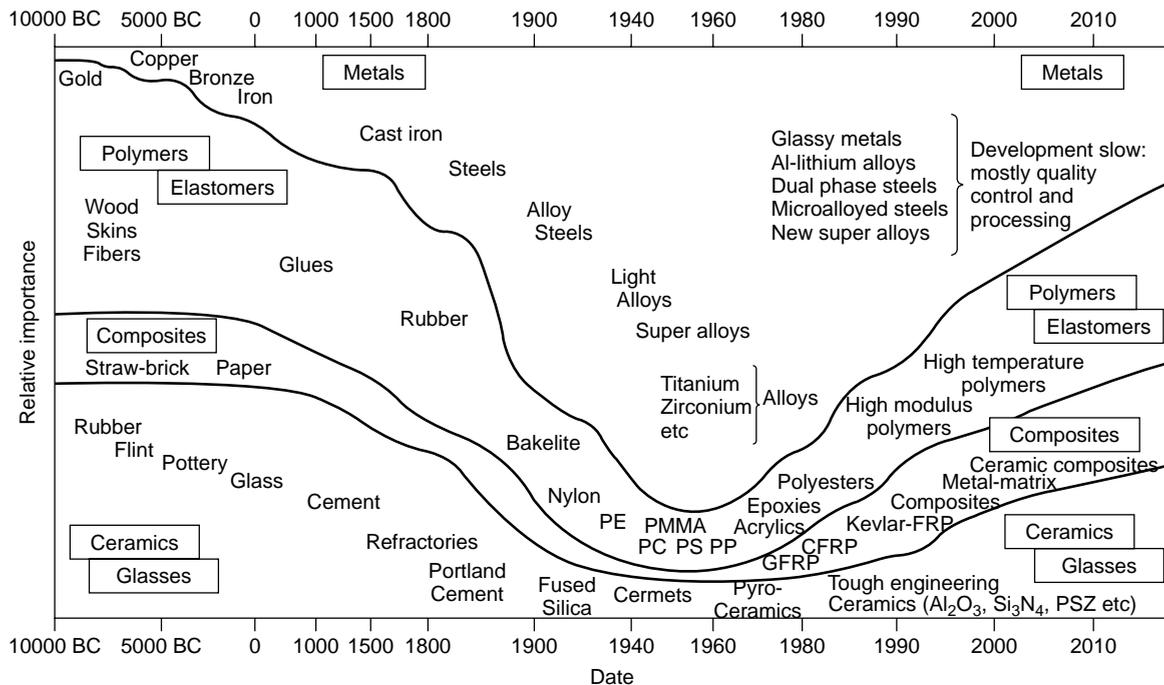


Figure 2. Relative importance and evolution of civil and mechanical engineering materials (1).

from a modest 50°C (120°F), as exhibited by early 1900s trainers, to in excess of 1000°C (1800°F), as exhibited by the space shuttle. Many of the aircraft and aerospace powerplant temperatures have risen by similar magnitudes. In both of these examples, the boundaries of functionality were obviously broadened and more ingenuity in the application and consolidation of materials was required.

Bulk metals and metallic alloys possess high elastic moduli and can be made extremely strong through various heat-treating and mechanical processes. They also remain relatively ductile, permitting shaping operations by deformation. However, they are the most susceptible materials to corrosion attack, and their ductility exposes them to failure by fatigue.

Ceramics and some glasses have relatively high elastic moduli, but unlike bulk and alloyed metallics, they are very brittle. The strengths of structural ceramics and glasses are statistically dependent on volume, consequently these materials are termed “net shape” meaning little tolerance is available for changes in geometry without manufacturing an entirely new component. They do function very well in hostile environments where temperatures, wear, and corrosion are excessive, but brittle behavior tends to foster catastrophic failure from concentrated stresses.

Elastomeric and polymeric materials have elastic moduli 40 to 50 times lower than metals, but they are as strong, if not stronger than metals. They usually exhibit much higher strain to failure and properties that are much more dependent on temperature with a useful limit of 200°C (390°F). Some thermosets and thermoplastics can extend the limit by an additional 100°C (200°F), but this is typically accompanied by increased brittle behavior. There are added benefits associated with processing and designing with polymers in that additional surface finishing is usually not necessary, resistance to corrosion is favorable, more complicated shapes are easy to form, and large deflections foster component designs that are flexible yet geometrically stable.

Contrary to the other material classes, composites tend to exhibit microscopic and macroscopic heterogeneity and behave anisotropically. That is, composite mechanical properties vary from point to point due to the intrinsic properties of the reinforcements and their orientations. From the standpoint of design, predicting responses to external mechanical and thermal loads becomes more complicated. With the exception of mainly thermosetting polymer or nonpolymer matrices specifically tailored for elevated temperature environments, composites are also of limited use structurally above 250°C (480°F). Given this, composites do avoid some of the drawbacks associated with the other material classes while necessarily combining the attractive properties.

By incorporating various materials from principally diverse classes into the synthetically derived composite materials, several essential characteristics may be drawn upon: the corrosion resistance of polymers, the high strength and ductility, the light weight and lower cost of fabrication, the high temperature performance of ceramics and the thermal-electrical conductivity of metals (3). As shown by Fig. 3, there is merit in grouping materials that show commonalities in properties and process.

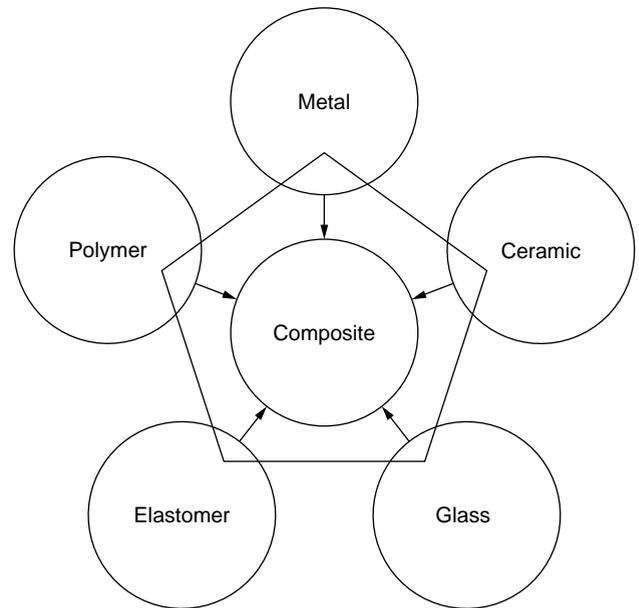


Figure 3. Commonality of composite material processing and properties to the principal classes of engineering materials (1).

COMPOSITE MATERIALS

What are composite materials? If we adopt the generality that a composite material is any material consisting of some combination of two or more phases, then almost every material in the universe that exists naturally or synthetically may be referred to as a composite. This generality leads to a very broad classification of materials that are composites. For example, wood can be classified as a composite, since its honeycomb microstructure consists of arrays of fibers (cellulose) encompassed and held together by a polymer matrix (lignin) as shown in Fig. 4. At the macroscale, numerous combinations of the 60 native and 30 imported species of wood are possible for tailoring performance to suit a particular need (4). In many cases, the macroscale wood composites serve as excellent bulk engineering materials (plywood, particle board, etc.).

Other, less-tailorable, natural composite materials and biomaterials might also include bamboo, bone, and muscular tissues. Bamboo, referred to as “nature’s own fiberglass,” has an aligned fibrillar structure and exhibits a “broomed” fracture morphology similar to a failed glass fiber composite (5). The microstructure of muscular tissues is such that flexibility occurs along with a high degree of strength. Fast- or slow-twitch fibers (collagen) are aligned in the general directions of loading along the bone axes and surrounded by a continuous matrix having lower stiffness, thus allowing neighboring fibers to slide past one another.

Bulk engineering materials such as metal alloys and plastics are traditionally not thought of as composite materials. However, on the microscale, these materials do lie well within the bounds given by the previous generality. Consider a heavily utilized bulk engineering material like plain carbon, hypoeutectoid steel (up to 0.8% carbon). Although the separable phases require some form of microscopy to resolve, they exist as fine dispersions that

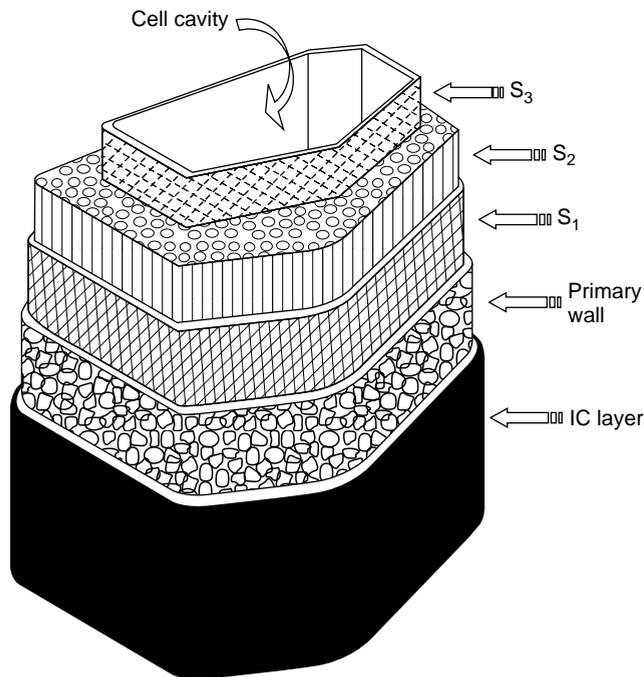


Figure 4. Lamellar composite structure of wood cell composed of the primary wall cellulose fibers, a linear polymer in an irregular network, and the 3-layer secondary wall: S_1 —crisscross network, S_2 —spiral-type network, S_3 —irregular network. The matrix is formed from the lignin and hemicellulose deposits.

act synergistically with temperature to control structural behavior under equilibrium conditions. Upon cooling to below 800°C (1472°F), pearlite forms, which is a *lamellar* composite of the polyphases α + carbide. The layers of soft ductile alpha iron are interleaved with the layers of hard brittle iron carbide (cementite or Fe_3C), yielding a tough, strong steel. Similarly, an engineering plastic like acrylonitrile-butadiene-styrene (ABS) is a prominent copolymer with rubber particles finely dispersed within the styrene-acrylonitrile (SAN) phase and, thus, is a microscale composite material. The acrylonitrile provides surface hardness and heat resistance, while styrene contributes to strength and butadiene improves toughness and resistance to impact.

A succinct definition in the classical sense of what composite materials are is a prerequisite to understanding the structure-performance relationships that are possible and benefiting from the ability to tailor material performance through processing to suit end-use requirements. Essentially, particulate and filamentary, namely fibrous, composite materials may be viewed as selective arrangements of hybrid materials of which their performance in structural and nonstructural applications is controlled through calculated variations of internal compositions and architectures. More adequately defined, acceptable composites possess the following characteristics (5):

- Consist of two or more physically distinct and mechanically separable phases.
- Can be dispersed or mixed in a controlled manner to achieve a desired behavior.

- Have properties that are necessarily superior to those of the constituents and possibly unique.

Clearly, from the foregoing definition, differentiating between the constituents (e.g., particulates, short fibers, and continuous fibers), individual composite layers (laminae), and laminated composite structures ultimately becomes a matter of scale.

Fundamental to the treatment of composite material and structure behaviors, it is often convenient to study the material interactions in terms of three scales, namely the microscale (fibers, matrices, fiber-matrix interface), mesoscale (lamina architectures), and macroscale (geometric arrangement of laminae). At the macroscale, the laminated composite and structural engineering or nonstructural components are often synonymous. What follows is a review of composite principles and behaviors in terms of the material interactions occurring at each of these scales, beginning with the microscale interactions.

MICROSCALE BEHAVIOR

Fibers, Fillers, and Matrix

Composites are nominally a synergistic combination of two phases, the reinforcing phase and the matrix phase, with likely combinations of polymer matrix and polymer (e.g., epoxy-Kevlar), polymer matrix and ceramic (e.g., fiberglass), ceramic matrix and ceramic (e.g., Al_2O_3 -SiC), or metal matrix and ceramic (e.g., WC-Co). The matrix phase is a continuous medium wherein the reinforcing phase is uniformly dispersed. The majority of reinforcements are in the form of either continuous aligned fibers or chopped random fibers and dispersions or particles. Other, less frequently used reinforcing materials include ribbons and flakes. Reinforcements vary in cross-sectional shape, with characteristic sizes ranging from circa $1.0\ \mu\text{m}$ (3.94×10^{-2} mils) to $20.0\ \mu\text{m}$ (0.78 mils). Since their diameters are small, the fibers, ribbons, or particulates cannot form a structure singularly unless bounded by a matrix that is geometrically stable.

Composites are often categorized according to the form of the reinforcements: *dispersion-strengthened*, *particle-reinforced*, and *fiber-reinforced*. Dispersion-strengthened composites have small particles dispersed in a matrix which is the primary load-bearing constituent. Particle-reinforced composites have larger particles incorporated within the matrix and the load is shared equally between the particles and matrix. Fiber-reinforced composites have either continuous (perhaps practically continuous) or chopped fibers incorporated in a matrix and the fibers are the primary load-bearing constituent.

Several materials are common to each form of composite reinforcement. For fiber-reinforcements the available selection of traditionally utilized materials includes tungsten, titanium, aluminum, boron, carbon [graphite from polyacrylonitrile (PAN) precursors], aramid [Kevlar (polyphenylene terephthalamide)], and glass (E, S, C, and ECR grades) with compositions of SiO_2 , Al_2O_3 , Fe_2O_3 , CaO, MgO, Na_2O , K_2O , BaO, and Ba_2O_3 . For ribbon, flake, and particle reinforcements, the available selection of

Table 1. Selected Material Properties for Several Materials Commonly Used as Constituent Reinforcing Phases in MMCs, CMCs, and PMCs

Reinforcing Material	Units	Tensile Modulus (E_{11})	Tensile Strength (σ_{11})	Units	Density (ρ)
S. Steel	GPa (Msi)	203.0 (29.4)	2.1 (0.31)	kg/m ³ (lb/ft ³)	7900.0 (492.9)
Be	GPa (Msi)	315.0 (45.7)	1.3 (0.19)	kg/m ³ (lb/ft ³)	1800.0 (112.3)
Mo	GPa (Msi)	343.0 (49.7)	2.1 (0.31)	kg/m ³ (lb/ft ³)	10,300.0 (642.7)
W	GPa (Msi)	350.0 (50.8)	3.9 (0.57)	kg/m ³ (lb/ft ³)	19,300.0 (1,204.3)
β -Ti	GPa (Msi)	119.0 (17.3)	2.3 (0.33)	kg/m ³ (lb/ft ³)	4600.0 (287.0)
Al ₂ O ₃	GPa (Msi)	470.0 (68.2)	2.0 (0.29)	kg/m ³ (lb/ft ³)	3960.0 (247.1)
B	GPa (Msi)	385.0 (55.8)	7.0 (1.02)	kg/m ³ (lb/ft ³)	2600.0 (162.2)
BN	GPa (Msi)	90.0 (13.05)	1.4 (0.20)	kg/m ³ (lb/ft ³)	1900.0 (118.6)
Carbon (PAN-Type 2)	GPa (Msi)	250.0 (36.3)	2.7 (0.39)	kg/m ³ (lb/ft ³)	1750.0 (109.2)
Graphite	GPa (Msi)	490.0 (71.1)	3.2 (0.46)	kg/m ³ (lb/ft ³)	1900.0 (118.6)
Kevlar 49	GPa (Msi)	125.0 (18.1)	3.2 (0.46)	kg/m ³ (lb/ft ³)	1450.0 (90.5)
S-glass	GPa (Msi)	72.0 (10.4)	6.0 (0.87)	kg/m ³ (lb/ft ³)	2500.0 (156.0)
E-glass	GPa (Msi)	84.0 (12.2)	4.6 (0.67)	kg/m ³ (lb/ft ³)	2550.0 (159.1)
SiC	GPa (Msi)	380.0 (55.1)	2.8 (0.41)	kg/m ³ (lb/ft ³)	2700.0 (168.5)
Si ₃ N ₄	GPa (Msi)	380.0 (55.1)	1.0–10.0 (0.15–1.5)	kg/m ³ (lb/ft ³)	3800.0 (237.1)

Sources: Data after Courtney (p. 224); Hull (5, p. 14).

materials includes glass, boron and graphite films, silicon carbide, and mica and aluminum of boride. The list of usable materials for selection of a constituent matrix phase is vast, and composites are usually grouped into the material classes in which the constituent matrix belongs.

Many polymer matrix composites (PMCs) incorporate the following thermoplastic and thermoset polymers as constituent matrices: polyethylenes (PE), polyurethanes (PU), polyamides (PA-Nylon 6,6), polyether-etherketones (PEEK), polycarbonates (PC), epoxies, polyesters, and polyimides (PI), and bismaleimides (BMI). Ceramic matrix composites (CMCs) may incorporate the following advanced structural ceramics as constituent matrices: silica oxides (SiO₂), alumina (Al₂O₃), metallic carbides (W, Ta, Hf, Zr, Mo, V, Ti, Cr), nonmetallic carbides (B, Si), borides (Ta, V, Cr, W, Ti, C), metallic nitrides (Zr, Y, Mo, Ti, Cr), and nonmetallic nitrides (Al, B, Si). For metal matrix composites (MMCs), materials incorporated as constituent matrices include titanium, copper, aluminum, and nickel.

Specific material properties for some of the most common reinforcing and matrix constituents are given in Tables 1 and 2, respectively.

Fiber-reinforced PMC's are achieving widespread acceptance, particularly in the aerospace industry. Composites based on polymer epoxy matrices currently account for about 4% by weight of commercial aircraft and about 10% of military aircraft with forecasts spanning the next decade of 65% by weight (6). These composites possess a high specific strength, stiffness, and very large endurance to fatigue. The room temperature specific strengths of epoxy PMCs are higher than aluminum, titanium alloys, and some super alloys. The upper service temperature of these materials is limited to approximately 150°C (302°F). In many aerospace PMC applications such as aircraft skins and engines, however, temperatures often exceed 200°C (392°F) and could reach 427°C (800°F). The strength of epoxy-based PMCs would begin to diminish far below these temperatures, hence there is a growing demand for other

Table 2. Selected Material Properties for Several Materials Commonly Used as Constituent Matrix Phases in MMCs, CMCs, and PMCs

Matrix Material	Units	Tensile Modulus (E_{11})	Tensile Strength (σ_{11})	Units	Density (ρ)
Al ₂ O ₃	GPa (Msi)	470.0 (68.2)	2.0 (0.29)	kg/m ³ (lb/ft ³)	3960.0 (247.1)
BN	GPa (Msi)	90.0 (13.05)	1.4 (0.20)	kg/m ³ (lb/ft ³)	1900.0 (118.6)
SiC	GPa (Msi)	380.0 (55.1)	2.8 (0.41)	kg/m ³ (lb/ft ³)	2700.0 (168.5)
Si ₃ N ₄	GPa (Msi)	380.0 (55.1)	1.0–10.0 (0.15–1.5)	kg/m ³ (lb/ft ³)	3800.0 (237.1)
Epoxy	MPa (ksi)	4500.0 (652.7)	35.0–100.0 (5.1–14.5)	kg/m ³ (lb/ft ³)	1250.0 (78.0)
Polyester	MPa (ksi)	3250.0 (471.4)	40.0–90.0 (5.8–13.1)	kg/m ³ (lb/ft ³)	1350.0 (84.2)
Nylon 66	MPa (ksi)	2100.0 (304.6)	60.0–75.0 (8.7–10.9)	kg/m ³ (lb/ft ³)	1140.0 (71.1)
Polycarbonate	MPa (ksi)	2300.0 (333.6)	45.0–70.0 (6.5–10.2)	kg/m ³ (lb/ft ³)	1130.0 (70.5)
Polypropylene	MPa (ksi)	1200.0 (174.0)	25.0–38.0 (3.6–5.5)	kg/m ³ (lb/ft ³)	900.0 (56.2)
Polyethylene-HD	MPa (ksi)	827.4 (120.0)	27.6 (4.0)	kg/m ³ (lb/ft ³)	950.0 (59.3)
Urethane	MPa (ksi)	—	34.5 (5.0)	kg/m ³ (lb/ft ³)	1200.0 (74.9)
Polyvinylchloride	MPa (ksi)	2757.9 (400.0)	41.4 (6.0)	kg/m ³ (lb/ft ³)	1400.0 (87.4)
Polyetheretherketone	MPa (ksi)	3650.0 (529.4)	92.0 (13.3)	kg/m ³ (lb/ft ³)	1320.0 (82.4)
ABS (Acrylonitrile-butadiene-styrene)	MPa (ksi)	2068.4 (300.0)	27.6–48.3 (4.0–7.0)	kg/m ³ (lb/ft ³)	1060.0 (66.1)
Acrylic	MPa (ksi)	2895.8 (420.0)	55.2 (8.0)	kg/m ³ (lb/ft ³)	1190.0 (74.3)
Polyimide PMR-15	MPa (ksi)	4000.0 (580.2)	39.0 (5.7)	kg/m ³ (lb/ft ³)	1350.0 (84.2)

Sources: Data after Flinn and Trojan, (4, pp. 560–573); Hull (5, pp. 29–33); NASA LeRC; DuPont AMS.

candidate PMCs with much higher upper service temperature capabilities.

PMCs consolidated with thermosetting polyimides such as PMR (polymerization of monomer reactants), Avimid-R, AMB-21 and VCAP-75 are replacing epoxies for elevated temperature applications. PMR (e.g., PMR-15, PMR-II-50) and the other aforementioned polyimide-based composites have extremely good specific properties and are stable up to temperatures approaching 360°C (680°F).

Some of the more common applications for advanced PMR compositions include aer propulsion system components such as turbine engine thruster rings, reversers, compressor casings, bearings, and external structural surfaces for supersonics. Materials used in these applications must possess a good balance of mechanical properties over a wide temperature range, withstand thermal shock and experience a low percentage of weight loss (thermal-oxidative stability, TOS) even after long-term exposure (60,000 h) at or near upper service temperatures. To maintain sufficient mechanical properties, the polymer glass transition temperature (T_g) must be at least 25°C (77°F) higher than its intended use temperature. Currently, the PMR compositions are limited for use in applications operating in the 200 to 300°C (392 to 572°F) temperature range (7).

Fiber-Matrix Interface

In any composite system where one phase is completely dispersed within another phase, continuity between phases is extremely important. When a continuum does not exist, the benefits associated with the combination of phases may not be fully realized. In fiber-reinforced and many particulate-reinforced systems, the matrix acts as a "conduit" for transporting load and sharing stresses equally within the reinforcing phase. Whether or not an equivalent distribution occurs or concentrations arise becomes highly dependent on the structure and properties of the interphase, namely the fiber-matrix interface (particulate is referred to as a fiber from this point on).

The interface acts as the medium by which stresses in the matrix are transported to the fiber, or alternatively, the significant differences between the elastic properties of the fibers and the matrix are communicated through the interface (5). How efficiently this communication occurs is governed primarily by the degree of bonding that takes place between the fibers and the matrix. Fiber pullout may result if the interfacial bonding is poor and if the interfacial bonding is too good, then energy absorption subsequent to the debonding process will be insufficient.

Ideally, debonding of the fiber-matrix interface should follow fracturing of the matrix (8). The nature and efficiency of bonding between a fiber and matrix will be dependent on the molecular conformation and chemistry of both constituents. Even though the interface is specific to each fiber-matrix system, insight can be given toward the origins of bond strength by way of the principal mechanisms governing adhesion, chiefly wetting, interdiffusion, electrostatics, chemical bonding and mechanical interlocking.

Contact and Wetting. Central to the study and application of adhesion wetting mechanisms, the Dupre equation

defines the work of adhesion in terms of the respective surface energies of two materials in intimate contact and is given by

$$W_A = \gamma_1 + \gamma_2 - \gamma_{12},$$

where W_A is the work of adhesion, γ_1 is the surface energy of material 1, γ_2 is the surface energy of material 2, and γ_{12} is the interfacial energy, that is, the energy necessary to create a unit area of interface (9). The surface energy γ of a material is highly dependent on the magnitude of intermolecular forces and can be defined as

$$\gamma = \frac{\pi n^2 A}{32r_0^2},$$

where A is a constant which scales the attractive interactions, n is the surface molecular density on either side of the interface, and r_0 is the equilibrium distance between the mating surfaces near intimate contact. For fiber-matrix interactions, it is also useful to consider the relationship between the stiffness of the material and its surface energy in terms of the isothermal Young's modulus:

$$E = \frac{32\gamma}{r_0}.$$

The isothermal Young's modulus E of a material may also be expressed by the following relationship:

$$E = r_0 \left(\frac{\partial \phi^{L-J}(r)}{\partial r} \right)_{T, r=r_0},$$

where ϕ^{L-J} represents the potential energy according to Lennard-Jones proposed interactions between atoms and molecules. While the work of adhesion W_A can be regarded as a thermodynamic parameter that describes the work necessary for two mutual surfaces to remain in intimate contact, the work of cohesion W_{COH} may be regarded as the work required to create two new surfaces of the same material as in the case of crack origination. If a monolithic material is broken such that the separated materials are completely nonenergy dissipating nor absorbing (perfectly brittle), then the work of cohesion is simply 2γ .

One might now discern that cohesion is adhesion with consideration for the respective materials in contact and their interfacial energy existing prior to a separation. By extending the concept of surface energies further, the wetting adhesion mechanism can be understood in terms of the Dupre equation and the Young equation. The Young equation relates the physical occurrence of a liquid drop on a solid surface. It states that

$$\gamma_{LV} \cos \theta = \gamma_{SV} - \gamma_{SL},$$

where γ_{ij} are the appropriate interfacial tensions between the liquid L , solid S , and vapor V and θ is the contact angle as shown in Fig. 5. The solid-vapor interfacial energy is not the true surface-free energy of the solid, but it is related

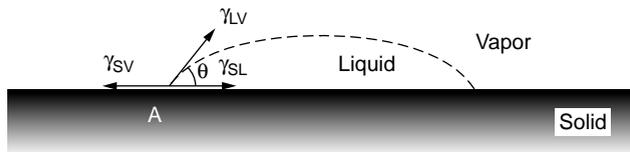


Figure 5. Schematic of contact angle representation where a liquid drop is perfectly placed on a smooth, rigid solid. The contact angle θ is measured at A: the three points between the liquid, the solid, and the vapor.

by the following equation:

$$\gamma_{SV} = \gamma_S - \pi_e.$$

Here γ_S is the true surface free energy and π_e is referred to as the equilibrium spreading pressure, a measure of released energy through adsorption of the vapor onto the solid surface. If combined, the Dupre equation with the Young equation yield a simple expression (Dupre-Young) that relates a thermodynamic parameter to two readily determined quantities:

$$W_A = \gamma_{LV}(1 + \cos \theta).$$

Knowing the liquid–vapor interfacial tension and the contact angle, the work of adhesion is easily calculated. In considering the wetting of a fiber by a matrix, only liquids with γ_{LV} less than the critical surface tension of wetting will spontaneously spread on the fiber. For example, epoxy has a surface energy of $.042 \text{ J m}^{-2}$ ($0.100 \times 10^{-1} \text{ cal in}^{-2}$), glass a surface energy of 0.560 J m^{-2} ($1.337 \times 10^{-1} \text{ cal in}^{-2}$), and graphite a surface energy of 0.077 J m^{-2} ($0.184 \times 10^{-1} \text{ cal in}^{-2}$). Although epoxy will wet both glass and graphite, the glass will wet more readily due to a much lower contact angle between the epoxy and glass. When fibers are considered for use with matrices having a greater surface energy, coatings and sizings are often introduced via polymers with lower energies to promote better wetting and adhesion characteristics.

Interdiffusion. The mechanism of adhesion by interdiffusion is thought to occur when the two surfaces that are brought into intimate contact are at least partially soluble in one another. If we adopt the philosophy that “like dissolves like,” then only a limited number of adhesive examples are applicable to a mechanism of adhesion by interdiffusion, unless coatings are used that are very similar as in some fiber-reinforced composite applications. In any case, diffusive bonding implies that an interphase forms between two materials that is a mixture of both materials as shown in Fig. 6. The extent to which the interphase is equally apportioned and continuous between the two materials depends on the solubility parameter criterion, defined as

$$\delta = \sqrt{\frac{E_{COH}}{V}},$$

where V is the molar volume and the cohesive energy E_{COH} , in this case, is the energy required to separate all of the

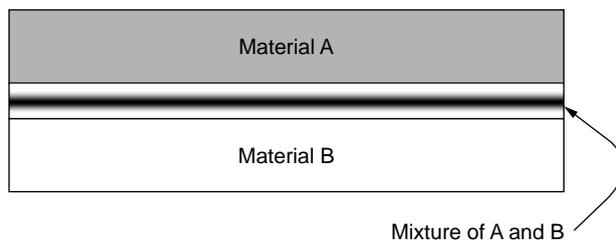


Figure 6. Schematic of diffusive bonding showing that material A and material B are at least partially soluble in each other, resulting in formation of an interphase mixture of both A and B.

atoms or molecules in a material an infinite distance (9). In terms of the enthalpy of vaporization ΔH_{vap} , the gas constant R , and the absolute temperature T , the cohesive energy is given by the following relationship:

$$E_{COH} = \Delta H_{vap} - RT.$$

This equation implies that the extent of energy of cohesion is indicated by the energy required for vaporization. Spontaneous formation of a solution is dependent upon the magnitude and sign of the Gibbs free energy of mixing:

$$\Delta G_{mix} = \Delta H_{mix} - T\Delta S_{mix}.$$

Polymer materials are limited to a few states that exist because of their high molecular weight. Therefore the entropy change ΔS is small and the enthalpy change ΔH is either zero or positive, meaning that dissolution and mixing between polymers is not likely. Complete interphase formation requires a larger entropy change leading to a negative Gibbs free energy of mixing. Perfect adhesion by diffusion requires two materials that are completely soluble in one another, giving the most negative value for change in enthalpy of mixing which would be zero. In polymer materials, autohesion (interdiffusion) may be assisted by the addition of plasticising agents or solvents ultimately affecting the degree of molecular entanglement and bond strength.

Electrostatic Interactions. Adhesion by electrostatic interaction appears to be an unlikely contributor to direct bond strength in composite materials. However, this mechanism of adhesion provides a reasonable justification for acid–base interactions and the use of coupling agents. In theory, it can be stated that surfaces that are basic are electropositive in character, whereas surfaces that are acidic are electronegative in character. Accordingly, there should be some strength of attraction between these two surfaces, with the greatest adhesion by electrostatic interaction occurring when one material is the most basic while the other material is the most acidic. Coupling agents such as silane can be introduced to couple an organic phase with an inorganic phase. These agents possess chemical functions that are twofold: they are reactive with both the inorganic and organic phases. Fiberglass utilizes such coupling agents to protect the glass fiber–matrix interface against attack by moisture and subsequent debonding. A monolayer is formed between the glass fiber and styrenated

polyester that bonds covalently, and it is here where the electrostatic interactions are most relevant and may lead to adhesion by chemical bonding mechanisms.

Chemical Bonding. The chemical bonding mechanism offers a solid explanation for using coupling agents on glass fibers and for the bond strength between carbon fibers and polymer matrices (5). Covalent bonding at interfaces is desirable because of kinetics limitations in the presence of water. That is, adhesion due to primarily covalent bonding mechanisms offers resistance to debonding at the fiber–matrix interface when exposed to moisture, since water must first hydrolyze the bonds. When a chemical bond is formed between compatible chemical groups on the fiber surface and matrix, the bond strength will depend highly on the number of bonds as well as the types. The energy required to separate the two materials at their interface will need to be greater than the covalent bond energy. The bond energy is attributable not only to the activity of the chemical groups but also to the surface area, since a larger area can provide a greater number of chemical bond sites.

The silane coupling agents R–SiX₃ added to the surface of glass fibers ensure good wetting of the fibers in the presence of a hydrated water layer, which tends to greatly reduce the glass fiber surface energy. These agents also provide a strong chemical link between the glass surface oxides and the polymer matrix, leading to a strong, water-resistant bond. Strong bonds between carbon fibers and matrices occur due to the highly reactive surface structure and large surface area (a large surface microroughness) of carbon fibers. If the carbon fiber surface is treated in thermal and acidic environments, ranges of functional groups can be produced that bond directly with unsaturated matrix groups and unsaturated thermoplastic matrix groups.

Mechanical Interlocking. Adhesion by the mechanism of mechanical interlocking is possible, provided that a sufficient surface roughness exists in conjunction with a high degree of wetting. The higher the surface wetting, the more likely the matrix is to conform with the existing fiber surface topography. The lower the surface wetting, the more likely the matrix is to fill interstices lending a smooth surface and little adhesion by mechanical interlocking. The existence of sharply transitioning topography does two things to promote adhesive strength between materials: roughness increases the effective area of contact, and developing cracks are forced to propagate along a tortuous path. Whether or not the matrix fills the pores and smooths topography can be described by Poiseuille's law, which states that

$$x \left(\frac{dx}{dt} \right) = \frac{r^2 P}{8\eta},$$

where x is the distance of pore penetration by the matrix, t is time, r is the pore radius, P is the capillary pressure, and η is the polymer viscosity. The capillary pressure can also be related to the liquid–vapor interfacial energy and the wetting contact angle in terms of the pore radius by

the following relationship:

$$P = \frac{2\gamma_{LV} \cos \theta}{r}.$$

MESOSCALE BEHAVIOR

In composites, the reinforcing constituent materials (fibers) take up nearly all of the applied load as previously outlined. Since the reinforcements cannot support a load directly, the matrix must act as a conduit for transferring the load to the reinforcements. The overall response of individual composite laminae is influenced by the strengths and stiffnesses of the constituent materials. At the micro- and mesoscales, the transfer of load, and elastic and failure behaviors of composites involve many factors besides the constituent material properties. Generally speaking, the primary factors that influence lamina behavior are (10):

- The properties of the constituent reinforcement and matrix materials.
- The properties of the interphase bonding the constituent materials (previous section).
- The volume fractions of the constituent materials.
- The geometry of packing (aligned, random, textile, etc.).
- The cross-sectional shapes and aspect ratios of the reinforcements.

A single composite lamina may have mesoscale properties and exhibit behavior that are far removed from the properties and behavior of the constituents. For example, the constituents can be isotropic while the lamina exhibits orthotropy of strength and modulus. As a result composite mechanical behavior is often categorized as being either fully isotropic or plane orthotropic with transverse anisotropy. That is, the properties are considered to exhibit orthotropic behavior in the principal reinforcing plane and to be equivalent in planes transverse to the reinforcing plane. Because of their complexity, they are usually never considered to be fully anisotropic as 21 constants are required to completely describe fully anisotropic, elastic behavior. The number of possible composite architectures (classes), with textile classes defined to a large extent by NASA Langley Research Center's (LaRC) Advanced Composites Program (ACT), fall into one of the categories and are presented in Fig. 7 (11).

When evaluating orthotropic composite behavior, it is useful to establish a coordinate system relative to a Cartesian or polar coordinate system and coincident to the reinforcing direction. This principal material coordinate system (1–2–3) serves to establish the nine elastic constants that apply for orthotropic behavior, namely the longitudinal tensile modulus, the shear modulus, and Poisson's ratio (E_{11} , G_{12} , ν_{12}) and the transverse tensile moduli, shear moduli, and Poisson's ratios (E_{22} , E_{33} , G_{13} , G_{23} , ν_{13} , ν_{23}). However, by assuming orthotropic symmetry, one can reduce the number of effective elastic constants required to describe orthotropic behavior from nine to five as shown in Fig. 8 for a unidirectional (UD) composite lamina. This

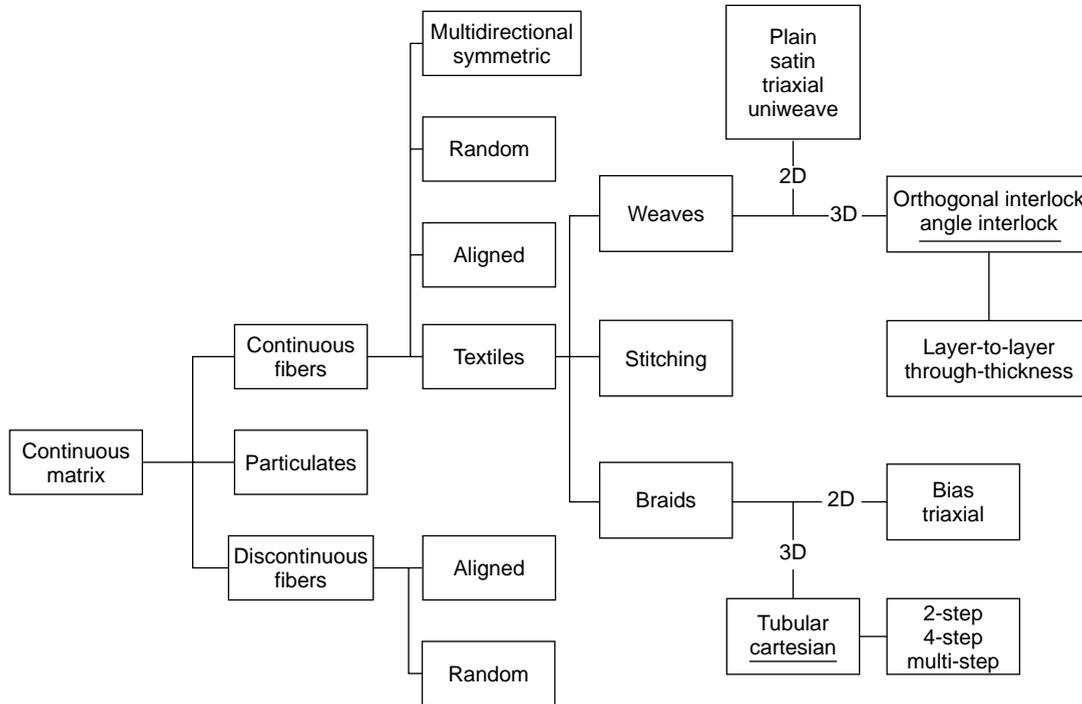


Figure 7. Possible fiber and textile composite architectures (classes) for general and advanced structural high-performance use categorized, in large part, by NASA LaRC's ACT program established in 1989.

leads to the following relations:

$$E_{22} = E_{33}, \quad \nu_{12} = \nu_{13},$$

$$G_{23} = \frac{E_{22}(\text{or } E_{33})}{2(1 + \nu_{23})}, \quad G_{12} = G_{13}.$$

Continuous Fiber Reinforcement

The elastic behavior of a composite lamina having continuous fiber reinforcements is evaluated on the basis of the constituent material properties and their associated volume fractions. In a continuous fiber, UD composite lamina (Fig. 8), all fibers are assumed to be aligned in parallel and extend to the boundaries of the matrix. For the purposes of theoretical analysis, the fibers are considered to be arranged in an ideal square or hexagonal array (analogous to the Kepler conjecture) giving a fiber volume fraction V_f according to

$$V_f = 0.25\pi \left(\frac{r}{R}\right)^2 : \text{square}, \quad V_f = 0.29\pi \left(\frac{r}{R}\right)^2 : \text{hexagonal},$$

where $2R$ is the center-to-center fiber spacing and $2r$ is the fiber diameter. In practice, however, the fiber volume fraction is usually determined by ASTM standardized acid digestion techniques or image analysis using some form of microscopy.

On the premise of linear elasticity, when a load is applied parallel to the lamina fibers, the strain experienced by both the fibers and matrix is the same (*isostrain* condition). Since the fiber stress is $\sigma_f = E_f \varepsilon_1$ and the matrix

stress is $\sigma_m = E_m \varepsilon_1$, it follows that the fiber stress is greater in the fibers than in the matrix ($E_f > E_m$). If we assume that a total load P_c is applied and the bonding between constituents is continuous, then $P_c = P_f + P_m$, where the subscripts c, f, and m denote the composite, fiber, and matrix, respectively. Knowing that the strain in the composite is the same as the strain in the fibers and matrix, and the stresses are a function of the applied load and area fractions, the isostrain condition can be explained by

$$\frac{\sigma_c}{\varepsilon_c} = \frac{\sigma_f}{\varepsilon_f}(V_f) + \frac{\sigma_m}{\varepsilon_m}(V_m).$$

Upon substitution of the definitions for the fiber and matrix tensile moduli in this equation, the traditional "rule of mixtures" equation based on an isostrain loading condition is defined:

$$E_{11} = E_{\parallel} = E_f V_f + E_m (V_m \text{ or } (1 - V_f)).$$

Loads applied to the composite lamina that are perpendicular or transverse to the fiber direction follow a similar derivation, but the loading is referred to as an isostress condition (see Fig. 9 for isostrain vs. isostress). Under the isostress assumption, the stresses are assumed to be the same in the lamina, fibers, and matrix. However the strains become a function of the area fractions of fibers and matrix to that of the total composite. It follows that the isostress condition can be explained by

$$\varepsilon_c = \frac{\sigma_2}{E_m}(V_m) + \frac{\sigma_2}{E_f}(V_f).$$

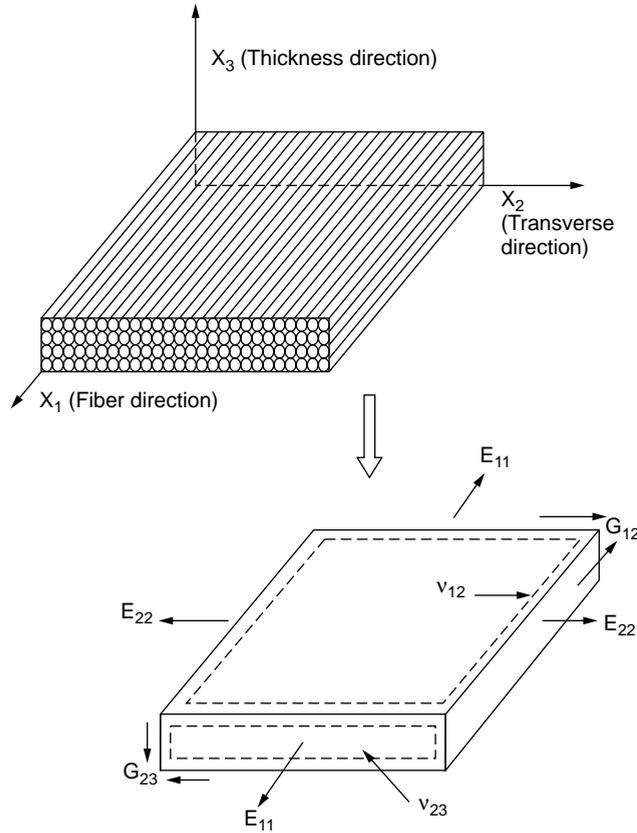


Figure 8. Representation of the principal material coordinate system in a UD composite lamina exhibiting orthotropic symmetry and the effective elastic constants derived from the basic loading schemes.

Upon substitution of the definition for stress (Hooke's law) in the preceding equation, a more traditional equation for the transverse tensile composite modulus is defined:

$$E_{22} = E_1 = \frac{E_f E_m}{E_f(1 - V_f) + E_m V_f}$$

Composite lamina elastic properties derived on the basis of the isostress loading condition provide only marginal agreement with actual results. The isostress derivation does not take into account the effects of the Poisson contraction. Therefore other mesomechanics approaches are necessary for evaluating lamina elastic constants from constituent materials that exhibit highly anisotropic elastic behavior.

According to more realistic assumptions that account for the Poisson contraction and nonuniform stress distributions due to strain magnification between neighboring fibers, additional solutions obtained from elasticity theory and finite element analysis have been offered for predicting transverse and longitudinal elastic properties. One such set of solutions that are generally more applicable are the Halpin-Tsai equations (12). These equations can be written in the general form as:

$$\frac{(\bullet)_c}{(\bullet)_m} = \frac{1 + \xi \eta V_f}{1 - \psi \eta V_f}$$

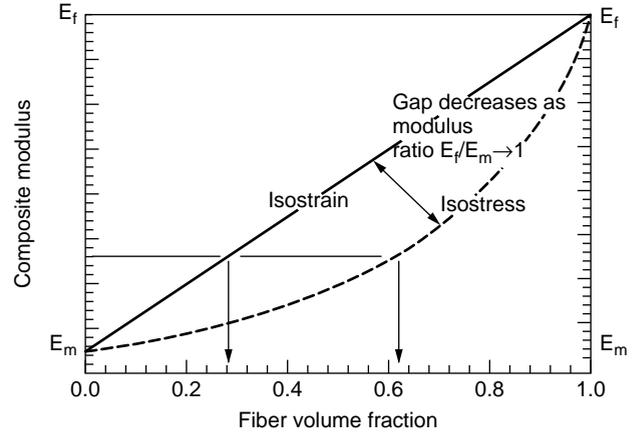


Figure 9. Schematic of isostrain versus isostress loading for a composite material, indicating that less fiber is required according to isostrain assumption for yielding the same elastic modulus. The isostress assumption approaches the isostrain assumption as the ratio of elastic constituent moduli $E_f/E_m \rightarrow 1$.

where (\bullet) is the elastic constant, ξ depends on the characteristics of the reinforcing phase such as fiber aspect ratio, packing geometry, shape, and loading conditions, and ψ is a factor that accounts for the maximum packing possible. The other factor η is given by

$$\eta = \frac{r - 1}{r + \xi}, \quad r = \frac{(\bullet)_f}{(\bullet)_m}$$

When the factor ψ is properly evaluated for maximum packing fraction, the general form of the Halpin-Tsai equations for UD lamina elastic properties can be given in expanded notation as

$$\begin{aligned} E_{11} &= E_{\parallel} = E_f V_f + E_m(1 - V_f), \\ \nu_{12} &= \nu_{\parallel+} = \nu_f V_f + \nu_m(1 - V_f), \\ \frac{M}{M_m} &= \frac{(1 + \xi \eta V_f)}{(1 - \eta V_f)}, \quad \eta = \frac{(M_f/M_m) - 1}{(M_f/M_m) + \xi}, \end{aligned}$$

in which M is the elastic modulus of interest, E_{22} , G_{12} , or ν_{23} ; M_f the constituent reinforcing fiber modulus E_f , G_f , ν_f ; and M_m the constituent matrix modulus E_m , G_m , ν_m . The reinforcement factor ξ is determined empirically by curve-fitting the experimental results. When $\xi \rightarrow \infty$, the Halpin-Tsai formulas lead to the classical rule-of-mixtures relation, and when $\xi \rightarrow 0$, the Halpin-Tsai formulas lead to the inverse form of the rule-of-mixtures:

$$(\bullet)_c = V_f (\bullet)_f + (1 - V_f) (\bullet)_m,$$

$$\frac{1}{(\bullet)_c} = \frac{V_f}{(\bullet)_f} + \frac{(1 - V_f)}{(\bullet)_m}$$

For reinforcing fibers such as carbon or Kevlar, which exhibit anisotropic behavior but are considered transversely isotropic, the following expressions from Chamis (13)

describe the elastic properties of a UD lamina composed of anisotropic fibers in an isotropic matrix:

$$E_{11} = E_{1f}V_f + E_m V_m, \quad E_{22} = E_{33} = \frac{E_m}{1 - V_f [1 - (E_m/E_{2f})]},$$

$$G_{12} = G_{13} = \frac{G_m}{1 - V_f [1 - (G_m/G_{12f})]},$$

$$G_{23} = \frac{G_m}{1 - V_f [1 - (G_m/G_{23f})]},$$

$$\nu_{12} = \nu_{13} = \nu_{12f}V_f + \nu_m V_m,$$

$$\nu_{23} = \frac{E_{22}}{2G_{23}} - 1,$$

where the subscripts f , m , ij_f and ij_m ($i, j = 1, 2$) denote the elastic properties of the reinforcing fibers and matrix.

Composite cylinder assemblage (CCA) theory also provides simple, closed-form analytical solutions for the effective composite moduli and the fiber and matrix are also considered to be transversely isotropic (14–16). CCA yields close-bounded solutions for the transverse tensile and shear moduli and models the composite as an assemblage of long, circular fibers surrounded by concentric matrix shells.

The dependence of the effective elastic moduli on fiber volume fraction is presented in Fig. 10 and 11 for two different composite systems, a E-glass-epoxy composite lamina and a carbon-epoxy lamina. It is apparent from the results that packing densities and constituent material elastic behaviors can readily be used as parameters in the design of composite materials when tailoring the performance characteristics.

Textile Architectures. Many of the woven, braided, and stitched fabric-reinforced (textile) composite laminae also

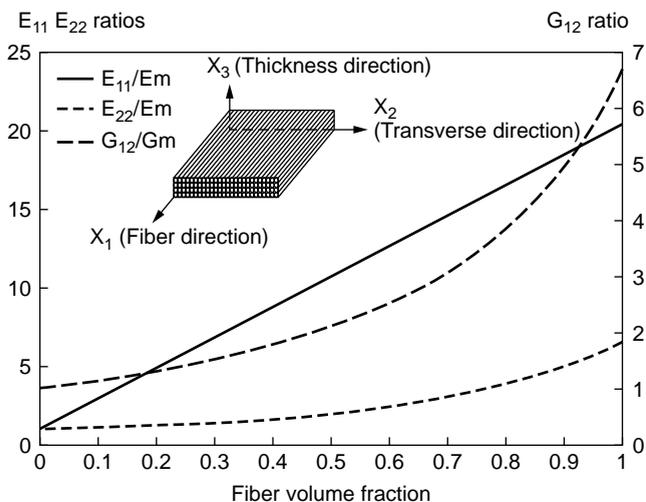


Figure 10. Effect of fiber volume fraction (packing density) on the effective elastic modulus ratios for a E-glass-epoxy composite system (solution of Halpin-Tsai). Constituent properties: $E_{11f} = E_{22f} = 72.0$ GPa (10.4 Msi), $G_{12f} = 27.7$ GPa (4.02 Msi), $\nu_{12f} = 0.30$, $E_{11m} = E_{22m} = 3.5$ GPa (0.51 Msi), $G_{12m} = 1.3$ GPa (0.19 Msi), $\nu_{12m} = 0.35$.

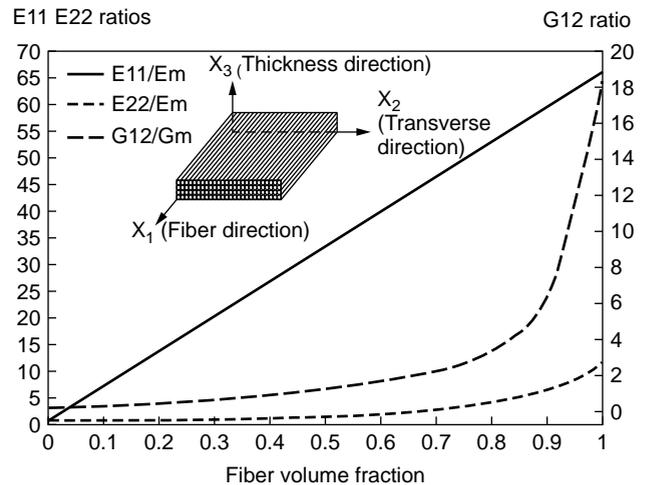


Figure 11. Effect of fiber volume fraction (packing density) on the effective elastic modulus ratios for a carbon-epoxy composite system (solution of Chamis). Constituent properties: $E_{11f} = 230.0$ GPa (33.4 Msi), $E_{22f} = 40.0$ GPa (5.8 Msi), $G_{12f} = 24.0$ GPa (3.48 Msi), $G_{23f} = 14.3$ GPa (2.07 Msi), $\nu_{12f} = 0.30$, $E_{11m} = E_{22m} = 3.5$ GPa (0.51 Msi), $G_{12m} = 1.3$ GPa (0.19 Msi), $\nu_{12m} = 0.35$.

fall under the umbrella of continuous fiber reinforcements. However, nearly all of the textile architectures are not aligned within the reinforcing plane. These fiber architectures provide reinforcement in two or three directions, with the disadvantage of having somewhat lower in-plane stiffnesses due to the textile geometries. Textile reinforcing phases do afford better shaping capabilities over curved surfaces (drape) and the possibility for through-thickness reinforcement as with three-dimensional architectures.

In contrast to aligned fiber, UD laminae, textile reinforcements consist of many filaments grouped together to act as primary braiders, weavers, and fillers (warp and weft tows). Fiber bundling influences not only the packing arrangement and density but also the cross-sectional shapes and aspect ratios. Such variations in these factors combined with different textile architectures require alternate methods to evaluate the mechanical response of the composites. For example, consider the eight harness satin (8HS) weave architecture shown in Fig. 12. The bundled filamentary tow cross-sectional shapes are biconvex (lenticular); that is, thickness variation exists. Additionally, the warp tows interlace at every eighth mutually orthogonal weft or fill tow (hence, 8HS). The periodic interlacings and orthogonal tows permit shapability, but increase the in-plane compliances. Because of these factors and others, it is necessary to evaluate the effective composite properties on the basis of transformed reduced stiffnesses and through-thickness homogenization schemes that smear thickness properties into an effective continuum.

The complete evaluation of elastic and thermoelastic behavior of woven and braided composite materials is involved and beyond the scope of this review. (However, the reader is encouraged to study references 17 to 23 for thorough analytical and numerical treatments on the subjects of textile composite elastic and nonlinear behaviors).

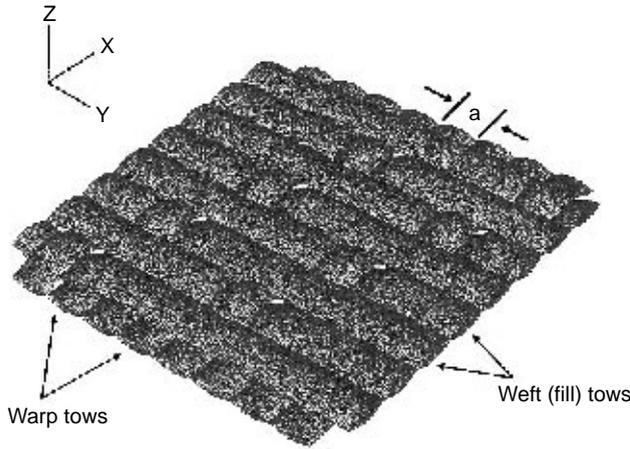


Figure 12. Unit cell of an 8-harness satin (8HS) woven reinforcing architecture consisting of weaver (warp) tows interlaced with filler (weft) tows. Maximum width of the filamentary bundles is given by a and the cross-sectional shape is lenticular.

Randomly Oriented Fiber Reinforcement

Continuous Fibers. As with laminae constructed using continuously aligned fibers, a lamina may also be made by impregnating a polymer matrix into continuous and discontinuous fiber mats having fibers at various orientations. For the case of continuous fibers, lamina elastic properties can be considered isotropic in plane so long as a uniform probability exists over the entire range of fiber angles from $-\pi/2$ to $+\pi/2$. Neglecting fiber end effects, it is only necessary to determine the average composite longitudinal tensile and shear moduli. Akasaka derived the following expressions for the averaged effective elastic constants of composites based on randomly oriented, continuous fibers (24):

$$\begin{aligned}\bar{E}_c &= \left[\frac{E_{11} + E_{22} + 2\nu_{12}E_{22}}{1 - \nu_{12}\nu_{21}} \right] \\ &\times \left[\frac{E_{11} + E_{22} - 2\nu_{12}E_{22} + 4(1 - \nu_{12}\nu_{21})G_{12}}{3(E_{11} + E_{22}) + 2\nu_{12}E_{22} + 4(1 - \nu_{12}\nu_{21})G_{12}} \right], \\ \bar{G}_c &= \left[\frac{E_{11} + E_{22} - 2\nu_{12}E_{22}}{8(1 - \nu_{12}\nu_{21})} \right] + \frac{1}{2}G_{12}; \\ \bar{\nu}_c &= \frac{\bar{E}_c}{2\bar{G}_c} - 1.\end{aligned}$$

A simpler approach that produces similar results can be used the above equations for random fiber orientation in two dimensions:

$$\begin{aligned}\bar{E}_c &= \frac{3}{8}E_{11} + \frac{5}{8}E_{22}, \\ \bar{G}_c &= \frac{1}{8}E_{11} + \frac{1}{4}E_{22},\end{aligned}$$

where E_{11} and E_{22} are the longitudinal and transverse tensile moduli along and across the fibers in a UD composite having the same fiber volume fraction. These two approaches are equivalent to assuming that there is an equal

probability of fiber angles (θ) ranging from 0 to 360° in a composite such that the averaged equivalent longitudinal tensile modulus can be obtained by integrating the expression $\bar{E} = 2\pi \int E(\theta) d\theta$ over the range from 0 to $+\pi/2$. The orientation dependence of the tensile modulus $E(\theta)$ of a UD lamina with the same fiber volume fraction is expressed as

$$\begin{aligned}E(\theta) &= \left[\frac{1}{E_{11}} \cos^4 \theta + \left(\frac{1}{G_{12}} - \frac{2\nu_{12}}{E_{11}} \right) \cos^2 \theta \sin^2 \theta \right. \\ &\quad \left. + \frac{1}{E_{22}} \sin^4 \theta \right]^{-1}.\end{aligned}$$

These sets of equations infer the possibility of tailormaking composite laminae with fiber distributions known *a priori* so that mechanical, thermal, and other physical properties can meet the specified requirements of a given design.

Discontinuous Fibers. Behavior of composite laminae made from randomly oriented, discontinuous (chopped) fibers present a more complex problem in so far as evaluating mechanical response because of the existence of fiber-end effects. The assumption of isostrain does not hold as with continuous fiber systems because the strain transfer from the fiber-end-matrix interface is not the same as the transfer borne by continuous fibers along their length. Under an applied tensile force, the matrix displaces relative to the fiber along the fiber-matrix interface resulting in interfacial shear stress. The relative displacement is zero at the fiber midpoint and maximum at either end of fiber termination. The relationship between the fiber tensile and shear stress is $(d/d_x)\sigma_x = (4\tau_m/d_f)$, and the variation with distance along the fiber is shown for the case of elastic tensile deformation in Fig. 13.

The load transfer from the matrix to the fiber occurs over a given length of fiber called the critical length (l_c) that is just enough to allow for complete transfer. This critical length is such that complete transfer takes place without fiber breakage or matrix failure and is defined as:

$$l_c = d_f \left(\frac{E_f \varepsilon_m}{2\tau_{ym}} \right),$$

where the subscripts f, m, and ym denote the fiber, matrix, and yield of matrix, respectively. The critical aspect ratio of fiber length-to-fiber diameter (l_c/d_f) increases with composite stress and strain according to the previous equation. If we define longer fibers to be $l > l_c$ and shorter fibers to be $l < l_c$, then the shorter fiber, composite effective longitudinal tensile, and shear elastic moduli become

$$\begin{aligned}E_c &= \frac{1}{3}V_f E_f \left(1 - \frac{l_c}{2l} \right) + (1 - V_f)E_m, \\ G_c &= \frac{1}{8}V_f E_f \left(1 - \frac{l_c}{2l} \right) + (1 - V_f)E_m,\end{aligned}$$

for the two-dimensional consideration. For random orientations of longer fibers in two-dimensions, the composite

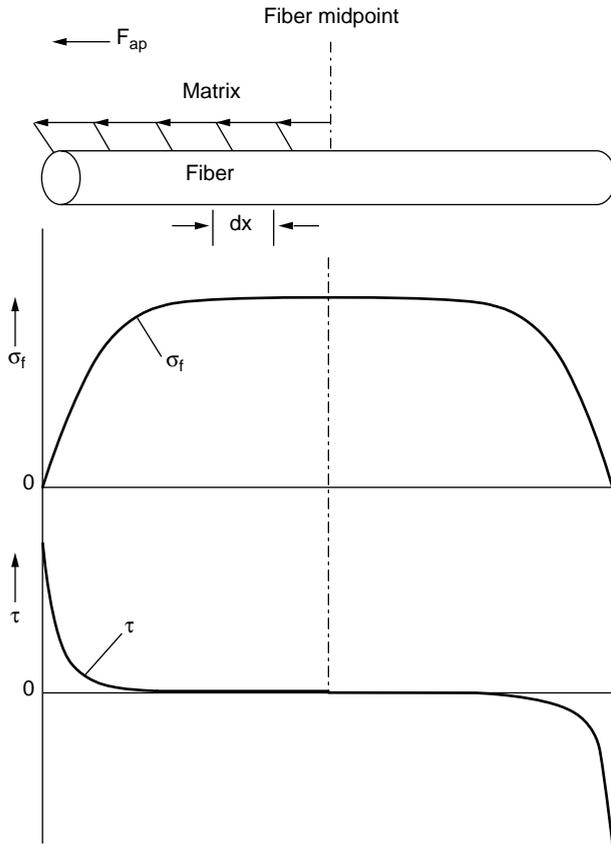


Figure 13. The tensile load transfer process by interfacial shear and the resulting variation of tensile and shear stresses along the length of a discontinuous fiber.

effective moduli are given by

$$E_c = \frac{1}{3}V_f E_f \left(\frac{l}{2l_c} \right) + (1 - V_f)E_m,$$

$$G_c = \frac{1}{8}V_f E_f \left(\frac{l}{2l_c} \right) + (1 - V_f)E_m.$$

Particulate Reinforcement

Composite materials based on particle reinforcements usually contain fillers that are spherical. It is reasonable to consider these composites as mesoscopically isotropic. Therefore only two effective elastic properties are necessary to fully describe composite elastic behavior. In evaluating these elastic properties, the modified Halpin-Tsai solution set is applicable:

$$\frac{(\bullet)_c}{(\bullet)_m} = \frac{1 + \xi \eta V_f}{1 - \psi \eta V_f}, \quad \eta = \frac{r - 1}{r + \xi}$$

$$r = \frac{(\bullet)_f}{(\bullet)_m}, \quad \xi = \frac{7 - \nu_m}{8 - 10\nu_m}$$

$$\psi = 1.$$

Again, (\bullet) is the elastic constant of interest, ξ depends on the characteristics of the reinforcing phase such as filler

aspect ratio, packing geometry, shape, and loading conditions, and ψ is a factor that accounts for the maximum packing possible. For particle reinforcement, the filler and matrix are both assumed to be isotropic and the packing factor is taken as 1.

Statistics and Strength Theories

Up to this point, the assumption was that stress-induced fiber failure would occur at the same level for all fibers within a composite lamina, suggesting that a simple rule-of-mixtures (ROM) approach would give a sufficient first-order approximation to composite longitudinal tensile strength according to

$$\sigma_{11} = E_f \varepsilon_{11} V_f + E_m \varepsilon_{11} (1 - V_f).$$

In practice, the stress levels reached to cause fiber failure will vary from point to point because of the various distributions of inherent flaws that exist within the constituent matrix and on the surface of the reinforcing fibers as well as the distributions of fiber lengths. For a given tensile stress, weaker fibers within a group may fail, giving rise to perturbations in the local stress fields near the fiber termination and resulting in the high interfacial shear stresses as previously discussed in regard to discontinuous fiber reinforcements. The interfacial shear stresses transfer the load across the interfaces and may introduce stress concentrations into adjacent, unbroken fibers.

Several possible failure events may arise due to the stress distribution at each broken fiber end. If the interfacial bond strength is weak, separation of the matrix and fiber at the termination will rapidly advance along the fiber–matrix interface. If the interfacial bonding is strong or the matrix is a soft ductile material, shear stresses are redistributed and fibers may fail before the surrounding matrix fails completely. Progression of undesirable modes of failure results in the development of damage zones (see Fig. 14), eventually leading to a strength reduction such that the composite strength is no greater than that of the weakest fiber or “link.”

Prior to composite material failure, increased loading produces a statistical accumulation of dispersed damage zones until a sufficient number interact to provide a weakened surface. At the point of failure, the weakened surface becomes quite large in comparison to the undamaged material and the composite can no longer support additional tensile loading.

The experimentally measured tensile strengths of brittle fibers are subject to significant amounts of scatter that can be described by the Weibull distribution function. A schematic illustration of this is given in Fig. 15, where the fraction of fibers of uniform length that fail are plotted against the given failure stress. The Weibull cumulative probability distribution function $G(\sigma)$ represents the probability of fiber and fiber bundle fracture at a stress level that is equal to or less than σ , and it is given by

$$G(\sigma) = 1 - \left[1 - \left(\frac{\sigma - \sigma_I}{\sigma_o} \right)^m \right]^\omega,$$

where σ_o is the difference between the upper strength limit

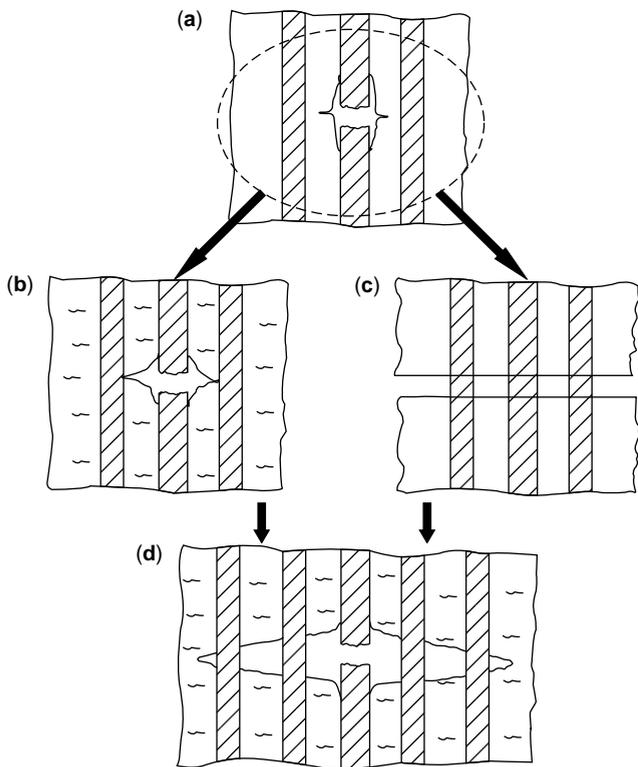


Figure 14. Schematic of the evolution of tensile fiber damage in aligned fiber composites: (a) fiber break with interfacial debonding, (b) fiber break expanding matrix crack, (c) matrix crack with fiber bridging, and (d) a compilation of a, b, and c resulting in a damage zone.

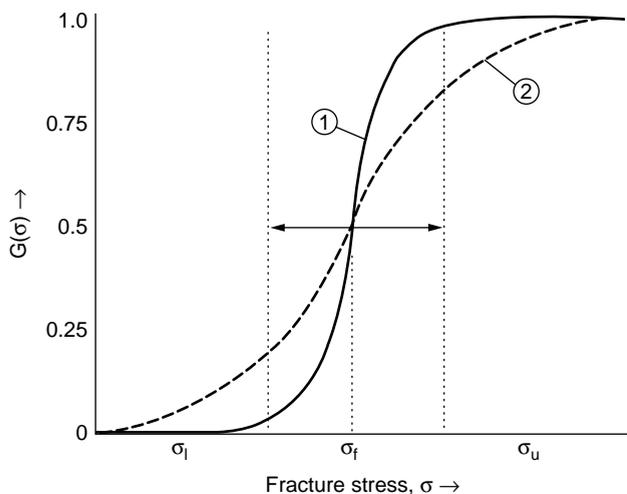


Figure 15. Weibull cumulative probability distribution function $G(\sigma)$ describing variations in fiber strength: (1) Fibers do not exhibit a wide variability in fracture strength between 0 and 1, where 0.5 is the occurrence of tensile failure in 50% of fibers, and (2) a wide variation exists and is statistically described by a standard deviation as indicated with vertical lines.

σ_u and the lower strength limit σ_l , ω is a function of the test sample aspect ratio and m depends on the amount of scatter. The exponent m is approximately $1.2\bar{\sigma}/s$, where $\bar{\sigma}/s$ is the inverse of the coefficient of variation given by

$$\frac{\bar{\sigma}}{s} = \left(\frac{\sum_{i=1}^N \sigma_i}{N} \right) \left(\left[\frac{\sum_{i=1}^N (\sigma_i - \bar{\sigma})^2}{N} \right]^{1/2} \right)^{-1}$$

Below the lower strength limit, all fibers undergo the same amount of elongation and remain unbroken. As the lower strength limit is exceeded, the weakest fibers (those containing internal flaws leading to reduced effective cross sections) break in succession, and the load must be transferred to the surviving fibers. Complete fracture of the bundle occurs when the upper strength limit is reached.

Several tensile dominated failure modes adopted to consider the fiber-and-fiber bundle failure processes when they are bound by a matrix include the weakest link failure mode, cumulative weakening failure mode, fiber break propagation failure mode, and cumulative group failure mode. The weakest link failure mode associates catastrophic failure with the occurrence of a single or isolated small number of independent fiber breaks. Realistically, this mode of failure is an unlikely characterization because the stress level at which weakest link events occur would not be sufficient enough to invoke composite material failure.

The cumulative weakening failure mode is necessarily an extension of the weakest link failure mode. Within characterization of this mode, a fiber fracture site inhibits redistribution of stress near the site. As more sites develop along a fiber, they tend to have a statistical strength distribution that is equivalent to the distribution of flaws along the fiber. Failure is thought to occur when a layer across the section of a lamina is weakened to the point of not being able to support any further increments in load. A critical argument to acceptance of this mode entirely as a characterization of failure is that no consideration is given to the effects on neighboring fibers and flaws.

It is well known that the effects of stress perturbations at terminations are significant to neighboring fibers. The fiber break propagation failure mode is more realistic in the sense that the effects of perturbations on the progressive weakening of adjacent fibers are considered. As redistribution of stress occurs, the stresses on adjacent fibers are magnified, increasing the probability that failure will occur in these fibers. With increased loading, the failure probability increases until sequential fiber failure occurs.

Under auspices of the fiber break propagation model, it is difficult to achieve a meaningful strength estimate, and lamina tensile strength predictions generally depend on the micromechanisms of deformation and fracture at fiber termination points. For the smaller damaged volumes of material, strength predictions are acceptable, but predicted failure stresses are lower for larger volumes. The cumulative group mode failure model considers the effects of variability in fiber strengths, stress concentrations in adjacent fibers arising from stress redistributions, and the interfacial debonding process due to increased matrix shear stresses. It is more likely that fiber breaks

will progressively accumulate in groups between the stress level necessary to initiate the first fiber break to the stress level necessary to cause composite material failure. Composite failure will occur when the distributed groups of damage zones are of a sufficient number and size that their cumulative effect reduces the material stiffness by an amount sufficient enough to prohibit any additional load-carrying capability. Weakening mechanisms by this mode could be thought of in a couple of different manners. In one way, the number of developed damage zones would grow to such a number that the summed interactions exceeded the critical material stress. In another, the size and number of zones would reach such magnitude that catastrophic and rapid crack propagation ensue due to the lack of both intact material and crack tip blunting mechanisms between zones. Although the cumulative group model suggests a generalization of the cumulative weakening model, the practicality of use is complicated by its complexity in considering mostly all of the singular fiber longitudinal tensile failure mechanisms.

The longitudinal compressive strength, like the longitudinal tensile strength, is highly dependent on many factors and is particularly sensitive to constituent matrix properties and fiber volume fraction. Several failure mechanisms have been proposed, but the most dominant mechanism is microbuckling, analogous to the buckling of a beam on an elastic foundation. The surrounding matrix resists fiber microbuckling, but there are several factors that can lead to a reduction in the support given by the matrix and neighboring fibers. At a low fiber volume fraction, the out-of-phase or extensional buckling mode is suggested with the lamina compressive strength predicted by the following equation:

$$\sigma_{11,c}^{cr} = 2V_f \sqrt{\frac{V_f E_m E_f}{3(1 - V_f)}}.$$

At higher, more industrially practicable fiber volume fractions, the in-phase or shear buckling mode is suggested with the lamina compressive strength predicted by the following equation:

$$\sigma_{11,c}^{cr} = \frac{G_m}{(1 - V_f)}.$$

Given a constant fiber volume fraction, any factors contributing to reduction in the matrix shear modulus will lead to a reduction in composite compressive strength, since the mode is in-phase. More specifically, the identified factors that influence reduced support from the surrounding media include: (25)

- Fiber bunching and waviness, which leads to preferential buckling, local matrix rich regions and matrix instability.
- The presence of voids, which tend to have a greater effect than the matrix rich regions.
- Interfacial debonding, due to circumferential tensile stresses that arise principally from a difference in Poisson's ratios between the fibers and surrounding matrix or the opposite effect induced by thermal curing stresses.

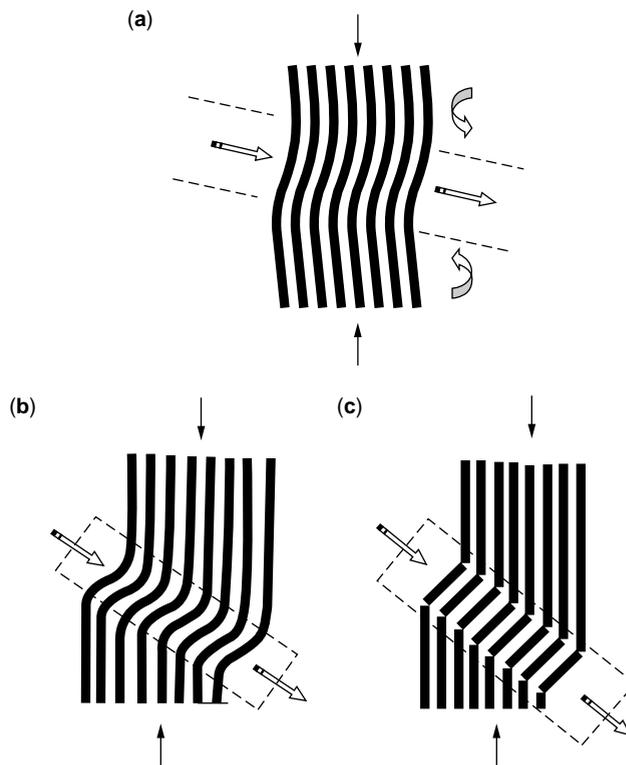


Figure 16. Progression of compressive fiber failure resulting from longitudinal compressive in-phase buckling (a). In polymeric aramid fibers, compressive yielding is common (b) during formation of a kink zone, while more pronounced kinking often leads to fiber fracture at two locations (c) after (25).

- A lower effective matrix shear modulus, compared to the instantaneous matrix shear modulus, as a result of viscoelastic deformation processes.

Another longitudinal compressive failure mechanism specific to the structurally oriented, wholly aromatic polyamide polymer fiber (Kevlar aramid) and carbon/graphite fiber families, is the formation of kink-bands as illustrated in Fig. 16. The highly anisotropic behavior of these fibers leads to massive fiber rotation at one zone and counter-rotation at another zone. In the extreme case, compressive failure at the kink zones results in complete fiber fracture at two locations. Compressive yielding without complete failure is more typical of the polymeric Kevlar aramids such as Kevlar 49.

The transverse tensile, compressive, shear, and longitudinal shear strengths can be regarded as matrix dominated, so the failure modes can be thought of as matrix-modes of failure. Transverse tensile strength is governed by the same factors as longitudinal compression, but with one added detail. Unlike longitudinal tension where composite strength is prescribed primarily on the basis of fiber strength, the presence of fibers in transverse tension have a negative effect. Transverse strength is often lower than the strength of the constituent neat matrix material because of the stress magnification effects from fibers. Without regard to the presence of stress magnification from fiber ends and

matrix voids, the transverse strength is dictated primarily by the interfacial bond strength. The interface is made weaker when cohesive failure occurs prior to the cohesive failure in either the constituent matrix or fibers.

Where interface bonding is weak, stress magnification from fiber ends and voids tends to promote transverse cracking more readily along the common edges of adjacent fiber ends. These same factors also affect the transverse and longitudinal shear strengths, depending on the direction of shear displacements and the viscoelastic properties of the matrix. The only real differences here may be the direction of crack propagation and the failure mode(s) of the matrix, unless the fiber volume fraction is sufficiently higher. If a large number of fibers are present and the interfacial bonding is good, the fibers will offer reinforcement, provided the shearing plane is normal to the fibers. If the shearing plane contains the fibers, then little fiber reinforcement is available and the strength is determined by the properties of the matrix.

Identification of a predominant failure mechanism, whether a fiber or matrix mode, is important from the perspective of designing composite structures. Knowledge of the different failure mechanisms and the nature of single-stress component damage initiation can be used to evaluate the predominant mode of failure through formulation of practical failure criteria. In establishing the failure criterion, a fundamental assumption is that a failure criterion exists to characterize failure in a UD composite and is of the following form:

$$F(\sigma_{11}, \sigma_{22}, \tau_{12}) = 1,$$

where some function F is defined in terms of the principal stresses. A suitable failure criterion generally takes the form of a quadratic polynomial because this is the simplest form that has been found to adequately describe experimental data. The advantages are that several failure criteria can be defined in terms of uniaxial strengths, and a predominant mode of failure can be identified from the criterion that is initially satisfied.

If a certain mode of failure is identified and deemed undesirable for a given load, the designer can tailor the composite properties and re-evaluate the failure criteria until some other mode is predicted that is less detrimental to the design. For UD fiber composites, the general quadratic failure criterion is a two-dimensional version of the Tsai-Wu criterion given by

$$\begin{aligned} & \left(\frac{1}{S_1^t S_1^c} \right) \sigma_{11}^2 + \left(\frac{1}{S_1^t} - \frac{1}{S_1^c} \right) \sigma_{11} + \left(\frac{1}{S_2^t S_2^c} \right) \sigma_{22}^2 \\ & + \left(\frac{1}{S_2^t} - \frac{1}{S_2^c} \right) \sigma_{22} - \frac{1}{2} \left(\left(\frac{1}{S_1^t S_1^c} \right) \left(\frac{1}{S_2^t S_2^c} \right) \right)^{1/2} \sigma_{11} \sigma_{22} \\ & + \left(\frac{1}{(S_{12}^s)^2} \right) \tau_{12}^2 = 1, \end{aligned}$$

where the S_{ij} denote the single-component strengths and the superscripts t, c, and s denote tension, compression, and shear, respectively. The biggest drawback of this criterion is that it ignores the diversity in the possible failure modes.

Each of the failure modes previously mentioned can be modeled as a specific criterion and, as such, evaluated and identified independently. The following set of equations provides a reasonable set of criteria for each of the dominant fiber and matrix failure modes (26):

- Tensile Fiber Failure

$$\left(\frac{\sigma_{11}}{S_1^t} \right)^2 + \left(\frac{\tau_{12}}{S_{12}^s} \right)^2 = 1.$$

- Compressive Fiber Failure

$$\left(\frac{\sigma_{11}}{S_1^c} \right)^2 + \left(\frac{\tau_{12}}{S_{12}^s} \right)^2 = 1.$$

- Tensile Matrix Failure

$$\left(\frac{\sigma_{22}}{S_2^c} \right)^2 + \left(\frac{\tau_{12}}{S_{12}^s} \right)^2 = 1.$$

- Compressive Matrix Failure

$$\left(\frac{\sigma_{22}}{2S_{23}^s} \right)^2 + \left(\left(\frac{S_2^c}{2S_{23}^s} \right)^2 - 1 \right) \left(\frac{\sigma_{22}}{S_2^c} \right) + \left(\frac{\tau_{12}}{S_{12}^s} \right)^2 = 1.$$

Since the transverse shear strength S_{23} is difficult to obtain without performing thickness shear tests, the matrix shear strength is used as an approximation. Upon evaluating each of the failure criteria for a given circumstance, the predominant mode or modes of failure can be determined. Necessarily, no biaxial tests are required, and a mode of failure is identified by the criterion that is satisfied first.

MACROSCALE BEHAVIOR

On the macroscale, the effective composite elastic properties are evaluated on the basis of a composite laminate that is composed of several laminae bonded together at various orientations to one another. It was previously stated that the composite structure or component and the laminae may, in some cases, coincide on the same structural scale. This being the case, tailoring laminate properties will also coincide directly with influencing component behavior.

One of the most important aspects relating to the effective design of composite laminates and structures is knowledge of composite lamina off-axis behavior and associated limitations with particular fiber orientations. Aligned fiber composite laminae are highly anisotropic in-plane, and commonly varying degrees of coupling between extension and shear occur when the direction of loading is not coincident with a principal material direction. The designer must have some knowledge *a priori* of the lamina response to off-axis loading conditions in order to determine a suitable lamina lay-up sequence that provides optimum reinforcement. An accurate prediction of laminate elastic properties, which are highly dependent on the orientation, properties, and distribution of individual laminae, is essential for understanding the response of the resulting structure to external loading and environmental conditions.

Elastic Behavior Off-Axis

Hooke's law can be generalized using a contracted form of tensor notation and expressed concisely by the following equation:

$$\sigma_i = \sum_{j=1}^6 C_{ij} \varepsilon_j,$$

where $i, j = 1, \dots, 6$, σ_i are the components of stress, C_{ij} is the stiffness matrix, and ε_j are the components of strain. Since the stiffness constants are symmetrical (i.e., $C_{ij} = C_{ji}$), the expanded form of the previous equation is given in matrix notation by

$$\begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \sigma_3 \\ \tau_{23} \\ \tau_{31} \\ \tau_{12} \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} & C_{13} & C_{14} & C_{15} & C_{16} \\ & C_{22} & C_{23} & C_{24} & C_{25} & C_{26} \\ & & c_{33} & C_{34} & C_{35} & C_{36} \\ & & & C_{44} & C_{45} & C_{46} \\ & & & & C_{55} & C_{56} \\ & & & & & C_{66} \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \gamma_{23} \\ \gamma_{31} \\ \gamma_{12} \end{bmatrix}.$$

(SYM)

The constitutive relations that link stress to strain in terms of the stiffness matrix may also be inverted to relate strain to stress in terms of the compliance matrix. The constitutive relations for a UD composite lamina, which exhibits orthotropic symmetry and transverse isotropy in the x_2 - x_3 material principal coordinate plane, can be simplified if the dimension in the x_3 (thickness) direction is considered to be sufficiently smaller than both of the in-plane dimensions. This consideration reduces the problem to two dimensions, either of the plane stress or plane strain form. Clearly, the implication is that the nonzero stresses are arbitrarily restricted to in-plane; hence the nonzero quantities are not functions of x_3 ($\sigma_3 = \tau_{23} = \tau_{31} = 0$). For this, the stress-strain relation for a UD lamina given in terms of the matrix of mathematical moduli $[Q_{ij}]$ becomes

$$\begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \sigma_6 \end{bmatrix} = \begin{bmatrix} Q_{11} & Q_{12} & 0 \\ Q_{12} & Q_{22} & 0 \\ 0 & 0 & Q_{66} \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_6 \end{bmatrix},$$

where Q_{11} , Q_{22} , Q_{12} , and Q_{66} are identified as the *reduced stiffnesses*.

The equation above suggests that no coupling exists between tensile and shear strains; that is, orthotropic composite materials exhibit no shearing strains when applied loads act coincident to the principal material directions. The Q_{ij} components of the reduced stiffness matrix from this equation are given in terms of the engineering constants as

$$Q_{11} = C_{11} = \frac{E_{11}}{1 - \nu_{12}\nu_{21}},$$

$$Q_{22} = C_{22} = \frac{E_{22}}{1 - \nu_{12}\nu_{21}},$$

$$Q_{66} = \frac{1}{2}(C_{11} - C_{12}) = G_{12},$$

$$Q_{12} = C_{12} = \frac{\nu_{12}E_{22}}{1 - \nu_{12}\nu_{21}} = \frac{\nu_{21}E_{11}}{1 - \nu_{12}\nu_{21}}.$$

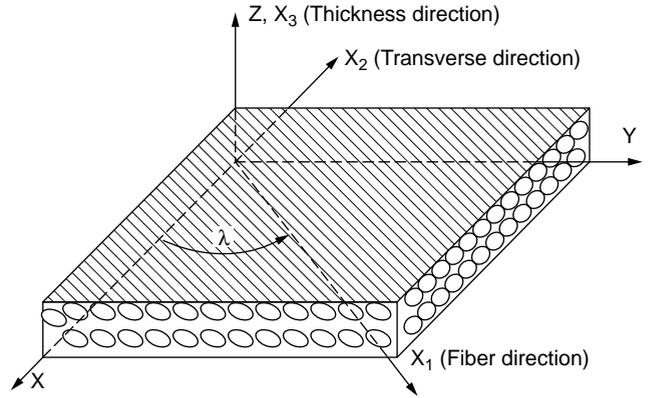


Figure 17. Representation of a UD composite lamina with the principal material direction (fibers) oriented at some arbitrary in-plane angle λ to the Cartesian coordinate X - Y plane.

When the direction of applied load does not coincide with a principal material direction, then coupling between tensile and shear strains exists. Consider the sufficiently thin, UD lamina with fibers oriented at an angle λ to the principal coordinate axis shown in Fig. 17. From classical theory of elasticity, the stress-strain relation becomes

$$\begin{bmatrix} \sigma_x \\ \sigma_y \\ \tau_{xy} \end{bmatrix} = \begin{bmatrix} \bar{Q}_{11} & \bar{Q}_{12} & \bar{Q}_{16} \\ \bar{Q}_{12} & \bar{Q}_{22} & \bar{Q}_{26} \\ \bar{Q}_{16} & \bar{Q}_{26} & \bar{Q}_{66} \end{bmatrix} \begin{bmatrix} \varepsilon_x \\ \varepsilon_y \\ \gamma_{xy} \end{bmatrix},$$

where the \bar{Q}_{ij} components of the matrix are referred to as the *transformed reduced stiffness components*. In terms of the reduced stiffness matrix components and λ , the transformed reduced stiffness components have the following values:

$$\bar{Q}_{11} = Q_{11} \cos^4 \lambda + 2(Q_{12} + 2Q_{66}) \sin^2 \lambda \cos^2 \lambda + Q_{22} \sin^4 \lambda,$$

$$\bar{Q}_{22} = Q_{11} \sin^4 \lambda + 2(Q_{12} + 2Q_{66}) \sin^2 \lambda \cos^2 \lambda + Q_{22} \cos^4 \lambda,$$

$$\bar{Q}_{66} = (Q_{11} + Q_{22} - 2Q_{12} - 2Q_{66}) \sin^2 \lambda \cos^2 \lambda + Q_{66}(\sin^4 \lambda + \cos^4 \lambda),$$

$$\bar{Q}_{12} = (Q_{11} + Q_{22} - 4Q_{66}) \sin^2 \lambda \cos^2 \lambda + Q_{12}(\sin^4 \lambda + \cos^4 \lambda),$$

$$\bar{Q}_{16} = (Q_{11} - Q_{12} - 2Q_{66}) \sin \lambda \cos^3 \lambda + (Q_{12} - Q_{22} + 2Q_{66}) \sin^3 \lambda \cos \lambda,$$

$$\bar{Q}_{26} = (Q_{11} - Q_{12} - 2Q_{66}) \sin^3 \lambda \cos \lambda + (Q_{12} - Q_{22} + 2Q_{66}) \sin \lambda \cos^3 \lambda.$$

If the local elastic properties of the UD composite lamina are known with respect to the material coordinate system, the engineering elastic constants can be determined for the Cartesian coordinate system as follows:

$$E_x = \left[\frac{1}{E_1} \cos^4 \lambda + \left(\frac{1}{G_{12}} - \frac{2\nu_{12}}{E_1} \right) \sin^2 \lambda \cos^2 \lambda + \frac{1}{E_2} \sin^4 \lambda \right]^{-1},$$

$$E_y = \left[\frac{1}{E_1} \sin^4 \lambda + \left(\frac{1}{G_{12}} - \frac{2\nu_{12}}{E_1} \right) \sin^2 \lambda \cos^2 \lambda + \frac{1}{E_2} \sin^4 \lambda \right]^{-1},$$

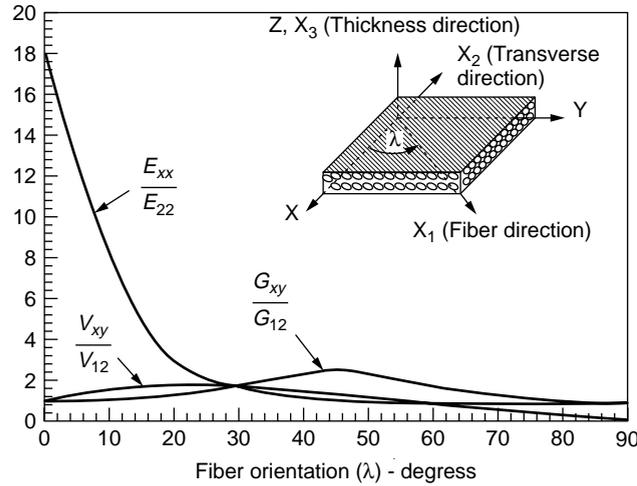


Figure 18. Variations of the engineering elastic constants E_x , G_{xy} , and v_{xy} with the fiber orientation angle λ for a UD carbon-epoxy composite of the following elastic properties: $E_{11} = 139.4$ GPa (20.2 Msi), $E_{22} = 7.7$ GPa (1.1 Msi), $G_{12} = 3.0$ GPa (0.44 Msi), $v_{12} = 0.3$, and $V_f = 0.6$.

$$G_{xy} = \left[2 \left(\frac{2}{E_1} + \frac{2}{E_2} + \frac{4v_{12}}{E_1} - \frac{1}{G_{12}} \right) \sin^2 \lambda \cos^2 \lambda + \frac{1}{G_{12}} (\sin^4 \lambda + \cos^4 \lambda) \right]^{-1}$$

$$V_{xy} = E_x \left[\frac{v_{12}}{E_1} (\sin^4 \lambda + \cos^4 \lambda) - \left(\frac{1}{E_1} + \frac{1}{E_2} - \frac{1}{G_{12}} \right) \sin^2 \lambda \cos^2 \lambda \right]$$

The variations of E_x , G_{xy} , and v_{xy} that result from these equations, with the fiber orientation angle λ relative to the principal material direction, are shown in Fig. 18 for a UD carbon-epoxy composite. It is possible in some cases that the predicted value of E_x may exceed the values of E_{11} and E_{22} depending on the differences among between G_{12} , E_{11} , and E_{22} . By carefully examining Fig. 18, one could envisage how the engineering elastic constants of a composite laminate might be modified according to the orientations of stacked laminae, hence allow performance tailoring characteristics with composites.

Classical Lamination Theory

The most established theory for analysis of laminates takes the form of the Kirchhoff hypothesis for thin plates or classical, linear, thin plate theory. Following the adaptation of this theory for analysis of composite laminates, commonly referred to as classical lamination theory (CLT), the subsequent four assumptions are made:

- Upon application of a load to a plate with a through-thickness, lineal element normal to the plane of the plate, the element undergoes at most a translation

and rotation with respect to the initial coordinate system, but remains normal to the plate.

- The plate resists in-plane and lateral loads only by in-plane action, bending and transverse shear stress, and not by through-thickness, blocklike tension or compression.
- There is a neutral plane, on which extensional strains may not be zero but bending strains are zero in all directions.
- The laminate midplane is analogous to the neutral plane of the plate.

According to the foregoing assumptions for adaptation of the Kirchhoff hypothesis for thin plates, the strain components can be derived from the midplane strains and curvatures. The midplane strains are expressed as $\varepsilon^{\circ xx} = \partial u^{\circ} / \partial x$, $\varepsilon^{\circ yy} = \partial v^{\circ} / \partial y$ and $\gamma^{\circ xy} = (\partial u^{\circ} / \partial y) + (\partial v^{\circ} / \partial x)$, where u° and v° are expressed in terms of the x and y coordinate directions. The midplane curvatures are expressed as $\kappa_{xx} = -\partial^2 w^{\circ} / \partial x^2$, $\kappa_{yy} = -\partial^2 w^{\circ} / \partial y^2$, and $\kappa_{xy} = -\partial^2 w^{\circ} / \partial x \partial y$ and are related to the z coordinate direction. Here κ_{xy} refers to the curvature of twist about the plane of the plate. The strain components are expressed in matrix form as

$$\begin{Bmatrix} \varepsilon_x \\ \varepsilon_y \\ \gamma_{xy} \end{Bmatrix} = \begin{Bmatrix} \varepsilon_{x,0} \\ \varepsilon_{y,0} \\ \gamma_{xy,0} \end{Bmatrix} + z \begin{Bmatrix} -\frac{\partial^2 w}{\partial x^2} \\ -\frac{\partial^2 w}{\partial y^2} \\ -2\frac{\partial^2 w}{\partial x \partial y} \end{Bmatrix}$$

$$\{\varepsilon\} = \{\varepsilon\}_0 + z\{\kappa\}_0$$

The equation above implies that the strains vary linearly with z , meaning that through-thickness sections remain plane and normal after deformation relative to the original coordinate system with its origin at the midplane. If the strains vary linearly, then lamina (ply) stresses must vary in proportion to lamina stiffnesses. In terms of the laminate, the ply stress components are given by

$$\begin{aligned} \{\sigma\}_k &= [Q_{xy}]_k \{\varepsilon\}_k \\ &= [Q_{xy}]_k \{\varepsilon\}_0 + z_k [Q_{xy}]_k \{\kappa\}_0 \end{aligned}$$

where the subscript k denotes the contribution from the k th ply within the composite laminate. According to the plate shown in Fig. 19, the forces and moments have a lineal distribution. In reference to the stress components for the k th ply in the previous equation, force and moment equilibrium are considered. The forces and moments that are responsible for producing in-plane ply stresses are denoted by N_x , N_y , N_{xy} , M_x , M_y , and M_{xy} , where the N 's are the ply-level forces and the M 's are the ply-level moments. For force equilibrium, the integrated, through-thickness laminate stress must be equivalent to the corresponding force that produces it. The total force and moment, determined from contributions of all plies within the laminate,

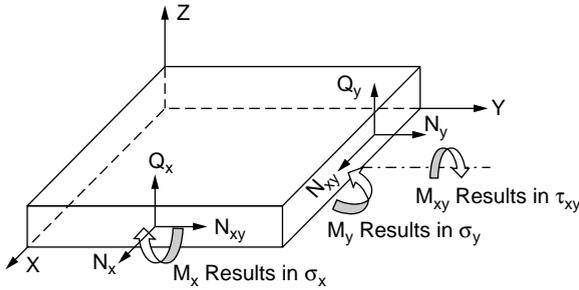


Figure 19. In-plane force and the moment resultants of a laminated plate subjected to extensional forces and bending moments.

can be expressed as

$$\begin{Bmatrix} N_x \\ N_y \\ N_{xy} \end{Bmatrix} = \sum_{k=1}^n \int_{h_{k-1}}^{h_k} (\sigma_{x,k}, \sigma_{y,k}, \tau_{xy,k}) dz,$$

$$\begin{aligned} \{N\} &= \begin{bmatrix} \sum_{k=1}^n [Q_{xy}]_k \int_{h_{k-1}}^{h_k} dz \\ \sum_{k=1}^n [Q_{xy}]_k \int_{h_{k-1}}^{h_k} z dz \end{bmatrix} \{\varepsilon\}_0 + \begin{bmatrix} \sum_{k=1}^n [Q_{xy}]_k \int_{h_{k-1}}^{h_k} dz \\ \sum_{k=1}^n [Q_{xy}]_k \int_{h_{k-1}}^{h_k} z dz \end{bmatrix} \{\kappa\}_0 \\ &= \begin{bmatrix} \sum_{k=1}^n (h_k - h_{k-1}) [Q_{xy}]_k \\ \sum_{k=1}^n \frac{1}{2} (h_k^2 - h_{k-1}^2) [Q_{xy}]_k \end{bmatrix} \{\varepsilon\}_0 \\ &\quad + \begin{bmatrix} \sum_{k=1}^n [Q_{xy}]_k \int_{h_{k-1}}^{h_k} dz \\ \sum_{k=1}^n [Q_{xy}]_k \int_{h_{k-1}}^{h_k} z dz \end{bmatrix} \{\kappa\}_0 \\ &= [A]\{\varepsilon\}_0 + [B]\{\kappa\}_0, \end{aligned}$$

$$\begin{Bmatrix} M_x \\ M_y \\ M_{xy} \end{Bmatrix} = \sum_{k=1}^n \int_{h_{k-1}}^{h_k} (\sigma_{x,k}, \sigma_{y,k}, \tau_{xy,k}) z dz,$$

$$\begin{aligned} \{M\} &= \begin{bmatrix} \sum_{k=1}^n [Q_{xy}]_k \int_{h_{k-1}}^{h_k} z dz \\ \sum_{k=1}^n [Q_{xy}]_k \int_{h_{k-1}}^{h_k} z^2 dz \end{bmatrix} \{\varepsilon\}_0 + \begin{bmatrix} \sum_{k=1}^n [Q_{xy}]_k \int_{h_{k-1}}^{h_k} z dz \\ \sum_{k=1}^n [Q_{xy}]_k \int_{h_{k-1}}^{h_k} z^2 dz \end{bmatrix} \{\kappa\}_0 \\ &= \begin{bmatrix} \sum_{k=1}^n \frac{1}{2} (h_k^2 - h_{k-1}^2) [Q_{xy}]_k \\ \sum_{k=1}^n \frac{1}{3} (h_k^3 - h_{k-1}^3) [Q_{xy}]_k \end{bmatrix} \{\varepsilon\}_0 \\ &\quad + \begin{bmatrix} \sum_{k=1}^n [Q_{xy}]_k \int_{h_{k-1}}^{h_k} z dz \\ \sum_{k=1}^n [Q_{xy}]_k \int_{h_{k-1}}^{h_k} z^2 dz \end{bmatrix} \{\kappa\}_0 \\ &= [B]\{\varepsilon\}_0 + [D]\{\kappa\}_0. \end{aligned}$$

The peculiar mechanical behavior of composite laminates can be discerned by examining the two previous equations. The first equation implies that changes in curvature (bending strains), stretching and squeezing are brought about by the tensile forces and compressive forces given by $\{N\}$. Also the second equation implies that the moments given by $\{M\}$, in addition to changes in curvature, can produce squeezing and stretching strains. From the

force and moment equilibrium analysis, the constitutive relations for laminated composites can be expressed in a condensed form as follows:

$$\begin{Bmatrix} N \\ M \end{Bmatrix} = \begin{bmatrix} A & B \\ B & D \end{bmatrix} \begin{Bmatrix} \varepsilon_0 \\ \kappa_0 \end{Bmatrix},$$

where the A , B , and D matrices are the extension, extension-bending coupling and bending stiffnesses, respectively. Upon expansion of the condensed form, the solution to the stiffnesses can be written in terms of summations of transformed, reduced stiffnesses belonging to individual laminae having h_k th thicknesses:

$$[A] = \sum_{k=1}^n (h_k - h_{k-1}) [Q_{xy}]_k,$$

$$[B] = \sum_{k=1}^n \frac{1}{2} (h_k^2 - h_{k-1}^2) [Q_{xy}]_k,$$

$$[D] = \sum_{k=1}^n \frac{1}{3} (h_k^3 - h_{k-1}^3) [Q_{xy}]_k.$$

Evaluation of the extension, extension-bending coupling and bending stiffnesses, or more simply, the $[ABD]$ matrix serves many purposes in the analysis of composite laminates. This matrix has many uses from the standpoint of designing composite laminates and engineering structures, and it may be used for the following (27):

- Calculating the effective composite laminate elastic properties.
- Calculating the ply-level stresses and ply-level strains for a given load on the laminate.
- Calculating the ply-level stresses and laminate load for a given mid-plane strain.
- Evaluating whether bending strains would result from an extensional load, and vice versa.
- Comparative evaluations of different lay-ups followed by optimization.
- Determining the variation of laminate properties along different directions.
- Calculating the thermal expansion and swelling coefficients of the laminate.
- Estimating the laminate residual stresses due to curing.
- Calculating the ply-level hygral and thermal stresses.

Effects of Orientation and Stacking

The derivation of the $[ABD]$ matrix suggests that the elastic behavior of a composite laminate made from UD laminae is influenced by the constituent fiber and matrix properties as well as the orientations and locations of individual laminae with respect to the geometric midplane of the laminate. The extensional $[A]$ matrix relates the stress resultants with the midplane strains, and the normal stress resultant-to-midplane shear strain coupling and shear stress resultant-to-midplane normal strain coupling

are due to the A_{16} and A_{26} components, respectively. The B_{16} and B_{26} components of the extension-bending coupling $[B]$ matrix relate the normal stress resultants with laminate twisting, and the $[B]$ matrix also suggests the coupling between the moment resultants and the in-plane strains. Finally, the interaction between the laminate bending moment and laminate twisting are related through the D_{16} and D_{26} terms of the bending $[D]$ matrix (28). A physical sense of the coupling effects that exist in relation to the laminate midplane can be seen in Fig. 20(a–b).

If an isostrain condition is assumed for the laminae shown in Fig. 20(a), different stresses will result normal and transverse to the laminae due to their orthotropic behavior. Then, upon bonding and releasing of applied stress, the laminate will distort and bend favorably toward the lamina with higher in-plane stiffness. For the laminate to remain flat, an additional force normal to the plane would be necessary. Similarly, if a uniaxial stress were applied to a laminate having laminae oriented at $+/-\lambda$ and lacking end constraints as shown in Fig. 20(b), twisting about the axis would result due to the extensional-shear coupling arising from anti-symmetry about the midplane.

From a practical standpoint, it is useful to minimize or eliminate these coupling effects, since most engineering structures are required to maintain dimensional stability for long periods of time under various loading and environmental conditions. According to the premises of the $[ABD]$ matrix, coupling can be minimized by selecting the appropriate sequences in which to lay-up individual laminae having various materials, thicknesses, and orientations. This may be referred to as the design of composite laminates and engineering structures.

Two of the most important classes of composite laminate designs from an engineering perspective are symmetric laminates and quasi-isotropic laminates. In symmetric laminates, laminae (plies) on opposing sides of the laminate geometric midplane have the same material, thickness, and orientation. Symmetry about the midplane eliminates the undesirable effects of extension-bending coupling; that is, all of the elements in the $[B]$ matrix become zero and unknown residual stresses from warping deformation are avoided. Except for the cases of cross-ply, all 0° , or all 90° , bending moments in symmetric laminates still produce torsional deflections ($[D]$ matrix). However, the magnitudes can be reduced by increasing the number of plies, for example, in cross-ply configurations.

The notation often adopted in describing a lay-up that is symmetric is as follows: a six-layered stacking sequence expressed as $[0^\circ/-45^\circ/+45^\circ_2/-45^\circ/0^\circ]$ is equivalent to the sequence denoting symmetry expressed as $[0^\circ/-45^\circ/+45^\circ]_s$ provided that the thicknesses and materials are matched below the midplane. The term “quasi-isotropic” as used to describe laminate behavior suggests the same $[A]$ matrix in all directions. Quasi-isotropic laminates exhibit very little variation in apparent elastic moduli with direction, and this becomes useful when the load-direction is unknown or variable.

From the perspective of designing laminates, a laminate can be made isotropic, or nearly isotropic, by having a number of plies greater than four that are equal in thickness and oriented by $2\pi/n$ (n is the total number of plies) to

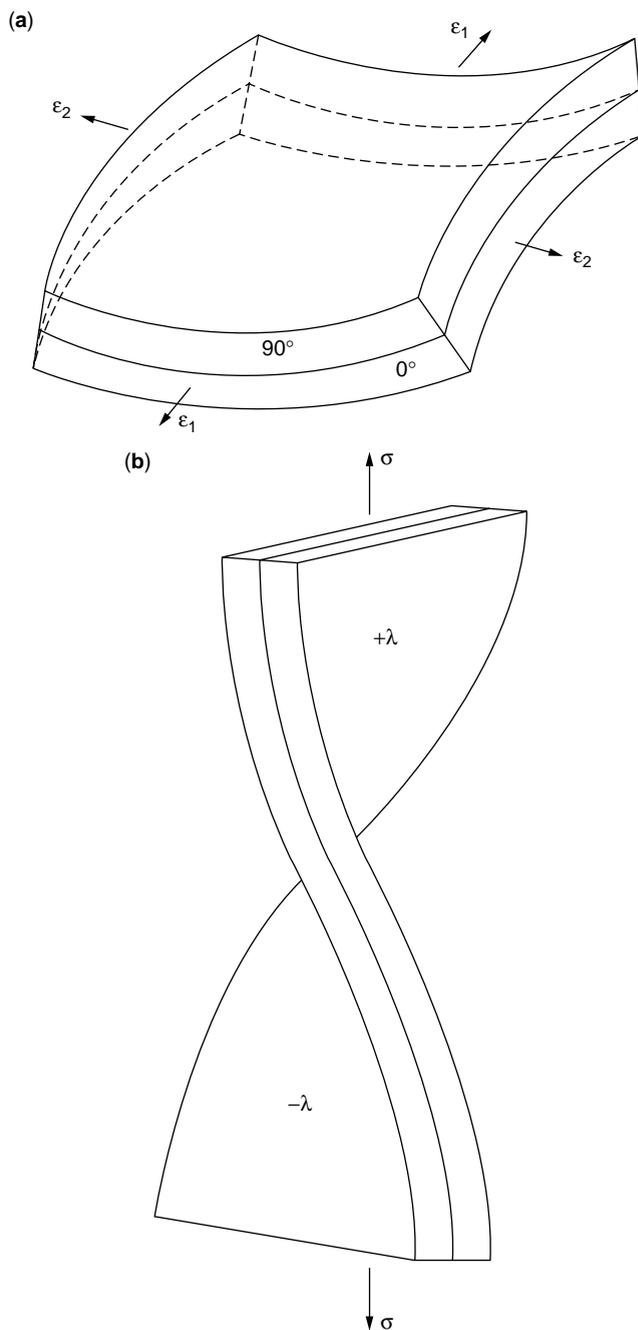


Figure 20. Interpretation of the coupling effects between two bonded composite laminae at various orientations with respect to the geometric midplane: (a) Extensional-bending coupling in well-bonded laminae oriented at 0 and 90° under isostrain conditions, and (b) extensional-shear coupling, which produces twisting in well-bonded laminae oriented at $+\lambda$ and $-\lambda$ to the principal material axis.

adjacent plies. Ideally, quasi-isotropic laminates are symmetric, and symmetric or unsymmetric laminates are at least balanced in thickness, since these designs will tend to be most well-behaved structurally and at least somewhat predictable in response. Examples of symmetric and unsymmetric composite laminate lay-up sequences are shown in Fig. 21.

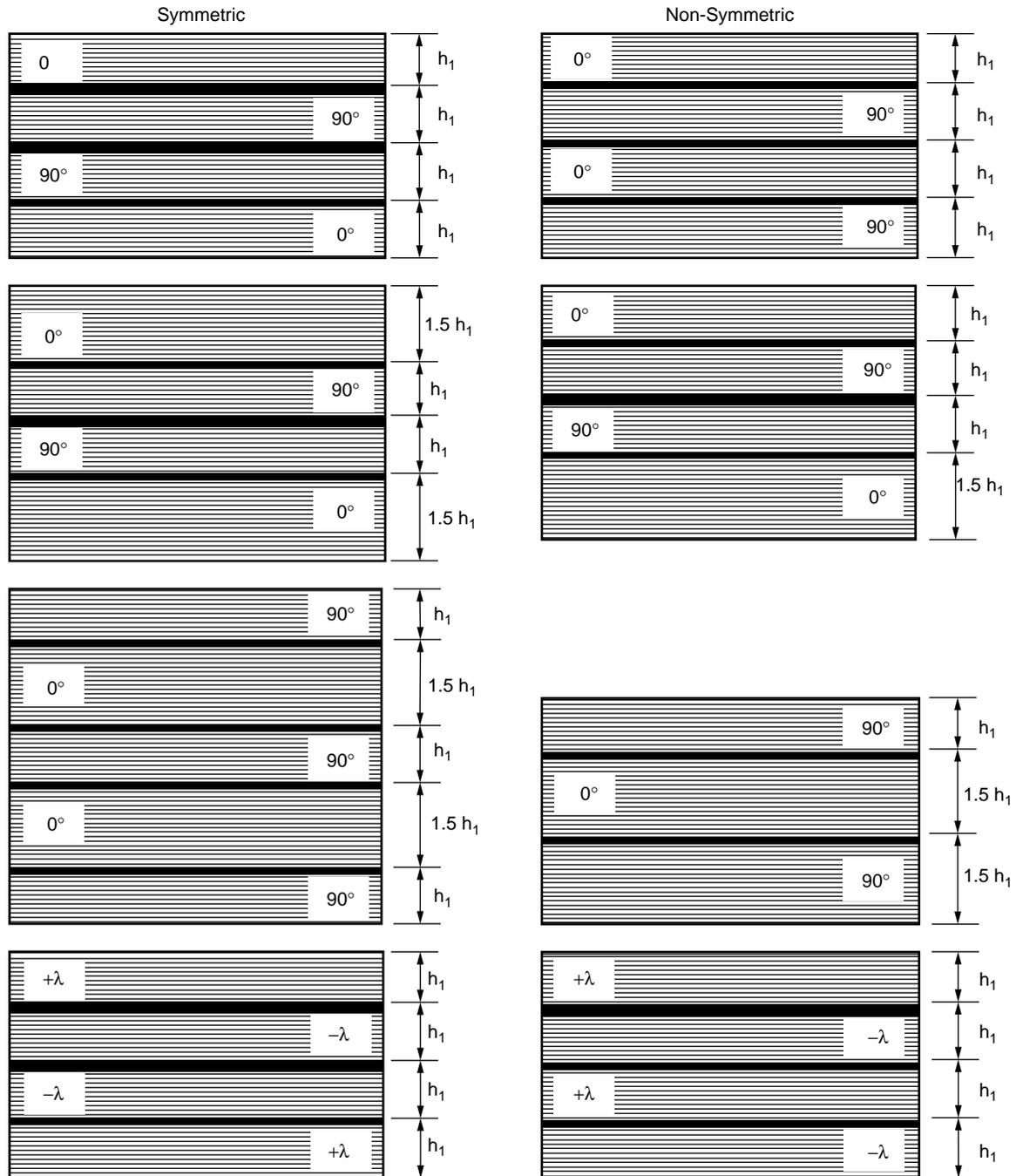


Figure 21. Examples of symmetric and nonsymmetric laminates for the general $0^\circ/90^\circ$ cross-ply and $+\lambda/-\lambda$ angle-ply configurations.

Laminate Failure

Identification of the precise manner in which a composite may fail depends not only on the composite architecture but also on the conditions to which it is exposed. For the purposes of engineering design, it is somewhat less of an arduous task to at least estimate when the composite may fail rather than how it will fail. Failure of a composite may be restrictively considered when failure of the first lamina occurs or more realistically considered when the

composite can no longer support any additional load. The first situation is often referred to as the first-ply-failure (FPF) philosophy, and the second situation is referred to as the ultimate-laminate-failure (ULF) philosophy. With FPF, the inverted $[ABD]$ matrix is used to evaluate the midplane strains and curvature changes in accordance with the applied load vector. Upon evaluating the strains, the stresses in the principal material coordinate system can be calculated and used with any of the composite failure criteria to determine if the applied load vector satisfies a failure

condition. Knowledge of when the first ply failure occurs can lead to appropriate choices for laminate safety factors in design.

ULF extends the application of FPF to the entire laminate. Rather than considering the composite as "failed" once the FPF load is reached, the properties of the failed ply are reduced to values incapable of sustaining load. The "new" composite is re-evaluated, whereby the process is repeated in an iterative fashion until the plies remaining can no longer support any load. At this point, the composite is considered to have failed. Although less conservative than the FPF approach, the ULF approach does offer merit in the sense of capturing the progressive stiffness changes that occur prior to ultimate failure. In this manner, the ULF approach is similar to the classical techniques available for metals.

OTHER CONSIDERATIONS

The particular mechanical behavior associated with composite laminates and structures involves the interactions of many materials on distinct geometric scales. Principles fundamental to the treatment of composite performance in the elastic regime have been presented, notwithstanding considerations for environmental conditions and that new material technologies must also be ascertained. Many applications that are emerging where composite materials may be employed as suitable replacements involve long-term durability in hot and wet conditions. Here knowledge of the hygrothermal effects in a specific composite becomes critical to the design process.

Stresses can be developed in individual plies when they are constrained by neighboring plies against dimensional changes due to thermal and hygroscopic expansions. The distribution of stresses from hygrothermal effects are a function of ply orientation, and the resulting deformation due to these effects may be evaluated by considering the total strain minus the mechanical strain. Since thermal diffusion takes place in composites at a much faster rate than moisture diffusion, the nonmechanical strains due to thermal and moisture exposure may be treated as component effects.

In addition to the continued development of techniques for evaluating the behavior of composites exposed to various environmental conditions, further understanding of the peculiarities with composites is also necessary for future growth toward that of "smarter" structures. That is, such composite structures would not only receive external stimuli in a positive manner but also provide predictable and measurable feedback to those stimuli. To capitalize on the benefits from these structures, designers must explore many of the unresolved issues within the regimes of understanding nonlinear behavior, new (hybrid) material interactions, and constitutive material relations. For example, if we want a material that exhibits piezoelectric, electrostrictive, or magnetostrictive characteristics, then we would introduce phases that exhibit these behaviors. However, the presence of these phases could also result in more complicated predictions of composite behavior due to their interactions and resulting stress redistributions. Since these phases might be incorporated to inhibit some

type of linear or nonlinear response to external stimuli in the first place, the current framework of linear elastic theory may not offer reasonable answers. Consequently, much greater opportunity now exists to offer new theories and ideas to the already established and rapidly progressing comprehension of composite material behavior.

BIBLIOGRAPHY

1. M.F. Ashby. *Materials Selection in Mechanical Design*. Pergamon Press, Oxford, 1992, pp. 1–15.
2. W.D. Compton and N.A. Gjostein. *Sci. Am.* 255: 92–100 (1986).
3. T.W. Chou. *Microstructural Design of Fiber Composites*. Cambridge University Press, Cambridge, 1992, pp. 10–11.
4. R.A. Flinn and P.K. Trojan. *Engineering Materials and Their Applications*, 4th ed. Houghton Mifflin, Boston, 1990, pp. 703–709.
5. D. Hull, *An Introduction to Composite Materials*. Cambridge University Press, Cambridge, 1981, pp. 1–5.
6. M.A. Meador, P.J. Cavano, and D.C. Malarik. *Proc. 6th Ann. ASM/ESD Advanced Composites Conference*. Detroit, Michigan, 1990, pp. 529–539.
7. R.D. Vannucci. *Proc. 32nd Int. SAMPE Symp.* Anaheim, CA, April 6–9, 1987.
8. L.H. Sperling. *Introduction to Physical Polymer Science*, 2nd ed. Wiley, New York, 1992, p. 527.
9. A.V. Pocius. *Adhesion and Adhesives Technology: An Introduction*. Hanser Munich, 1997, p. 81.
10. S.I. Krishnamachari. *Applied Stress Analysis of Plastics: A Mechanical Engineering Approach*. Van Nostrand Reinhold, New York, 1992, p. 355.
11. B.N. Cox and G. Flanagan. *Handbook of Analytical Methods for Textile Composites*. NASA Contractor Report 4750 (1–33). NASA Langley Research Center, Hampton, VA., March 1997, pp. 2:1–2:4.
12. J.C. Halpin and S.W. Tsai. Environmental factors in composite materials design. Air Force Materials Laboratory Technical Report AFML-TR-67-423, 67-423 (1967).
13. C.C. Chamis. *Proc. 38th Ann. Conf. Society of Plastics Industry (SPI)*. Houston, TX, February, 1983.
14. Z. Hashin and B.W. Rosen. *J. Appl. Mech.* 31: 223–232 (1964).
15. Z. Hashin. *J. Appl. Mech.* 46: 543–550 (1979).
16. Z. Hashin. *J. Appl. Mech.* 50: 481–505 (1983).
17. T. Ishikawa and T.W. Chou. *J. Mat. Sci.* 17: 3211–3220 (1982).
18. T. Ishikawa, M. Matsushima, and Y. Hayashi. *J. Comp. Mat.* 19: 443–458 (1985).
19. N.K. Naik and P.S. Shembekar. *J. Comp. Mat.* 26: 2196–2225 (1992).
20. N.K. Naik, P.S. Shembekar, and M.V. Hosur. *J. Comp. Tech. Res.* 13: 107–116 (1991).
21. T.J. Walsh and O.O. Ochoa. *Mech. Comp. Mat. Struct.* 3: 133–52 (1996).
22. K.H. Searles. *The Elastic and Failure Behaviors of 8HS Woven Graphite Fabric Reinforced Polyimide Composites*. Ph.D. thesis. Oregon Graduate Institute of Science and Technology, 1999.
23. K. Searles, G. Odegard, and M. Kumosa. *Mech. Comp. Mat. Struct.* 1999, in press.
24. T. Akasaka. *Comp. Mat. Struct. Jpn* 3: 21–22 (1974).
25. D. Hull. *An Introduction to Composite Materials*. Cambridge University Press, Cambridge, 1981, pp. 156–157.

26. Z. Hashin. *J. Appl. Mech.* 47: 329 (1980).
27. S.I. Krishnamachari. *Applied Stress Analysis of Plastics: A Mechanical Engineering Approach*. Van Nostrand Reinhold, New York, 1992, p. 419.
28. T.W. Chou. *Microstructural Design of Fiber Composites*. Cambridge University Press, Cambridge, 1992, pp. 44–45.

COMPUTATIONAL TECHNIQUES FOR SMART MATERIALS

MANUEL LASO
 JUAN L. CORMENZANA
 ETSII / Polytechnic University of Madrid
 Madrid, Spain

INTRODUCTION

In the following sections, we will use the term “design” in a rather restricted sense. Specifically, we will refer to the calculations, simulations, or in general to any quantitative approach necessary to specify a structure, part, mechanism, processing operation, or function, in which a smart material is used.

In a large number of cases, the design with smart materials relies on well-known and established principles of thermodynamics and continuum mechanics, such as the theories of elasticity (1), fluid mechanics (2), classical electromagnetic field theory (3), chemical equilibrium and kinetics, and solid state physics (4). These theoretical frameworks typically result in a consistent set of equations, of which at least one relates the stimulus and the response of the system. The design task consists often in specifying dimensions of structures or operating conditions of devices that guarantee satisfactory function. A typical example is the design of a smart structure that, under changes in temperature, deforms in a controlled way, possibly operating a valve or tripping a relay switch. The design of such a component, involves a straightforward application of the laws of thermoelasticity, provided that the thermomechanical properties of the material are known.

The controlling principles can often be expressed as very concise and elegant partial differential equations (PDEs) that must be satisfied in domains of complicated shape that have rather involved boundary and initial conditions. This combination of highly nonlinear PDEs, boundary, and initial conditions makes an analytical approach impossible in most cases. Approximate numerical techniques like finite differences (FD), finite elements (FE), finite volumes (FV), spectral methods (SM) and the like are then resorted to often with spectacular success in mechanical and electrical engineering and fluid mechanics (5–7).

In other cases, the difficult part of the design task is not the structural, fluid-mechanical, optical or thermal design itself, but the description of the behavior of the smart material (8). The behavior of a material has been traditionally represented by a so-called constitutive equation (CE) that, put in very broad terms, links stimulus and response. Constitutive equations are used daily in design tasks, sometimes even without our realizing it. For example, one

of the simplest CEs is the linear relationship between the tensorial magnitudes strain and stress for a linearly elastic material, which in its more general form, that is, valid also for anisotropic materials and using the convention of summation over repeated indexes (1), has the following aspect:

$$\sigma_{ik} = \lambda_{iklm} u_{lm}. \quad (1)$$

This expression basically makes the deformation of a material proportional to the cause (stress) and includes, as a special case, Hooke’s law

$$u_{zz} = \frac{\sigma_{zz}}{E}, \quad (2)$$

where E is Young’s modulus.

This very simple CE can be said to be the basis of the vast majority of isothermal linear elastic structural designs. Similarly, most of computational fluid dynamics (CFD) makes use of Newton’s relationship between stress and a velocity gradient:

$$\underline{\tau} = -\eta \dot{\gamma}, \quad (3)$$

where $\dot{\gamma} = (\nabla \underline{v}) + (\nabla \underline{v})^T$ is the strain rate and η is the viscosity.

Again, this very simple CE has extremely wide applicability, smart materials included. A key point worth emphasizing in this context is the fact that constitutive equations are postulates and therefore have a certain degree of arbitrariness. They do not follow from general fundamental principles, as conservation laws do, although they have to conform to certain deeply rooted requirements. Thus, a design problem involves typically a set of fundamental laws, expressed in one of several possible and more or less general ways (thermodynamic, chemical or mechanical equilibrium, conservation of energy, mass and momentum, minimization of action, and minimization of free energy) together with one or more CEs that characterize the material used. The fundamental conservation or variational laws are universal and have to be obeyed by any material we care to consider (Fig. 1).

By way of example, consider now the design of an isothermal flow process of a smart material that behaves as an incompressible memory or viscoelastic fluid. In this case, the fundamental laws that must be satisfied so that the design has physical sense are just two:

$$(\nabla \cdot \underline{v}) = 0 \text{ (conservation of mass),} \quad (4)$$

$$\rho \frac{\partial}{\partial t} \underline{v} + [\nabla \cdot \rho \underline{v} \underline{v}] + [\nabla \cdot \underline{\pi}] = 0 \text{ (conservation of linear momentum),} \quad (5)$$

where $\underline{\pi}$ is the total momentum-flux or total stress tensor which can be split in the following way:

$$\underline{\pi} = -p \underline{\delta} + \underline{\tau},$$

where p is the pressure and $\underline{\tau}$ is the stress tensor due to the fluid. $\underline{\tau}$ is as yet unspecified. We need a CE to define it and close the system of equations. The goodness of our

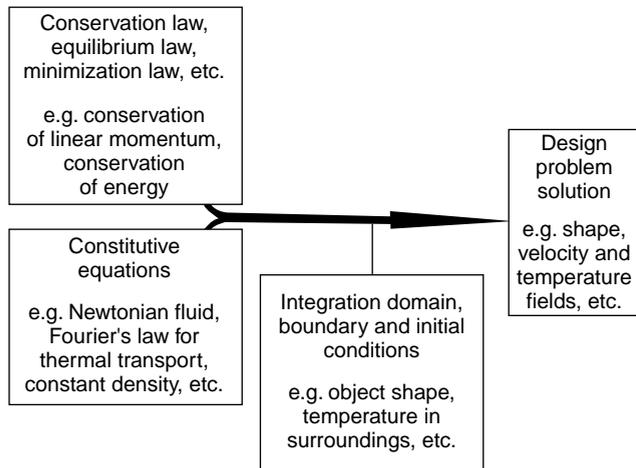


Figure 1. Basic structure of the design task. Conservation laws form an incomplete set of equations that must be closed by one or more constitutive equations.

design will depend critically on the precise form we give to τ , that is, on the way the stress in the fluid is going to depend on its velocity or on the history of the flow.

This CE, on the other hand, is entirely specific for our material. For example, if the material seems to behave as a Newtonian fluid during processing, the most reasonable CE would be Eq. (3). But this is not, of course, our only choice. Within certain boundaries, we are free to specify any relationship between stimulus and response (between strain rate and stress in this case). Even the roles of stimulus and response are interchangeable under some circumstances.

It is often useful to think of the CE as a kind of “calculator” that, when fed a value of the stimulus (say, a strain rate), gives us back the response of the material (say, a stress). This calculator typically contains one or more free parameters that are obtained from a fit to experimental measurements and are specific for the material under consideration. These free parameters are often referred to as “material properties.” Typical material properties are viscosity, elastic modulus, diffusion coefficient, thermal diffusivity, and optical linear and nonlinear susceptibility. In this discussion, it is assumed that material properties are known. In many cases, this assumption entirely eliminates the difficulty of the problem, which often is the characterization of the material. We are dealing here with design problems in which the material is perfectly known, but its behavior in a complex situation has to be determined.

The conservation, equilibrium, or minimization laws form a consistent but incomplete set of equations that require so-called “closure” to become solvable. The CE is the closure. CEs are so often taken for granted, that their very special nature is easily forgotten. It suffices to think of the Navier–Stokes equations, on which most CFD is based: the Navier–Stokes equations are almost automatically taken for granted as the foundation of fluid mechanics. But in fact they are already a combination of the momentum conservation equation and the CE for the Newtonian fluid: they can be obtained by plugging Eq. (3) into Eq. (5) and assuming that the fluid has constant density. Consequently,

whenever we use them to design a flow system, we are automatically and tacitly assuming that the flowing material obeys a very special and simple Newtonian CE. Furthermore, looking beyond the fact that different Newtonian fluids have different numerical values of viscosity, there is only one Newtonian fluid. The same is true for a perfectly elastic solid. All Newtonian fluids, all linear elastic solids, all linear optical materials behave in essentially one and the same way. Therefore, as soon as it is postulated that a smart material obeys one of these simple CEs, the design task becomes relatively simple. It will require only the same standard techniques used for non-smart materials. Such techniques may, of course, be very involved themselves (think of turbulent CFD), but they do not differ fundamentally from the techniques used to design for less smart materials. We will informally call such “standard” cases “design problems of the first kind.” They probably constitute 75% of all design tasks in which smart materials are involved. Because the techniques used for problems of the first kind are the same as those for non-smart materials, they will not be dealt with here in any depth.

In the remaining 25% of the design problems for smart materials, the sophisticated numerical machinery developed during the last four decades is not sufficient to provide reliable solutions in a reasonable time. We will call these “design problems of the second kind.”

The coming sections will be devoted to the two main aspects in which the design and calculation for smart materials departs significantly from standard design techniques. Both aspects are intimately related to the CE or, somewhat ironically, to its nonexistence. We have already seen that the conservation equations are the same for smart and less smart materials. It is the additional complication brought about by the CE that distinguishes these special design or calculation problems.

The first aspect specific of CEs for smart materials has to do with the existence of memory effects. As a matter of fact, some of the most spectacular effects that smart materials display are related to what is somewhat vaguely called memory. The next section discusses some general aspects of memory in materials and its mathematical formulation. In the following section, we consider the question how to handle materials that have memory in practical calculations. Finally, the subsequent section deals with the more fundamental question how to postulate a constitutive equation for smart materials. These last two sections reflect some recent developments in fields that are rapidly developing. A tentative outlook into the future of designing smart materials is presented in the closing section.

SMART MATERIALS, MEMORY EFFECTS, AND MOLECULAR COMPLEXITY

Frequently, complicated material behavior is closely related to the concept of memory, a key word very often heard in the context of smart materials. For example, form or shape memory materials constitute one of the best known classes of smart materials mainly due to the spectacularity of some of its applications (9). Less widely known, but also capable of displaying a stunning range of

nontrivial and often counterintuitive behavior, the large family of complex fluids has found numerous applications as smart materials. Smart materials based on complex fluids include gels, polymeric melts and solutions, colloidal dispersions, electro- and magnetorheological fluids, and liquid crystalline materials among others (10–13). All of these materials share some or all of the following characteristics, which are in many cases responsible for their smart behavior:

- They possess the ability to react to external excitations (fields, pH changes, strain) in a highly nonlinear way and often undergo a phase change.
- They display their most interesting behavior when driven strongly away from equilibrium.
- They typically possess either a very wide spectrum of relaxation times or else a main relaxation time, whose order of magnitude is comparable with the timescale of the physical process in which they find application (14).
- They often have a complicated small-scale structure, either at the molecular level (polymers) or at some mesoscopic scale (dispersions, emulsions, polycrystalline materials).

Although there is no single mechanism responsible for memory, several of these characteristics are responsible for phenomena such as hysteresis and memory. In some cases, a material appears to have a fading memory of past events because its internal structure (e.g., molecular) requires a certain time to adapt to a changing environment. Thus, memory is also a question of the timescale in which the relevant material property is observed. We can take water as an example. It is a low molecular weight fluid that behaves as a simple Newtonian fluid in most cases because its characteristic molecular relaxation time is very short compared with the characteristic times of flow in everyday life. Thus, it can adapt instantly to changes in the velocity field and therefore displays no memory effect. On the other hand, a high molecular weight polymer has a spectrum of relaxation times that can reach well into seconds. Any stimulus, for example, a change in an electric field, that tries to change its orientation will be followed by an observably slow response, during which the material retains information about the past state.

In other cases, memory is due to a kinetically frozen-in state, for example, due to a martensitic–austenitic phase transition, which can be unlocked by applying an external stimulus. The material is then forced to revert to a previous state and thus appears to possess memory.

There have been several attempts to capture these phenomena in a mathematical formulation. At this point, instead of addressing the question in an all-encompassing and general way, we will rather continue with our specific example, which is a representative example of materials that have complex behavior. We will address the family of high molecular weight polymers, which are considered by many as memory fluids par excellence.

Polymers display strong memory effects that are a consequence of their non-Newtonian nature and ultimately of their complex molecular structure and of the

entanglements they form, either in solution or as melts. Whereas there is just one CE for Newtonian fluids, literally dozens of CEs for non-Newtonian fluids have been proposed (13,15). Most of them directly or indirectly attempt to take into account memory effects. One of the simplest CEs that attempts to take into account both viscous and elastic behavior is that of the so-called Oldroyd-B fluid (16):

$$\underline{\tau} + \lambda_1 \underline{\dot{\tau}} = -\eta_0(\dot{\gamma}_{(1)} + \lambda_2 \dot{\gamma}_{(2)}), \quad (6)$$

in which $\dot{\gamma}_{(1)}$ and $\dot{\gamma}_{(2)}$ are kinematic tensors defined by

$$\dot{\gamma}_{(1)} = (\nabla \underline{v}) + (\nabla \underline{v})^T \text{ (rate of strain tensor)}, \quad (7)$$

$$\dot{\gamma}_{(2)} = \frac{\partial}{\partial t} \dot{\gamma}_{(1)} + \{\underline{v} \cdot \nabla \dot{\gamma}_{(1)}\} - \{(\nabla \underline{v})^T \cdot \dot{\gamma}_{(1)} + \dot{\gamma}_{(1)} \cdot (\nabla \underline{v})\}, \quad (8)$$

where η_0 is the zero-shear-rate viscosity, λ_1 is the relaxation time, and λ_2 is the retardation time. The first remarkable aspect of this CE is its complication, although it is one of the simplest CEs for non-Newtonian fluids. It is also worth noting that, written in this differential way, the memory aspect is not very obvious. It is possible, however, to rewrite this CE in a different but entirely equivalent way:

$$\underline{\tau}(r) = \eta \dot{\gamma}_{(1)} + \int_{-\infty}^t \left\{ \frac{nkT}{\lambda_H} e^{-\frac{(t-t')}{\lambda_H}} \right\} \dot{\gamma}_{(1)}(r, t, t') dt', \quad (9)$$

in which $\dot{\gamma}_{(1)}(t, t')$ is the finite strain tensor defined by

$$\dot{\gamma}_{(1)}(r, t, t') = \underline{\delta} - \left(\frac{\partial}{\partial r'^r} \underline{r} \right)^T \left(\frac{\partial}{\partial r'^r} \underline{r} \right), \quad (10)$$

and $\underline{r} = \underline{r}(r', t, t')$ tells us where (\underline{r}) a particle is located at time t knowing that it was located at \underline{r}' at time t' . This superficially harmless last sentence is notoriously perverse: first of all, the instantaneous value of the stress no longer depends in any explicit way on the velocity. Nowhere in Eq. (9) or in Eq. (10) is the velocity to be seen (compare this to Eq. (3) where the stress and velocity gradient appear explicitly in the same equation). Instead, the stress depends on the whole history of the deformation of the fluid, as expressed in a deviously indirect way by Eq. (10). Second, to determine the present value of the stress, we must know the entire past of the flow. But we will know the past history of the flow only if we can compute previous stresses also, that in turn requires the knowledge of *their* flow past, and so on. This kind of infinite regress is unheard of in nonmemory materials: given a strain rate, the Newtonian fluid produces a given stress that depends only on that instantaneous strain rate and not on any other aspect of the past. The Newtonian fluid reacts infinitely fast to an external stimulus and consequently has no memory. Our memory fluids react to the present strain rate in a way that depends on their whole flow history through equations like Eqs. (9) and (10) or even more complicated ones.

This alternative integral way of writing the CE, although not much more transparent, does show how memory effects can be formulated mathematically: the stress at any given time and position $\underline{\tau}(r)$ depends on all of the previous history of the flow (through the term $\dot{\gamma}_{(1)}(r, t, t')$ and through the integration). Recent flow or deformation

events influence more the resulting $\tau(r)$ than long past ones, due to the exponential memory function. Thus, the material has a fading memory.

Flow properties and memory of smart materials are characterized by the numerical values of the parameters in this and other similar but more complex CEs. More complex CEs make more physically correct predictions of material behavior but at the cost of greater complexity. It is useful now to realize that the whole field of Newtonian CFD is based on the mass, energy, and momentum conservation equations closed by the very simple Newtonian CE given in Eq. (3). The fluid mechanics of non-Newtonian memory fluids is controlled by the same conservation laws but augmented by a CE similar to or more complicated than Eq. (9). In practice, this complication makes any calculations of memory fluids in realistic three-dimensional geometries quite complex and extraordinarily time-consuming. In the previous example of the Oldroyd-B fluid, the conservation equations can be expressed in a FE calculation in a weak Galerkin form together with the CE either in differential Eq. (6) or in integral Eq. (9) forms; the latter is generally more cumbersome. Other CEs admit only an integral form (17), and until now, its use in practical calculations has been very limited.

Furthermore, the very mathematical nature of the problem is changed by the CE, so that unexpected and fundamental mathematical difficulties appear that often represent an insurmountable barrier. To the shock of the rheological community, many deceptively simple non-Newtonian flow problems still resist all attempts at solution.

The reader may get the impression that the mathematical complexity of the CE for memory fluids must reflect some deep-lying complicated physical behavior. The astonishing fact is that a mathematical formulation as intricate as Eq. (9) and those even much more complex can be deduced from or correspond to any one of several extremely simple molecular descriptions of the material (one of which we will show in the next section). In the following, we will refer to this game of postulating a molecular picture of a material and extracting a CE from it, as “solving the kinetic theory of the material.”

Although the previous paragraphs refer to a specific type of smart material behavior, namely, memory fluids, the discussion has general validity. Any material property (mechanical, optical, or chemical) that somehow depends on the past history of the excitation will lead to a similar situation. For example, for a shape memory material, the elastic constants will depend on the history of the strain or of the temperature or both.

Because smart materials can be expected to be complicated or structured, the natural question now arises what happens when we want to put more realism into the underlying molecular picture. If an extremely simple model leads to quite a complex CE, what kind of CE will we obtain for a more physically correct molecular picture of the smart material? The answer is that almost immediately the kinetic theory of the material becomes unsolvable. In other words, it is very easy to develop not too sophisticated a molecular model of a material for which we cannot obtain a CE, no matter how hard we try. Lacking a CE, that is, if

one of the equations is missing, how can we possibly expect to solve the set of equations that describe smart material behavior? Smart materials easily outsmart us if we follow the strategy of a frontal attack. But not everything is lost: there are alternative and very straightforward ways to bypass the difficulty of the nonexistent CE. The next section is devoted to them.

CONTINUUM AND MOLECULAR DESCRIPTIONS OF SMART MATERIALS

In the previous section, we introduced the idea that smart material behavior almost always stems from underlying microscopic or mesoscopic complexity. Expressing this complexity in mathematical form, that is, solving the kinetic theory of the material, is very often impossible. Even in those cases where it is possible, predicting material behavior is a tough challenge. We will now further use our example of one of the simplest memory fluids, the Oldroyd-B fluid, to show how the apparent conundrum presented in the previous section can be approached.

Thermoplastic polymers are a class of materials whose behavior can be approximately represented by a CE like Eq. (9). These polymers are made of very long molecules, and have a backbone comprising several thousand atoms bonded covalently. These bonds have the possibility of rotating at the cost of some torsional energy, either by spontaneous thermal agitation or by the application of some external field (e.g., electrical) or deformation. Once the external effect disappears, these huge molecules tend to regain their average shape by releasing the torsional energy stored in the backbone of the molecule and adopting molecular conformations similar to the statistical coil (18). This tendency to go back to states of minimum free energy results in an approximately linear restoring force that acts on the whole molecule. This spring-like force opposes molecular stretching. If suspended in a liquid, that is, if the polymer molecule is in solution, it will also be subjected to random thermal bombardment by small solvent molecules and the effect of any velocity field of the solvent. This additional effect of the solvent is a double one: a tendency to deform the polymer molecule and a drag due to the relative motion between the molecule in solution and the solvent.

Treating all of these effects and the chemical structure of the molecule in a fully detailed way is completely beyond our current capabilities. Instead, a coarse-graining procedure is invoked: most of the details of the molecule are discarded, and only those most relevant are kept. A very coarse-grained picture of the molecule is shown in Fig. 2. The whole macromolecule in its fully detailed chemistry is represented by a dumbbell that consists of two masses joined by a Hookean spring. This dumbbell is fully characterized by specifying only three numbers: the components of the connector vector Q . The dynamics of this simplified mesoscopic object, which we will almost unjustifiably still call “molecule” can be written quite easily. The differential change in the connector Q in a very short time dt is

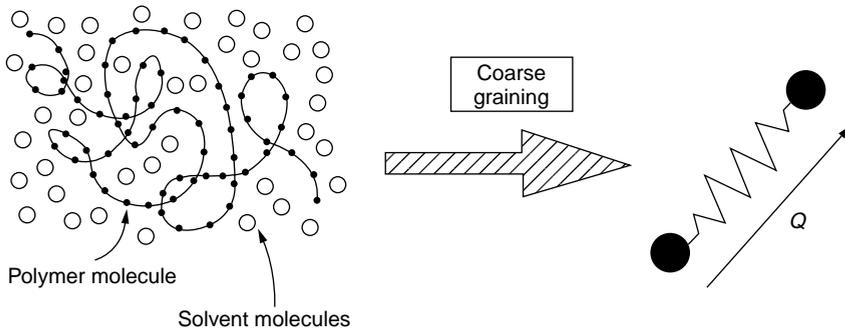


Figure 2. Coarse graining a polymer molecule in solution.

given by

$$d\mathbf{Q} = \left([\boldsymbol{\kappa} \cdot \mathbf{Q}] - \frac{2H}{\zeta} \mathbf{Q} \right) dt + \sqrt{\frac{4kT}{\zeta}} d\mathbf{w} \quad (11)$$

(see the end of the article for a list of symbols). In Eq. (11), we can recognize the spring restoring force (term containing $2H\mathbf{Q}/\zeta$), the effect of the flow field of the solvent ($[\boldsymbol{\kappa} \cdot \mathbf{Q}]$), the random buffeting by solvent molecules ($\sqrt{\frac{4kT}{\zeta}} d\mathbf{w}$), and the solvent drag ($\frac{2H}{\zeta}$). Thus, in a coarse-grained sense, our polymer solution consists of a suspension of these highly simplified objects in a Newtonian solvent.

The obvious question now is what has this absurdly simple picture of a polymer solution got to do with a CE like Eq. (9), that predicts non-Newtonian behavior, memory included. The unexpected answer is that if we rigorously solve the kinetic theory of a suspension of dumbbells in a Newtonian solvent, we end up with a relationship between stress and strain history that is precisely our Eq. (9) (15). We were referring to this exact equivalence in the previous section when it was stated that extremely simple molecular models lead to very complicated and often intractable CEs. Therefore, for example, when designing a polymer solution process using Eq. (9), we know we face serious computational difficulties. But we are willing to invest additional time and effort in the hope that a complex CE will reflect a very complex material structure and dynamics. However, whether or not we are aware of it, we are implicitly assuming that the polymer is nothing else than a suspension of Hookean dumbbells. In spite of their extreme simplicity, it is the presence of the dumbbells that endows the solution with memory. Dumbbells are not that dumb after all.

This interesting equivalence between simple mesoscopic molecular models and extremely complicated CEs can be established for most of the CEs used nowadays. But there still remains the immensely larger class of not-so-simple molecular models for which there never will be a CE, but which, nevertheless, are much better at capturing smart material behavior. We are naturally interested in tapping the resources of these more advanced molecular models. But how can we use advanced models for smart materials, if it is not even possible to write a CE for them?

The answer lies in a further connection between the micro- or mesoscopic molecular picture and the macroscopic response of the system. This missing ingredient is actually the simplest. In Eq. (11), the dynamics of the simplified molecular model was written in full detail. If we knew the initial state (its initial \mathbf{Q}) of a given molecule, we could predict its evolution, that is, its future states, by integrating Eq. (11) for as many time steps as we like. In this way, we would know how a single molecule evolves. This is clearly not enough because a polymer solution contains a very large number of such objects swimming in a Newtonian fluid. Furthermore, the stress $\boldsymbol{\tau}$ is a collective property of this large number of molecules (we call this population of molecules an “ensemble”). A single molecule does not allow us to determine the stress. What if we had a large, ideally infinite, collection of different dumbbells? Would it then be possible to obtain the stress from this ensemble? Fortunately enough, the answer is yes: the following simple formula tells us how to compute the macroscopic response of the material, the stress in this case, for an ensemble of molecules:

$$\boldsymbol{\tau} = -nkT \overline{(\mathbf{Q}\mathbf{Q} - \delta)}, \quad (12)$$

where the overbar means an ensemble average, that is, we compute the simple arithmetic average $nkT [\frac{1}{N} (\sum_{i=1}^N \mathbf{Q}_i \mathbf{Q}_i) - \delta]$ using an ensemble of N molecules and, within the statistical error bars due to finite ensemble size, the result is precisely $\boldsymbol{\tau}$. It is quite unexpected at first that we can obtain the same result for the stress of the material either using the idea of the ensemble or integrating Eq. (9). The connection between these two ways of obtaining the macroscopic response of the material is completely rigorous and stands on sound mathematical footing. But even without going into its details, this connection is not totally unexpected. Because Eq. (9) is an exact result of the kinetic theory of Hookean, noninteracting dumbbells, its predictions should be identical to those obtained from a direct simulation of a large number of such objects (Fig. 3).

Now, we know how to describe the behavior of a memory material at both the continuum level, Eq. (9), and at the microscopic level, Eqs. (11) and (12). We have also gained a great deal of insight into the mechanisms of memory or smart material behavior: in the continuum mechanical version, memory is introduced as an integral across the

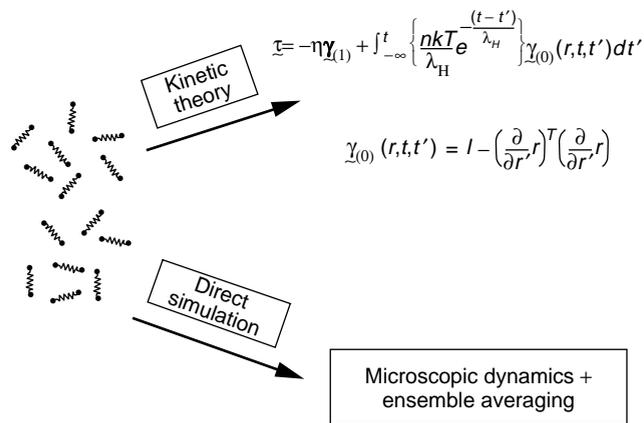


Figure 3. Correspondence between analytical constitutive equation for a memory fluid and direct simulation of an ensemble.

history of the flow, but it does not shed any light on the molecular mechanism responsible for memory. In the microscopic dynamic view, memory is stored in the present configurations of the dumbbells. If the memory fluid has been at rest, the \underline{Q} will be distributed isotropically and will yield, on average, a value of $\underline{\tau} = 0$. If the polymer has been subjected to a long period of shear or elongation, again on average, its molecules will be stretched and oriented. Their connectors \underline{Q} will be large and highly oriented and thus will make an important contribution to $\underline{\tau}$. If a change in the flow field takes place, for example, if flow ceases, the molecules will still be in predominantly stretched configurations for a period of time controlled by the typical relaxation time of the molecules (e.g., λ_1 for the Oldroyd-B fluid). The stress will drop to zero, not instantly, but over a timescale of the order of λ_1 . This gradual loss of information about previous events (shear or elongational flow in this example) is what we recognize as memory in a material. Thus, we see that the very simple fact that the internal structure of the molecule needs time to adapt to a new external stimulus is already sufficient to produce memory effects. Other, more advanced molecular models for smart materials have notoriously more complicated microscopic dynamics which take into account other relaxation processes of the individual polymer molecule and also of the surrounding molecules. The basic ingredient of memory, namely, the existence of a comparatively slow relaxation mechanism, is not very different, however.

Although we have seen some of the niceties of the equivalence between the continuum and microscopic pictures of our memory fluid, we still have to address the question how to solve design problems for smart materials in complex situations (shapes, boundary conditions).

Because there is no CE for the vast majority of consistent microscopic material models, with which we can close the design equations (4) and (5), it seems that there would be no hope of ever performing a design using a reasonably advanced CE. Recently, however, a number of approaches have been proposed that use the equivalence between the continuum and microscopic pictures of a material (19–21). Continuing with our example of the flow of a memory fluid, assume that we want to determine the amount of swelling

this material experiences upon exit from a cylindrical channel. The unknowns of the problem in this case are the velocity field and the free surface. We will further assume that we will be using a finite-element technique. In this method, the unknown fields are discretized on a mesh, and the solution sought consists of the values of the velocity at the nodes of the mesh and the coordinates of the free surface.

Solving the problem means obtaining a velocity field and a shape of the boundary that satisfies $(\nabla \cdot \underline{v}) = 0$ and $\rho \frac{\partial}{\partial t} \underline{v} + [\nabla \cdot \rho \underline{v} \underline{v}] + [\nabla \cdot \underline{\pi}] = 0$ where $\underline{\pi} = -p \underline{\delta} + \underline{\tau}$ and $\underline{\tau}(r)$ and comes from the CE [e.g., Eq. (9)]. There are additional free surface boundary conditions for the velocity and the stress which are not important for our discussion. In a standard FE (Finite Element) formulation, a large nonlinear set of equations is set up in which some of the unknowns are the components of the velocity and the coordinates of the boundary. The stress is evaluated at several integration points within each element of the grid by using Eq. (9). This value of the stress is then used to complete the momentum conservation equation, which is solved for velocity. This is fine so long as we have a CE, but what are we to do when there is no CE to describe the material? In the light of the correspondence between the continuum level and the microscopic levels, an alternative suggests itself naturally: we can fill all of the elements that make up the integration domain with a large number of molecules, dumbbells in our example, and use them to compute the stress in each element by using Eq. (12). The dumbbells in an element form a local ensemble. This local ensemble serves as a stress calculator that closes the mass and momentum conservation equation, just as an analytical CE does. Dumbbells are entrained and deformed by the fluid. The strong coupling between macroscopic flow and microscopic molecules is then very obvious: the macroscopic flow carries and distorts the dumbbells, which in turn produce the correct response (stress) that modifies the velocity field. This cycle is repeated as many times as we wish or until we reach a steady state. The general scheme of such a micro/macro method is illustrated in Fig. 4.

For our example of the flow of a memory material out of a cylindrical pipe, Fig. 5 illustrates a typical FE grid, a schematic representation of the “molecules” and the solution given as values of the velocity vector at the nodes of the grid.

Following this idea, we can have as complicated a molecular model as we want without worrying about its kinetic theory, that is, whether or not we can obtain a CE for it. This basic idea of combining a macroscopic formulation of the conservation equations with a direct simulation of a large ensemble of microscopic molecules is extremely powerful. It opens the way to the development and practical use of much more realistic molecular models than was possible until now. There is, of course, a price to pay for this extra power: because our ensembles can never be infinitely large, the computed $\underline{\tau}(r)$ will contain statistical noise and so will the velocity field. Besides, the calculation will be more expensive now, because we have to follow the dynamics of hundreds of thousands or millions of simple molecules. Some very recent advances in the area aim precisely

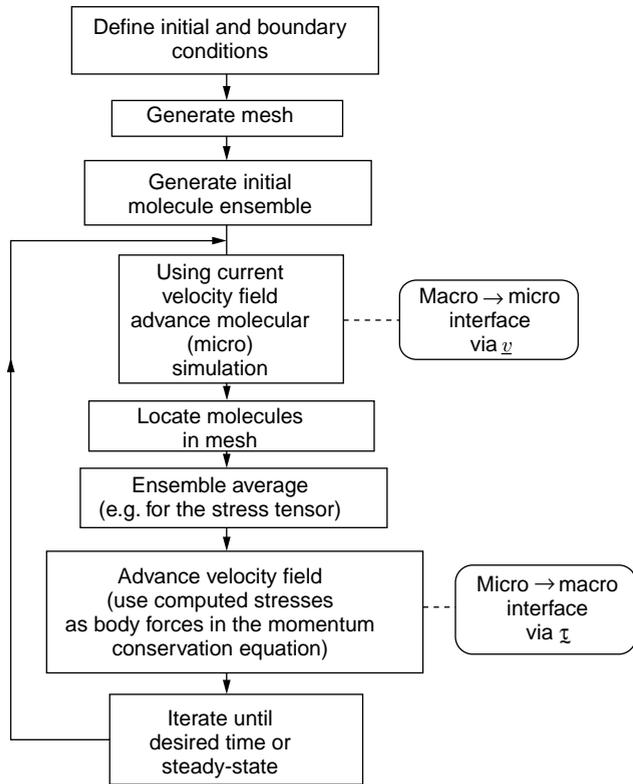


Figure 4. Basic time-marching scheme in a micro/macro method.

at reducing both the statistical noise and the computational effort by resorting to variance reduction techniques (20).

The basic idea behind all of these schemes is actually very simple and has been around under several disguises for quite some time (22). The degree of sophistication and the range of applications of such combined methods are truly phenomenal. In (23), an atomistic simulation taken from the field of fracture mechanics at a very basic level (density functional theory actually solves Schrödinger's equation for a fairly large collection of atoms of the material being investigated) has been successfully combined with higher level methods to predict crack propagation very satisfactorily. Needless to say, the microscopic model need not reside at the most basic level: some very significant and also spectacular applications come from the field of solidification and casting of metallic alloys (24). In this kind of micro/macro model, the microscopic level resolves metallic crystalline structures such as dendrites, and the macroscopic level has a typical length scale of centimeters. Clearly, the term "microscopic" in alloy casting picks up where the macroscopic level of fracture propagation leaves off. In both cases, the basic idea is the same: try and bridge the gap in time- and space scales by hierarchical modeling.

In spite of this additional extra cost, micro/macro methods are starting to find widespread application whenever the behavior of a material is too complex to be tackled by standard continuum mechanical techniques. Coupling very detailed microscopic descriptions of the

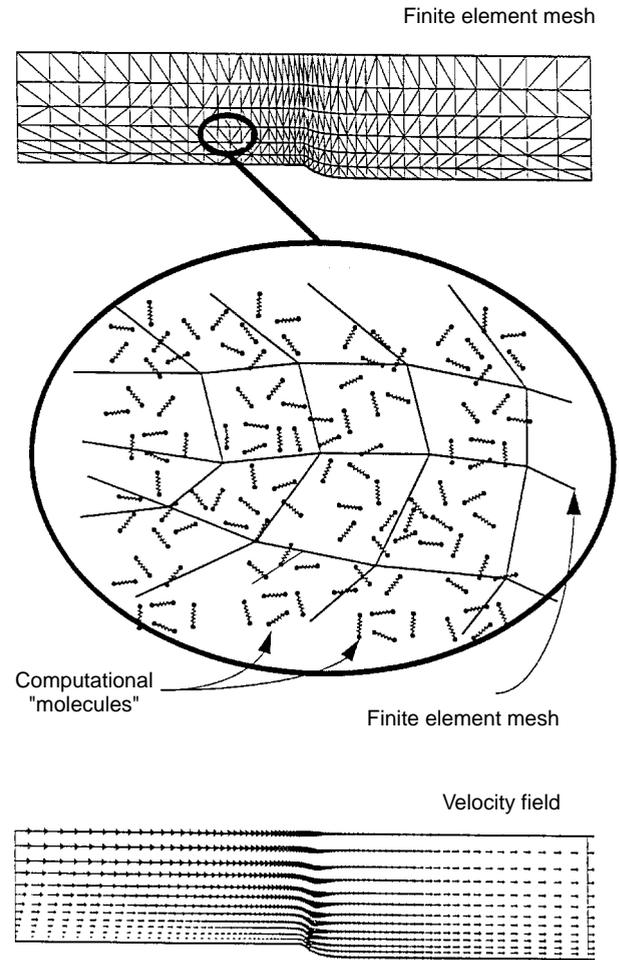


Figure 5. Calculation of the flow of a memory fluid using finite elements in an integration domain. The conservation and CE equations are discretized and solved on the grid. The solution is the velocity field shown and the shape of the domain (free liquid surface).

material with macroscopic methods allows making design calculations that were unthinkable as recently as a decade ago.

SMART MATERIALS AND NONEQUILIBRIUM THERMODYNAMICS

As already mentioned, one of the key features of smart materials is that they frequently have to operate far away from equilibrium. There is considerable freedom in the process of establishing a microscopic model of the smart material and extracting a relevant macroscopic property from it [for example, when obtaining the stress from an ensemble of dumbbells via Eq. (12)]. This freedom is not complete, however, because any micro- or mesoscopic model that we set up must comply with the rules of nonequilibrium thermodynamics. Major developments in the field of nonequilibrium thermodynamics or nonequilibrium statistical mechanics have been few and far apart (25). The application of nonequilibrium thermodynamics to complex materials is by no means obvious. At present, there

are a staggering number of different theories and a wide variety of approaches: classical nonequilibrium thermodynamics with internal variables, Lagrangian methods, bracket formulations, continuum or rational thermodynamics, variational formulations, extended irreversible thermodynamics, the matrix model, network thermodynamics, and the recent GENERIC formalism. In the last few years, however, there is a growing consensus that it should be possible now to combine the various approaches in some kind of common generalized theory. The final goal is to bring nonequilibrium thermodynamics to the same level of clarity and usefulness as equilibrium thermodynamics.

Unless it can be shown that the microscopic models we develop for a smart material satisfy the rules of equilibrium thermodynamics, there is no guarantee that they will be consistent, let alone able to make consistent predictions. However, given the lack of a unified formulation of nonequilibrium thermodynamics, microscopic models have been proposed based largely on intuition and have been used in practical work. Dynamic equations have been postulated directly with varying degrees of success. The unavoidable consequence is that the required self-consistency is missing in some of these models.

Given the fundamental importance of thermodynamic consistency in smart material modeling, this chapter would not be complete without at least mentioning these very recent advances. We will limit ourselves to summarily sketching one of the most recent theories of nonequilibrium thermodynamics (GENERIC is an acronym for general equation for nonequilibrium reversible-irreversible coupling). Although it is not possible to go into a detailed discussion of any of the current frameworks of nonequilibrium thermodynamics, it is important to give at least an idea of their structure, their main building blocks, and the kind of predictions they can make.

For example, in the GENERIC framework (26,27), the temporal evolution of any isolated thermodynamic system is given in the form,

$$\frac{dx}{dt} = L(x) \cdot \frac{\delta E(x)}{\delta x} + M(x) \cdot \frac{\delta S(x)}{\delta x}, \quad (13)$$

where x represents a set of independent state variables required to describe a given nonequilibrium system completely. $\delta/\delta x$ is to be understood as a functional derivative and the application of the operators implies summations over discrete labels and also integrations over continuous labels. The functionals E and S represent the total energy and entropy expressed in terms of the state variables x , and L and M are certain matrices. The two contributions to the temporal evolution of x generated by energy E and entropy S in Eq. (13) are called the reversible and irreversible contributions, respectively. Using the energy as the generator of reversible dynamics is inspired by Hamilton's description of a conservative system. Using the entropy as the generator of irreversible dynamics is inspired by the Ginzburg–Landau formulation of relaxation equations. The use of these two generators is a key aspect

of GENERIC. It has special importance and makes it capable of treating systems far from equilibrium.

In GENERIC, Eq. (13) is supplemented by the two degeneracy requirements:

$$L(x) \cdot \frac{\delta S(x)}{\delta x} = 0$$

and

$$M(x) \cdot \frac{\delta E(x)}{\delta x} = 0. \quad (14)$$

Furthermore, the matrix L is required to be antisymmetric, whereas M is symmetrical and positive-semidefinite. The positive-semidefiniteness of M together with the first degeneracy condition imply a strong implementation of the second law of thermodynamics. Both the complementary degeneracy requirements and the symmetry properties of L and M are extremely important in formulating proper L and M matrices when modeling nonequilibrium materials. Finally, it is assumed that the Poisson bracket $\{\dots, \dots\}$ associated with the antisymmetric matrix L ,

$$\{A, B\} = \frac{\delta A}{\delta x} \cdot L(x) \cdot \frac{\delta B}{\delta x}, \quad (15)$$

satisfies the Jacobi identity,

$$\{\{A, B\}, C\} + \{\{B, C\}, A\} + \{\{C, A\}, B\} = 0, \quad (16)$$

for arbitrary functionals A , B , and C . The Jacobi identity severely restricts convection mechanisms for structural variables and expresses the time-structure invariance of the reversible dynamics implied by L .

The power of any such formulation of nonequilibrium thermodynamics resides precisely in the ability to define which micro- or mesoscopic models are permissible and consistent. Until recently, due to the lack of a general framework, any ad hoc or intuitively proposed microscopic dynamics run the risk of violating the degeneracy requirements or the Jacobi identity. Using a consistent framework such as GENERIC to constrain our microscopic material models automatically guarantees consistency.

The GENERIC formulation of nonequilibrium thermodynamics has led, among others, to fully consistent generalized reptation models and to new models for liquid crystal polymers, both of great importance for applications in the area of advanced materials. The tremendous advantage of having a framework of nonequilibrium thermodynamics at our disposal in which to formulate microscopic models is that consistency is guaranteed by construction. Furthermore, such a formalism acts as a helpful guide in improving and refining microscopic material models, that is, ultimately in the quality of the resulting CEs (when it is possible to write one) or of the resulting stimulus–response behavior.

Nonequilibrium thermodynamics is probably an even more important tool for engineers than equilibrium thermodynamics, for example, in connection with the design

and processing of all kinds of memory materials. Nonequilibrium thermodynamics is obviously important if phase transitions, such as phase separation of crystallization, take place under deformation or flow. The successful application of a formalism such as GENERIC to such problems depends on the possibility of obtaining the four building blocks of GENERIC: the two generators, energy E and entropy S , and the matrices L and M . Thermodynamic modeling in terms of these basic building blocks is strongly advocated, rather than the direct formulation of temporal-evolution equations. The situation is analogous to that in equilibrium thermodynamics: it is more advantageous to work with a thermodynamic potential as a basic building block than with several equations of state. Experience with empirical expressions for the GENERIC building blocks for different design and material cases needs to be collected by reformulating and generalizing existing theories. Although microscopic expressions for the building blocks do exist, they will become useful only when the numerical methods for handling these formal expressions are developed. Just as Monte Carlo simulations allow us an atomistic understanding of equilibrium physics, molecular and stochastic dynamics will be the key to applying non-equilibrium thermodynamics to practical cases. The basic tools for understanding structure–property relationships are available now.

OUTLOOK

To summarize, the “smartness” of smart materials almost always has its origin in a complex structure at the microscopic or mesoscopic level. Attempts to capture this complexity mathematically are successful in some simple cases but soon run into fundamental difficulties. Having no constitutive equation to close the system of equations that are the basis of any design calculation, the task of describing smart material behavior seems to be hopeless. Fortunately, recent methodological and computational advances have reopened the road to successful design and prediction of smart material behavior.

On one hand, micro/macro methods are reaching a state now where they can compete with classical methods based on a continuum mechanical description of the material. Micro/macro methods make it possible to model a smart material at a very high level of sophistication without worrying about the solvability of the corresponding kinetic theory.

On the other hand, very recent advances in nonequilibrium thermodynamics are starting to yield their first truly groundbreaking results. Due to them, we can now postulate and develop microscopic models for smart materials certain that they are correct and consistent at the most fundamental level.

These two avenues of research are essential and complementary in process or part design for smart materials: nonequilibrium thermodynamics is the tool of choice to guide the development and the validation of the microscopic models used in micro/macro calculations.

Micro/macro methods and nonequilibrium thermodynamics are two of the most promising paths along which future advances in smart material design are likely to come.

NOTATION

E	Young’s modulus	(Pa)
$E(x)$	Energy functional	
H	Spring constant	(N/m)
k	Boltzmann’s constant	(J/K)
N	Number of molecules in an ensemble	(-)
n	Number density	(m ⁻³)
p	Pressure	(Pa)
Q	Connector vector	(m)
$S(x)$	Entropy functional	
T	Temperature	(K)
t	Time	(s)
\underline{u}	Strain tensor	(-)
v	Velocity	(m/s)
w	Three-dimensional Wiener process	(-)
x	State variables	

Greek Letters

δ	Unit tensor	
η	Viscosity	(kg/m/s)
$\gamma^{(n)}$	Nth rate of strain tensor, where $\tau^{(n)}$ denotes the n th convected time derivative (codeformational derivative using contravariant components):	(s ⁻ⁿ)

$$A_{(1)} = \frac{\partial}{\partial t} A + \{v \cdot \nabla A\} - \{(\nabla v)^T \cdot A + A \cdot (\nabla v)\}$$

$$A_{(n)} = (A_{(n-1)})_{(1)}$$

κ	Transposed velocity gradient $(\nabla v)^T$	(s ⁻¹)
λ_1	Relaxation time constant of Oldroyd-B model	(s)
λ_2	Retardation time constant of Oldroyd-B model	(s)
λ_{ijkl}	Elastic moduli	(m ² /N)
π	Total stress tensor	(Pa)
ρ	Density	(kg/m ³)
σ	Stress tensor	(Pa)
τ	Polymer contribution to the extra-stress	(Pa)
ζ	Friction coefficient	(Ns/m)

BIBLIOGRAPHY

1. L.D. Landau and E.M. Lifschitz, *Lehrbuch der theoretischen Physik*, Vol. 7. Akademie Verlag, Berlin, 1991.
2. G.K. Batchelor, *An Introduction to Fluid Dynamics*, Cambridge University Press, London, 1967.
3. L.D. Landau and E.M. Lifschitz, *Lehrbuch der theoretischen Physik*, Vol. 2. Akademie Verlag, Berlin, 1992.

4. N.W. Ashcroft and N.D. Mermin, *Solid State Physics*, Saunders College Publishing, Orlando, FL, 1976.
5. P.M. Gresho and R.L. Sani, *Incompressible Flow and the Finite Element Method*, Wiley, Chichester, 1998.
6. O. Zienkiewicz and R. Taylor, *The Finite Element Method. Basic Formulation and Linear Problems*, McGraw-Hill, London 1989.
7. G. Strang and G. Fix, *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
8. J.F. Nye, *Physical Properties of Crystals*, Oxford University Press, Oxford, UK, 1995.
9. K. Otsuka and C.M. Wayman, eds., *Shape Memory Materials*, Cambridge University Press, New York, 1998.
10. J.W. Goodwin, ed., *Colloidal Dispersions*, Royal Society of Chemistry, London, 1982.
11. M.P. Allen, *NATO ASI Ser.* pp. 557–591 (1995).
12. R.B. Bird, R.C. Armstrong, and O. Hassager, *Dynamics of Polymeric Liquids*, Vol. I. Wiley, New York, 1987.
13. J. Brandrup and E.H. Immergut, *Polymer Handbook*, Wiley, New York, 1989.
14. M. Doi and S.F. Edwards, *The Theory of Polymer Dynamics*, Oxford University Press, Oxford, UK, 1986.
15. R.B. Bird, C.F. Curtiss, R.C. Armstrong and O. Hassager, *Dynamics of Polymeric Liquids*, Vol. 2. Wiley, New York, 1987.
16. J.G. Oldroyd, *Proc. R. Soc. London, Ser., A* **200**, 523–541 (1950).
17. C.F. Curtiss and R.B. Bird, *J. Chem. Phys.* **74**, 2029–2042 (1981).
18. P.J. Flory, *Statistical Mechanics of Chain Molecules*, Hanser Publishers, Munich, 1988.
19. M. Laso and H.C. Öttinger, *J. Non-Newtonian Fluid Mech.* **47**, 1–20 (1993).
20. M.A. Hulsen, A.P.G. van Heel, and B.H.A.A. van den Brule, *J. Non-Newtonian Fluid Mech.* **70**, 79–101 (1997).
21. P. Halin, G. Lielens, R. Keunings, and V. Legat, *J. Non-Newtonian Fluid Mech.* **79**, 387–403 (1998).
22. R.W. Hockney and J.W. Eastwood, *Computer Simulation using Particles*, McGraw-Hill, New York, 1981.
23. F.F. Abraham, J.Q., Broughton, N. Bernstein, and E. Kaxiras, *Comput. Phys.* **12**, 538 (1998).
24. C.-A. Gandin and M. Rappaz, *Acta Mater.* **45**, 2187–2198 (1997).
25. L.E. Reichl, *A Modern Course in Statistical Physics*, University of Texas Press, Austin, 1980.
26. M. Grmela and H.C. Öttinger, *Phys. Rev. E* **56**, 6620–6632 (1997).
27. H.C. Öttinger and M. Grmela, *Phys. Rev. E* **56**, 6633–6655 (1997).
28. H.C. Öttinger, *Stochastic Processes in Polymeric Fluids*, Springer-Verlag, Berlin, 1995.
29. M. Grmela, *Phys. Lett. A* **102**, 81–86 (1984).
30. M. Grmela, *J. Phys. A* **22**, 4375–4394 (1989).
31. A.N. Beris and B.J. Edwards, *Thermodynamics of Flowing Systems*, Oxford University Press, New York, 1994.
32. W. Muschik and H.C. Öttinger, *An Example for Comparing GENERIC with Modern Conventional Non-Equilibrium Thermodynamics*. In preparation.
33. R.J.J. Jongschaap, K.H. de Haas, and C.A.J. Damen, *J. Rheol. (N. Y.)* **38**, 769–796 (1994).

CONDUCTIVE POLYMER COMPOSITES WITH LARGE POSITIVE TEMPERATURE COEFFICIENTS

C. P. WONG
SHIJIAN LUO
Georgia Institute of Technology
Atlanta, GA

INTRODUCTION

Conductive polymer composites (1,2) that contain conductive fillers such as metal powder, carbon black, and other highly conductive particles in a nonconductive polymer matrix have been widely used in electrostatic dissipation (ESD) and electromagnetic interference shielding (EMIS). A special group among electrically conductive polymer composites are conductive polymer composites that have large positive temperature coefficients (PTC), which in some cases are called positive temperature coefficient resistance (PTCR). The resistivity of this kind of composite increases several orders of magnitude in a narrow temperature range, as shown in Fig. 1. The transition temperature T_t is defined by the intersection of the tangent to the point of inflection of the resistivity versus temperature curve which is horizontal from the resistivity at 25°C (ρ_{25}). This kind of smart material can change from a conductive material to an insulating material or vice versa upon heating or cooling, respectively. The smartness of this kind material lies in this large PTC amplitude (defined as the ratio of maximum resistivity at the peak or the resistivity right after the sharp increase to the resistivity at 25°C), and also in its reversibility, its ability to adjust the transition temperature, its low-temperature resistivity, and high-temperature resistivity. PTC behavior in a polymer composite was first discovered by Frydman in 1945 (3), but not much attention was paid to it originally. Because Kohler obtained a much higher PTC amplitude from high density polyethylene loaded with carbon black in 1961 (4), this kind of temperature-sensitive materials has aroused wide research interest and also led to many very useful applications. In this review, the general theories

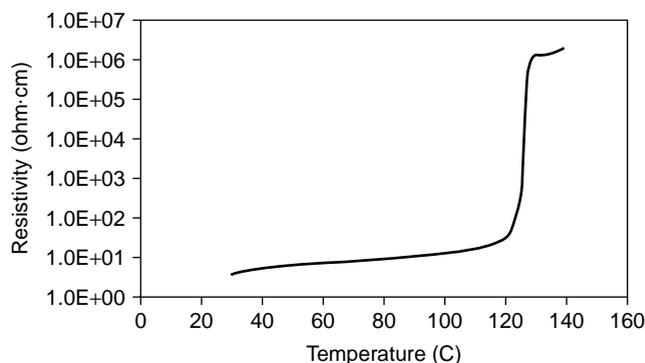


Figure 1. Resistivity versus temperature behavior of a conductive polymer composite that has a large positive temperature coefficient.

of PTC conductive polymer are introduced. Carbon-black-filled conductive polymer composites and their PTC behavior are discussed in more detail, in regard to the effects of fillers, the polymer matrix, and processing conditions, and additives. At the end, applications of this kind of smart temperature-sensitive material are presented.

BASIC THEORY OF CONDUCTIVE POLYMER COMPOSITES AND PTC BEHAVIOR

The conductivity of polymer composites that contain conductive particles dispersed in a polymer matrix may result from direct contact between conductive particles and electron tunneling. The conductivity of a highly filled conductive adhesive is due to the former mechanism. Matsushige used atomic force microscopy (AFM) to study the conduction mechanism in a PTC composite on a nanoscale (5). It was proposed that electron tunneling or hopping through the conductive carbon particles in the polymer matrix might be the governing mechanism for organic PTC materials.

There are two very simple mechanisms for small PTC behavior of conductive polymer composites: reduction of the contact area of neighboring conductive particles and an increase in the junction distance in electric tunneling when heated. Although the large PTC phenomenon is well known, its mechanism has not been fully understood. Different theories have been proposed (4,6,7) to explain the large PTC behavior.

Kohler (4) suggested that the PTC is due to the difference in thermal expansion of the materials. His theory was supported by some other researchers in percolation theory (1). The conductivity of conductive polymer composites increases as the volume fraction of the conductive filler increases. For a polymer filled with conductive particles, a critical volume fraction of filler may exist, which is called the percolation volume fraction. The resistance of the conductive polymer composite whose filler volume fraction is higher than the percolation volume fraction is several orders of magnitude less than that of the composite whose filler volume fraction is less than the percolation volume fraction. In the region of low filler concentration, the filler particles are distributed homogeneously in the insulating polymer matrix. There is no contact between adjacent filler particles. The resistance decreases slowly as the volume fraction of filler particles increases. As filler concentration increases further, filler particles begin to contact other particles and agglomerate. At a certain filler concentration, the growing agglomerates form a one-, two-, or three-dimensional network of the conducting phase within the insulating polymer matrix. At this range, the resistivity of the mixture shows a deep decrease to the low value of the conductive network. After the formation of the continuous conductive network, the resistivity of the mixture increases slowly as the filler content increases due to the slightly improved quality of the conductive network.

Many models have been proposed (8) to explain the electrical conductivity of mixtures composed of conductive and insulating materials. Percolation concentration is the

most interesting of all of these models. Several parameters, such as filler distribution, filler shape, filler/matrix interactions, and processing technique, can influence the percolation concentration. Among these models, the statistical percolation model (9) uses finite regular arrays of points and bonds (between the points) to estimate percolation concentration. The thermodynamic model (10) emphasizes the importance of interfacial interactions at the boundary between individual filler particles and the polymeric host in network formation. The most promising are structure-oriented models, which explain conductivity on the basis of factors determined from the microlevel structure of the as-produced mixtures (11).

Because the thermal expansion coefficient of a polymer matrix is generally higher than that of the conductive particles, the volume fraction of conductive filler in a conductive polymer composite decreases as temperature increases; thus, the resistivity increases. If a conductive polymer composite is made of semicrystalline polymer as an insulator and a filler of conducting particles, whose concentration is just above the percolation volume fraction, the relatively large change in specific volume of the polymer at its melting temperature may bring the volume fraction of the conductive filler down below the critical volume fraction when the composite is heated beyond the melting temperature of the polymer crystal. Thus, the resistivity increases greatly. Kohler's theory cannot explain the very small rise in resistance exhibited by such filled polymer systems when they are strained to an amount equivalent to that found at the crystalline melting point. And the PTC amplitude should be a direct function of volume change according to Kohler's theory; however, it is not the case in reality.

Ohe proposed a more complex theory (6). He stated that PTC phenomenon could be explained by the increasing intergrain gap among the carbon black particles caused by thermal expansion. He visualized that the distribution of the intergrain gaps in a conductive composite is rather uniform at low temperature, and the gap is small enough for extensive tunneling to occur, but the distribution at high temperature becomes random due to thermal expansion. Although the average gap distance does not change greatly, the presence of a significant amount of gap distance too large to allow electron tunneling will result in a great increase in resistance.

Meyer's theory (7) was based on the assumption that a thin (300 Å) crystalline film of polymer is much more conductive than an amorphous film of polymer. It was shown that carbon black particles remain in the amorphous region between crystallites in a conductive composite. The high conductivity at low temperature is due to tunneling through the thin crystallite, and the PTC phenomenon is caused by a preliminary change in state of these crystallites just before the crystalline melting point that leads to a sharp reduction in the ease of tunneling and thus much higher resistivity.

The authors of this article propose a new theory for PTC behavior. Large thermal expansion during crystal melting surely will contribute to a large amplitude of PTC behavior. But it contributes only to a limited level. Ohe's

vision of the change from uniform distribution of carbon particle to random distribution is groundless and cannot be justified. Actually, as is shown later, cross-linking can eliminate the redistribution of carbon black after melting and stabilizes PTC behavior. PTC behavior takes place at the same time as melting, rather than before crystalline melting, as stated by Meyer. It probably is true that tunneling is easier in a crystalline region than in an amorphous region. The difference is probably due to polymer chain mobility. The work function of the conductive particle at the interface between the conductive particle and polymer matrix may increase after crystal melting due to the high mobility of the polymer chain. The same theory can explain PTC behavior in conductive polymer composites as well as the conductivity phenomenon in an electrically conductive adhesive after curing (12). Before the conductive adhesive is cured, the resin matrix has high mobility and prohibits tunneling between conductive particles. After curing, the mobility of the polymer chain is greatly reduced and thus allows tunneling between conductive particles.

EFFECT OF CONDUCTIVE FILLERS ON PTC CONDUCTIVE POLYMER

Different conductive fillers have been used as in PTC conductive polymers. Metallic powders that are stable at high temperature, such as tin, gold, and silver were suggested as conductive fillers in PTC materials (13a). In addition, ceramic powder such as tungsten carbide was also used as a filler in PTC conductive polymer composites (13b). It was found that V_2O_3 has several phase transitions (1). At 160 K, it is transformed from an antiferromagnetic insulator (AFI) to a paraelectric metallic conductor (PMC), accompanied by a resistivity change from $10^5 \Omega\cdot\text{cm}$ to $10^{-2} \Omega\cdot\text{cm}$. At 400 K, it changes from a PMC to a paraelectric insulator (PI) whose resistivity is $10^3\text{--}10^4 \Omega\cdot\text{cm}$. Most interestingly, low density polyethylene (LDPE) filled with V_2O_3 shows a square well in the resistivity versus temperature profile by combining a sharp negative temperature coefficient (NTC) around -110°C and a sharp PTC around 100°C (1,14–16). A PTC transition temperature of conductive polymer composites filled with V_2O_3 was also reported in other polymer systems (17). The T_i of a V_2O_3 -filled system changes in the following manner: LDPE (100°C) < polypropylene (150°C) < polytetrafluoroethylene (260°C). However, the fillers mentioned before are expensive. Work has been done to develop alternative less expensive PTC conductive polymer composites. Most of the conductive polymers for ESD and EMIS applications are thermoplastics filled with carbon black or carbon graphite because of their very low cost. Carbon black is also one of the major fillers used in so-called PTC conductive polymers (18–20).

There are several important parameters of carbon black (21): particle size (surface area), aggregate structure (carbon black particles aggregate to form a grapelike structure), porosity, crystallinity, and surface functionality. Small particle size and high structure lead to more difficult dispersion. The initial grapelike structure of carbon black formed during the manufacturing of carbon black

is highly stable and can be destroyed only by very intensive processing such as grinding in a ball mill. For a given loading of carbon black, a smaller particle size would add more particles to the composite than that using carbon black of larger particle size. Thus, carbon blacks of smaller particle size would produce a composite that has a smaller separation between carbon particles (as well as the probability of more carbon particles in contact), resulting in greater conductivity. Small particle size gives a low critical volume for a carbon-black-filled polymer system (1). However, for fiberlike conductive fillers, large filler particles favor the formation of conducting paths at a low percolation concentration. High-structure carbon blacks tend to produce a larger number of aggregates in contact, as well as, smaller separation distances, that result in greater conductivity. For a given carbon black loading, the more porous carbon black generally provides a larger number of aggregates to the composite. This results in a smaller interaggregate distance and higher conductivity. The increase in the degree of carbon black structuring is found more efficient than the increase of the specific surface of carbon black in conductive polymers. Carbon particles that have higher oxygen content have higher resistance. Removal of the surface oxides increases the conductivity of the original carbon black much more than heat treatment to produce graphitization. Higher graphite content in carbon black leads to higher electrical conductivity.

Although the small particle size and the highly aggregated structure of carbon black (such as BP2000 manufactured by Cabot Corp.) can give polymer composites that have low resistance, this kind of composite does not show a large PTC amplitude because the aggregated structure cannot be broken down by the thermal expansion of the polymer (Fig. 2). On the contrary, a polymer composite filled with carbon black that has a large particle size and low aggregate structure (such as N660 manufactured by Columbia Chemicals) shows high room temperature resistance but high PTC amplitude (22). To obtain a PTC

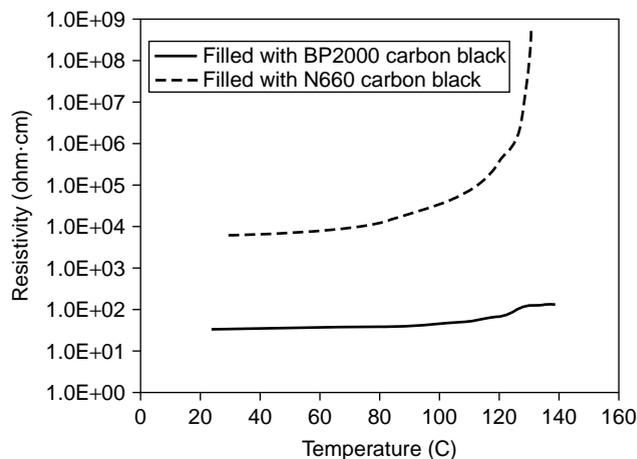


Figure 2. Resistivity versus temperature of HDPE filled with different carbon blacks (the loading is 30% by weight). BP2000: small particle size and high aggregated structure; N660: large particle size and low aggregated structure.

conductive polymer composite that has both lower room temperature resistance and high PTC transition amplitude, porous carbon black is much better than nonporous carbon black. Ueno et al. reported that etching a carbon black at an elevated temperature to remove the less crystalline portion and therefore to increase the surface area can improve the PTC characteristics of a conductive polymer filled with carbon black, and this material was suitable for use as a resettable fuse (20).

The PTC amplitude depends on the loading of carbon black. It was shown that for different carbon blacks, a different loading exists at which the composite has a maximum PTC amplitude (23). The carbon black concentration that gives the optimum PTC intensity can be predicted approximately from room temperature data (17).

EFFECT OF POLYMER MATRIX ON PTC BEHAVIOR

Polymers used as the matrix in electrically conductive polymer composites can vary from elastomers to thermoplastics and thermosets that have crystallinity varying between 0 and 80%. As mentioned in the previous paragraph, the large PTC anomaly is due to the large thermal expansion of the polymer matrix, especially during melting of a polymer crystal. The PTC transition temperature is determined by the melting point of the polymer matrix. Because polymers that have low and high melting points are available for use in conductive polymer composites, the transition temperature can be controlled by selecting and compounding the matrix polymer for different applications that require different transition temperatures (24). A PTC conductive composite based on high-density polyethylene whose melting peak temperature is 129–131°C and whose specific volume increases by approximately 10% due to melting across a narrow temperature range, showed maximum resistivity as a matrix at 129–131°C (22). The transition temperature can be slightly adjusted by using a copolymer or polymer blend that has more than one homopolymer. A compound of 40 parts by weight of carbon black, 60 parts of a melted olefin copolymer (ethylene-ethyl acrylate copolymer) (EEA), and an organic peroxide, had a T_i at 82°C (5). Another reported recipe (25) is a composite of carbon black dispersed in high-density polyethylene (HDPE) and poly(ethylene vinyl acetate) (EVA), whose T_i is 120°C. Ultra high molecular weight polyethylene (UHMWPE) reportedly enhances PTC behavior (18). Thermosetting material such as thermosetting polyester resin that was cross-linked by a free radical reaction, was also reportedly used as a polymer matrix for a PTC conductive polymer composite (25).

The PTC amplitude depends on crystallinity. Meyer showed (26) that crystalline *trans*-polybutadiene filled with carbon black has low room temperature resistivity and a significant anomaly, whereas the amorphous *cis*-polybutadiene filled with same amount of carbon black has much higher resistivity and no anomaly. Within a polymeric family, a polymer that is more crystalline has higher PTC amplitude. But also note that different classes of polymers that have the same crystallinity do not exhibit identical PTC behavior and no relationship was correlated. PTC

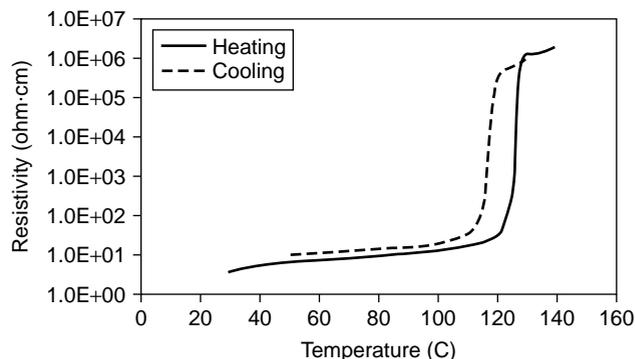


Figure 3. Resistivity vs. temperature for a PTC conductive polymer during heating and cooling cycles.

amplitude depends on polymer type. PTC amplitude increases in the following order with respect to the matrix polymer: nylon 66 < polypropylene < polyethylene oxide < low-density polyethylene < high-density polyethylene.

As mentioned before, the mechanism for the PTC anomaly in semicrystalline polymer composites is accompanied by a relatively large change in the specific volume of the polymer at its melting temperature. The resistivity versus temperature curve can be well matched by the specific volume–temperature curve. Crystallization during cooling of a polymer is the reverse of melting of a polymer crystal during heating. The PTC transition of this kind of smart material is reversible. During cooling, the same material shows a sharp decrease in resistivity, as shown by Fig. 3. The thermal expansion of a polymer depends on its heating and cooling cycle. Because the melting temperature of a polymer crystal is always higher than the recrystallization temperature, the PTC transition of a conductive polymer composite is always higher in the heating cycle than that during the cooling cycle (1). The difference is about 18 K for polyethylene, 34 K for polyoxymethylene, and 50 K for polypropylene. Actually, all factors that affect the melting and recrystallization behavior such as pressure and heating and cooling rates influence the PTC behavior of a conductive polymer composite. Meyer showed that the PTC transition temperature increases and PTC amplitude decreases as pressure increases (26).

In some conductive polymer composites, the negative temperature coefficient (NTC) effect follows, for example, the resistivity decreases as the temperature increases further after a PTC transition. The NTC effect is probably due to the reorientation, reaggregation, or reassembling of carbon black. Initially dispersed particles may become mobile in the temperature range of polymer melting to repair the broken percolation network. The measurement of resistance versus the temperature behavior of the conductive composite was repeated for the same sample (27,28). Tang (27) observed that the PTC intensity and the base resistance decrease with thermal cycles. The reason is obviously reorganization of carbon black at the high temperature.

Radiation was used to cross-link a carbon-black-filled conductive polymer composite. The NTC effect can be alleviated or reduced by cross-linking, and the PTC

amplitude is stabilized (the resistance at lower temperature is stabilized) after cross-linking (1, 27). The organization of carbon black is hampered in a cross-linked polymer network.

EFFECT OF PROCESSING CONDITION AND ADDITIVES

The effect of mixing time on the resistance-temperature behavior was also investigated (27). Both the PTC effect and the reproducibility were improved greatly when the sample was mixed for a long time. It was suggested that the improvement is due to increasing absorption of the polymer on the carbon black surface and this absorption forms a polymer layer outside the carbon black particle. The room temperature resistivity increases with mixing time at constant carbon black concentration. It can also be explained that structures are broken down during mixing, thus the resistivity increases. If the power consumption during mixing is too great, the composition would have too high resistivity at a low temperature and have unsatisfactory electrical stability on aging at elevated temperature. If the power consumption is too low, it can also result in a composition that has low PTC amplitude.

Tang (27) studied the effect of the interaction between carbon black and the polymer on electrical behavior. The absorption of polymer on the carbon black surface may be physical or chemical. The latter is caused by free radical reaction between the polymer and carbon black, and it can occur during radiation or the preparation of the composite. In carbon-black-filled HDPE, the cross-linked network of the polymer restrict the freedom of movement of carbon black. The free radical reaction enhances the binding force between the polymer and carbon black.

Polymeric materials may be broken down under high voltage. The voltage stability of a cross-linked PTC conductive polymer is improved by incorporating a polymerizable monomer such as triallylisocyanurate before it is cross-linked (28). Antimony oxide, which does not degrade PTC resistance, can be used as a flame retardant (13a). A semiconductive inorganic substance such as silicon carbide or boron carbide was used to improve the high voltage stability (29). Alumina trihydrate can be added to a PTC conductive polymer composite to prevent dielectric breakdown, arcing, and carbon tracking under high voltage (19).

APPLICATION OF PTC CONDUCTIVE POLYMER COMPOSITE

There are many applications for PTC conductive polymer composites, including thermistors (13b), circuit protection devices (30), and self-regulating heaters (31). Because the material both heats and controls the temperature, it can be used to manufacture a self-regulating heating device. As the temperature increases, the resistance increases, and thus the power decreases. This kind of self-regulating heater can be used to prevent freezing of water and pipes used in chemical processing. It has also been used to manufacture a heater for heating a hot-melt adhesive to seal a

cable splice case (32) and a hair curler (33). Self-regulating heaters can be manufactured into different forms. The blank form of PTC conductive polymer composite can allow precise temperature control across larger areas. This kind of device has been used to repair thermally complex aircraft structures (31,34).

Another application of PTC conductive polymers is in over-temperature and over-current protection. A device manufactured from a PTC conductive polymer composite has low resistance and much less resistance than the rest of the circuit at normal temperature; thus it has no influence on normal performance. But at high temperature, these devices become highly resistant or insulators; thus, they dominate the circuit, reduce the current, and protect the circuit. For large abnormal current, the device can rapidly self-heat to a high resistance state and thus reduces the current. The smartness lies in the over-temperature and over-current protection and also in its resetability. After the current drops and the temperature of the device decreases, the device returns to a low resistance state and allows current to pass. A resettable fuse made from a PTC conductive polymer has been on the market. This kind of resettable fuse has been used in battery chargers to terminate the charging function based on the battery temperature and protect the battery from overheating. It is also used in telecommunication equipment, computers, and power supplies.

SUMMARY

A temperature-sensitive PTC conductive polymer composite is a true smart material. Its property can also be tailored by selecting the filler, polymer matrix, and processing conditions. Its transition temperature is determined by the melting point of the polymer matrix. Its room temperature resistivity, high-temperature resistivity, and PTC transition amplitude can be adjusted by the filler and its combination with the polymer matrix. PTC transition behavior can be stabilized by cross-linking the polymer matrix. This kind of smart material can be used in many temperature-sensitive applications such as thermistors, self-regulating heaters, and circuit protection devices.

BIBLIOGRAPHY

1. X. Yi, G. Wu, and Y. Pan, *Polym. Int.* **44**: 117 (1997).
2. V.E. Gul, in *Structure and Properties of Conducting Polymer Composites*, VSP BV, The Netherlands, 1996.
3. UK Pat. 604, 695, E. Frydman.
4. US Pat. 3,243,753, F. Kohler.
5. K. Matsushige, K. Kobayashi, N. Iwami, T. Horiuchi, E. Shitamori, and M. Itoi, *Thin Solid Films* **273**: 128 (1996).
6. K. Ohe and Y. Naito, *Jpn. J. Appl. Phys.* **10**: 99 (1971).
7. J. Meyer, *Polym. Eng. Sci.* **14**: 706 (1974).
8. F. Lux, *J. Mater. Sci.* **28**: 285 (1993).
9. R. Zallen, in *The Physics of Amorphous Solids*, Wiley, NY, 1983, Chap. 4.
10. M. Sumita, K. Sakata, S. Asai, K. Miyasaka, and H. Nakagawa, *Polym. Bull.* **25**: 265 (1991).

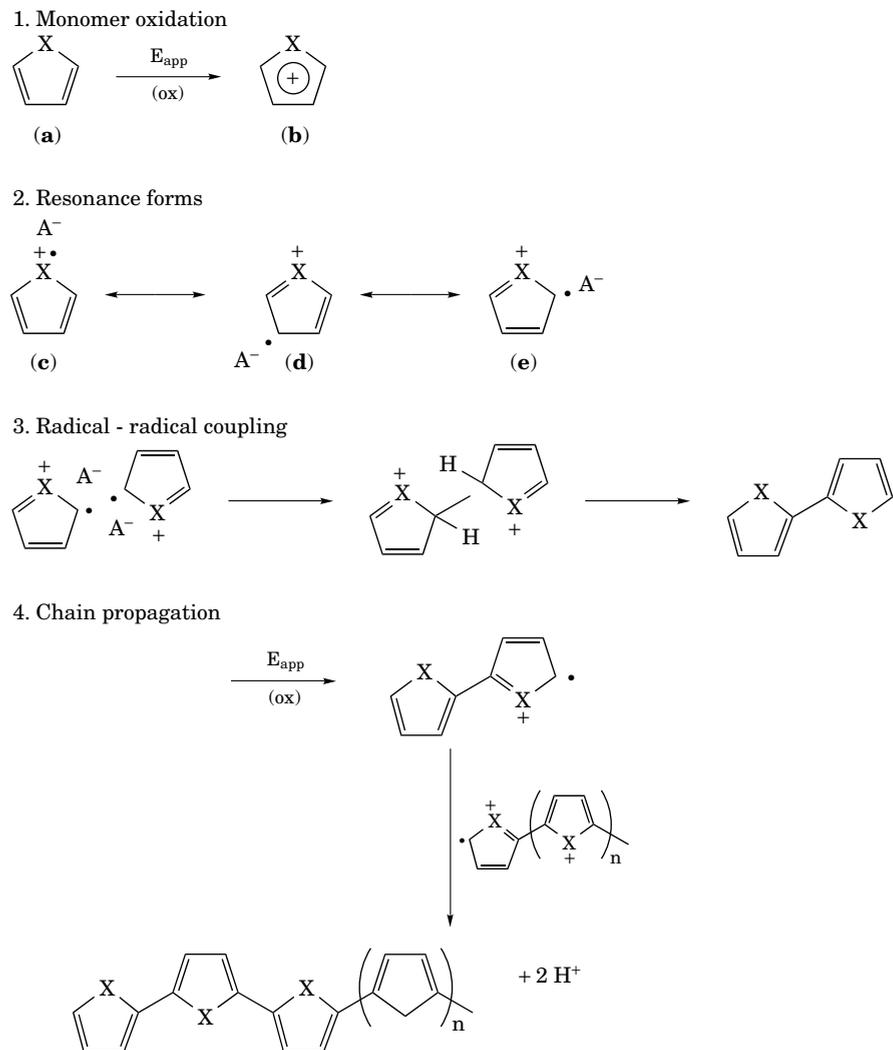


Figure 2. Steps involved in polypyrrole formation.

The polymerizations involve formation of lower molecular weight oligomers that are further oxidized (at lower potentials than the initial monomer) to form a polymer that eventually precipitates or electrodeposits as a critical molecular weight is exceeded. For example, the steps involved in polypyrrole formation are shown in Fig. 2 ($X=NH$).

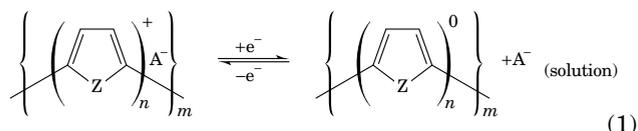
A counterion (A^-) is incorporated during synthesis to balance the charge on the polymer backbone. Common chemical oxidants are $FeCl_3$ and $(NH_4)_2S_2O_8$, which provide Cl^- and HSO_4^-/SO_4^{2-} , respectively, as the dopant anions. Electrochemical oxidation provides greater flexibility in terms of the anion that can be incorporated from the electrolyte (MA salt or HA acid) added to the polymerization medium. Acidic conditions (pH generally ≤ 3) are required for aniline polymerization in order to solubilize the monomer, while pyrrole and thiophene polymerizations may be carried out in neutral solution.

In general, chemical oxidation provides ICPs as powders, while electrochemical synthesis leads to films deposited on the working electrode. A wide range of working electrodes may be employed, including platinum, gold, reticulated vitreous carbon (RVC), and indium-tin-oxide coated glass. The last-mentioned ITO-glass electrodes,

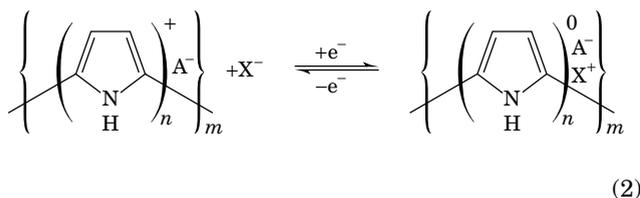
being transparent in the visible near-infrared region, are very useful for recording absorption or circular dichroism spectra for the deposited conducting polymer films. Both the chemical and electrochemical oxidation approaches can be modified to produce soluble or dispersible conducting polymers.

An important additional feature is that all of these polymer structures are amenable to facile oxidation/reduction processes that can be initiated at moderate potentials. For polypyrroles, and polythiophenes, two oxidation states can be reversibly switched, as shown in Eq. (1) where $Z=NH$ or S, respectively. The doped oxidized forms exhibit good electrical conductivity ($\sigma = 1 - 100 \text{ s cm}^{-1}$), while the reduced forms have very low conductivity ($\sigma \sim 10^{-8} \text{ s cm}^{-1}$). This ability to conduct electrons is important in that information can be readily relayed within an intelligent polymer system. However, this feature alone is not sufficient for intelligent performance. The dynamic character of these polymer systems is equally important, with chemical, physical, and mechanical properties being a function of applied potential. Both polypyrroles and polythiophenes can be reversibly reduced/oxidized according to Eq. (1) and (2). Polyaniline undergoes similar transitions, although its

redox chemistry is slightly more complex.



If the dopant anion (A^-) is small and mobile (e.g., Cl^-), and the polymer has a high surface area to volume ratio, then upon reduction, the anion will be efficiently ejected from the polymer. However, extensive studies with polypyrroles have shown (3) that if the dopant is large and immobile (e.g., if A^- is a polyelectrolyte such as polystyrene sulfonate), then an electrically induced cation exchange process occurs, according to Eq. (2); where the cation (X^+) is incorporated from the supporting electrolyte solution. This reduction process has a dramatic effect on the physical and chemical properties of the polymer. For example, conductivity will decrease, color will be altered, anion exchange capacity will diminish, cation exchange capacity may increase, and hydrophobicity will be altered in a manner determined by which ion exchange process predominates (2). Some of the changes accompanying these redox reactions are summarized in Fig. 3. These changes are important in determining the sensing, information processing, and actuation capabilities of the systems.



The situation with polyanilines is more complex, with the polymer able to exist in three different oxidation states: leucoemeraldine, emeraldine and pernigraniline (Fig. 4) (4,5). In addition, protonation/deprotonation equilibria occur for two of these oxidation states, depending on the pH. Thus, the emeraldine salt form (ES), which is the only electrically conducting form of polyaniline, is typically dedoped at $\text{pH} > 4$ to give nonconducting emeraldine base (EB). Reversible redox and pH switching between these different forms of polyaniline leads to important changes in their physical and chemical properties (Fig. 3), which may be exploited in a range of applications for intelligent polymer systems.

Besides responding to direct electrical stimulation or a change in the redox environment, ICPs may be sensitive to other stimuli such as temperature, humidity, and inorganic and organic vapors. For example, substituted polythiophenes, show a marked blue shift of the highest wavelength absorption band when films or solutions are heated (6–8). These reversible color changes have been attributed to a twisting of the polymer backbone to a less ordered nonplanar conformation. Less dramatic thermochromic effects have also been reported for polyanilines (9–11) where circular dichroism studies of chiral polyaniline salts (9), as with related studies on chiral polythiophenes (12), have provided further insights into the nature of the thermochromism.

The response of ICPs to exposure to various compounds has been reported by many authors. Among these are investigations of changes in the conductivity of polyaniline doped with hydrochloric and sulfonic acids (13). The conductivity of both polyaniline films was at a minimum in dry air and increased over time with increasing humidity. The interaction between water and polypyrrole has also been employed as the driving force in an ICP motor, where a band of this polymer is alternately wetted and dried in order to generate motion (14). Volume changes were reported in polypyrrole perchlorate when it was exposed to ammonia gas and in substituted polythiophenes exposed to iodine vapor (15). These volume changes are utilized as sensors or chemically driven bending beam actuators.

CHEMICAL AND PHYSICAL STIMULI

Chemical Stimuli

Nature has developed chemical recognition systems that are able to discriminate on the basis of highly specific molecule–molecule interactions generating a unique signal. Alternatively, nature utilizes arrays of less specific chemical sensors to collect information that is deciphered using pattern recognition processes carried out in the brain. Both approaches have also been pursued using ICPs in the development of synthetic sensors.

Specific Molecular Recognition Approach

Immobilization of Biologically Active Agents on ICPs.

Chemical specificity has been induced in conducting polymers by borrowing elements of the sensor from nature. For example, the ICP may be used as an immobilization platform for enzymes (16–20), antibodies (21,22) or even whole living cells (23,24). Generally, the bioactive component is

Polymer reduced	\rightleftharpoons	Polymer oxidised
<ul style="list-style-type: none"> • Low ion exchange capacity • Hydrophobic • Transparent • Resistive • Low capacity • Greater elongation to break • Contracted 		<ul style="list-style-type: none"> • High ion exchange capacity • Hydrophilic • Coloured • Conductive • High capacity • Brittle • Expanded

Figure 3. Some changes accompanying polymer oxidation/reduction.

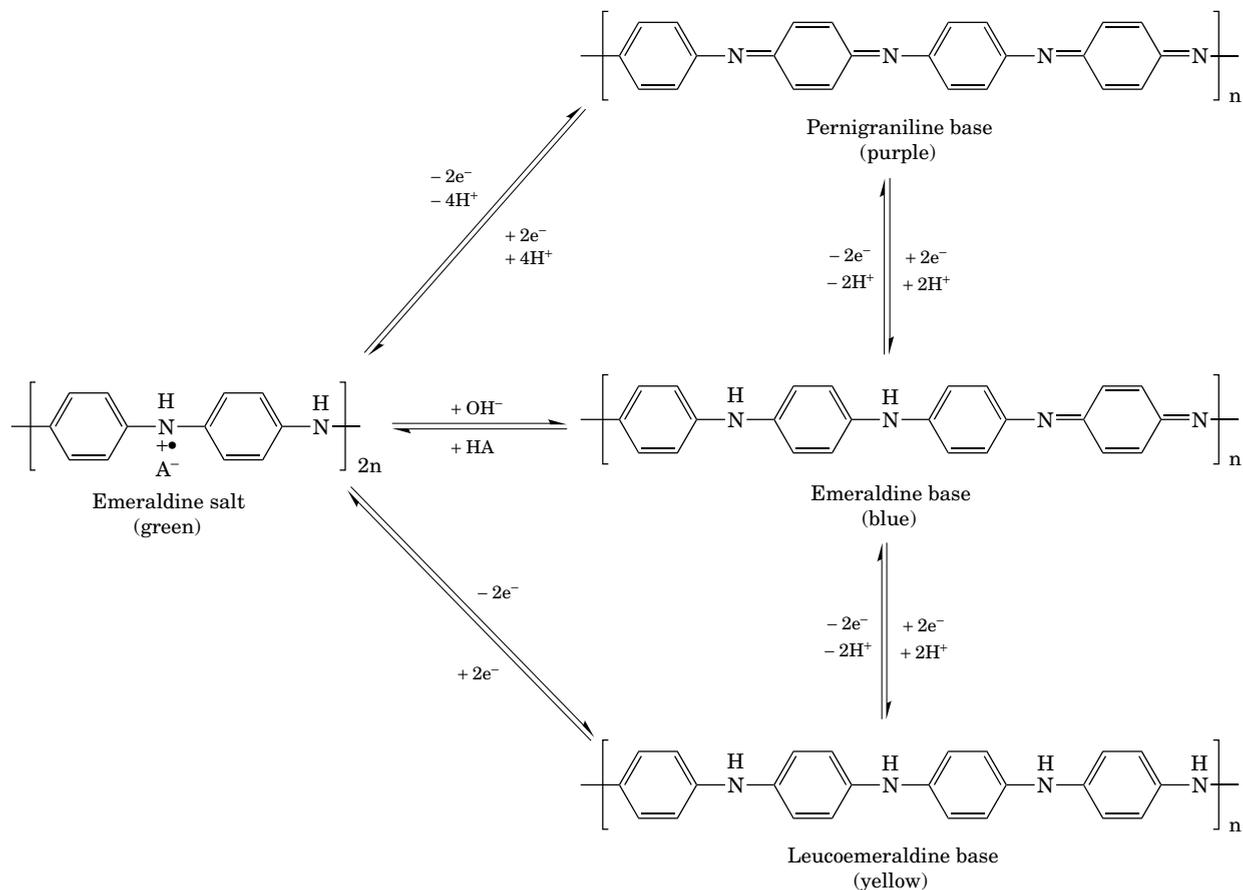


Figure 4. Chemical transitions observed in the oxidation/reduction of polyaniline.

simply incorporated during the polymerization process. The ICP plays a dual role in that it provides a biocompatible matrix that does not destroy the bioactivity of the incorporated species and also provides signal transduction and transmission capabilities.

The majority of enzyme-containing ICP sensors generate a signal because of the increase in concentration of an electroactive product (e.g., H_2O_2) generated by the enzymatic reaction, or because of the decrease of an electroactive product (e.g., O_2) consumed by this reaction. Some others utilize the fact that the bioevent triggers a change in pH that results in a change in resistance of the polymer (25).

The mechanism of signal generation with antibody-containing conducting polymer sensors (21,22) is not so clear. This is also the case for sensors that rely on antibody-antigen reactions that occur on mammalian red blood cell membranes immobilized on a conducting polymer (23). However, both amperometric (21,22) and resistometric (23) measurement techniques have been explored and give rise to useful analytical signals.

Selectivity in the detection process can be increased further by the use of an appropriately designed sensor configuration (26), for example, the use of an electromembrane sensor (Fig. 5).

With this configuration the conducting polymer serves as an immobilization platform for the enzyme. The potential applied to one side of the membrane (the reaction zone) can be independently optimized to ensure maximum

enzyme activity. The electroactive product generated, H_2O_2 , is then selectively transported across the support (PVDF) membrane for detection in an electrically insulated detection zone. Potential interferents such as ascorbic acid and glutathione are not transported and hence not detected. The area covering the use of conducting polymers as biosensors has recently been reviewed (27).

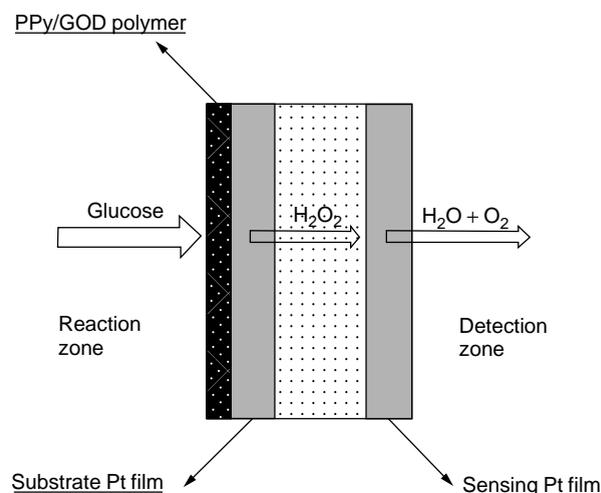
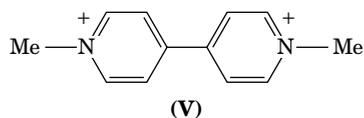
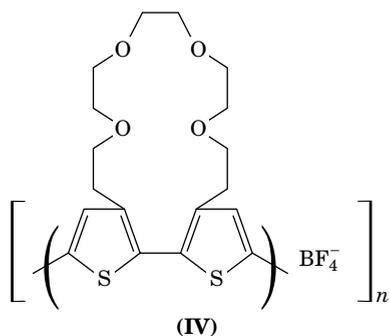


Figure 5. The Electromembrane Sensor

Covalent Attachment of Specific Binding Groups to ICP Backbones. The selective detection of a wide range of target analytes, both chemical and biochemical, has also been achieved by the covalent attachment of molecular recognition moieties to the ICP backbone. The usual approach has been to synthesize a monomer or dimer containing the appropriate recognition group, and this is then oxidized to produce the conducting polymer (28). Subsequent binding of the target analyte to the recognition site is usually accompanied by an electrochemical response in the ICP (e.g., a change in resistance) although other physicochemical responses such as a change in color are sometimes employed.

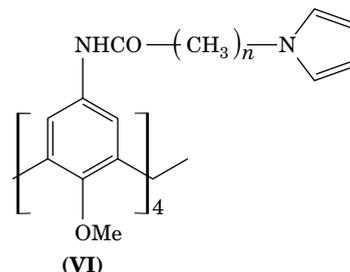
A drawback with this approach to functionalized ICPs is that the synthesis of the initial substituted monomer may be complex and time-consuming. In addition subsequent oxidation to the desired polymer may prove difficult because of steric hindrance by the functional group or electronic effects that shift the oxidation potential of the monomer. A significant recent development, therefore, is a route involving the facile modification of pre-formed polypyrroles containing good leaving groups such as N-hydroxysuccinamide (29). Using this approach, crown-ethers and electroactive groups such as ferrocene have been covalently attached to the pyrrole rings. This generic approach should be extendable to analogous polythiophenes and polyanilines.

Metal Ion and Organic Cation Sensing. Swager et al. (30) have prepared polythiophenes containing crown-ethers and calix[4]arenes, covalently bound to the bithiophene repeat units (e.g., structure IV below), that exhibit tunable selectivities toward Li^+ , Na^+ , and K^+ ions. Electrochemical detection is facilitated in these cases by electrostatic and conformational perturbations caused by the coordination of the metal ions to these chemical receptors. Swager's group has also developed real-time sensors for viologen dications such as paraquat (V), using conducting polymers based on poly(bi- and trithiophene)s to which cyclophane receptors have been covalently attached to the thiophene rings (31). These form self-assembled pseudorotaxane complexes with paraquat, causing a large decrease in conductivity for the polymer system as a sensitive sensor signal.

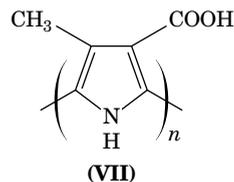


More recently, pyrrole monomers VI ($n = 2$ or 4) have also been prepared with calix[4]arenes grafted to the N

atom of the pyrrole ring (32). Electropolymerization to give films of the corresponding N-substituted polypyrroles was more facile for the monomer with the longer alkyl spacing arm length ($n = 4$) (33). However, no change was observed in the electroactivity of these films upon cyclic voltammetry in the presence of Li^+ , Na^+ , or K^+ ions (34).



Sensing Biological Molecules. Covalent attachment of simple functional groups has also been used to promote electron transfer with biocomponents in solution. For example, Cooper et al. (35) produced the functionalized polypyrrole (VII shown below) and demonstrated the ability to mediate in the electrochemical oxidation/reduction of cytochrome c.



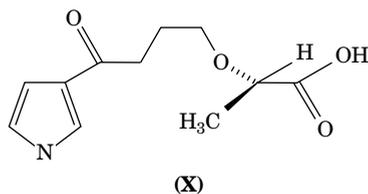
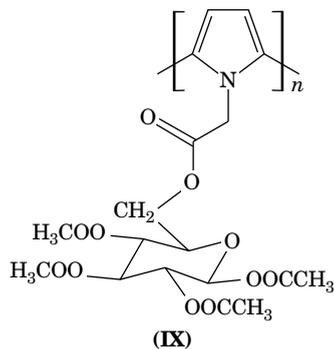
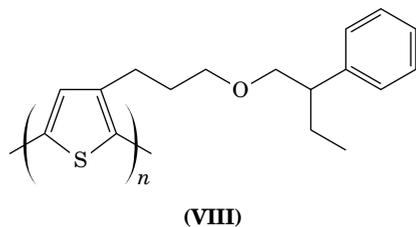
Garnier et al. (36) have similarly oxidized pyrrole monomers bearing a range of dipeptide substituents, producing substituted polypyrrole materials with specific recognition properties toward proteolytic enzymes.

A recent exciting advance involves the incorporation of oligonucleotide chains into polypyrrole backbones. This approach has been used to produce DNA-biochips (37–39) applicable in a number of important sensing applications.

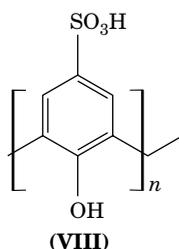
Sensing Enantiomeric Molecules/Ions—Chiral Recognition. ICPs with chiral discrimination capabilities have recently been developed via the covalent attachment of optically active groups, such as amino acids, to the β -ring position of pyrrole and thiophene monomers prior to oxidation to the corresponding polymers. The presence of the chiral substituents is believed to induce a one-handed helical structure on the polymer backbones of the polypyrroles and polythiophenes formed. Polypyrroles with main chain chirality have also been generated via the covalent attachment of chiral groups (e.g., sugars) to the N center of the pyrrole monomer prior to polymerization.

In some cases, (e.g., polymers VIII–X) these optically active polypyrroles and polythiophenes were shown to possess the ability to discriminate between the enantiomers of chiral molecules and ions (40–43). The first demonstration of such chiral discrimination was with polymer (VIII).

Cyclic voltammetric studies on the (2*S*)-(+)-enantiomer of VIII in the presence of (+)- and (-)-camphorsulfonic acid (HCSA) showed substantially higher doping with the (+)-HCSA enantiomer (40). The reverse discrimination was shown by the (2*R*)-(-)- polymer.



Use of Dopant Anions Containing Molecular Recognition Groups. An alternative and often facile route to appropriately functionalized ICPs, which avoids the synthetic problems outlined above, is the use of sulfonated species containing the desired molecular recognition/receptor site as the dopant anion for the conducting polymer chains. For example, calixarene-containing polypyrroles (44) and polyanilines (45) for selective metal ion detection have been recently prepared via the use of the sulfonated calixarenes (VIII, $n = 4$ or 6) as dopant anions. We have similarly found that the incorporation of metal complexing agents such as sulfonated 8-hydroxyquinoline as dopants in polypyrroles provides a simple route to metal ion-selective ICPs (46).



Similarly, chiral recognition properties have been introduced into polyanilines via the use of optically active acids such as (+)- or (-)-camphorsulfonic acid (HCSA) (47–49) or (+)- or (-)-tartaric acid (50) to acid dope emeraldine base, producing optically active emeraldine salts. The chiral dopant anions are believed to induce a preferred one-handed helical chirality into the polyaniline chains.

We (51) and others (52) have recently shown that films of optically active polyaniline salts such as PAN(+)-HCSA, or the optically active emeraldine base (EB) derived from them, exhibit chiral discrimination toward chiral compounds such as the enantiomers of CSA⁻ and amino acids.

Specificity can also be controlled to some extent by the introduction of electrocatalytic properties in the polymer. For example, the incorporation of Fe(CN)₆⁴⁻ as a counterion (53) or the use of a prussian blue coating on a conducting polymer inner layer (54) can promote electron exchange with Cytochrome C.

Electrocatalytic dopants (heteropolyanions) have also been incorporated into conducting polymers with a view to developing sensors for detection of nitrite (55) or nitrogen monoxide (56).

Pattern Recognition Approach

Microsensing arrays of ICPs have been assembled with a view to collecting less specific data and using pattern recognition software to decipher it. The so-called electronic noses are based on this principle (57–62). A range of conducting polymers with differing molecular selectivity respond to a complex mixture to produce a unique pattern of responses. This approach has been used to differentiate beers (57) and olive oils (58), as well as to detect microorganisms (59,61) among other applications.

Normally the sensing chips are produced using microlithography and a four-point measuring technique is used to achieve high accuracy in measurement (63). Signal generation is achieved via changes in resistance (increase or decrease), since the sensor works in the dry state. The origin of a particular ICP film's response is uncertain and could include factors such as changes in polymer conformation, volume, and/or screening between the counterions and carriers induced by the analyte.

The means of polymer deposition on the tip of the four-track sensing device is critical. Electrodeposition of the ICP is preferred over chemical polymerization, since the device to be coated is small and deposition can be localized using the electrochemical method. While growth on the gold tracks is readily initiated, lateral growth across the insulating silicon surface is necessary to form a thin coherent film—the structure required for optimal sensitivity. Such lateral growth can be encouraged by silanizing the nonconducting substrate to render it more hydrophobic, enhancing polymer deposition.

The array approach has also been developed for amperometric sensing when used in solution. This has been used by us recently to discriminate between simple ions (64) and even proteins (65). The approach used is similar to the electronic nose in that none of the sensing elements is specific, however, each polymer has a different selectivity

series giving rise to a unique pattern of responses for any given protein. The electronic tongue may not be far away.

ACTUATORS

Generation of electrical stimuli (e.g., a change in potential) as a result of the sensing process can be used to initiate an appropriate response—actuation. Consideration of the processes discussed earlier indicates that substantial changes occur when conducting polymers are oxidized/reduced. These changes have been studied at the molecular level using a technique known as inverse chromatography, both in solution (66–68) and in the gas phase (69,70).

These studies revealed that at least when the dopant (A^-) is small and mobile, reduction of the polymer decreases the anion exchange capacity of the polymer and increases the hydrophobicity of the polymer backbone. These changes inevitably affect the way the polymer interacts with the immediate environment.

Molecular Actuators

The reduction of ICPs also results in exclusion of small dopant anions from the polymer backbone. The dopants can be chosen so that the release process has the desired effect on the chemical composition of the immediate environment. For example, Miller described the triggered release of glutamate (71) and salicylate (72) among other compounds. We have similarly demonstrated the ability to release quinones (73) and metal complexing agents, dithiocarbamates (74). Devices/structures based on this principle will have a reservoir capacity of active ingredients determined by the original doping level of the conducting polymer. If this reservoir of active ingredients is not sufficient, then controlled release devices that utilise the unique properties of conducting polymer membranes can be configured. It has been shown that the transport properties of ICPs are dependent on the oxidation state of the membrane. This has been demonstrated both in solution (75,76) for transport of dissolved ions or molecules and in the dry state (77,78) for transport of volatiles. Extraordinary selectivity factors have been reported for the separation of some volatiles, for example, selectivity factors of 3590 for H_2/N_2 , 30 for O_2/N_2 and 336 for CO_2/CH_4 (79). With membranes operating in solution, the controlled transport of simple ions (80), metal ions (81), small organic molecules (82), and even proteins (83) has been demonstrated.

One could further envisage structures containing packages of active materials (capsules, hollow fibres) wrapped in an ICP membrane, with the capacity determined by the internal volume rather than the dopant capacity of the ICP itself.

Mechanical Actuators

The transitions that occur within the conducting polymers result in dramatic changes in physical properties. For example, upon reduction, the resistance increases markedly (84,85) and the materials became more transparent (86,87).

These incorporation/exclusion events at the molecular level result in changes in the mechanical properties of the bulk material. For example, both tensile strength (88,89) and Young's modulus (89) decrease dramatically as does the overall volume (dimensions) of the polymer change (90). It was these volume changes that led Baughman and colleagues to the concept of electromechanical actuators based on conducting polymers (90). By producing simple laminated structures containing the ICP, they were able to generate a force upon the oxidation/reduction of the active polymers, causing movement. A number of other studies into the effects of polymer composition, supporting electrolyte, and rate of stimulation on the forces generated have been carried out (91–97).

Since the process of force generation depends on the ability of ions to diffuse into/out of the polymer, systems with improved transport properties will give enhanced performances. This has led Hutchison et al. to employ polypyrrole-coated fiber bundles in actuator devices to increase the active surface area and, hence, the rate of diffusion of ions (98). This approach not only substantially increases the surface area exposed to the ion source/sink, it more than doubles the proportion of the actual component in the device. This has led to an increase in force density from 1.25 N/mm² for the laminated device to 5.1 N/mm² for a fiber bundle device (98).

Smela and co-workers (99,100) have demonstrated that elegant performance can be obtained from polymer devices of three dimensions as long as they are small—the folding boxes video is a technological treat (101). Others have demonstrated microcantilevers based on conducting polymers fabricated on silicon substrates (102).

Other Triggers

All of the above-mentioned responses (actuators) are initiated by creating an appropriate electrical potential that causes the polymer to change form. For demonstration purposes this is usually achieved by the imposition of a potential from an external source. However, this potential may be generated by configuring the active electrode as one part of a galvanic cell that is charged/discharged directly. Alternatively, the system/device could be configured so that the ICP sensor acts as a switch, gating the actuation mechanism. Changes in the conductivity or patterns of conductivity can be used to complete circuits and activate responses.

INFORMATION PROCESSING

We are familiar with the use of inorganic, silicon-based materials in the development of complex circuitry that is capable of processing vast amounts of information. Following on from this, the concept of molecular electronics has attracted considerable interest in recent years (103). Within this field ICPs have attracted attention for use as components in polymer-based diodes (104,105), transistors (106,107), and even amplifiers (108).

It is conceivable that eventually more complex information processing, storage, and transport based on polymeric devices will be incorporated into intelligent polymer

Table 1. Summary of Photovoltaic Performance of Some Conducting Polymers

Polymer System	Device	Voc (V)	Isc (mA/cm ²)	FF	Y (%)
Polyacetylene (PA)	n-Si/p-PA	0.53	18.18	0.32	4.3
Polymethylthiophene	Al/PMT/Au	0.23	0.16 ($\times 10^{-3}$)	0.30	—
Polypyrrole (PP)	PP/n-Si	0.29	9.0	0.46	1.2
Poly (N-vinyl carbazole) (PVK)	Al/PVK/Au	1.0	0.18 ($\times 10^{-3}$)	0.23	0.028
Polyaniline (PAn) (9)	Al/PAn/ITO	0.43	—	0.70	0.8 ^a
Polythiophene (3)	Photoelectrochemical cell	0.41	0.35 ($\times 10^{-3}$)	—	0.6 ^b
Polyparavethylene (PPV) (4,5)		1.00	4 ($\times 10^{-3}$)	0.60	6.0 ^c

Source: Adapted from (114).

Note: (Voc) open circuit voltage, (Isc) short circuit current, (FF) fill factor, (Y) engineering conversion efficiency.

^aPower conversion efficiency

^bMonochromatic photon to current efficiency

^cQuantum efficiency

systems. This will require the development of innovative approaches to polymer processing and device fabrication.

ENERGY CONVERSION/STORAGE

The functions/performance required of intelligent polymer systems, namely sensing, information processing, and actuation, will most likely expend energy. Consequently, the intelligent polymer system should preferably be capable of converting energy from a natural source, such as sunlight, and storing it until required. Fortunately, inherently conducting polymers appear to be capable of these functions also.

It has been demonstrated that ICPs are capable of functioning as the active layer in photovoltaic devices (109,110). Since p-n junctions are readily created using conducting polymers, a number of photovoltaic cell designs are possible. Initially, poly(p-phenylene vinylene)-based polymers were used (111,112). More recently, polyaniline (32) and polythiophene (113)-based polymers have been employed. Although these polymers are currently low in efficiency (Table 1), the attachment of light-harvesting molecules or moieties that enhance charge separation and transport within the polymer structure should improve this (114). The fact that they can be fabricated in different forms means that the efficiencies attainable in \$ per square meter would be achievable.

Inherently conducting polymers have also been used in polymer battery fabrication, and an all-polymer battery structure has been developed (115). A specific charge capacity of 22 mA hg⁻¹ and a cell potential of 0.4 V were obtained. The cells showed no loss in capacity when cycled 100 times. Others have utilized conducting polymers as just one of the electrodes in a battery setup (116).

In addition, ICPs have been used in the development of so-called "super capacitors" (117-119).

POLYMER PROCESSING

The above-mentioned applications of ICPs have significant commercial potential. However, in most cases, this has not been exploited because of the lack of convenient polymer processing and device fabrication protocols. For example, while polypyrroles exhibit the desirable properties

mentioned above, polymerization usually results in the formation of an insoluble, infusible material not amenable to subsequent fabrication. Conducting polyaniline and polythiophene salts are similarly intractable. Several approaches have recently been employed to overcome this problem of intractability.

Use of Ring-Substituted ICPs

Solubility has been induced in polypyrroles by attaching alkyl (120,121) or alkyl sulfonate (122) groups to the pyrrole monomer prior to polymerization. This results in markedly enhanced solubility in organic or aqueous media, respectively. For example, we have shown that the electrochemical method (123) can be used to produce alkylated polypyrroles with high (400 g/L) solubility in organic solvents and reasonable (1-30 S cm⁻¹) conductivity.

Both electrochemical and chemical oxidation have been used to produce 3-substituted alkylsulfonated pyrroles (124). Electrochemical polymerisation was achieved using acetonitrile as solvent to form a solid deposit on the electrode. Alternatively, FeCl₃ was used as oxidant. Conductivities in the range 0.001 to 0.500 s cm⁻¹ were obtained, with lower conductivity products obtained from chemical polymerisation. Others (125,126) have prepared homopolymers and copolymers of polypyrroles with alkyl sulfonate groups attached via the N-group. This N-group substitution decreases the polymers' inherent conductivity.

Polythiophenes can also be rendered either organic solvent soluble (127) or water soluble (128) using these derivitization approaches. Similarly, the incorporation of ionizable sulfonic acid groups onto the aniline rings or the aniline nitrogen atom, either pre- or postpolymerization, has provided routes to self-doped, water-soluble polyanilines (129-131). Water-soluble sulfonated polyanilines have been recently synthesized under high pressure to obtain products with higher molecular weight (132). The polymerization of aniline monomers containing alkyl or alkoxy ring substituents leads to polymers with improved solubility in organic solvents (133,134).

Formation of Colloidal ICPs

The improvements in solubility achieved via ring substitution generally result in significant loss in electrical

conductivity of the final polymer. Consequently, formation of colloidal dispersions is an attractive alternative route to solution processing in water, as this allows for postsynthesis handling while retaining reasonable conductivity.

Conducting polymer colloids can be produced by chemical (135,136) or electrochemical (128,137,138) oxidation of monomer in the presence of a steric stabilizer. Colloids produced electrochemically are formed by interrupting the polymer deposition on the electrode surface utilizing hydrodynamic control. This is facilitated by the presence of a steric stabilizer in solution that coats the insoluble polymer upon formation, preventing deposition. The electrochemical approach is advantageous in that the polymer properties can be altered by accurate control of the oxidation potential during polymerisation. This technique also allows a wide range of dopants to be incorporated into the polymer to give different properties. For example, proteins can be incorporated into conducting polymers while retaining their biological integrity (138).

Armes et al. (139–142) have shown that polypyrrole and polyaniline colloids can be successfully prepared via chemical oxidation using fine colloidal silica as a dispersant. The colloids have a low percentage of conducting polymer but still have reasonable conductivity. Zeta potential measurements (139) suggest that stabilization is actually provided by formation of “raspberry” morphologies with the inorganic oxide on the outer layer. The colloids obtained have significant microporosity (140,143). Silica nanocomposites containing polyanilines had particle sizes in the range 300 to 600 nm and typical conductivities of about $6 \times 10^{-2} \text{ s cm}^{-1}$. We have generated similar silica-stabilized colloidal polyanilines via the electrohydrodynamic route, including optically active PAn.(+)-HCSA/silica (144).

Use of Surfactant-like Dopant Anions

For organic solvent solubility, an alternative approach to solubilizing polyanilines and polypyrroles, without sacrificing high electrical conductivity, is the use of surfactant-like dopant anions. With polypyrrole this has recently been achieved via oxidation of the pyrrole monomer with ammonium persulfate in the presence of dodecylbenzene

sulfonate (145,146). Similarly, the conducting emeraldine salt form of PAn.HA can be readily solubilized in a range of organic solvents via the use of camphorsulfonic acid or dodecylbenzenesulfonic acid as the dopant, HA (147,148).

DEVICE FABRICATION

For the fabrication of a practical device, the functional properties of ICPs must be integrated within a host structure that provides the mechanical/physical properties required.

Components with improved mechanical properties can be produced by mixing the above-noted processable polymers with other polymers. For example, conducting polymer colloids have been mixed with water-based latex paints to form conductive, electroactive paints with excellent adhesion to a range of metals (149). Interestingly, the paint-metal adhesion was actually increased by addition of the conducting polymer colloid.

ICPs have also been assembled inside a number of host polymers including polyacrylonitrile (150) and polyvinyl alcohol (151). The inert (insulating) host matrix is first cast onto a suitable electrode. After imbibing the monomer into the host polymer and supporting electrolyte (to provide the dopant), electropolymerization to form a conducting polymer network is initiated. ICPs have also been assembled inside hydrogels retaining both the electronic properties of ICPs and the water adsorption properties of gels (152,153). The growth of conducting polymers inside these host structures provides an opportunity to make all polymer devices since the host may function as a polymer electrolyte.

Composites have been prepared using emulsion polymerization approaches (154,155) or by co-precipitation (156). Others have coated polymethyl methacrylate (PMMA) spheres (157) with conducting polymers. The resultant particles can then be pressed to form films. A similar approach was used by DSM Pty. Ltd. (158) to produce water-borne polyurethane dispersions that could be simply cast as films. The solubility of polyanilines containing selected dopants facilitates their use in formation of polymer blends (159).

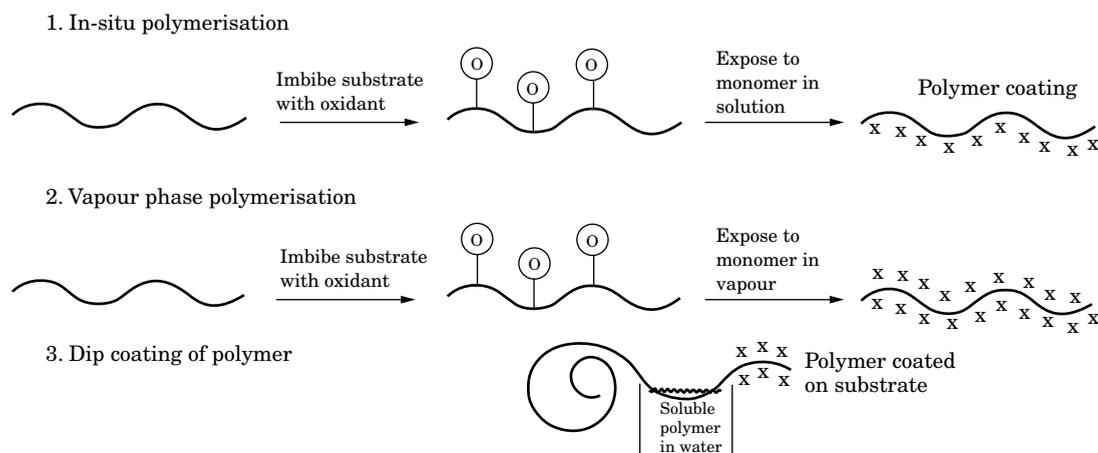
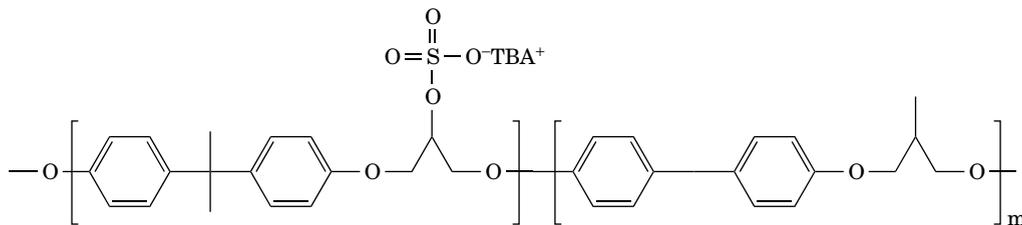


Figure 6. Schematic representation of approaches used to coat fabrics.



*TBA = Tributylammonium

Figure 7. Structure of PBHE.

Conducting polymers have been prepared as coatings on both natural and synthetic fibres and fabrics. For example, silk, wool (160,161), and nylon (162) have been successfully coated with thin, uniform, adherent coatings. The Milliken Corporation developed the first commercial process for producing conducting polymer-coated fabrics (163) by an in situ chemical polymerization method that utilizes dopants that promote adhesion to the fabric. Detailed studies in our laboratories have shown that the properties of the ICP coating are intimately related to the surface chemistry of the textile substrate used as well as the polymerization condition. A vapor phase polymerization method (Fig. 6) has also been used to coat a range of fabrics.

The physical properties of ICPs can be manipulated by incorporating polyelectrolytes (PEs) as dopants. For example, sulfated poly(β -hydroxy) ethers (PBHE) shown in Fig. 7 have been incorporated with dramatic effects on the mechanical properties—increasing the elongation to break to greater than 200% (117,164).

Recently (165), this polyelectrolyte has been incorporated into polythiophene using bithiophene as the monomer. Again, remarkable physical properties (tensile strength = 120 MPa) were obtained.

Others (166) have shown that polyanilines can be rendered “water soluble” by incorporation of appropriate polyelectrolytes such as polystyrenesulfonate. Electrically conducting gels (materials with high water content/good conductivity) are also formed by incorporation of polyelectrolytes as dopants (167,168).

A number of functions require creation of p–n junctions. To date this is usually achieved with an ICP-metal, the metal being predeposited on a suitable substrate or sputter coated onto the polymer. Thin metal layers may also be necessary for highly conducting interconnects. More recently, p–n junctions have been created using the same polymer with different dopants (169).

The ability to provide more processable inherently conducting polymers, as described previously, enables new approaches to device fabrication, including ink jet printing (170) and screen printing (171). Both approaches utilize well-proven technologies to produce patterns on surfaces. As conducting polymer formulations are refined for these purposes, the production of polymer circuits on a wide range of surfaces will be realized. Photolithography has also been used to produce ICP patterns (172–174), while spin coating has been used to produce thin, even films (175).

CONCLUSIONS

The discovery of inherently conducting polymers just over 20 years ago provided materials that could be utilized as an organic alternative in areas previously limited to the use of inorganic materials. The use of inherently conducting polymers in areas such as sensors, electrochemical actuators, photovoltaic materials, electronic components, and light-emitting coatings has subsequently been pursued. In many instances, the polymers were coupled with the new processing and fabrication approaches and attained great performance.

The discovery of inherently conducting polymers has provided a class of materials that will feed the imagination of scientists and engineers in pursuit of intelligent material systems and structures. No other class of materials possesses the inherent properties necessary to function as sensors, information processors and actuators, as well as the possibility of providing an energy conversion and storage system.

BIBLIOGRAPHY

1. G.G. Wallace, G.M. Spinks, and P.R. Teasdale, *Conductive Electroactive Polymers: Intelligent Material Systems*. Technomic Publishing, Lancaster, PA, 1997.
2. G.G. Wallace, *Chem. Britain* (Nov.): 967 (1993).
3. X. Ren and P.G. Pickup, *J. Phys. Chem.* **97**: 5356 (1993).
4. A.A. Syed and M.K. Dinesan, *Talanta* **38**: 815 (1991) (and references cited therein).
5. L.H.C. Mattoso and A.G. MacDiarmid. In J.C. Salamone, ed., *Polymeric Materials Encyclopedia*. CRC Press, Boca Raton, FL, p. 5505.
6. O. Inganas, W.R. Salaneck, J.E. Osterholm, and J. Laakso, *Synth. Met.* **22**: 395 (1998).
7. M.J. Winokur, D. Spiecel, Y. Kim, S. Hotta, and A.J. Heeger, *Synth. Met.* **28**: C419 (1989).
8. M. Leclerc, M. Frechette, J.-Y. Bergonen, M. Ranger, I. Levesque, and K. Faid, *Macromol. Chem. Phys.* **197**: 2077 (1996) (and references cited therein).
9. I.D. Norris, L.A.P. Kane-Maguire, and G.G. Wallace, *Macromol.* **31**: 6529 (1998).
10. F. Zuo, R.P. McCall, J.M. Ginder, M.G. Roc, J.M. Leng, A.J. Epstein, G.E. Asturias, S.P. Ermer, A. Ray, and A.G. MacDiarmid, *Synth. Met.* **29**: E445 (1989).

11. H.-T. Lee, K.-R. Chuang, S.-A. Chen, P.-K. Wei, J.H. Hsu, and W. Fann, *Macromolecules* **28**: 7645 (1995) (and references cited therein).
12. B.M.W. Langeveld, E. Peeters, R.A.J. Janssen, and E.W. Meijer, *Synth. Met.* **84**: 611 (1997).
13. T. Laka, *Synth. Met.* **55-57**: 5014 (1993).
14. H. Okuzaki and T. Kunugi, *Funct. Mater.* **17**: 14 (1997).
15. Q. Pei and O. Inganas, *Synth. Met.* **55-57**: 3730 (1993).
16. M. Umana and J. Waller, *Anal. Chem.* **58**: 2979 (1986).
17. N.C. Foulds and C.R. Lave, *Anal. Chem.* **60**: 2473 (1988).
18. H.S. Shinohara, T. Chiba, and M. Aizawa, *Sens Actuators* **13**: 79 (1988).
19. S.B. Adeloju, S.J. Shaw, and G.G. Wallace, *Anal. Chim. Acta.* **281**: 621 (1993).
20. S.B. Adeloju, J.N. Barisci, and G.G. Wallace, *Anal. Chim. Acta.* **332**: 145 (1996).
21. O.A. Sadik and G.G. Wallace, *Anal. Chim. Acta.* **279**: 200 (1993).
22. D. Barnett, O.A. Sadik, M.J. John, and G.G. Wallace, *Analyst* **119**: 1997 (1994).
23. T.E. Campbell, A.J. Hodgson, and G.G. Wallace, *Electroanal.* (1999), in press.
24. M.V. Deshpande and E.A.H. Hall, *Biosens. Bioelectron.* **5**: 431 (1990).
25. M. Nishizawa, T. Matsue, and J. Uchida, *Anal. Chem.* **64**: 2642 (1992).
26. W. Lu, D. Zhou, and G.G. Wallace, *Anal. Comm.* **35**: 245 (1998).
27. G. Bidan. In G. Harsany, ed., *Polymer Films in Sensor Applications*. Technomic Publishing, Lancaster, FA, p. 206.
28. S.J. Higgins, *Chem. Soc. Rev.* **26**: 247 (1997).
29. H. Korri-Yousouffi, P. Godillot, P. Srivastava, A. El Kassmi, and F. Garnier, *Synth. Met.* **84**: 169 (1997).
30. T.M. Swager, *Acc. Chem. Res.* **31**: 201 (1998) (and references cited therein).
31. M. Marsella, P. Carroll, and T. Swager, *J. Am. Chem. Soc.* **117**: 9832 (1995).
32. Z. Chen, P.A. Gale, and P.D. Beer, *J. Electroanal. Chem.* **393**: 113 (1995).
33. A. Buffenoir, G. Bidan, L. Chalumeau, and I. Soury-Lavergne, *J. Electroanal. Chem.* **451**: 261 (1998).
34. A. Buffenoir and G. Bidan. In *Int. Conf. Sci. Tech. Synth. Metals*. Montpellier, France, 1998, p. 146.
35. J.M. Cooper, D.G. Morris, and K.S. Ryder, *J. Chem. Soc. Chem. Commun.* 697 (1995).
36. F. Garnier, H. Youssoufi, P. Srivastava, and A. Yassar, *J. Am. Chem. Soc.* **116**: 8813 (1994).
37. G. Bidan, M. Billan, K. Galasso, T. Livache, G. Mathis, A. Reget, L.M. Torres-Rodriguez, and E. Vieil, *Appl. Biochem.* (1999) (and references cited therein), in press.
38. F. Garnier, H. Korri-Yousouffi, P. Srivastava, B. Mandrand, and T. Delair, *Synth. Met.* **100**: 89 (1999) (and references cited therein).
39. T. Livache, B. Fouque, A. Roget, J. Marchand, G. Bidan, R. Toule, and G. Mathis, *Anal. Biochem.* **255**: 188 (1998) (and references cited within).
40. M. Lemaire, D. Delabouglise, R. Garreau, A. Guy, and J. Roncali. *J. Chem. Soc. Chem. Commun.* 658 (1988).
41. D. Delabouglise and F. Garnier, *Synth. Met.* **39**: 117 (1990).
42. J.C. Moutet, E. Saint-Aman, F. Tran-Van, P. Angibaud, and J. P. Utille, *Adv. Mater.* **718**: 511 (1992).
43. S. Pleus and M. Schwientek, *Synth. Met.* **95**: 233 (1998).
44. G. Bidan and M.-A. Niel, *Synth. Met.* **84**: 255 (1997).
45. M. Davey, S. Ralph, and G. Wallace, unpublished results.
46. V. Misoska. In *Chemistry*. University of Wollongong: Wollongong, 1998.
47. M. Majidi, L.A.P. Kane-Maguire, and G.G. Wallace, *Polymer* **35**: 3113 (1994).
48. M. Majidi, L.A.P. Kane-Maguire and G.G. Wallace, *Polymer*. **18**: 3597 (1995).
49. E.E. Havinga, M.M. Bouman, E.W. Meijer, A. Pomp, and M.M. Simenon, *Synth. Met.* **66**: 93 (1994).
50. S.A. Syed, L.A.P. Kane-Maguire, M.R. Majidi, S.G. Pyne, and G.G. Wallace, *Polymer* **38**: 2627 (1997).
51. M.R. Majidi, S.A. Ashraf, L.A.P. Kane-Maguire, and G.G. Wallace. In *RACI Symp. Adv. Polymers III—Designer Polymers*. Melbourne, Australia, 1996, p. 12.
52. H. Guo, C.M. Knobler, and R.B. Kaner. In *Int. Conf. Sci. Tech. Synth. Metals*. Montpellier, France, 1998, p. 176.
53. W. Lu, H. Zhao, and G.G. Wallace, *Electroanal.* **8**: 248 (1996).
54. W. Lu, G.G. Wallace, and A. Karaykin, *Electroanal.* **10**: 472 (1998).
55. B. Fabre and G. Bidan, *J. Chem. Soc. Farad. Trans.* **973**: 591 (1997).
56. B. Fabre, S. Burlet, G. Bidan, and R. Cespuaglio. **426**: 75 (1997).
57. H.V. Shurmer, P. Corcoran, and J.W. Gardner, *Sens. Actuators* **B4**: 29 (1991).
58. R. Stella, N. Barisci, G. Serra, G. Wallace, and D. De Rossi, *Sens. Actuators* (1999), in press.
59. P.K. Namdev, Y. Alroy, and V. Singh, *Biotechnol. Prog.* **14**: 75 (1998).
60. M.C. Lonergan, E.J. Severin, B.J. Doleman, S.A. Beaber, R.H. Grubbs, and N. S. Lewis, *Chem. Mater.* **8**: 2298 (1996).
61. T.D. Gibson, O. Prouser, J.N. Halbert, R.W. Marshall, P. Corcoran, P. Lowery, E.A. Ruck-kerene, and S. Heron, *Sens. Actuators* **B44** (1997).
62. J.W. Gardner, T.C. Pearce, and S. Friel, *Sens. Actuators* **B118-9**: 240 (1994).
63. A.C. Partridge, P. Harris, and M.K. Andrews, *Analyst* **121**: 1349 (1996).
64. R. John, D. Ongarato, and G.G. Wallace, *Electroanal.* **8**: 623 (1996).
65. W. Lu, T. Nguyen, and G.G. Wallace, *Electroanal.* **10**: 1101 (1998).
66. P.R. Teasdale and G.G. Wallace, *Polym. Int.* **35**: 197 (1994).
67. H. Chriswanto and G.G. Wallace, *Chromatographia* **42**: 191 (1996).
68. H. Chriswanto and G.G. Wallace, *J. Liq. Chrom.* **19**: 2457 (1996).
69. M.M. Chehimi, in N.P. Cheremisinoff, and P.N. Cheremisinoff, eds., *Handbook of Advanced Materials Testing*. Dekker: New York.
70. M.M. Chehimi, M.L. Abel, and Z. Sahraoui, *J. Adhesion Sci. technol.* **10**: 287 (1996).
71. B. Zinger and L.L. Miller, *J. Am. Chem. Soc.* **196**: 6861 (1987).
72. Q.X. Zhou, L.L. Miller, and J.R. Valentine, *J. Electroanal. Chem.* **261**: 147 (1989).

73. Y. Lin and G.G. Wallace, *J. Cont. Release* **30**: 137 (1994).
74. Y. Lin, P.J. Riley, and G.G. Wallace, *Anal. Lett.* **22**: 669 (1989).
75. P. Burgmayer and R.W. Murray, *J. Phys. Chem.* **88**: 2515 (1984).
76. H. Zhao, W.E. Price, C.O. Too, G.G. Wallace, and D. Zhou, *J. Memb. Sci.* **119**: 199 (1996) (and references cited therein).
77. M.R. Anderson, B.R. Mattes, H. Reiss, and R.B. Kaner, *Science* **252**: 1412 (1991).
78. S. Kuwabata and C.R. Martin, *J. Memb. Sci.* **91**: 1 (1994).
79. R. Parhasarathy, V.P. Menon, and C.R. Martin, *Chem. Mater.* **9**: 560 (1997).
80. W.E. Price, G.G. Wallace, and H.J. Zhao, *J. Electroanal. Chem.* **334**: 111 (1992).
81. H. Zhao, W.E. Price, and G.G. Wallace, *Polymer* **34**: 16 (1993).
82. H. Zhao, W.E. Price, and G.G. Wallace *J. Memb. Sci.* **100**: 239 (1995).
83. D. Zhou, C.O. Too, and G.G. Wallace, *J. Int. Mater. Syst. Struct.* (1999), in press.
84. A. Talaie and G.G. Wallace, *Synth. Met.* **63**: 83 (1994).
85. R. John and G.G. Wallace, *J. Electroanal. Chem.* **354**: 145 (1993).
86. K. Hyodo, *Electrochimica Acta.* **39**: 265 (1994).
87. G.A. Sotzijing, J.R. Reynolds, and P.J. Steel, *J. Chem. Mater.* **8**: 882 (1996).
88. P. Murray, G.M. Spinks, G.G. Wallace, and R.P. Burford, *Synth. Met.* **97**: 117 (1998).
89. K. Yoshino, M. Tabata, M. Satoh, K. Kaneto, and T. Ohsawa, *Tech. Reports*, Osaka Univ., **35**: 231 (1985).
90. R. Baughman, L. Shacklette, R. Elsenbaumer, E. Plitcha, and C. Becht. In J. Bredas and R. Chance, eds., *Conjugated Polymeric Materials: Opportunities in Electronics. Optoelectronics and Molecular Electronics*. Klywer Academic, Dordrecht, p. 599.
91. M. Ghandi, P. Murray, G.M. Spinks, and G.G. Wallace, *Synth. Met.* **73**: 247 (1995).
92. T.W. Lewis, S.E. Moulton, G.M. Spinks, and G.G. Wallace, *Synth. Met.* **85**: (1997).
93. T.W. Lewis, G.M. Spinks, G.G. Wallace, D. De Rossi, and M. Pachetti, *Pol. Preprints* **38**: 520 (1997).
94. T.F. Otero and J.M. Samsinena, *Bioelectrochem. Bioenerg.* **42**: 117 (1997).
95. S. Shakuda, S. Morita, T. Kawai, and K. Yoshiro, *Jpn. J. Appl. Phys.* **32**: 5143 (1993).
96. Q. Pei and O. Inganas, *J. Phys. Chem.* **97**: 6034 (1993).
97. P. Chiarelli, A. Della Santa, D. De Rossi, and A. Mazzoldi, *J. Int. Mater. Syst. Struct.* **6**: 32 (1995).
98. A. Hutchison, T. Lewis, S. Moulton, G. Spinks, and G. Wallace, *Synth. Met.* (1999), in press.
99. E. Smela, O. Inganas, and I. Landstrom, *Adv. Mater.* **5**:630 (1993).
100. E. Smela, O. Inganas, and I. Landstrom, *J. Micromech. Microeng.* **9**: 203 (1993).
101. E. Smela, O. Inganas, and I. Landstrom, *Science* **268**: 1735 (1995).
102. A.P. Lee, K.C. Hong, J. Trevino, and M.A. Northrop, *Dyn. Syst. Contr.* **2**: 725 (1994).
103. F.T. Hong, ed, *Molecular Electronics: Biosensors and Biocomputers*. Plenum Press, New York, 1988.
104. R. Gupta, S.C.K. Misra, B.D. Malhorta, N.N. Beladakee, and S. Chandra, *Appl. Phys. Lett.* **58**: 51 (1991).
105. Y. Ohmori, H. Tahrahashi, H. Muroki, M. Uchida, T. Kawai, and K. Yoshino, *Jpn. J. App. Phys.* **30**: 610 (1991).
106. E.W. Paul, A.J. Ricco, and M.S. Wrighton, *J. Phys. Chem.* **89**: 1441 (1985).
107. G.P. Kittleson, H.S. White, and M.S. Wrighton, *J. Am. Chem. Soc.* **106**: 7389 (1984).
108. C.H. McCoy and M.S. Wrighton, *Chem. Mater.* **5**: 914 (1993).
109. M. Kaneko, In H.S. Nalwa, ed., *Handbook of Conducting Molecules and Polymers*. Wiley, New York.
110. G. Yu, J. Gao, J.C. Hummelen, F. Wudl, and A.J. Heeger, *Science* **270**: 1789 (1995).
111. R.N. Marks, J.J.M. Halls, D.D.C. Bradley, R.H. Friend, and A.B. Holmes, *J. Phys. Condens. Matter* **6**: 1379 (1994).
112. S.A. Chen and Y. Fang, *Synth. Met.* **60**: 215 (1993).
113. T. Yohannes, T. Solomon, and O. Inganas, *Synth. Met.* **82**: 215 (1996).
114. P. Dastoor, D. Officer, C. Too, and G. Wallace, *Chemical Innovations* (1999), in press.
115. J.G. Killian, B.M. Coffey, F. Gao, T.O. Poehler, and P.C. Searson, *J. Electrochem. Soc.* **143**: 936 (1996).
116. E.M. Genies, P. Hany, and C. Santier, *J. Appl. Electrochem.* **18**: 751 (1988).
117. H. Yamato, W. Wernet, M. Ohwa, and B. Rotzinger, *Synth. Met.* **55**: 3550 (1993).
118. M. Satoh, H. Ishikawa, K. Amano, E. Hasegawa, and K. Yoshino, *Synth. Met.* **71**: 2259 (1995).
119. A. Clemente, S. Panero, E. Spila, and B. Scrosati, *Sol. State Ionics* **85**: 273 (1996).
120. H. Masuda and K. Kaeriyama, *J. Mat. Sci.* **26**: 5637 (1991).
121. J. Ruhe, T. Ezquerra, and G. Wegner, *Makrom. Chem. Rapid. Commun.* **10**: 103 (1989).
122. A.O. Patil, Y. Ikenoue, F. Wudl, and A.J. Heeger, *J. Am. Chem. Soc.* **109**: 1858 (1987).
123. S.A. Ashraf, F. Chen, C.O. Too, and G.G. Wallace, *Polymer* **37**: 2811 (1996).
124. E.E. Havinga, W. ten Hoeve, E.W. Meijer, and H. Wynberg, *Chem. Mater.* **1**: 650 (1989).
125. N.S. Sundarasan, S. Basak, M. Pomerantz, and J.R. Reynolds, *J. Chem. Soc. Commun.* 621 (1987).
126. J.R. Reynolds, N.S. Sundaresen, M. Pomerantz, and C.K. Basak, *J. Electroanal. Chem.* **250**: 355 (1988).
127. M.R. Bryce, A. Chissel, P. Kathirgamanathan, D. Parker, and R.M. Smith, *J. Chem. Soc. Chem. Commun.* 466 (1987).
128. J.N. Barisci, J. Mansouri, G.M. Spinks, G.G. Wallace, C.Y. Kim, D.Y. Kim, and J.Y. Kim, *Colloids and Surfaces* **126**: 129 (1997).
129. S.-A. Chen and M.-Y. Hua, *Macromol.* **29**: 4919 (1996) (and references cited therein).
130. M.T. Nguyen, P. Katai, J.L. Miller, and A.F. Diaz, *Macromol.* **27**: 3625 (1994).
131. X.-L. Wei, Y.-Z. Wang, S.M. Long, C. Bobeczko, and A.J. Epstein, *J. Am. Chem. Soc.* **118**: 2545 (1997) (and references cited therein).
132. H.S.O. Chan, A.J. Neuendorf, S.C. Ng, P.M.L. Wong, and Y.D.J, *Chem. Commun.* 1327 (1998).
133. Y. Wei, R. Hariharan, and S.A. Patel, *Macromol.* **23**: 758 (1990).
134. W.A. Gozotti and M. De Paoli, *Synth. Met.* **80**: 263 (1996).
135. E.C. Cooper and B. Vincent, *J. Appl. Phys.* **22**: 1580 (1989).
136. M. Aldissi and S.P. Armes, *Progr. Org. Coat.* **19**: 21 (1991).

137. H. Eisazadeh, G.M. Spinks, and G.G. Wallace, *Polymer* **35**: 3801 (1994).
138. J.N. Barisci, A.J. Hodgson, L. Liu, and G.G. Wallace, *React. Polym.* **39**: 269 (1999).
139. M.D. Butterworth, R. Cavvadi, J. Johal, S.F. Lascelles, S. Maeda, and S.P. Armes, *J. Coll. Sci. Interfac. Sci.* **174**: 510 (1995).
140. R. Flitton, J. Johal, S. Maeda, and S.P. Armes, *J. Coll. Sci. Interfac. Sci.* **173**: 135 (1995).
141. G.P. McCarthy, S. P. Armes, S. J. Greaves, and J. F. Watts, *Langmuir*. **13**: 3686 (1997).
142. J. Stejskal, P. Kratochvil, S.P. Armes, S.F. Lascelles, A. Riede, M. Helmstedt, A. Riede, J. Prokes, and I. Krivka, *Macromolecules*. **29**: 6814 (1996).
143. C. Perruchot, M.M. Chehimi, M. Delamer, F. Lascelles, and S. P. Armes, *J. Coll. Sci. Interfac. Sci.* **193**: 190 (1997).
144. V. Aboutanus, L.A.P. Kane-Maguire, and G.G. Wallace, *Synth. Met.* **Submitted**, (1999).
145. J.Y. Lee, D. Y. Kim, and C.Y. Kim, *Synth. Met.* **74**: 103 (1995).
146. J.Y. Lee, K.T. Song, S.Y. Kim, Y.C. Kim, D.Y. Kim, and C. Y. Kim, *Synth. Met.* **84**: 137 (1997).
147. Y. Cao, J. Qui, and P. Smith, *Synth. Met.* **69**: 187 (1995).
148. K.G. Neoh, M.Y. Pun, E.T. Kang, and K. L. Tan, *Synth. Met.* **73**: 209 (1995).
149. H. Eisazadeh, G.M. Spinks, and G.G. Wallace, *Materials Forum*. **16**: 111 (1992).
150. Y.H. Park and M.J. Hee Han, *J. Appl. Polym. Sci.* **45**: 1973, (1992).
151. H.L. Wang and J.E. Fernandez, *Macromol.* **26**: 3336 (1993).
152. K. Gilmore, A.J. Hodgson, B. Luan, C.J. Small, and G.G. Wallace, *Polym. Gels and Networks* **2**: 135 (1994).
153. C.J. Small, C.O. Too, and G.G. Wallace, *Polym. Gels and Networks*. **5**: 251 (1997).
154. E. Ruckenstein and J. Park, *J. Appl. Polym. Sci.* **42**: 925 (1991).
155. E. Ruckenstein and L. Hang, *Synth. Met.* **66**: 249 (1994).
156. E. Ruckenstein and J. H. Chen, *Polym.* **32**: 1230 (1991).
157. K. Yoshino, S. Marita, X. H. Yin, M. Onada, H. Yamamoto, T. Watanuki, and I. Isa, *Synth. Met.* **55**: 3562 (1993).
158. A.E. Wiersma, V. Steeg, and T.J.M. Jongeling, *Synth. Met.* **71**: 2269 (1995).
159. M. Regbu, C.O. Yoon, Y. Yang, D. Moses, P. Smith, A.J. Heeger, and Y. Cao, *Phys. Rev. B.* **50**: 13931 (1994).
160. S.N. Bhadani, S.K. Sen Gupta, G.C. Sahu, and M. Kumari, *J. Appl. Polym. Sci.* **61**: 207 (1996).
161. S.W. Tan and H. Ge, *Polymer*. **37**: 965 (1996).
162. D. Abraham, A. Bharati, and I. Subramanyam, *Polymer*. **23**: 5295 (1996).
163. R.V. Gregory, W.C. Kimbrell, and H.H. Kum, *Synth. Met.* **28**: 823 (1989).
164. H. Wernet, *Synthetic Metals*. **43**: 843 (1991).
165. J. Ding, W. Price, S. Ralph, and G. Wallace. (1999), in press.
166. Y.H. Liao and J. Levank, *Macromol. Rapid Commun.* **16**: 393 (1995).
167. A.J. Hodgson, K. Gilmore, C.J. Small, G.G. Wallace, I.L. Mackenzie, T. Aoki, and N. Ogata, *Supramol. Sci.* **1**: 77 (1994).
168. H. Zhao and G.G. Wallace, *Polym. Gels and Networks*. **6**: 233 (1998).
169. S. Jin, M. Ji, and G. Xue, *Appl. Phys. A.* **63**: 397 (1996).
170. D. Pede, G. Serra, and D. De Rossi, *Mat. Sci. Eng.* **C5**: 89 (1998).
171. J.P. Hart and S.A. Wrings, *TrAc.* **16**: 89 (1997).
172. A. Nannini and G. Serra, *J. Mol. Electron.* **6**: 81 (1990).
173. M.S.A. Abdou, G.A. Diaz-Guijada, M.I. Arroyo, and S. Holdcraft, *Chem. Mater.* **3**: 1003 (1991).
174. D.G. Lidzey, M.A. Pate, M.S. Weaver, T.A. Fisher and D.D.C. Bradley, *Synth. Met.* **82**: 141 (1996).
175. D. Chinn and J. Janata, *Thin Sol. Films.* **252**: 145 (1994).

CURE AND HEALTH MONITORING

TAKEHITO FUKUDA
TATSURO KOSAKA
Osaka City University
Sumiyoshi-ku, Osaka, Japan

SMART MONITORING SYSTEM

The smart monitoring system has evolved due to advances in sensors, sensing methods, and communication technologies, as shown in Fig. 1. The sensing function involves sensors, signal sources, and signal detectors. Measured data are analyzed by a signal processor using an appropriate analytical approach. The analyzed results are reported to operators or observers. In the old days, human beings performed these functions (Fig. 1a). Today, they are replaced by sensor devices, computers, and communication networks (Fig. 1b). For example, a fire alarm system consists of a temperature sensor, a microcomputer, and an alarm. Furthermore, the recent development in sensors and computers make these devices so small and light that they can be integrated into structures. An integrated monitoring system makes the materials and structures "smart", and then it can be called a "smart" monitoring system (Fig. 1c).

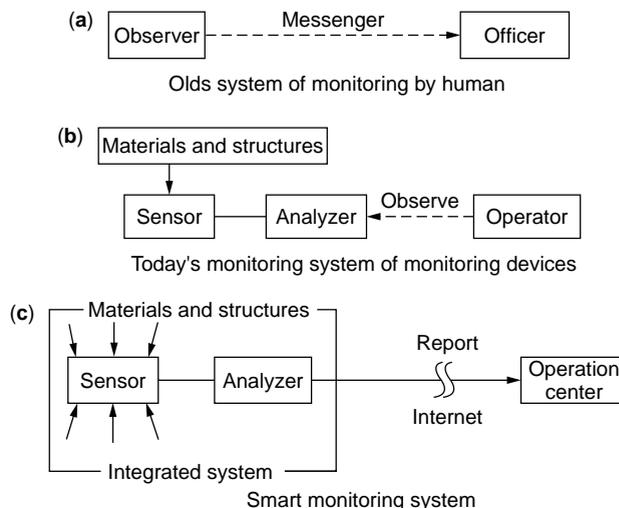


Figure 1. Concepts of monitoring systems in the past, present, and in the future.

The smart monitoring system was an imaginary technology several decades ago. You may see or read a science fiction story of a spacecraft that has a function for monitoring the state of all elements. Such a structure is still a potential, but it is becoming a real structure as the smart monitoring system advances. The elements of a smart monitoring system are integrated sensors, integrated microprocessors, analytical software, and a network. Small, lightweight sensors that are embedded in or attached to structures add the function of self-inspection to the host structures. Powerful microprocessors and analytical software make it possible to analyze data more precisely for identification or prediction in real time. An integrated network function enables remote monitoring everywhere in the world. The integrated monitoring system has much merit. The material and structure, where the monitoring system is integrated, can know and estimate the condition by itself, in other words, can feel the condition. Therefore, the maintenance cost of materials and structures can be reduced. The total weight of the monitoring systems can also be reduced because the sensors used in smart monitoring systems are smaller and lighter than traditional sensors. In addition, a new business service may develop to monitor structures in the future and ensure the safety and performance after construction. These advantages have motivated studies of smart monitoring techniques that increase from year to year.

The smart monitoring system has mainly two practical, commercial targets. One is "manufacturing process monitoring" and another is "health monitoring." The former is a technique for monitoring the state of materials in the manufacturing process. Monitoring of the molding process for thermosetting polymers, called "cure monitoring," has been especially focused on polymer matrix composites (PMCs). In manufacturing process monitoring, the chemical and physical state of materials and the integrity of products are monitored. The collected data can also be used for optimal control of the forming process to improve efficiency. Monitoring the integrity of the product makes it unnecessary to inspect after manufacture. Then, the manufacturing cost can be reduced by introducing this technique. Health monitoring, on the other hand is a technique for monitoring the condition of materials and structures in operation. In the last decade, health monitoring has become a hot issue especially in aerospace and civil structures. These structures suffer from mechanical and thermal fatigue, long exposure time, and natural disasters such as earthquakes and typhoons. Thus, periodic inspections are required to maintain the safety of these structures adequately. However, the inspection cost of large structures is very high, and operation must be interrupted during the inspection. Therefore, real-time health monitoring promises to reduce maintenance costs and to improve reliability. By combining these two monitoring techniques, it is possible to monitor the state of materials and structures from their beginnings to their ends. This means that the quality of products can be ensured continuously at low cost. Therefore, the idea is a very attractive solution for reducing the total cost of ownership, including the price, maintenance, and operating cost.

As described before, the smart monitoring technique consists of three technologies, a sensing technology, an analytical technology, and a network technology. In this

article, the sensing technology is particularly emphasized because integrated sensors make materials and structures smart. Small size, lightweight, durable sensors are desired for the smart monitoring technique. Suitable sensors are fiber-optic sensors, piezoelectric sensors, electrostrictive sensors, magnetostrictive sensors, and dielectric sensors. Some applications require special properties. For example, the sensors for a space structure must be insensitive to an electromagnetic field and temperature. Civil structures require long-term survivability of integrated sensors. Traditional sensors such as strain gauges and thermocouples are used only for compensation in laboratory studies. Some types of sensors, such as fiber-optic strain sensors, dielectric sensors, and piezoelectric sensors, can be used for both cure monitoring and health monitoring; such sensors can be employed to monitor materials from their beginning to their end. Measurement of material properties such as stiffness, density, damping, and conductivity is also available for health monitoring. The advantage of measuring the variation of material properties makes embedded sensors unnecessary. More details of these topics are given in the following subsections.

CURE MONITORING

Today, engineering plastics and polymer matrix composites (PMCs) are widely used in consumer and leisure products such as golf clubs, fishing rods, skis, and tennis rackets. Fiber-reinforced plastics (FRPs), particularly, are the most promising composite materials for airplanes, space structures, and military ships. Thermosetting and thermoplastic polymers are common materials used in FRPs. In the molding process for polymers, a liquid resin becomes solid. As for thermosetting polymers, a monomeric liquid resin becomes a cross-linked rigid solid and a tightly bound three-dimensional network is produced. This process is called "cure." Thermoplastics do not need to be cured because they are not cross-linked. In these advanced engineering materials, the integrity of the product is very important for certifying the performance. The quality of a product strongly depends on the profiles of the control parameters such as the temperature and pressure in the molding process. Therefore, many researchers have developed techniques for optimizing the molding process. The monitoring technique is essential for the optimal control system of the molding process. The cure monitoring technique, especially has been an important focus because the cure reaction of a thermosetting FRP is too complicated to predict. Techniques for monitoring the molding process of thermosetting FRPs are not discussed. However, it should be remembered that some of the techniques that monitor the mechanical, optical, and electrical properties, can also be applied to monitor the state of solidification in the process of molding thermoplastic polymers.

There are many molding processes for manufacturing PMCs. As shown in Fig. 2, they can be classified by the intermediate stage or the forming stage (1). In the intermediate stage, the manufacturing processes are divided into the manufacturing of molding compounds, the impregnation of resin, and the manufacturing of preforms. The molding methods in forming the final products are

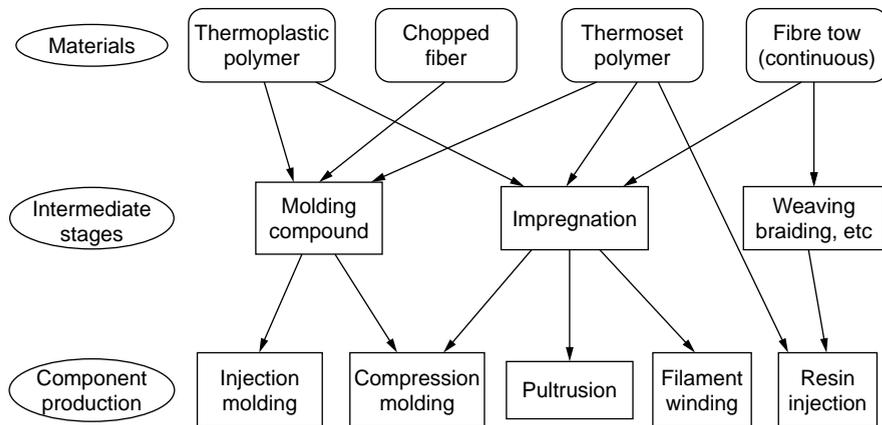


Figure 2. Schematic overview of the approaches employed in fabricating polymer matrix composites (1).

autoclave molding, hot-press molding, liquid molding, injection molding, pultrusion molding, and filament winding (FW) molding. Sheet molding compounds (SMCs) and bulk molding compounds (BMCs) are intermediate molding compounds that contain thermosetting resin, hardeners, thickening agents, and chopped fibers. SMCs are formed into sheets and used for hot-press molding. BMCs have no shape and are used for hot-press molding, injection molding, etc. Hand lay-up (HLU) or spray lay-up (SLU) methods are commonly used for manufacturing of large composite structures. The HLU is a forming method, in which prepregs are stacked up in a die. A prepreg is an intermediate product, which is formed into a tape or a sheet of continuous fibers impregnated by resin. In the SLU method, a liquid resin that contains chopped fibers is sprayed into the die. The resin that covers the die is cured at room or high temperature. The laminate FRPs reinforced by using continuous fibers are molded by applying the proper temperature and pressure to the stacked prepregs in autoclave molding or hot-press molding. Liquid molding methods such as resin transfer molding (RTM), vacuum assisted resin infusion (VARI), and resin infusion molding (RIM) are used to form complex-shaped FRPs. In liquid molding, a reinforcement preform is placed in the mold, and liquid resin is poured into it. FW molding is used to form a pipe or a tank. In FW molding, a mandrel is wound with reinforcement fiber strands that are impregnated by resin. Pultrusion molding is a method for manufacturing long FRPs. In pultrusion molding, continuous fibers impregnated by resin are forced through a die, in which the composites are cured by heat. In most cases, traditional noncontact monitoring cannot be applied to monitoring the cure of FRPs because the surfaces of composites are covered by mold release sheets. Therefore, *in situ* monitoring is essential to monitor the cure state of FRPs.

By the way, remember that cure monitoring is part of manufacturing process monitoring for thermosetting polymers and PMCs. The technology, including monitoring and optimal control of the manufacturing process, can be called “smart” manufacturing technology (Fig. 3). In this article, the manufacturing process includes setup, molding, and inspection. The setup is important for molding complex-shaped products. For example, in liquid molding, monitoring and optimal control of resin flow are desired to certify the integrity of the resin impregnation before curing. In the

curing process, *in situ* sensors monitor changes in chemical constitution, viscosity, elasticity, and conductivity to determine the cure state of the resin. The *in situ* temperature sensor is useful for controlling temperature more precisely by monitoring the temperature distribution in the material because thick or complex-shaped products have complicated temperature distributions during cure. Optimal control of the curing process minimizes processing time. In the cooling stage of the curing process, the residual stress and imperfections in products are monitored to ensure the quality of the final products because the residual stress affects their strength and durability. It is important to monitor residual strain distribution because unexpected residual strain causes defects and damage such as warping, matrix cracks, and delaminations. The smart manufacturing process does not need inspection after curing because the quality is already ensured by integrated sensors at the end of the curing process. Furthermore, product defects can be discovered in the early stages. These advantages have the potential for reducing manufacturing cost. Here, it is considered that the words “smart manufacturing” have another face, which is the manufacturing technique for smart composites. The manufacturing technique for smart composites has been developed to solve problems of interface and interphase between the sensor/actuator and the matrix and problems of connection between the sensor/actuator and the controller. Note that “smart manufacturing” means “smart” manufacturing technology in this article because the monitoring system is the main focus.

The best final product performance is achieved from optimum conditions of temperature and pressure in the curing process. However, these conditions have been obtained mainly by trial and error, and therefore, the curing process is inefficient. Furthermore, nondestructive inspection must be conducted for advanced composites such as carbon-fiber-reinforced plastics (CFRPs) because internal defects cannot be detected visually. The inefficiency, which is caused by the long processing time and nondestructive inspection, increases the manufacturing cost. Therefore, efficient (short time) and optimally controlled cure processing techniques are desired. The cure monitoring technology has an older history than that of smart materials and structures technology, which involves smart monitoring techniques. The traditional methods of cure monitoring

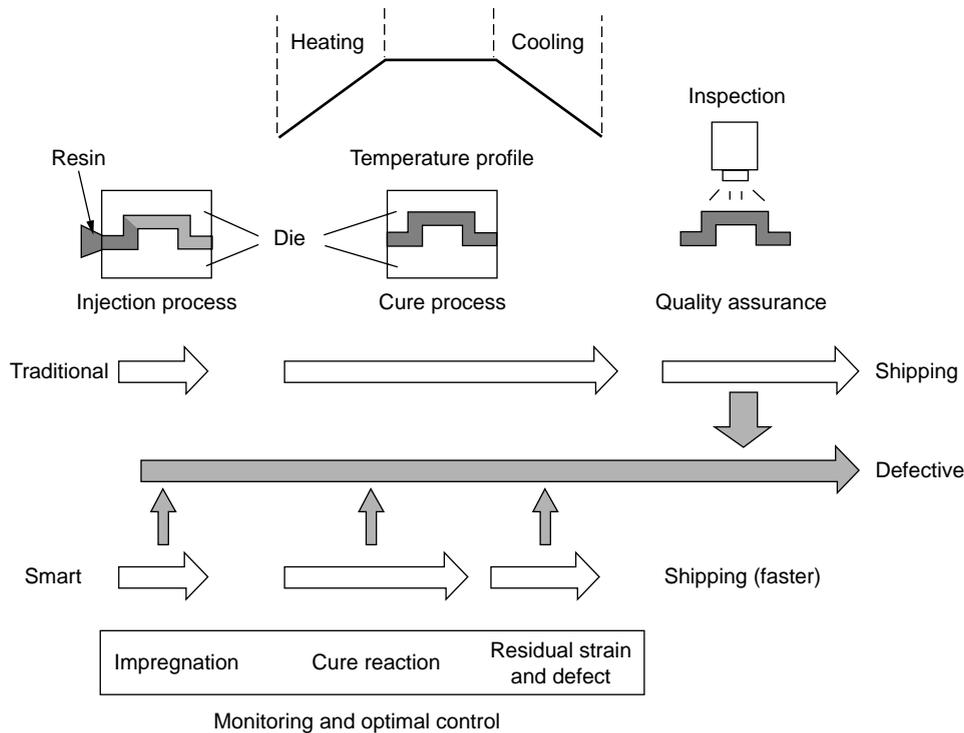


Figure 3. Comparison of the traditional manufacturing process and the smart manufacturing process.

are differential scanning calorimetry (DSC), dynamic mechanical analysis (DMA), infrared spectroscopy (IRS), and dielectric measurement. DSC and DMA cannot be used in molding of actual products. The DSC method measures the heat of calorific reaction in the curing process. The degree of cure (DOC) of the resin is generally defined as the ratio of the heat of initiation to the total heat of the curing reaction. The thermochemical requirements for predicting the DOC were presented by Loos and Springer (2). They also proposed a model to predict the behavior of resin flow in the curing process for FRP laminates. The DMA method measures the variation in the mechanical viscosity of the resin at various temperatures. IRS can be used to measure the variation in the chemical constitution of the resin. The changes in the electrical properties of the resin during cure can be measured by using a dielectric sensor.

The traditional techniques of cure monitoring are in a sense effective. However, *in situ* sensing, which is proposed in smart monitoring, is a more effective approach for

monitoring the cure state because local information is useful for identifying the DOC. Traditional monitoring techniques are employed to develop an analytical model and to supplement the data measured by *in situ* sensors. As shown in Table 1, fiber-optic sensors, dielectric sensors, piezoelectric sensors, and ultrasonic sensors have been investigated for monitoring the state in the manufacturing process. Some of them, such as IRS-based fiber-optic sensors, dielectric sensors, and ultrasonic sensors are based on traditional monitoring techniques. The recent monitoring techniques have come from smart materials and structures technology. Among these sensors, fiber-optic sensors are best for embedding, and they can be used to monitor the chemical, mechanical, and optical properties of the resin in the curing process. However, their cost is high because commercial optical devices are expensive. Dielectric sensors are used for measuring the electrical properties mainly in monitoring cure. The main purpose for using piezoelectric wafers for monitoring cure is their

Table 1. *In Situ* Sensors for Cure Monitoring

	Configuration	Monitored Value	Cost	Use in Operation
Fiber-optic sensors	Embed	Chemical constitution, refractive index, strain, temperature	High	Good
Dielectric sensors	Embed	Electrical permittivity	Middle	Poor
Piezoelectric sensors	Embed	Impedance	Middle	Good
Ultrasonic sensors	Attach to die	Sound velocity, attenuation	Cheap	No

^aBecause the sensors are integrated in the die not in the material

Table 2. Fiber-Optic Sensors for *In Situ* Cure Monitoring

	Monitored Value	Optical System	Sensor Cost	System Cost
NIRS-based	Chemical	Spectroscopy	Cheap	High
Fluorimetry-based	Chemical (curing agent)	Spectroscopy	Cheap	High
Index-based	Refractive index	Intensity	Cheap	Cheap
Strain/temperature (EFPI, FBG, etc.)	Strain, temperature	Interferometer, spectroscopy	High	Middle-high

application to vibrational control and health monitoring of structures in operation. Ultrasonic sensors do not need to be embedded because the materials propagate ultrasonic waves. In this article, sensing techniques are described from the viewpoint of sensor technology.

Fiber-Optic Sensors for Cure Monitoring

To apply a traditional optical technique such as IRS, the exposed face of the FRPs must be visible in the molding setup. The use of an optical fiber is the best solution for *in situ* optical monitoring because an optical fiber has the advantages of small size and high temperature resistance, and can be embedded. In addition, an embedded fiber-optic sensor can be used for health monitoring in operation after the product is manufactured. This technique has been developed in the last decade due to the advances in fiber-optic sensors. Most of the traditional optical cure-monitoring techniques can be applied to *in situ* monitoring by using an optical fiber guide. As shown in Table 2, three types of traditional optical sensing methods, near-infrared spectroscopy (NIRS), fluorimetry, and refractive index methods, are commonly used for monitoring cure by using optical fibers today. Nontraditional fiber-optic sensors, which are *in situ* strain and temperature sensors, can be employed to monitor strain and temperature in PMCs during cure. The NIRS fiber-optic sensor and the fluorimetry fiber-optic sensor can monitor the chemical constitution of the resin and the curing agent. The refractive index fiber-optic sensor is the cheapest solution for monitoring cure. Strain/temperature fiber-optic sensors provide information about internal status, and they can be used after manufacture. In this subsection, these four types of fiber-optic sensors for monitoring cure are described.

The NIRS-based sensing is a technique whereby light transmitting through a sample is absorbed at some

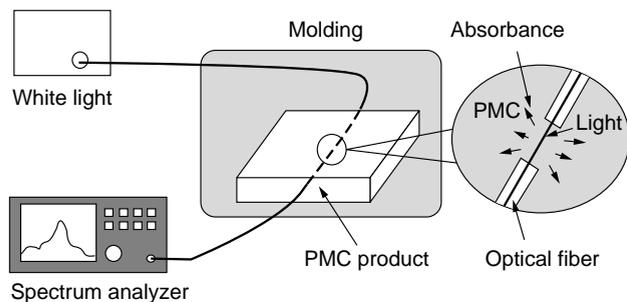


Figure 4. Schematic of the NIRS-based fiber-optic sensor system for monitoring cure.

wavelength in the range of the NIR due to the resonant vibration with the molecules of resin. In the process of curing polymers, some molecule groups appear or disappear, so the absorbance at the characteristic wavelength (or frequency) changes. Therefore, the NIRS-based sensor can measure the chemical constitution of the resin directly in the curing process. The NIRS-based fiber-optic sensor is an application of the technique that uses an optical fiber guide. The system has a white light source, a modified optical fiber, and a device for measuring an optical spectrum (Fig. 4). The wavelength used for NIRS-based cure monitoring is in the wavelength range of 1300–2300 nm (the range of NIR light). Note that optical fibers have peculiar transmission properties in the wavelength domain, as shown in Fig. 5 (3). Silica fibers are very cheap and have good transmission properties only in the range from visible light to NIR light. Sapphire and fluoride fibers have better transmission properties than silica fibers. Chalcogenide, AgCl, and AgBr fibers cover the mid-infrared region, but have lower transmission properties. Because optical fibers except silica fiber are much more expensive, silica fibers have been widely employed for monitoring cure. A spectral range of an FT-IR (Fourier Transform Infrared Spectroscopy) is in the near-infrared and that of a general grating-based optical spectrum analyzer is from visible to 1600 nm. Then, an FT-IR is preferred for NIRS-based sensing. Because the targeted wavelengths for monitoring cure depend on the kind of polymer, it is important

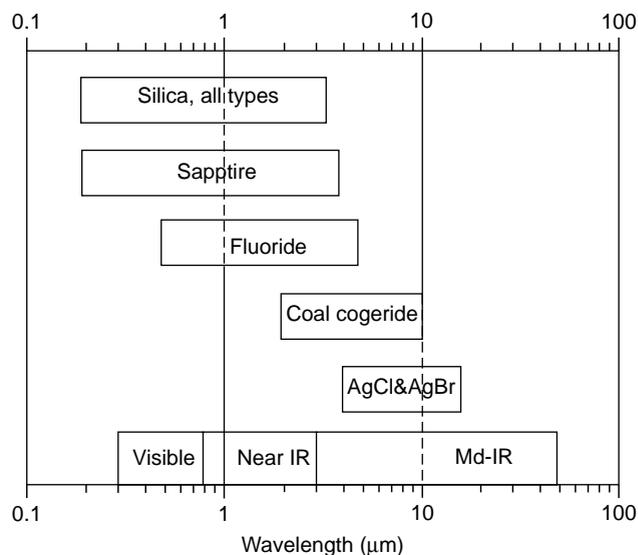


Figure 5. Transmission properties of optical fibers (3).

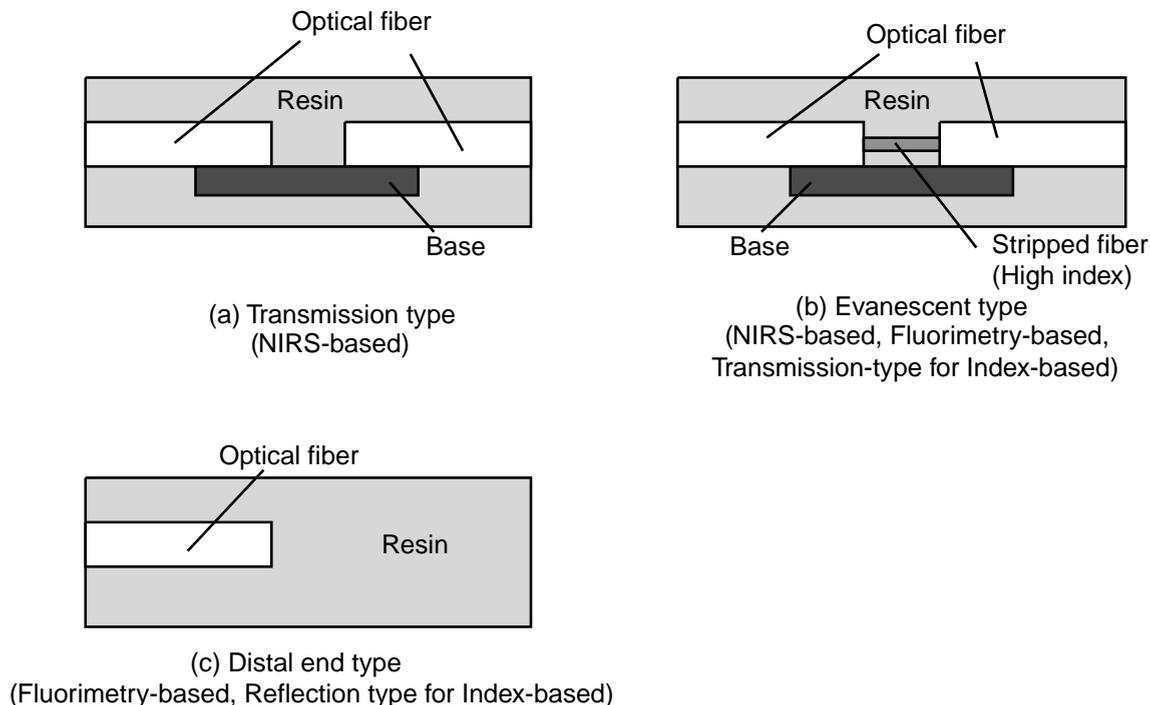


Figure 6. Constructions of sensing parts of NIRS-based, fluorimetry-based, and index-based fiber-optic sensors (a) transmission type; (b) evanescent type; (c) distal end type.

to determine the properties of the polymer's optical absorption to select a proper measuring instrument. For instance, monitoring the cure of the epoxy-amine resin system requires the 1500 to 1700-nm range, which includes the absorption bands of epoxy, amine, C-H, and O-H groups (4,5).

The sensing part of the optical fiber is fabricated so that the light propagates through the polymer. Two configurations of fiber-optic sensors are suggested (4–8). One is a transmission-type sensor, and the other is an evanescent-type sensor. The transmission-type sensor has a simple structure, in which the sensor has a gap, as shown in Fig. 6a. A configuration that uses a bore metal capillary is proposed to fix the input and output fibers (6). When the sensor placed in liquid polymer, the gap is filled by it. Then, light propagates through the polymer in the gap. The evanescent-type sensor consists of a fiber, which has a stripped cladding region, as shown in Fig. 6b. An evanescent wave is light transmitted in the cladding of the fiber. In the stripped region of the evanescent-type sensor, the evanescent wave transmits in the polymer instead of in the silica cladding of the fiber. The refractive index of the fiber core must be larger than that of the resin to propagate light in the stripped region (4). An example of the application of an NIRS-based sensor is shown in Fig. 7. The figure shows that the absorption peak of epoxy decreases due to a decrease in epoxy molecules from cross-linking in the epoxy-amine resin system in the curing process. Note that the behavior of absorption peaks is sometimes complex due to the overlaps of peaks related to different molecules. The use of neural network analysis has been proposed to improve the difficult quantitative analysis of spectra (9).

Fluorimetry is an optical spectroscopic technique that measures the molecular or atomic composition of a liquid, gas or solid by using ultraviolet (UV) light or X rays. This

technique is based on the photoluminescent phenomenon that incident light irradiates fluorescent materials. The fluorimetry-based fiber-optic sensor uses this phenomenon for monitoring the cure of the resin (7,8,10). When UV light is incident on a liquid resin mixed with a fluorescent curing agent, the curing agent absorbs the UV light and emits short-wavelength visible light (400–600 nm). The fluorimetry-based sensing system has UV light source and two wavelength-scanning filters for the excitatory light and the emission light, and a photo detector (Fig. 8). The emission spectra are scanned by fixing the excitatory wavelength at the absorption wavelength of the fluorescent material, and the excitation spectra have a fixed emission wavelength, which has maximum emission intensity. In the curing process, the peak position and the intensity of

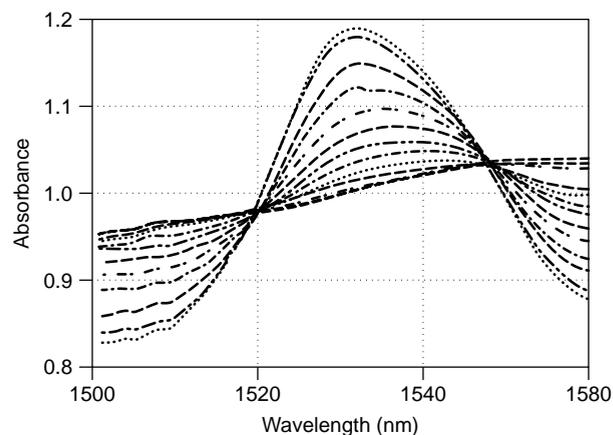


Figure 7. Overlaid optical fiber evanescent wave spectra obtained during the cure of Epikote 828 and hexanediamine at 40°C (6).

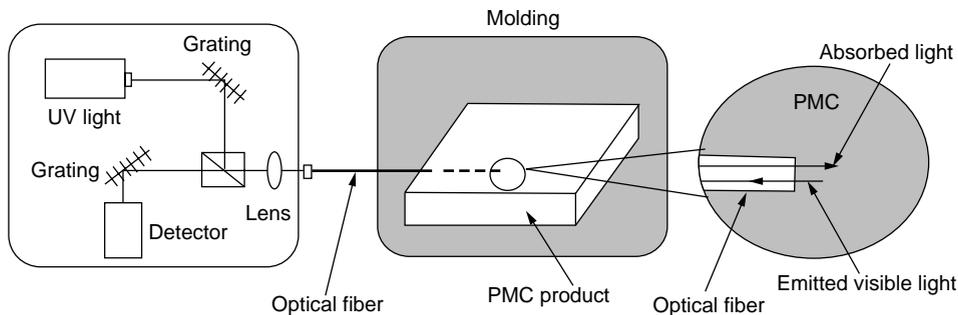


Figure 8. Schematic of fluorimetry-based fiber-optic sensor system for monitoring cure.

the spectrum are changed due to changes in the chemical structure of the fluorescent curing agent. The peak shifts of the spectra provide a quantitative measurement of the cure state. Fluorimetry-based sensors have an evanescent-type sensor and a distal end-type sensor as shown in Fig. 6 b,c (7,8,10). The construction of an evanescent-type sensor is similar to that of the NIRS-based sensor. A distal end-type sensor has a flat end where the light leaks out. An example of cure monitoring by using a fluorimetry-based fiber-optic sensor is shown in Fig. 9 (8). The figure shows the peak of excitatory spectra shifts in the curing process. The use of a sapphire optical fiber for the evanescent-type sensor has also been reported (7).

A refractive-index-based fiber-optic sensor measures changes in the refractive index of a polymer from the intensity of light. There are two types of construction for the sensor, a transmission-type sensor and a reflection-type sensor, as shown in Fig. 6b,c (5). The transmission-type sensor used in the index-based sensor is similar to the evanescent-type sensor used in the NIRS-based sensor and the fluorimetry-based sensor (4,5). The transmission-type sensor that uses a polymer core fiber has also been proposed since the late 1980s (11). A light propagates in the

fiber core by reflecting at the boundary between the fiber core and the resin in the stripped region. The reflection-type sensor uses Fresnel reflection at the cut end of the fiber, which contacts the polymer (5,12,13). The changes in the intensity of light result from changes in the reflection rate at the boundary between the fiber core and the resin. The reflection-type sensor requires a simple, low-cost optical system that uses a silica fiber for communication, so the cost is much lower than that of spectroscopic monitoring methods. However, note that the long-term stability of optical devices that include a light source and a detector is essential for stable and low S/N measurement during cure. Figure 10 shows the experimental measurements of the two types of refractive-index-based sensors during cure (6). The figure shows that the curve of the reflection-type sensor is inversely proportional to that of the transmission-type sensor.

Because most fiber-optic strain sensors are sensitive to temperature, they can also be used for measuring temperature. Several kinds of fiber-optic strain/temperature sensors are discussed later. An extrinsic Fabry–Perot interferometer (EFPI) sensor, a fiber Bragg grating (FBG) sensor, and an interferometric sensor are commonly used for *in situ* monitoring. These sensors were developed originally for health monitoring, and therefore, they can be used after the manufacture of products. An EFPI sensor is constructed from two optical fibers that are fixed in a capillary tube and have half-mirrors at the ends of the fibers (Fig. 11). The two mirrors comprise a multiple ray interferometer in the capillary tube, which is called a Fabry–Perot interferometer. There are two measurement systems for EFPI sensors. One uses a narrowband light source, and the other uses a broadband light source. The former is cheaper and is used for high-speed measurement but is strongly affected by the optical power loss in the fiber-optic guide. The loss is a problem for cure monitoring because high pressure is applied to PMCs in the manufacturing process. The latter is independent of the optical power loss due to the capability for absolute measurement of the cavity length in the wavelength domain (14). Therefore, the latter system is more suited to monitoring in the manufacturing process. Most of the commercial EFPI strain sensors have low thermal sensitivity because the gauge length is about 20 times as long as the cavity length. Then, the thermal effect on EFPI strain sensors is sometimes negligible for strain measurement. There are several applications for monitoring strain or temperature in the curing process. The residual strains in a pultruded composite rod in the

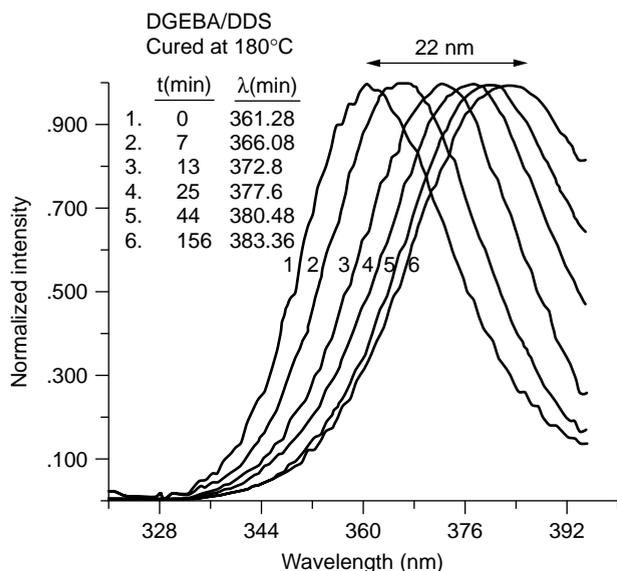


Figure 9. Excitatory spectra of DGEBA-DDS epoxy obtained *in situ* at 180°C as a function of cure time (spectra plotted without regard to intensity) (8).

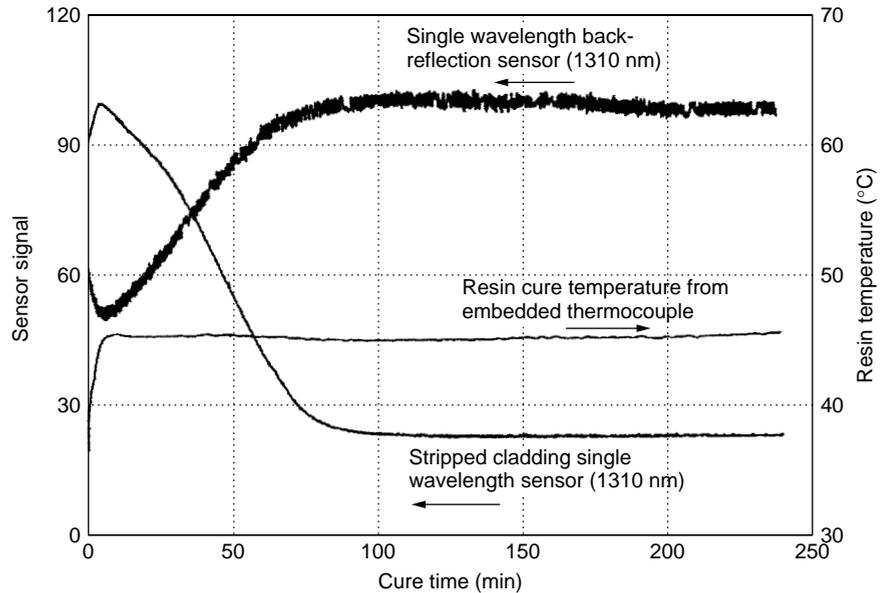


Figure 10. Cure data obtained from single-wavelength back-reflection (reflection-type) and stripped cladding (transmission-type) optical fiber sensors during the cure of Epikote 828 and hexanediamine at 45°C (6).

pultrusion molding process were evaluated in (14). It appeared that the strain measured in FRPs by using EFPI sensors could be used for cure monitoring in an autoclave molding and an FW molding (15,16). The thermal sensitivity of the sensor for temperature measurement can be maximized by bonding the capillary tube to a high CTE (coefficient of thermal expansion) material such as aluminum (17). An FBG sensor has a longitudinal periodic variation in its refractive index in the core of a single-mode fiber (Fig. 12a). The wavelength shift of the reflected light from the Bragg grating is proportional to the strain variation. This absolute measurement technique is affected by strain and temperature change. The effect of temperature on strain measurement by an FBG sensor cannot be negligible during cure at high temperature. It was reported that FBG sensors embedded in CFRP and GFRP composites can detect the onset of vitrification of the resin during cure (18). An FBG sensor for temperature measurement can be manufactured, so that it is sensitive only to temperature, by making a sensing part free from strains, as shown in Fig. 12b (19,20). Simultaneous measurement of temperature and strain by FBG sensors are of major interest, and the studies are described in the section on health monitoring.

Dielectric Sensors for Cure Monitoring

Most polymers are nonconductive but have a little conductivity in the liquid state. Therefore, the electric properties of polymers provide useful information about the cure state. Dielectric measurement techniques for polymers have been investigated since the 1960s. The application to monitoring cure started in the 1980s, and micro-dielectric sensors have been developed especially for *in situ* cure monitoring. This measurement technique is based on the method for measuring the complex dielectric constant of conductive materials. The real part ϵ' and the imaginary part ϵ'' are called relative permittivity and loss factor, respectively. The basic components of dielectric sensing are a voltage source and two electrodes. A micro dielectric sensor has an electrode pattern printed on a small, thin base plate, as shown in Fig. 13 (21). When the sensor is covered by resin, it can be assumed that the sensor and the resin comprise an equivalent RC electric circuit. Consequently, when a sinusoidal voltage is applied to the circuit, the sinusoidal current generates with a lag of phase angle δ . Then, the resin capacitance C , the resin resistance R , and $\tan \delta$ can be obtained simply from the current output. The complex dielectric constant is represented by the

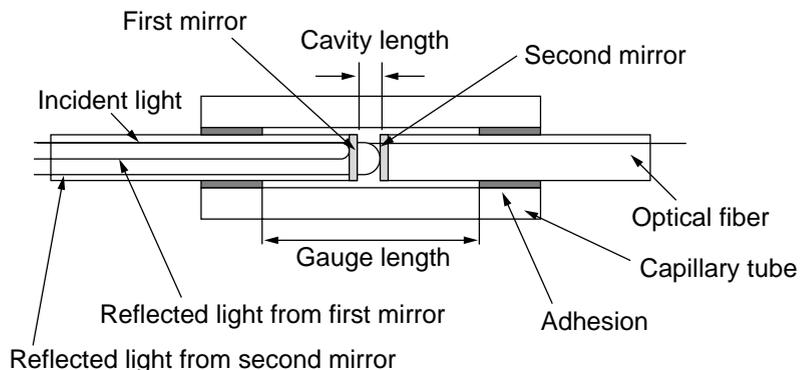


Figure 11. Schematic of an EFPI fiber-optic sensor.

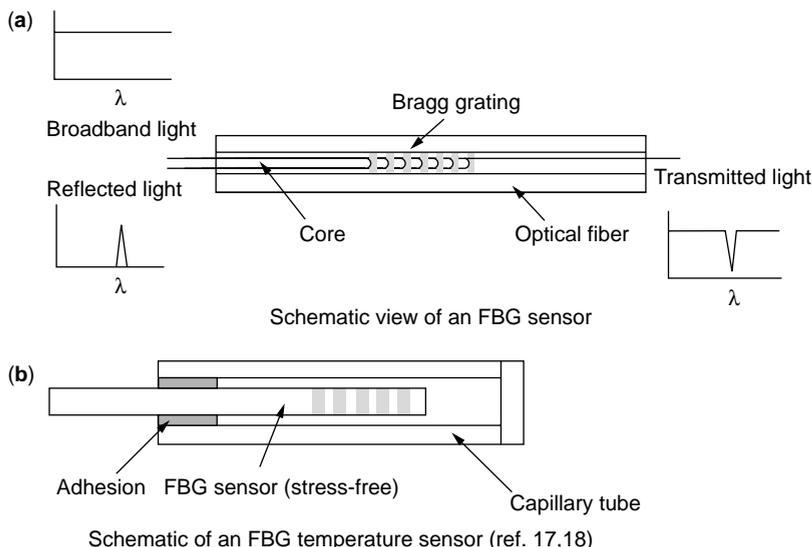


Figure 12. Schematic of an FBG fiber-optic sensor (19,20).

following simple form: $\epsilon' = C/C_0$, and $\epsilon'' = 1/R\omega C_0$, where C_0 is the capacitance of a free space capacitor and ω is the angular frequency of the voltage source. The previous relationship indicates that the loss factor depends on the frequency. The loss factor consists of both a dipole orientation and a free charge migration. Hence, the loss factor is expressed as a linear combination of the contribution of dipole polarization $(\epsilon_r - \epsilon_u)(\epsilon\tau)/(1 + \omega^2\tau^2)$ and the contribution of free charge migration $\sigma/\omega\epsilon_0$. Here, ϵ_r is the relaxed permittivity, ϵ_u is the unrelaxed permittivity, ϵ_0 is the permittivity in vacuum, τ is the relaxation time, and σ is the ionic conductivity defined as $\sigma = \epsilon_0 G/C_0$. The contribution of dipole polarization is negligible when $\omega\tau \ll 1$ at low frequency which is generally less than 1 kHz (22), and then the ionic conductivity can be calculated from the equation $\sigma = \omega\epsilon_0\epsilon''$. The ionic conductivity is convenient for estimating the cure state because it has a strong relationship to the mobility of ions in polymers. The resistance $1/\sigma$ is called the ion viscosity, and the logarithmic value is used also for the estimate. The behavior of the ion viscosity is similar to that of the viscosity before the gel point. Figure 14 shows that the behavior of the log ion viscosity of a graphite/epoxy composite is qualitatively similar

to that of the mechanical viscosity up to the gel point, but the difference increases after the point (23). A comparison of the DOC data from DSC and the dielectric measurement of an epoxy resin is shown in Fig. 15. It is evident that the DOC from the dielectric measurement does not have a linear relationship to that obtained by DSC measurement. The dielectric measurement of polymers is described in detail in the paper by Mijovic et al. (23).

Several new systems, new sensors, and new applications have been proposed in recent years for *in situ* cure monitoring by dielectric sensors. A comparative study of three types of commercial dielectric sensors was conducted (24). It was demonstrated that the dielectric sensors used for monitoring the cure of a polymer coating can monitor the degradation of performance properties during use in acid, at high temperature, and in water (25). This implies the feasibility of using embedded dielectric sensors in both cure and use. The dielectric parameters were measured at a high-frequency range (kHz–MHz) to monitor dipole rotational mobility (25,26). The new parameter was introduced to estimate the DOC from the measured dielectric parameters; the experimental data agreed well with simulation from using an analytical model and

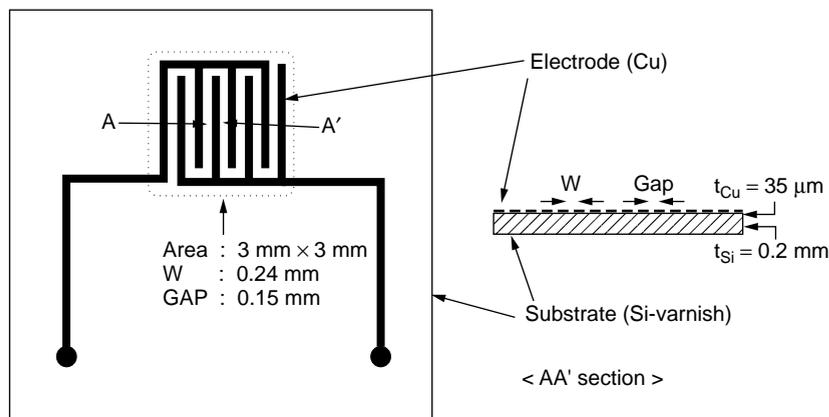


Figure 13. Schematic of a micro dielectric interdigital sensor (21).

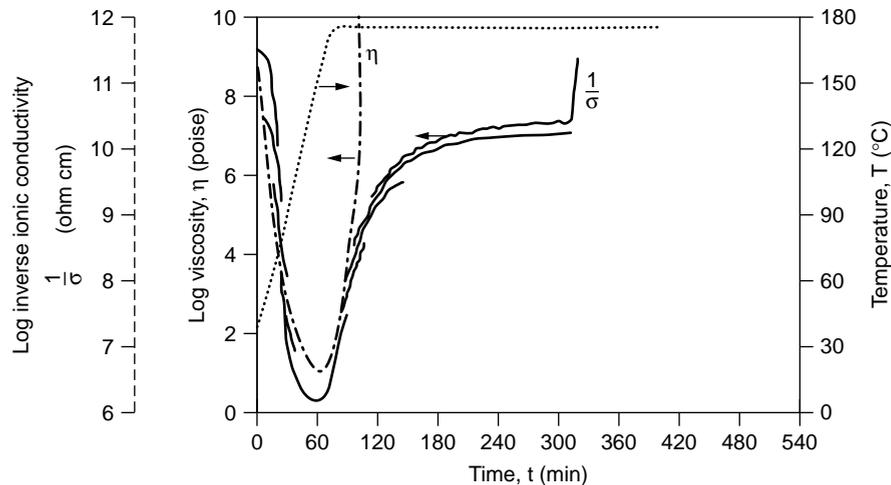


Figure 14. Measured resistivity and viscosity as a function of time during the cure of a graphite/epoxy composite (23).

DSC data from the various temperature profiles (21). The Dielectric sensing technique was applied to process monitoring in the SMC/BMC industry and involved cure monitoring and quality assurance/quality control (27). As for the impregnation process in liquid molding, it was shown that the dielectric sensors can be applied to monitoring the impregnation in resin infusion molding (28) and in RTM molding (22,29). The prediction method for the DOC using finite-element analysis from the results of dielectric measurement was also studied (30). Control of a curing process that had a dielectric sensing system was tried by using artificial intelligence (31).

Piezoelectric Sensors for Cure Monitoring

Piezoelectric ceramics wafers have been employed as sensors/actuators for monitoring and controlling structural vibration. Cure monitoring using a piezoelectric wafer actuator/sensor started in 1997. This cure monitoring technique uses the phenomenon that the piezoelectric wafer becomes

constrained by resin in the solidification during cure. Two types of measurement concepts were proposed. One is the measurement of viscosity using a PZT (lead zirconate titanate) laminate that sandwiches two PZT thin films in three insulating tapes (32). Another is the impedance measurement of an equivalent electromechanical circuit composed of a piezoelectric wafer and resin (33). The former has individual PZT sensor and PZT actuator parts, whereas the latter uses a piezoelectric wafer as both sensor and actuator. The former PZT sensor was applied to monitoring the cure of GFRP laminates in autoclave molding (32). The experimental results show that the output curve of the PZT sensor reflects the viscosity qualitatively and that gelation can be monitored.

For impedance measurement, the system composed of a piezoelectric wafer and resin can be modeled by a series of mass–spring–damper systems that comprise equivalent electric circuits (Fig. 16). In the process of curing the resin, changes in the shear modulus (spring) and viscosity (damper) affect the electric response of the piezoelectric wafer. The measurement of electric response in the resonant frequency region is carried out to monitor changes in electric admittance at the resonant frequency and the antiresonant frequency. The increase in the modulus and viscosity of the resin reduces the amplitude of the transfer function, which is the peak-to-peak value. An example of transfer functions of an epoxy resin measured at different curing times is shown in Fig. 17 (34). The temperature influences the capacitance of piezoelectric wafers and consequently, the magnitude of the transfer function. However, the peak-to-peak amplitude of the transfer function is more sensitive to changes in the mechanical properties of

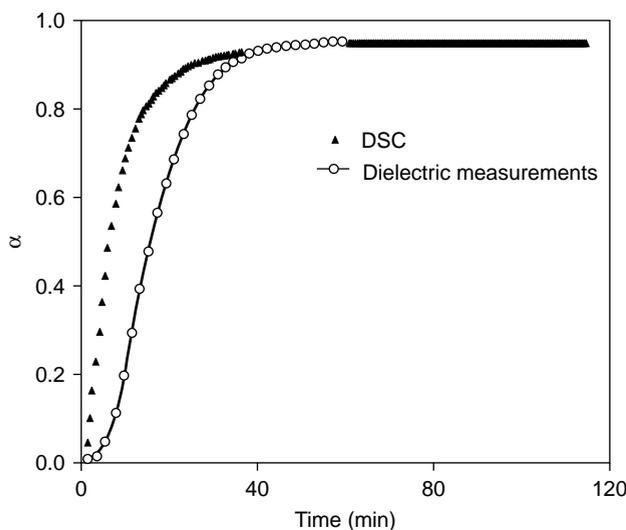


Figure 15. A comparison of the degree of cure from DSC and from dielectric measurements (normalized log resistivity) as a function of time during the cure of an epoxy resin at 200°C (23).

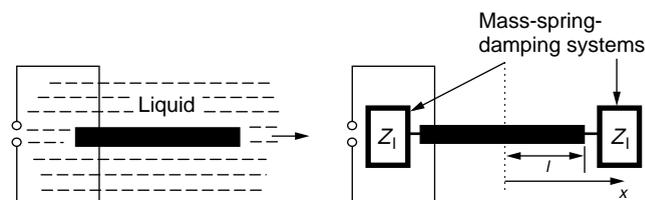


Figure 16. A simplified model of a piezoelectric wafer in a viscous liquid (34).

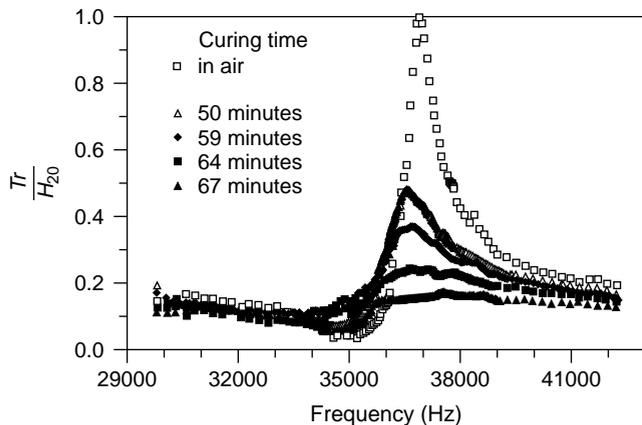


Figure 17. Transfer functions of a piezoelectric wafer embedded in epoxy taken at different curing times (34).

a liquid-state resin (35). Therefore, the measurement before the gelation of the resin is available. It is found that the resonance peak amplitudes of the transfer function of a piezoelectric wafer have a good relationship to the viscosity of the resin before gelation, whereas the resonant response is suppressed after gelation of the resin. Therefore, this sensor can be used only as an internal temperature sensor after gelation. This technique has the advantage that embedded piezoelectric ceramics can be used in operation as well as in cure. Because the peak-to-peak amplitude of the transfer function changes with respect to the contact area with liquid, it can be used for controlling the impregnation process in liquid molding such as RTM (35).

Ultrasonic Measurement for Cure Monitoring

The monitoring technique using an ultrasonic wave propagating in a material is a traditional nondestructive technique for measuring modulus, density, and viscosity. This technique is also widely used for nondestructive testing of products in inspection. The ultrasonic monitoring technique has been applied to *in situ* cure monitoring of polymers since the late 1980s. This cure monitoring technique is based on measuring the velocity and attenuation of an ultrasonic wave propagating in a viscoelastic and anisotropic material (36). Elastic wave propagation is affected by changes in the modulus, density, and viscosity of a resin in the curing process. In most cases, the size of the reinforcement of a composite is smaller than the wavelength of propagating elastic waves, so that the composite can be treated as homogeneous materials. There are two methods for generating ultrasonic waves in composites during the molding process. One locates ultrasonic transmitters and receivers in or on the mold, and therefore, this configuration has the advantage that internal sensors are not needed (37). Another method uses an acoustic waveguide that propagates an ultrasonic elastic wave (38). The wave velocity, the attenuation, and the reflection factor can be used to estimate the DOC. Sound velocity increases as the elastic modulus of a resin increases from liquid to solid in the curing process, whereas the attenuation decreases by the viscoelastic relaxation and the scattering factor. Sound velocity is convenient for evaluating the DOC because the influence of molding pressure on sound velocity is small.

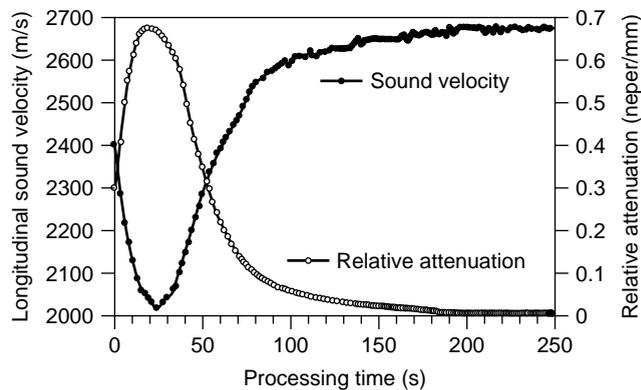


Figure 18. Longitudinal sound velocity and relative attenuation as a function of processing time of a phenolic-formaldehyde molding compound PF31 (37).

Figure 18 shows an example of measurements of longitudinal sound velocity and relative attenuation in the cure of a thermoset resin (37).

HEALTH MONITORING

Like the human body, structures deteriorate or are damaged in long-term use. The damages are generated by the initial defects, overload, and impacts. Structural performance such as modulus, strength, and damping is degraded by moisture, acid, and high temperature. The damage and the deterioration of structures are significant problems because they often cause catastrophic accidents. However, unlike the human body, the health of structures cannot be recovered automatically. Therefore, periodic inspection is essential to ensure the safe operation of structures. The most common inspection method is visual inspection by human eyes. It involves specimen inspection by microscopy and easier visible inspection techniques such as inspection by using a fluorescent dyestuff. However, damages generated in opaque materials cannot be found by visual inspection techniques. In addition, undetected small damages trigger accidents due to the rapid development of damages that result from their interaction. Nondestructive evaluation (NDE) techniques have been developed to detect internal or invisible damage. Traditional NDE techniques are ultrasonic scan, an eddy current method, X radiography, an acoustic emission method, and passive thermography. These NDE techniques are effective in detecting damages in materials and structures, but it is difficult to use them in operation due to the size and weight of the devices. This means that traditional NDE techniques require field operators and transporters for heavy, large testing machines. Then, the operation must be interrupted during traditional NDE testing. Because these facts increase operating cost, speedy and simple inspection techniques are desired.

Health monitoring is an attractive approach to solving the problems that occur in aged and degraded structures. The damage and performance degradation are checked for maintaining the health of materials and structures. The mechanical, thermal, and chemical states in and around structures provide useful information for predicting the

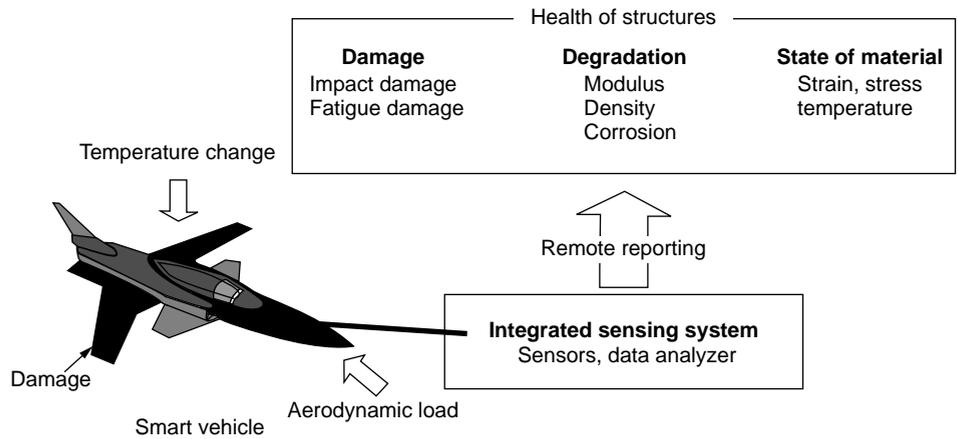


Figure 19. Concept of a smart vehicle that has a health monitoring system.

service life. These values are remotely monitored by a health monitoring automated system in real time. The need for health monitoring has been growing in the fields of aircraft, space structures, and civil structures since the 1980s. The accidents and the growth of maintenance costs of aging structures motivate the need for research into health monitoring systems. The structures located in space or in the deep sea especially require real-time and remote monitoring systems to improving safety and reliability because on-site maintenance sometimes costs more than manufacturing and installing new ones.

Here, we emphasize that the research area of health monitoring in smart materials and structures partially overlaps that of NDE. However, unlike NDE, a health monitoring system is naturally integrated into materials and structures by using small sensors and a powerful data analyzer. Remote monitoring is sometimes essential for practical applications. Figure 19 shows a schematic view of a vehicle that has a health monitoring system. The development of the health monitoring technique has been accelerated by advances in sensor technologies. Advanced computer technology is so powerful for analyzing monitored data in real-time and so small that it can be in a structure. The rapid development of the computer network, "Internet," enables remote monitoring on the www (World Wide Web) using software written in a network-friendly language like JAVA. These advanced technologies comprise an automated health monitoring system that can perform a self-inspection, a self-assurance of safety, and a self-report for the future. Nondestructive damage

detection techniques are employed for self-inspection. Safety assurance can be achieved by monitoring whether the measured values such as strain, load, or temperature go over the safety limit.

In recent years, many sensors and sensing techniques have been developed for health monitoring. Representative sensing techniques are shown in Table 3. They are fiber-optic sensors, piezoelectric sensors, a magnetostrictive tagging technique, and an electrical resistance technique. Fiber-optic sensors and piezoelectric sensors are so small that they are embedded in materials. Fiber-optic sensors are most suited for internal measurement by embedded sensors due to their size, weight, high flexibility and long-term durability. The magnetostrictive tagging technique and the electrical resistance technique do not need any embedded sensors for *in situ* monitoring because the material itself acts as a sensor. These four types of sensing techniques are available for detecting internal damages. Some types of fiber-optic sensors can detect internal damage directly without computational identification. To detect damage, piezoelectric sensors use diagnostic signals, which are generated by impact or actuators. The changes in magnetic and electrical properties of conductive materials such as carbon-reinforced composites reflect the presence and progress of damage. Note that detectable damage modes depend on the kind of sensors, sensing methods, and integrating configurations. Therefore, to select sensors and a sensing technique, it is important to understand the behavior of damage initiation and growth in materials and structures. Piezoelectric sensors can be used for

Table 3. *In Situ* Sensing Techniques for Health Monitoring

	Configuration	Monitored Value	Sensing Area	Cost	Sensor Networking
Fiber-optic sensors	Embed, Attach	Break, strain, vibration, temperature	Around fiber	High	Normal (depends on method)
Piezoelectric sensors	Embed, Attach	Dynamic strain, impedance	Middle-large	Middle	Easy
Magnetostrictive tagging technique	No sensor (tag)	Damage, static strain	Large	N/A	N/A
Electric resistance technique	No sensor (electrode)	Damage, static strain	Large	N/A	N/A

Table 4. Requirement and Purpose of Health Monitoring System in Engineering Fields

	Requirement	Purpose
Aircraft	Light weight, reliability	To maintain safe operation To reduce maintenance cost
Space structure	Light weight, reliability, insensitivity to electromagnetic field, temperature resistance, radiation resistance	To maintain performance
Civil structure	Long-term durability, chemical resistance, moisture resistance	To reduce maintenance cost

dynamic strain measurement, and magnetostrictive tagging and electrical resistance techniques for static strain measurement. Fiber-optic sensors can be used to measure both static and dynamic strain. Internal temperature can be measured by using fiber-optic sensors. Here, the sensing area of the sensing techniques should be considered. Magnetostrictive tagging and electrical resistance techniques provide large sensing areas. Size of popular piezoelectric sensors for health monitoring is several centimeters. Fiber-optic sensors have various gauge lengths according to the kind of sensor, but the sensing area is limited to the neighborhood of the optical fiber.

Applied studies of health monitoring techniques concentrate on aircraft, space structures, and civil structures. These fields have individual purposes for health monitoring systems, as shown in Table 4. The increase of aging aircraft motivates the development of health monitoring systems for aircraft to maintain safety and provide quick, low-cost maintenance. In the field of civil engineering, a health monitoring system is expected to reduce maintenance cost, which grows as large civil structures increase. The health monitoring of spacecraft is an approach to compensate for performance when the craft is damaged. As shown in Table 4, the requirements of the sensing techniques are different for each of the applied fields. Aircraft and space structures require lightweight sensors and measurement systems because of the additional weight introducing by the sensing system, which increases operating cost. A health monitoring system for aircraft and space

structures must be reliable because these engineering fields are conservative. The sensing techniques used for space structures must be insensitive to electromagnetic fields, temperature, and radiation. For civil structures, long-term durability, chemical resistance, and moisture resistance are required because of the long lifetimes of the structures.

In this section on health monitoring, four types of sensing techniques are described from the viewpoint of sensor technology. In additions, the application of health monitoring techniques to aircraft, space structures, and civil structures are also described.

Fiber-Optic Sensors for Health Monitoring

Fiber-optic sensors are the sensors most promising for monitoring the internal state of materials. Early studies of health monitoring of composites by using optical fibers can be seen in papers published in the 1980s (39–42). The simple sensing method in these studies was based on an optical power loss by a break in an optical fiber. The quantitative monitoring of internal strain and temperature using embedded fiber-optic sensors started in the early 1990s. There are many kinds of fiber-optic sensors for *in situ* health monitoring, including intensity-based sensors, interferometric sensors, polarimetric sensors, EFPI sensors, FBG sensors, long-period grating based (LPG) sensors, Raman scattering sensors, and Brillouin scattering sensors, as shown in Table 5. The sensors, except for Raman scattering

Table 5. Fiber-Optic Sensors for *In Situ* Health Monitoring

	Monitored Value	Multiplexing/ Distributing	Gauge Length	Sensor Cost	System Cost
Intensity-based	Break, microbend, strain, vibration	OTDR	Short/Long	Cheap	Cheap ^a
Interferometric	Strain, temperature, vibration	Switching	Long	Cheap	Middle
Polarimetric	Strain, temperature, vibration	Switching	Long	Cheap	Middle
EFPI	Strain, temperature, vibration	Switching / Frequency domain	Short	High	Middle-High
FBG, LPG	Strain, temperature, chemical property	Frequency domain (Easy multiplexing)	Short	High	High
Raman scattering	Temperature	OTDR (ROTDR)	Variable	Cheap	High ^b
Brillouin scattering	Strain, temperature	OTDR (BOTDR)	Variable	Cheap	High ^b

^aNot including OTDR.

^bIncludes OTDR.

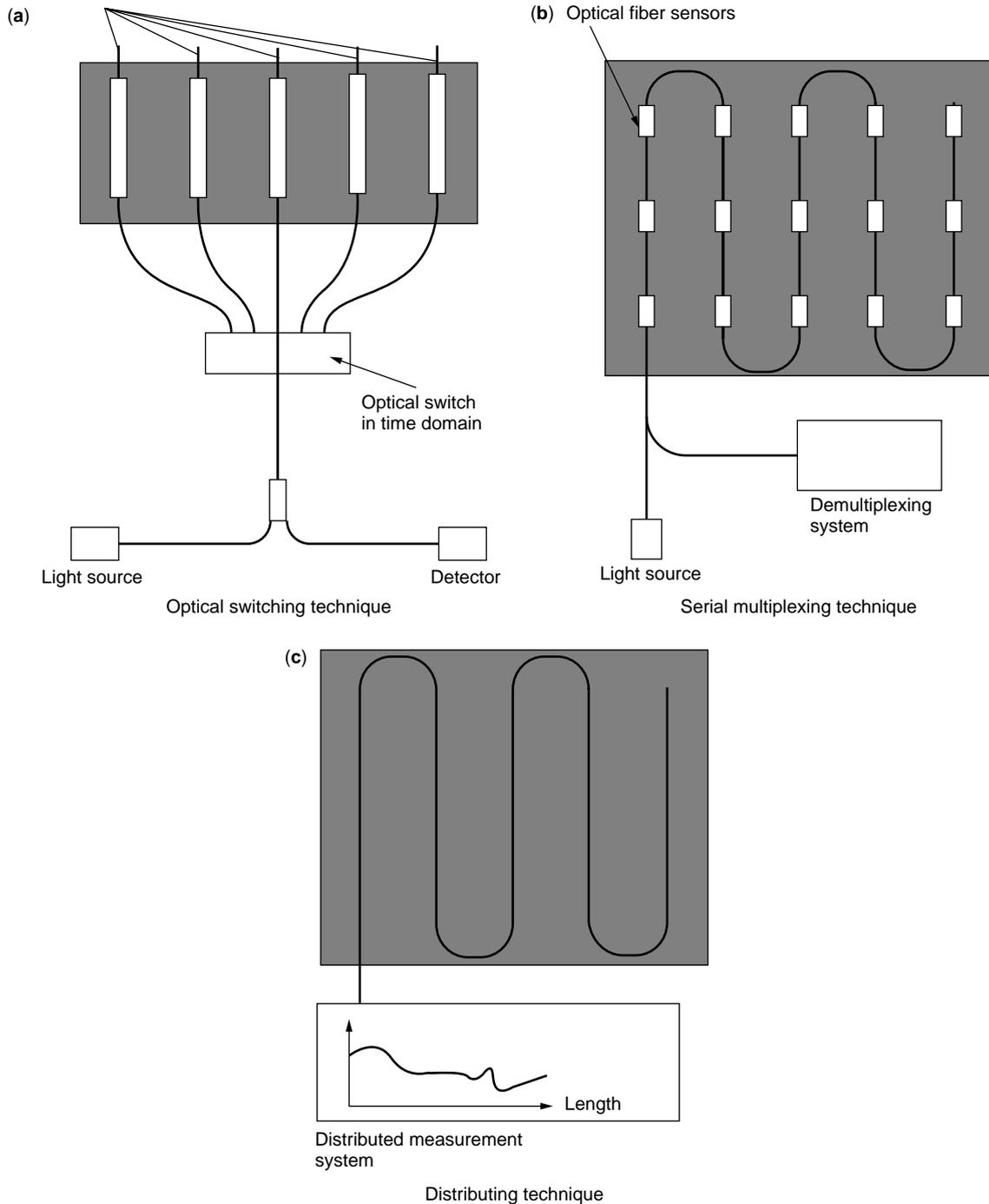


Figure 20. Configurations of distributing and multiplexing techniques.

sensors, can measure the static strain. It is difficult to apply sensing systems in the frequency domain such as absolute EFPI sensors, FBG sensors, LPG sensors, and Brillouin scattering sensors to measuring high-speed vibration. Most of the sensors can also measure temperature because the reflective index is sensitive to temperature. The distributing or multiplexing techniques for fiber-optic sensors are key techniques in making a health monitoring system practical. Three configurations, optical switching (parallel multiplexing), serial multiplexing, and

distributing, are available, as shown in Fig. 20. The optical switching system is the common method of measurement that uses multiple fiber-optic sensors, but the system is not ideal due to the low switching speed, (Fig. 20a). The serial multiplexing technique is ideal for short-gauge sensors such as intensity-based strain sensors, EFPI sensors, and FBG sensors (Fig. 20b). The total weight and cost of a serial multiplexed fiber-optic sensor system can be reduced compared to that of a system using optical switching devices due to the simple configuration and the short length

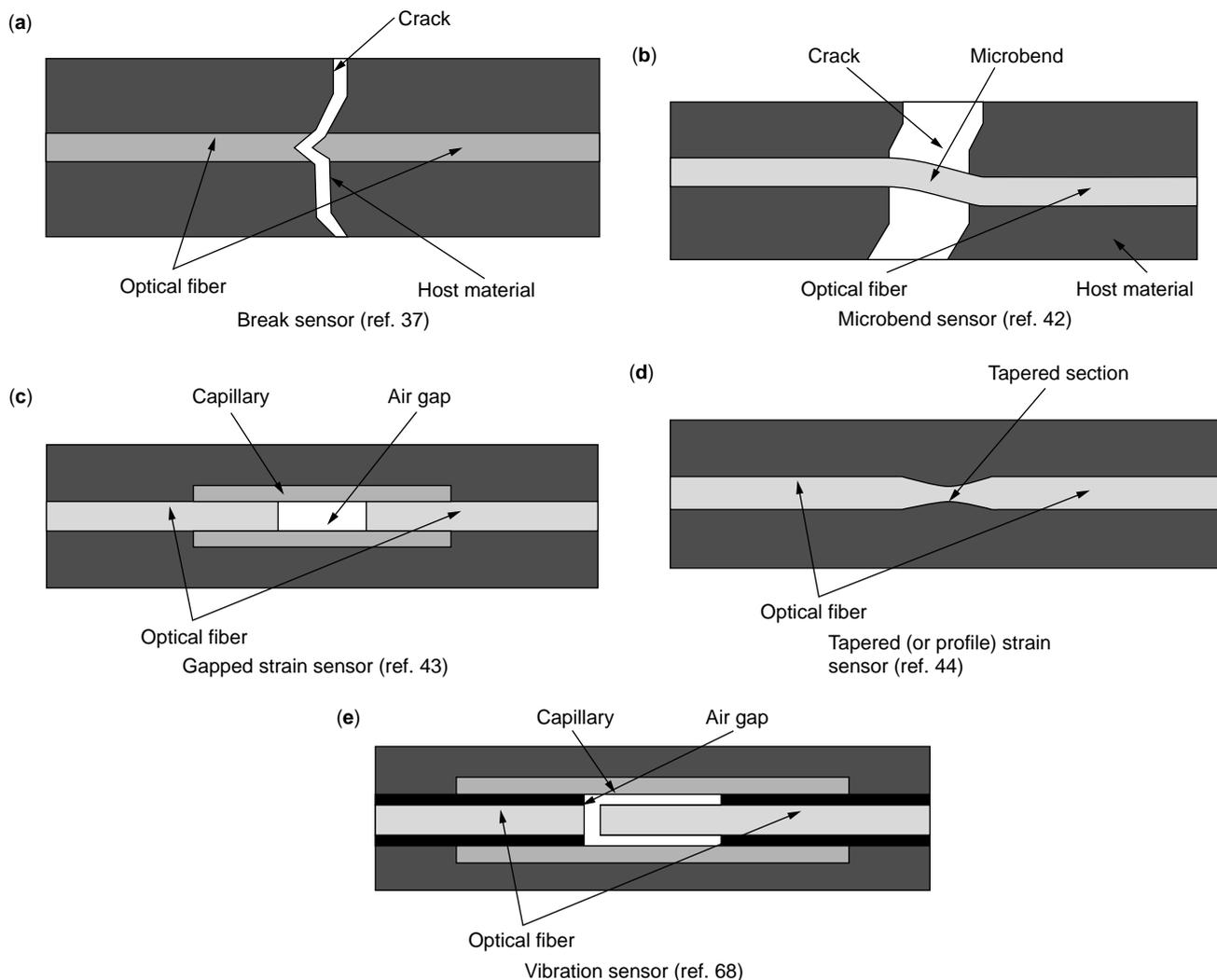


Figure 21. Configurations of intensity-based fiber-optic sensors.

of the optical fibers. EFPI sensors that measure the change in the cavity length in the frequency domain by using a broadband light source can be multiplexed in a single optical fiber (43). FBG sensors can also be easily multiplexed in a single fiber due to the frequency domain measurement. The number of FBG sensors in a single fiber is limited by the strain range and the dynamic range of the wavelength-scanning device. Intensity-based sensors can be multiplexed by an optical time domain reflectometer (OTDR). OTDR is a popular distribution sensing technique along a single fiber, as shown in Fig. 20c; it consists of a short-pulse laser and a high-speed detector. It scans the location of the reflection through a single fiber in the time domain. Because the OTDR is a reflectometer, any kind of reflected or back-scattered light can be detected. However, note that the resolution of OTDR depends on the pulse width that is from several hundred millimeters to several meters. Raman scattering sensors and Brillouin scattering sensors are generally used with an OTDR. Interferometric fiber-optic sensors and polarimetric fiber-optic sensors can be multiplexed by optical switching devices. Brief explanations of these sensors and sensing systems follow.

Intensity-based sensors are based on measuring the optical intensity of reflected or transmitted light. The sensing system is very simple and cheap due to the low cost of the fibers and devices; however, it becomes unstable when the guide section of the optical fiber is subject to an excessive bend. It consists of a cheap laser-emitting diode (LED) source, a cheap photodetector (PD), and a silica or plastic optical fiber used in communication. Intensity-based fiber-optic sensors have several configurations for measuring values, as shown in Fig. 21. Intensity-based break sensors and microbend sensors have no special sensing section, they are based on measuring optical power loss (39–42). The break sensor can detect damages that cut the optical fiber (Fig. 21a). They were investigated for detecting cracks in composites early in the development of health monitoring techniques that used optical fibers. Figure 22 shows that the transmission power rapidly drops when the optical fiber is fractured by impact (42). The local deformation of the optical fiber can be monitored by microbend sensors (Fig. 21b). Figure 23 shows that the optical power loss increases when the fiber is subject to a microbend by the crack opening (44). Intensity-based strain sensors have

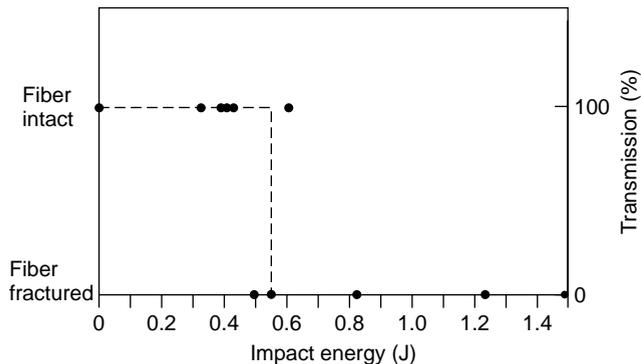


Figure 22. Transmission drop of optical fibers embedded in Kevlar/epoxy following impacts of various energies (42).

two configurations; gapped-type (Fig. 21c) and tapered-type (Fig. 21d). The gapped-type sensor is based on the optical power loss in air using broadband light sources (45).

The tapered-type sensor has a tapered-shape in the sensing part, from which some light leaks (46,47). Figure 24 shows that the output of the tapered-type sensor has a linear relationship to the strains (46). Interferometric fiber-optic sensors and polarimetric fiber-optic sensors are fundamental sensing methods that use optical fibers. In most cases, these sensors are not suitable for distributed measurement of internal values due to the long gauge length and the difficulty of multiplexing. Some applications that use interferometric and polarimetric fiber-optic sensors have been reported (48–50). The sensing methods that use EFPI and FBG sensors are described in the previous section on cure monitoring. The EFPI sensor is one of the interferometric fiber-optic sensors, but the most important difference is that it has a short sensing section in a single fiber. The FBG sensor is most promising for *in situ* strain and temperature measurement due to its strength, flexibility, and easy multiplexing. A number of demodulation systems for multiplexed FBG sensor systems are shown in the review paper by Rao (51). There are many applications of FBG sensors to concrete structures (52,53) and composite structures (54–56). Like an FBG sensor, a LPG sensor

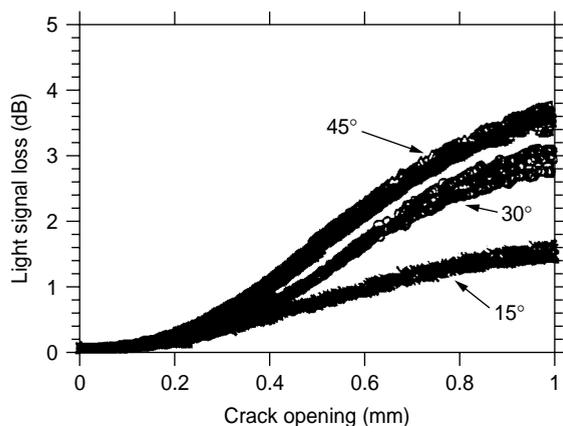


Figure 23. Signal loss vs. crack opening obtained from the crack simulator specimen (44).

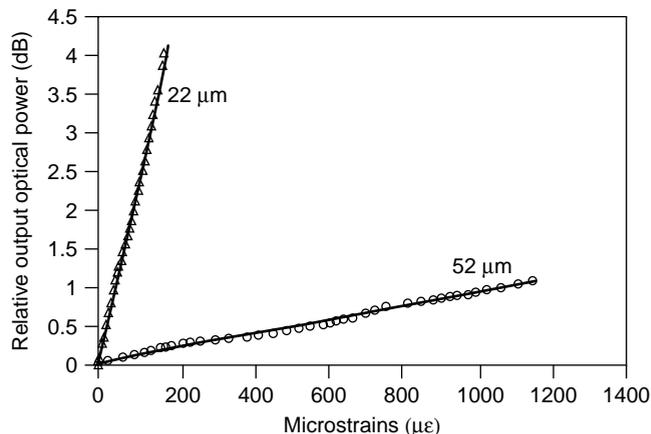


Figure 24. Relative output optical power vs. microstrains measured by resistive gauge for two different tapers of 52 and 22 μm waist diameters (46).

has a grating in the fiber core, but its period is much longer than that of an FBG sensor. As a forward-propagating mode in core is combined with several cladding modes in a LPG, the several valleys can be measured in the spectrum of transmitted broadband light, as shown in Fig. 25 (52). Because the wavelength shifts of the valleys have respective sensitivities to strain and temperature, the strain and temperature can be separated by selecting proper valleys (57). The LPG sensor can be used as a chemical sensor for corrosion monitoring because it is sensitive to changes in the refractive index of the fiber cladding (58). A Raman scattering OTDR (ROTDR) can measure temperature distribution from the frequency peak shift of the Raman back-scattered light (59). Like a ROTDR, a Brillouin scattering OTDR (BOTDR) uses the frequency peak shift of the back-scattered light, whereas the BOTDR can measure both

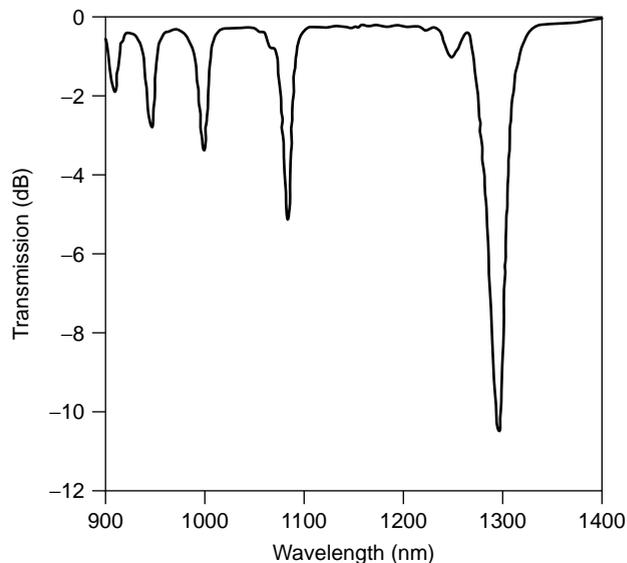


Figure 25. Transmission spectrum of a LPG whose period is 198 μm (52).

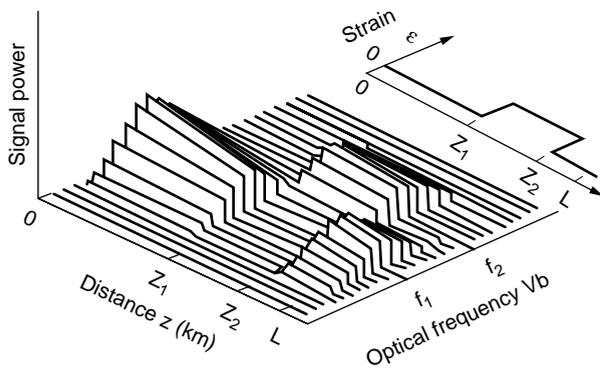


Figure 26. BOTDR waveforms along a fiber (60).

strain and temperature (60,61). Figure 26 shows waveforms measured by a BOTDR along a fiber (60). Strain was measured using a BOTDR that had a spatial resolution of 400 mm and a strain precision of 50 microstrains (61).

As described in the section on cure monitoring, fiber-optic strain sensors are sensitive to temperature. Therefore, temperature compensation is necessary to retain the precision of strain measurements if they are used in an environment that has large temperature variation such as space. The strain measured by an EFPI sensor is stable in a normal environment of small temperature variation due to its low temperature sensitivity. The LPG sensor can simultaneously sense strain and temperature (57). For other types of sensors, the simultaneous sensing technique is required if they are used in normal environments. The FBG sensor and the BOTDR sensor are sensitive to both strain and temperature like the EFPI sensor, but their temperature sensitivities are much larger than that of the EFPI sensor. Therefore, the simultaneous measuring strain and temperature is a major topic for the FBG sensor and the BOTDR sensor. One idea for simultaneous measurement uses several wavelengths reflected from a single sensor. Dual FBGs in a polarization maintaining (PM) fiber can simultaneously measure three-axis strains and temperature from four Bragg wavelengths (62). The FBG sensor using the first and the second diffractions can also be used for simultaneous measuring of strain and temperature from two Bragg wavelengths (63). The simultaneous measurement of strain and temperature by a BOTDR using a PM fiber has been demonstrated (64). Another idea is based on a combination of more than two types of strain/temperature sensors. The strain and temperature in composites were simultaneously measured by a combination of an EFPI sensor and an intrinsic rare-earth doped fiber (65). The FBG/EFPI combined sensor has been proposed for simultaneously monitoring strain and temperature (56). The idea of a combination sensor can be applied to the development of a multifunctional fiber-optic sensor. A multi-functional sensor that combines an FBG temperature sensor and an NIRS-based sensor for simultaneously monitoring both temperature and chemical cure was proposed (20).

Damage detection is one of the main purposes of health monitoring. The damage area and location can be

estimated analytically from the strain distribution measured by strain sensors. However, some types of fiber-optic sensors have the potential for directly detecting damages, which means that changes in signals directly indicate damage initiation and development without an analytical identification. These sensors must be placed near the damaged region because the sensing area of the fiber-optic sensors is limited to the neighborhood. Intensity-based break and microbend sensors can directly detect damage. The disadvantage of the break sensor is the difficulty of quantitatively measuring the damage mode, area, and location. Furthermore, the break sensor can detect only a few initial cracks (42,66). However, this concept is effective on irreparable parts such as composite parts of aircraft, and the sensing function of the fiber break is a final function of all types of fiber-optic sensors. The microbend sensor can detect cracks, which deform locally but do not break the fiber (44). Break or microbend sensors are available as distributed sensors combined with an OTDR technique to monitor the locations of damages (67,68). The FBG sensor can directly detect cracks by monitoring the optical spectral shape of reflected light, which is affected by the strain concentration at the crack tip. The spectral shape of reflected light is distorted when a nonuniform strain distribution occurs in the FBG sensing part. Studies of qualitative monitoring of strain concentration at a crack tip in a notched specimen (69) and at transverse cracks in a cross-ply laminate (54) were reported.

The location, size, and energy of impact damage applied to materials can be detected by fiber-optic vibrational sensors. Because high-speed measurement in the kilohertz to megahertz range is required for vibrational sensing, intensity-based sensors or interferometric sensors can be used for the purpose. Structural vibrational behavior measured by distributed fiber-optic vibrational sensors can be used to monitor the damages or the degradation of performance from changes in frequency responses. Ultrasonic waves from an impact can be caught by vibrational sensors to identify the location and energy of the impact. There are two concepts for vibrational sensors. One is based on dynamic strain measurement, and the other on the frequency property without a strain conversion. The latter is specialized for measuring the frequency property, as shown in Fig. 21e. An intensity-based vibrational sensor was proposed for detecting impact (6,70). For examples of dynamic strain measurement, impact damage detection in filament wound (FW) tubes using embedded intensity-based strain sensors was reported (47). An analytical model of wave propagation in a composite material was used to identify the damage from the responses of intensity-based fiber-optic vibrational sensors by neural network methods (70).

Piezoelectric Sensors for Health Monitoring

The use of piezoelectric elements in smart materials and structures has been studied since the 1980s. The research at this early stage was focused on the application to adaptive structures, which can control their properties of vibration, damping, and modal frequency. In these cases, the piezoelectric elements were used principally as actuators

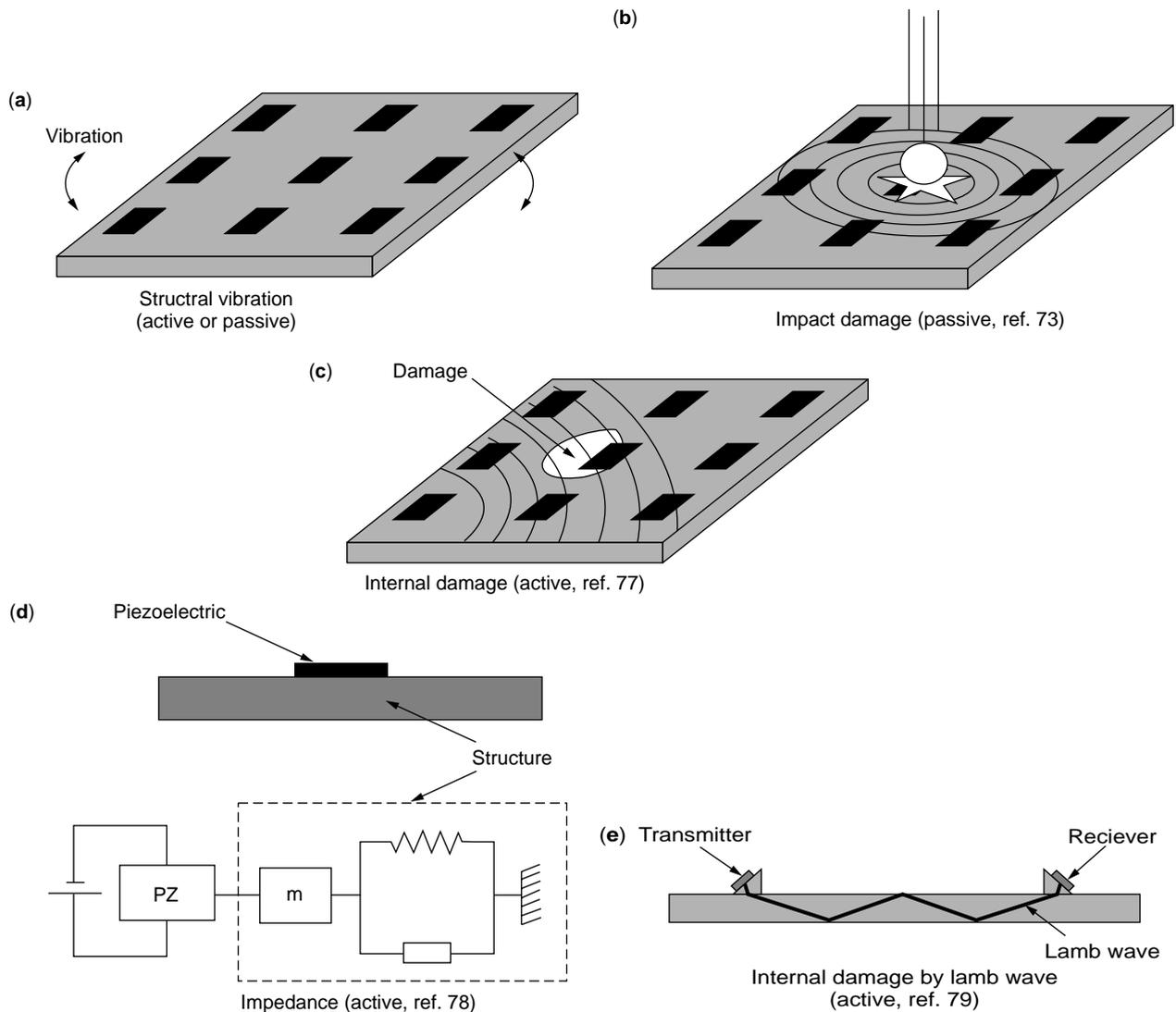


Figure 27. Configurations of piezoelectric elements for health monitoring.

and sensors in the control system. The application of piezoelectric elements to health monitoring started in the early 1990s. It is well known that the piezoelectric elements have a wider dynamic range than resistive strain gauges. Therefore, they are applied to health monitoring that uses the high-frequency range. It is a great advantage that a piezoelectric element can act as an actuator because the health monitoring technique using piezoelectric actuators and sensors can be used to detect tensile or fatigue damage as well as impact damage. Piezoelectric thin films such as polyvinylidene fluoride (PVDF) and PZT can be integrated in materials and structures due to their small size. PZT has better transmission efficiency and higher sensitivity than PVDF, but a PVDF thin film can be formed into any desired shape to be attached to the surface of complex structures due to its low stiffness and high flexibility (71). In most cases, piezoelectric sensors are distributed to monitor the overall region of materials and structures. Health monitoring by piezoelectric elements can be classified into structural vibrational monitoring, impact

damage monitoring, internal damage detection by diagnostic signals, structural impedance monitoring, and internal damage monitoring by Lamb wave, as shown in Fig. 27. The use of PVDF thin film for crack growth monitoring at low frequency was reported as an other interesting monitoring technique using piezoelectric elements (72).

Information on mode shapes and modal frequencies obtained from structural vibration properties can be used to evaluate structural health such as damage and performance degradation in structural vibrational monitoring, (Fig. 27a). A huge number of analytical models for evaluation from dynamic responses have been proposed. In passive health monitoring using modal analysis, piezoelectric elements are employed as a dynamic strain sensor substituting for a resistive strain gauge (71). This technique is aimed mainly at monitoring the health of large structures. Active health monitoring using actuators is more attractive for damage detection in materials because it is available without external vibration. Techniques for detecting

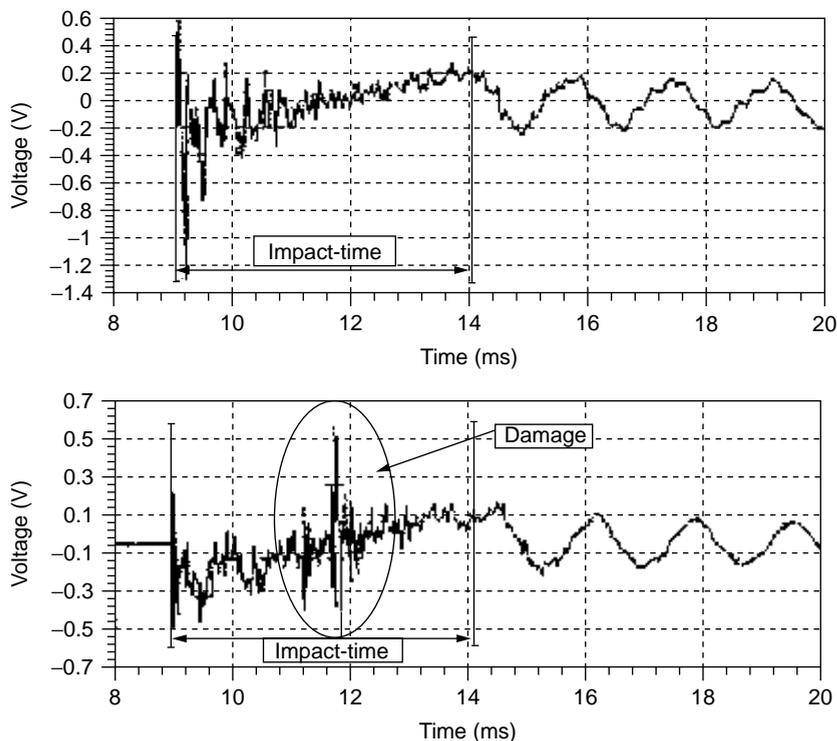


Figure 28. Piezosignals of elastic and injurious impacts (75).

delamination in composite materials using active health monitoring have been proposed (73,74).

For impact monitoring, impact damage, location, and energy can be evaluated by distributed piezoelectric elements, which catch the impacting signals transmitted in a material (Fig. 27b) (75–78). High-speed measurement by fiber-optic sensors is also available for impact monitoring. Figure 28 shows impacting, which is caught by a piezoelectric transducer bonded to a specimen. The signal from injurious impact involves a high-frequency signal generated by delamination (75). The identification of impact damage, location, and energy from the outputs of the distributed sensors is complicated because the process is a nonlinear and inverse problem. Therefore, a numerical identification technique is essential. A numerical code was developed and demonstrated for impact detection on a plate using piezoelectric sensors (79). Recently, an active sensing method, which uses a piezoelectric transducer as a transmitter and receiver, was proposed for internal damage detection (Fig. 27c) (79). The sensor configuration of the method is almost the same as that of the passive sensing method. The active sensing method has the advantage that it can detect damage without impact signals, that is, it is feasible to monitor the integrity generated by overload or fatigue as well as impact damage at any time. It was shown that the extent of impact damage could be predicted from the phase delay in transmitted diagnostic waves (Fig. 29; 79).

The impedance-based monitoring method using a piezoelectric transducer is based on measuring the coupled electromechanical impedance of a piezoelectric patch (80). The piezoelectric transducer and a host structure comprise an equivalent electric circuit (Fig. 27d). Therefore,

degradation of the structural performance is reflected in the impedance of the circuit. The admittance curves of a damaged experimental bridge joint were measured by an attached PZT, and the results represented the feasibility of qualitatively monitoring the damages in the range of the structural interactive frequency (80).

Some types of piezoelectric actuators/sensors are designed to generate and detect Lamb waves propagating through thin plates (Fig. 27e; 81–84). The Lamb wave has the capability of long-distance propagation and detecting internal damages such as delaminations of composite laminates. This technique is based on ultrasonic measurement, which is one of the traditional NDE techniques.

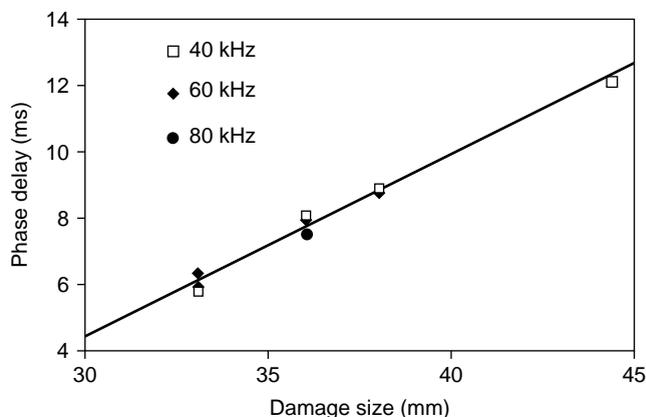


Figure 29. The relationship between impact damage size and phase delay in transmitted diagnostic waves measured by a piezosensor (79).

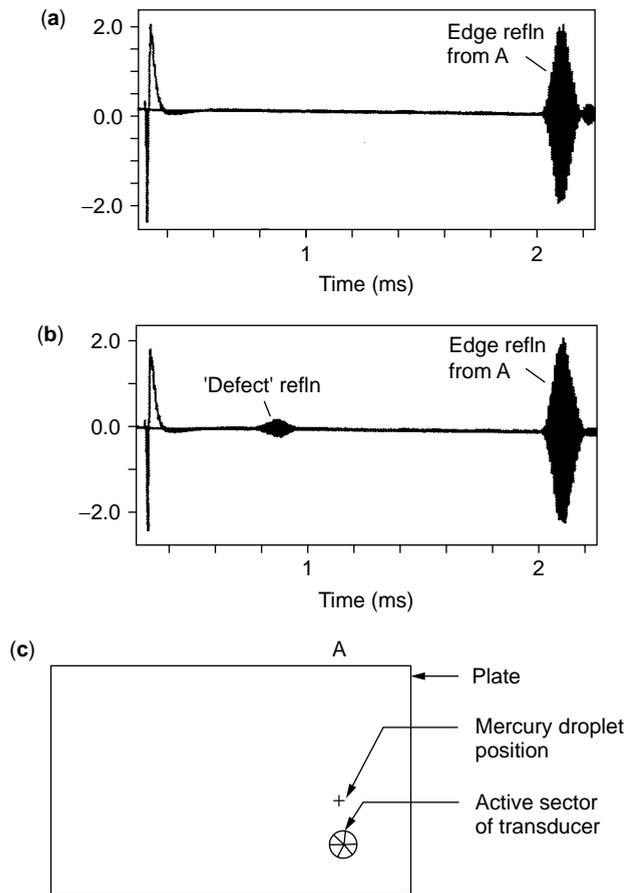


Figure 30. Measured signals from Lamb waves in the pulse-echo mode on an aluminum plate. (a) no defect; (b) defect; (c) defect position (83).

Miniaturized bonded PZT transducers were developed to produce Lamb waves (81,84). Interdigital PVDF transducers attached to a thin plate were investigated for generating Lamb waves (83). An embeddable PZT transducer was proposed for exciting Lamb waves in a composite plate (82). The experimental results for burst mode measurement of Lamb waves transmitted in a damaged composite laminate showed that attenuation changes in the damaged region (81). The pulse-echo mode of measurement, it was demonstrated, detects the reflection in the damaged region (Fig. 30; 83).

Magnetostrictive/Ferromagnetic Tagged Composites

Figure 31 illustrates a tagging technique that places functional material tags into the matrix of composites (85). The tag is small (mostly less than a micrometer) and has the shape of a particle or whisker. Mechanical properties of tagged composites are almost same as those of host materials due to their low volume fraction (mostly less than 10%). Magnetostrictive or ferromagnetic tagging techniques add a magnetic function to nonmagnetic composites. Magnetostrictive or ferromagnetic composites that have a high percentage have been developed since the early

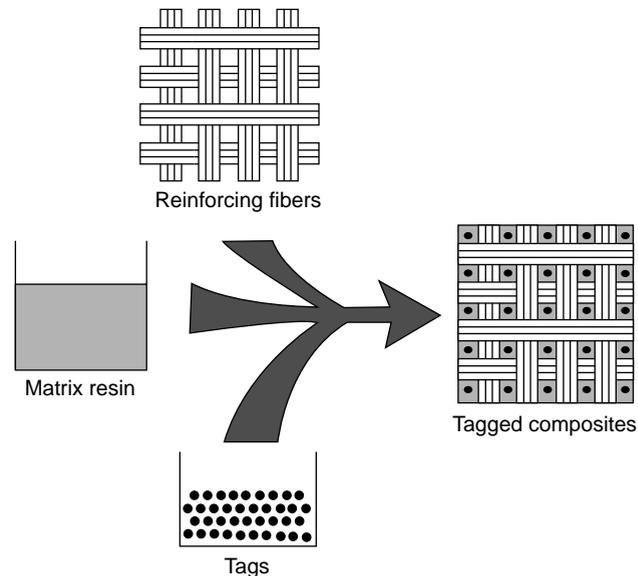


Figure 31. Concept of tagged composites (85).

1990s as actuators to improve the performance of magnetostrictive materials or to add an actuator function to polymers (86,87). This tagging technique has been used for monitoring the strain and internal damage in PMCs since the middle 1990s. Tagged composites have self-monitoring functions, so that embedded sensors in the materials are not necessary for *in situ* health monitoring.

A Terfenol-D magnetostrictive alloy particle (3–5 μm) is a representative magnetostrictive tag (87,88). The monitoring technique is based on the magnetostrictive effect, and therefore, the magnetic flux produced by the loaded material is measured to monitor the load or damage. The magnetic flux can be measured by magnetic probes such as a gauss meter probe or a Hall effect device (88). The transverse flux density produced is much larger than the axial flux density. It was reported that the magnetic flux has a nonlinear but monotonic relationship to the applied stress and the loading and unloading curves have a hysteretic loop (85). Figure 32 shows that the stress concentration around a hole affects the magnetic flux density.

Ferromagnetic elements such as nickel oxide (NiO), zinc oxide (ZnO), and ferrite (Fe_2O_3) are often employed in powder form (submicron–20 μm) (89). Health monitoring of ferromagnetic tagged composites is based on eddy current testing or the ferromagnetic effect. Eddy current testing is a traditional nondestructive technique for conductive materials. Therefore, carbon-reinforced composites do not need the tags for the test due to the electric conductivity of the carbon. Nonconductive PMCs such as glass-fiber-reinforced plastics (GFRPs) become conductive materials by tagging with ferromagnetic particles (or other conductive elements such as ferroelectric particles). It has been reported that eddy current testing was not so effective for monitoring internal damages (89). The ferromagnetic effect is a phenomenon whereby strain is generated in a ferromagnetic material when a magnetic field is applied. The ferromagnetic tagged composite vibrates in a periodic

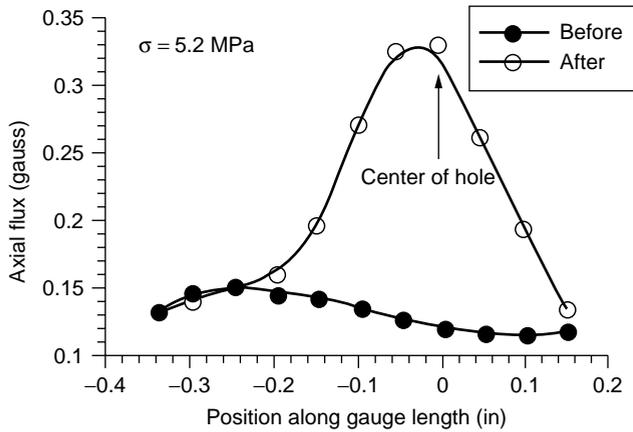


Figure 32. Axial magnetic flux readings for a through-hole specimen loaded in tension (85).

magnetic field. This means that the ferromagnetic tagged composite is used as the actuator itself to monitor damages from changes in the vibrational properties. It was reported that the frequency response is sensitive to cracks in/on the materials (89).

Electrical Resistance Measurement in Carbon-Reinforced Composites

The technique for health monitoring by measuring electrical resistance has become attractive since the late 1980s for carbon-reinforced composites (90). This technique measures changes in electrical resistance when strains or damages are applied to the composites. Like the tagging technique, the advantage of this technique is that there is no need for embedded sensors for *in situ* monitoring. In addition, the mechanical properties of the composites are not affected by using this monitoring technique because the carbon reinforcements work as sensors. Recently, applications have focused on three types of composites; carbon-fiber-reinforced concrete, carbon-fiber-reinforced polymers (CFRPs), and carbon fiber-carbon matrix (C/C) composites (91). The self-monitoring functions of carbon-reinforced composites are aimed at strain and damage monitoring. These functions result from changes in the electrical paths and in the conductivity of carbons. Figure 33 shows the electrical paths of these three types of composites. Short carbon-fiber-reinforced concrete consists of low conductive concrete and carbon fibers at a low volume fraction. In continuous carbon-reinforced polymers, the electrical paths

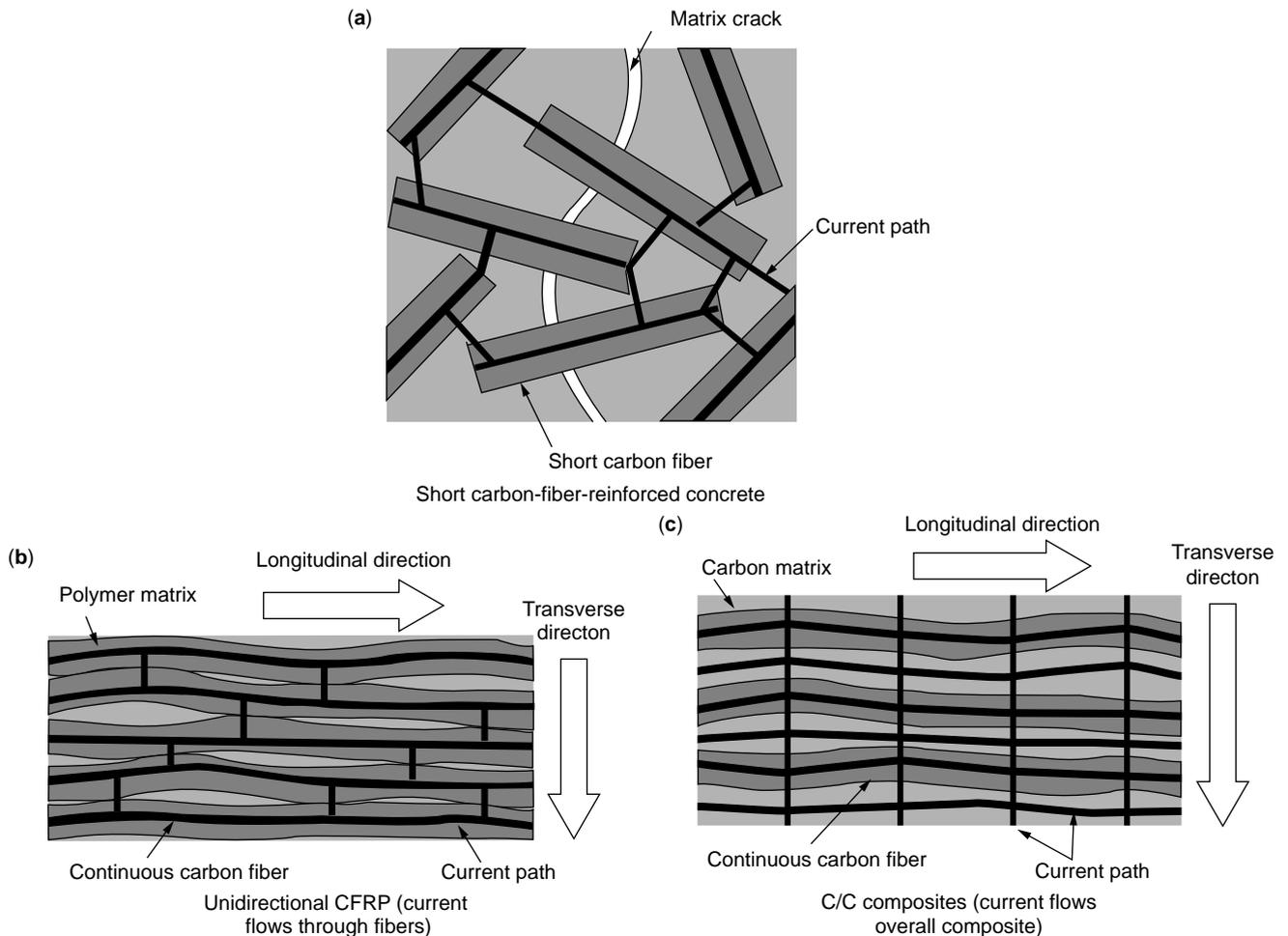


Figure 33. Electrical paths of current flowing through carbon-reinforced composites.

are composed of the carbon fibers due to the nonconductivity of polymers. A current flows overall in the C/C composite because it has conductivity in the fiber and matrix. The electrical paths in the composites are changed by damages such as fiber breaks, delaminations, matrix cracks, and debonding between the fiber and matrix. The mechanism of the variation of electrical resistance differs among these composites due to their different electrical paths.

The short carbon-fiber-reinforced concrete, which has high strength, high ductility, and low drying shrinkage, was developed for civil structures (91). Because the concrete matrix has low conductivity, contact between adjacent fibers is not essential in forming an electrical path (Fig. 33a). This means that composites that have a low volume fraction of carbon (less than 0.2 vol.%) have enough conductivity to monitor electric resistance (91). The effect can be seen in cement and mortar as well as in concrete. The electric resistance is reversibly proportional to the strain of the material, as shown in Fig. 34. This reversible and linear effect of the strain is driven by the variation in contact electrical resistivity between the fiber and the matrix (92). The figure shows the difference of the behavior in the first loading cycle due to matrix cracks and debonding between the fiber and matrix. After the first loading cycle, the fiber pull-out during loading and fiber push-in during unloading change the electrical resistance reversibly. The damage also affects the electric resistance, but the sensitivity to damage is less than that to strain sensing (92).

In unidirectional CFRPs, the carbon fibers comprise a complicated electrical network because neighboring fibers contact each other and the polymer matrix has no conductivity, as shown in Fig. 33b (93–96). The current flows along the fiber reinforcements in the longitudinal direction, and in the transverse direction through the contact area of the fibers. Therefore, unidirectional CFRPs have orthotropic electric conductivity. CFRPs without damages have the capability of reversible strain sensing due to variation of the conductivity of carbon fibers, and the residual strain that results from the alignment of carbon fibers can be observed in the first loading cycle. The electric paths change when damages such as fiber breaks, matrix cracks, and delamination occur under mechanical loading, as shown in Fig. 35 (93–95). Therefore, the electric resistance of CFRP laminates is sensitive to damages as well as strains. The breakage of carbon fiber is a principal damage mode that strongly affects the electric resistance of CFRP laminates. Figure 36 shows that the electrical resistance increases as fiber breakage grows (96). The residual resistance after unloading can be seen in the figure. This means that the history of damages can be recorded in the electrical resistance of CFRPs (97). This fact is very important for monitoring fatigue damage because fiber breakage occurs in a large number of loading cycles in the range of operational strain. Delamination in non-unidirectional CFRP laminates can be also detected by measuring the electrical resistivity due to the change in the electrical path in the transverse direction (95). Figure 37 shows that the delamination extent strongly affects the electrical resistance of CFRP cross-ply laminates.

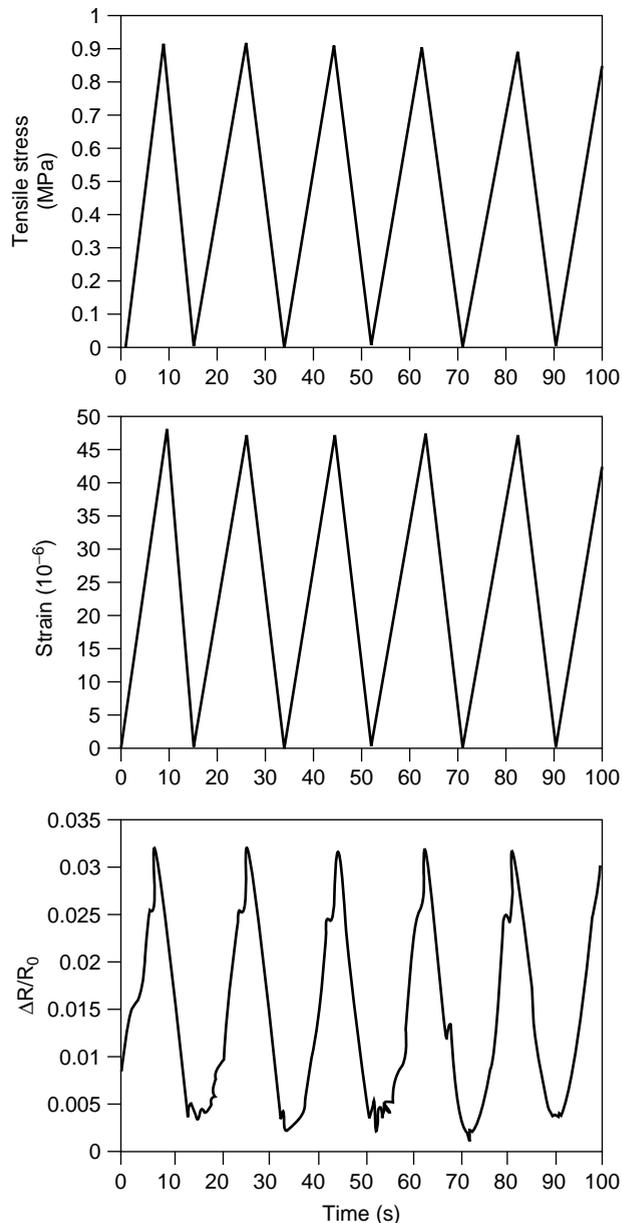


Figure 34. Changes in resistance, strain and stress during cyclic tensile loading of cement paste with ozone treated carbon fibers (91).

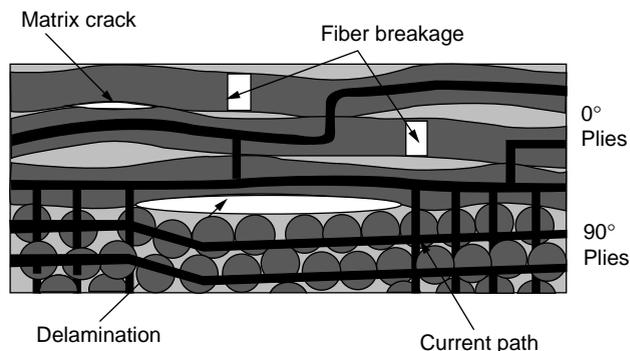


Figure 35. Electrical paths of damaged CFRP cross-ply laminates.

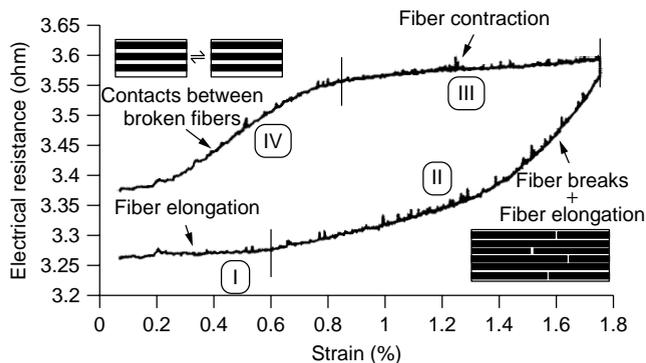


Figure 36. Schematic of the different processes that occur during a monotonic loading/unloading cycle below the strain to failure (96).

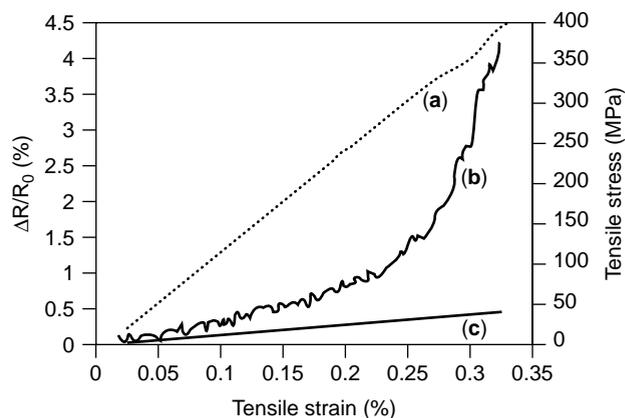


Figure 38. Plots of (a) tensile stress vs. strain, and (b) resistance vs. strain, obtained simultaneously during static tension up to failure for a C/C woven composite. Curve (c) is the calculated resistance based on dimensional changes (91).

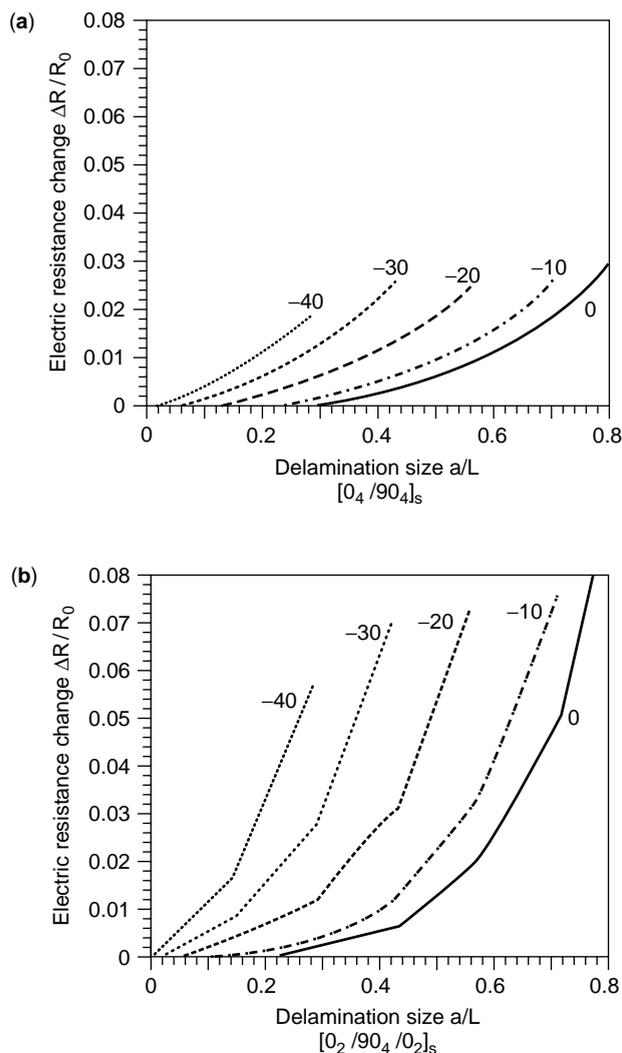


Figure 37. Relation between delamination extent and electric resistance of CFRP laminate. Numbers in the figures indicate the positions of delamination (95).

C/C composites are used for aerospace structures that operate at high temperature due to the high-temperature resistance of carbon. C/C composites are brittle and porous, and thus matrix cracks are easily generated under tension. Conventionally, monitoring the damage of C/C composites has been tried by using an acoustic emission technique. However, the wave propagation behavior in a C/C composite is very complicated, and attenuation of high-frequency waves is largely due to the porous matrix. Therefore, the electric resistance measurement is an effective technique for monitoring damages in C/C composites. Because they consist of continuous carbon fibers and a carbon matrix, a current flows overall through the composites (Fig. 33c). C/C composites have self-monitoring functions for strains and damages like CFRPs. The principal damage mode of C/C composites is a matrix crack, which affects the electrical resistance due to the conductivity of the carbon matrix. The electrical resistance of C/C composites is more sensitive to fatigue damage than that of CFRP laminates. Figure 38 shows that the electric resistance of a C/C woven composite increases nonlinearly under small strains due to generation of matrix cracks during a static tensile test (91).

Health Monitoring of Aircraft and Space Structures

On 28 April 1988, Aloha Airlines Boeing 737-200 cruising at 24,000 ft. over Hawaii suddenly lost an entire upper fuselage section. This accident resulted from fatigue damage, and then the health of aging aircraft that have undergone a high number of takeoff and landing cycles has been focused (98–100). A large number of flights degrades the structural performance of aircraft by mechanical and thermal fatigue and corrosion. Increases in aging aircraft in recent years and accidents caused by fatigue damage have become a serious problem of aircraft service (101). Two concepts have been applied to aircraft design to prevent accidents from fatigue damage. One is a fail-safe design, and the other is a damage tolerance design. The Fail-safe concept certifies the safe operation of an undamaged aircraft

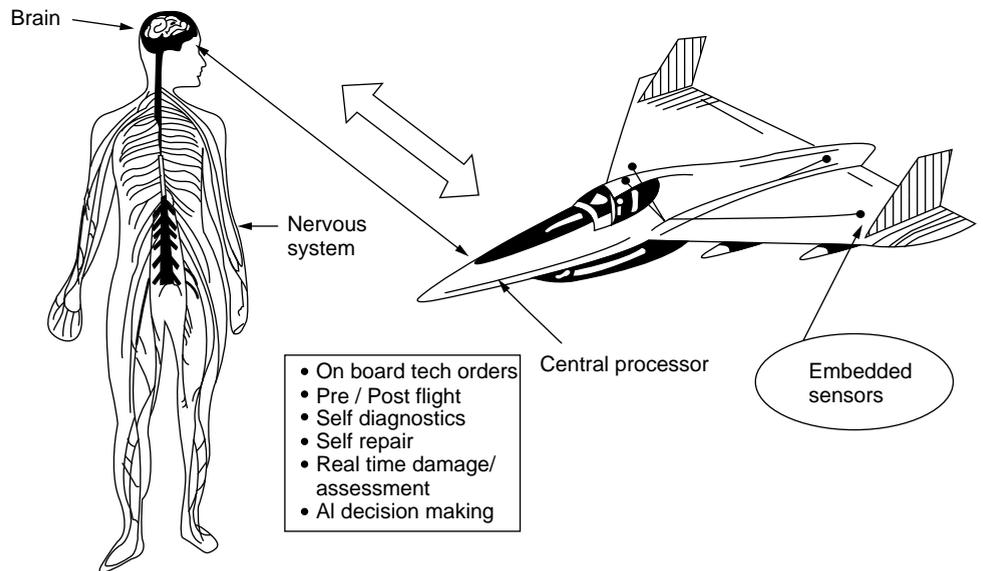


Figure 39. Concept of smart airplanes that have sensors (99).

under a limit load, and the damage tolerance concept aims at the survivability of a damaged aircraft under a limit load until the next inspection. In the latter concept, damages must be detected until they grow to a dangerous size at inspection. Multiple small damages, which are difficult to detect in periodic inspections, sometimes rapidly develop to large size damages because they combine, or they degrade damage tolerance due to their interaction (101,102). Therefore, structural inspection (overhaul) of aircraft must be performed to maintain the safety and reliability of aircraft. NDE techniques have been developed and employed to detect invisible damages during structural inspection. However, the cost of these inspections and repairs is very high because the overhaul require the airplane to be out of service. Such a time-consuming inspection is especially a problem for military aircraft. From this background, a new concept in the design of aircraft, called health monitoring aircraft, has emerged from the technology of smart materials and structures (99). Figure 39 shows the concept of smart airplanes that have sensors. Under the concept of health monitoring aircraft, an aircraft has a self-monitoring function provided by an integrated sensing system in the airframe and engines.

Recently, *in situ* sensor technologies for composites have been instituted by researchers in aerospace engineering because composite members used in aircraft are increasing due to the need for lightweight aircraft. Many types of composites, glass-fiber-reinforced polymers (GFRPs), CFRPs, and ceramic matrix composites (CMCs) are being considered as composite members of aircraft. CFRPs are promising as structural materials such as the frames and skins of the body or wings of the next generation because CFRPs have high specific stiffness, high specific strength, and high durability. However, the CFRP structural members that have invisible damages can cause tragic accidents due to their brittle behaviors of failure. Therefore, many demonstrations, in which health monitoring techniques are applied to composite aircraft frames, panels or wings,

have been conducted. Some applications of intensity-based fiber-optic sensor arrays, which were embedded in CFRP airframe skins, were proposed in the late 1980s (39–41). Recently, multiplexed or distributed fiber optic sensors have been applied to airframe components in laboratory studies. These sensors embedded in CFRP components can be used for monitoring strain, temperature, delaminations, transverse cracks, and impact (43,45,51,54,69,70). Piezoelectric sensors have been also employed for monitoring the health of CFRP components of aircraft, especially to detect impact damages such as delamination (73,74,77–79). For example, damage to an F/A-18 horizontal stabilizer was monitored by measuring the vibrational response using piezoelectric sensors, as shown in Fig. 40 (71). The impact damage in a large CFRP panel was detected by using embedded piezoelectric sensors (79). Electric resistance measurement of CFRP is a cost-effective approach to monitoring internal damages. However, these techniques are actually applied to small specimens. The health monitoring

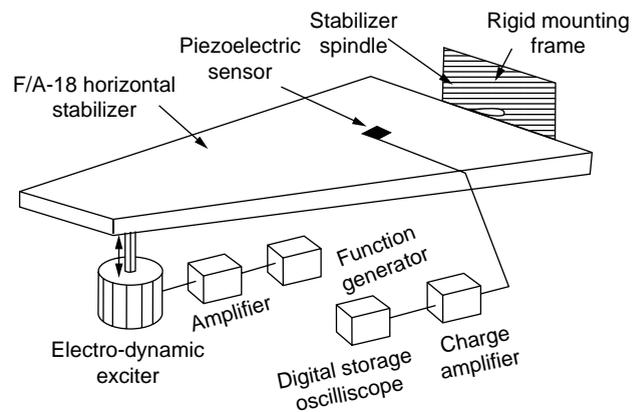


Figure 40. Damage monitoring of an F/A-18 horizontal stabilizer using a piezoelectric sensor (71).

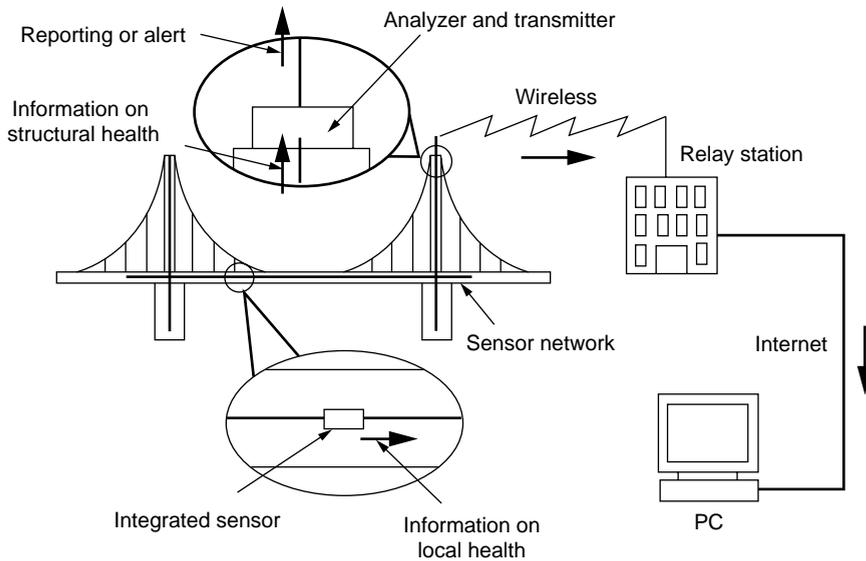


Figure 41. Concept of a smart bridge that has a health monitoring system.

techniques, which are proposed using smart materials and structures, have not yet been employed in actual aircraft, but experimental field tests are currently ongoing.

For space structures, degraded structures present different problems versus aircraft. The performance of space structures can be degraded by mechanical and thermal fatigue and damage by space debris. To compensate for errors in performance such as observation, monitoring, and communication, measurements of strain, deformation, temperature, and vibration are desired (103). Damage monitoring of orbital spacecraft will be required to monitor damage type, location, and size and to specify a repair method when a low-cost launch is realized in the future. The sensors for spacecraft require light weight; long-term durability to mechanical, thermal and radioactive fatigue; and immunity to electromagnetic interference. Therefore, the most suitable sensor is a fiber-optic sensor. Most of the fiber-optic sensing techniques for CFRP components can be applied to space structures, but smaller, lighter sensing devices are desired. The measurement of strain distribution in a composite plate element of a satellite and its antenna reflectors was demonstrated by using a multiplexed FBG sensor system (103).

Health Monitoring of Civil Structures

Civil structures such as bridges, highways, roads, large buildings, tunnels, and dams need periodic inspections because they deteriorate from fatigue, corrosion, and natural disasters such as earthquakes and typhoons. The number of civil structures is increasing, and therefore, the maintenance cost of civil structures, including inspection, repair, and renewal is increasing (104). The traditional inspection methods are visual and acoustic inspections by human operators, which are obviously inefficient methods for large structures. Therefore, low-cost, highly reliable inspection methods are desired in the field of civil engineering. Based on this background, health monitoring technology becomes an attractive approach for civil structures.

In Japan, the Kobe earthquake in 1995 accelerated the development of practical applications of health monitoring to civil structures (59). The key technologies for health monitoring of civil structures are long-lived distributed sensors, analytical modeling of structural behavior, and a remote monitoring system through a worldwide network, as shown in Fig. 41. Long-term survivability and distributed sensing are essential in civil structures due to their long-term continuous operation. In addition, it is important that the sensor system can be handled easily by workers or operators in construction areas. The structural materials of civil structures are steel, concrete, cement, mortar, and carbon-fiber-reinforced composites. Recently, CFRP composites were employed as structural members and wires. CFRP repair sheets were the most promising solution for repair of damaged concrete shoring and walls. Therefore, some of the health monitoring techniques used for CFRP composites can also be available for CFRP structures in civil structures.

A major monitoring technique, which is employed in health monitoring of civil structure, is a structural dynamic-based system. The structural dynamic-based system is an analytical approach to monitoring the damage and performance degradation of large structures by measuring the dynamic response. In this method, distributed sensor patches attached to the members of structures provide vibrational response such as mode shapes and modal frequency. Piezoelectric patches and fiber-optic vibrational sensors can be used for measuring the dynamic response. Optimizing the location of the sensors is important for the system to be cost-effective because the operating cost of health monitoring depends on the number of sensors (105). There are various techniques for damage identification for a structural dynamic-based system. They use a modal analysis technique that has a structural model or finite-element analysis (104), a neural network technique (106), etc.

Fiber-optic sensor-based health monitoring is an attractive idea for civil engineering because of its high durability, high strength, high sensitivity, nonperturbation by

electromagnetic interference, and ability to be embedded internally (49). However, because an optical fiber is fragile, easy installation of the fiber is required in construction areas. Furthermore, embedded fiber-optic sensors require modification to protect them from cracks that propagate in concrete or composites because the members of civil structures are difficult to replace when the sensor is broken. Some applied examples of the health monitoring of concrete and composite structures that use fiber-optic sensors are described following. Distributed or multiplexed fiber-optic sensors such as BOTDR, ROTDR, a multiplexed FBG sensor system, and a long-gauge fiber-optic sensor system (intensiometric and interferometric) have been employed for monitoring strain distribution and detecting damages in civil structures. It was proposed that ROTDR could be used for permanent monitoring of the soil temperature of an in-ground tank (59). FBG strain sensors, polarimetric extension sensors, and OTDR crack sensors were employed to monitor local strain change, 2.5 m long-gauge displacement, and crack generation and location in a full-scale destructive bridge test (107). The strain on a CFRP cable of a stay cable bridge in Winterthur, Switzerland, was monitored by using multiple FBG sensors (108). Intensity-based fiber-optic sensors were used to monitor the failure of concrete in the Stafford Medical Building in Vermont, U.S.A. (109). Structural monitoring of a concrete member was conducted by curvature analysis using an interferometric sensor (48). The health of a building at the University of Colorado was monitored using multiplexed FBG sensors and a remote sensing system through the Internet (53). Electric resistance measurement can be applied to the health monitoring of carbon- or steel-reinforced concrete or CFRP repair sheet. Electric resistance measurement of carbon or steel composite structures provides information about the matrix and the reinforcement condition such as breakage or corrosion. There are many laboratory studies of the resistance measurement technique (90–97).

A remote monitoring technique through a worldwide network has become practical because of the advance of the Internet in the late 1990s. This idea is very attractive to construction corporations because it produces a new business of low-cost maintenance service. This technique involves high-speed communication devices, wireless communication devices, and web-based technologies. Remote health monitoring on the Web has been proposed by Web-based software written in a network-friendly language (53). The advantage of Web-based remote monitoring is that special software installed in a local computer is not necessary. Wireless devices make it possible to collect data from integrated sensors without an on-line cable (110).

BIBLIOGRAPHY

1. D. Hull and T.W. Clyne, *An Introduction to Composite Materials*, 2e, Cambridge University Press, Cambridge, UK, 1996.
2. A.C. Loos and G.S. Springer, *J. Composite Mater.*, **17**: 135–169 (1983).
3. J.M. Fildes, S.M. Milkovich, R. Altkorn, R. Haidle, and J. Neatrou, *25th Int. SAMPE Tech. Conf.*, Philadelphia, PA, Oct. 1993, pp. 26–28.
4. G.R. Powell, P.A. Crosby, D.N. Waters, C.M. France, R.C. Spooncer, and G.F. Fernando, *Smart Mater. Struct.* **7**(4): 557–568 (1998).
5. P.A. Crosby, G.R. Powell, G.F. Fernando, C.M. France, R.C. Spooncer, and D.N. Waters, *Smart Mater. Struct.* **5**(4): 415–428 (1996).
6. C. Doyle, A. Martin, T. Liu, M. Wu, S. Hayes, P.A. Crosby, G.R. Powell, D. Brooks, and G.F. Fernando, *Smart Mater. Struct.* **7**(2): 145–158 (1998).
7. A. Fuchs and N.H. Sung, *53rd Soc. Plast. Eng. Annu. Tech. Conf. (ANTEC 95)*, Boston, MA, May 1995, Vol. 2, pp. 2437–2441.
8. H.J. Paik and N.H. Sung, *Polym. Eng. Sci.* **34**(12): 1025–1032 (1994).
9. B.P. Rice, *38th Int. SAMPE Symp.* Anaheim, PA, May 1993, pp. 1346–1356.
10. D.L. Woerdeman and R.S. Parnas, *Plast. Eng.* **5**(10): 25–27 (1995).
11. S.S.J. Roberts and R. Davidson, *Composites Sci. Technol.* **49**: 265–276 (1993).
12. J.P.H. Steele, D. Mishra, and C. Ganesh, *Proc. ASME Mater. Div.* **69**(2): 899–909 (1995).
13. Y.M. Liu, C. Ganesh, J.P.H. Steele, and J.E. Jones, *J. Composite Mater.* **31**(1): 87–102 (1997).
14. A.L. Kalamkarov, S.B. Fitzgerald, and D.O. MacDonald, *Composites: Part B* **30**: 167–175 (1999).
15. K. Osaka, T. Kosaka, Y. Asano, and T. Fukuda, *Proc. 2nd Asian-Australasian Conf. Composite Mater. (ACCM-2000)*, Kyongju, Korea, 2000, pp. 1117–1122.
16. T. Kosaka, K. Osaka, M. Sando, and T. Fukuda, *Proc. 9th US-Japan Conf. Composite Mater.*, Mishima, Japan, 2000, pp. 151–158.
17. L. Lai, G. Carman, S. Chiou, P. Kukuchek, and D. Echternach, *Smart Mater. Struct.* **4**(2): 118–125 (1995).
18. V.M. Murukeshan, P.Y. Chan, L.S. Ong, and L.K. Seah, *Sensors and Actuators: A phys.* **79**(2): 153–161 (2000).
19. R.C. Foedinger, D.L. Rea, J.S. Sirkis, C.S. Baldwin, and J.R. Troll, *Proc SPIE* **3670**: 289–301 (1999).
20. P.A. Crosby, C. Doyle, C. Tuck, M. Singh, and G.F. Fernando, *Proc SPIE* **3670**: 144–152 (1999).
21. J.S. Kim and D.G. Lee, *J. Composite Mater.* **30**(13): 1436–1457 (1996).
22. D.E. Kranbuehl, P. Kingsley, and S. Hart, G. Hasko, B. Dexter, and A.C. Loos, *Polym. Composites* **15**(4): 299–305 (1994).
23. J. Mijovic, J.M. Kenny, A. Maffezzoli, A. Trivisano, F. Bellucci, and L. Nicolais, *Composites Sci. Technol.* **49**: 277–290 (1993).
24. M.B. Buczek and C.W. Lee, *40th Int. SAMPE Symp.*, Anaheim, CA, May 1995, pp. 696–702.
25. D. Kranbuehl, D. Hood, J. Rogozinski, A. Meyer, and M. Neag, *Prog. Org. Coat.* **35**: 101–107 (1999).
26. D. Kranbuehl, D. Hood, Y. Wang, G. Boiteux, F. Stephan, C. Mathieu, G. Seytre, A. Loos, and D. McRae, *Polym. Adv. Technol.* **8**: 93–99 (1997).
27. D.D. Shepard, D.R. Day, and K.J. Craven, *J. Reinforced Plast. Composites* **14**: 297–308 (1995).
28. T. Krusche and W. Michaeli, *41st Int. SAMPE Symp.*, Anaheim, CA, March 1996, pp. 1542–1550.

29. S. Motogi, T. Itoh, T. Fukuda, K. Yamagishi, S. Kitade, and H. Morita, *12th Int. Conf. Composite Mater. (ICCM-12)*, Paris, France, July 1999, Paper 1260.
30. J.S. Kim and D.G. Lee, *Sensors Actuators B: Chem.* **30**(2): 159–164 (1996).
31. J.F. Maguire, M.A. Miller, and S. Venkatesan, *Eng. Appl. Artif. Intelligence* **11**(5): 605–618 (1998).
32. T. Fukuda, *Proc. US/Japan Workshop Collaborations Mater. Res. Nikko, Japan*, 1998, pp. 26–27.
33. X. Wang, L. Ye, and Y.W. Mai, *J. Intelligent Mater. Syst. Struct.* **8**(12): 1073–1078 (1997).
34. X. Wang, C. Ehlers, C. Kissinger, M. Neitzel, L. Ye, and Y.W. Mai, *Smart Mater. Struct.* **7**(1): 113–120 (1998).
35. X. Wang, C. Ehlers, C. Kissinger, M. Neitzel, L. Ye, and Y.W. Mai, *Smart Mater. Struct.* **7**(1): 121–127 (1998).
36. J.E. Eder and J.L. Rose, *ASME Appl. Mech. Div.* **188**: 179–186 (1994).
37. M. Rath, J. Döring, W. Stark, and G. Hinrichsen, *NDT & E Int.* **33**(2): 123–130 (2000).
38. N. Legros, C.-K. Jen, and I. Ihara, *Ultrasonics* **37**(4): 291–297 (1999).
39. B. Hofer, *Composites* **18**(4): 309–316 (1987).
40. S.R. Waite and G.N. Sage, *Composites* **19**(4): 288–294 (1988).
41. R.M. Measures, *Prog. Aerosp. Sci.* **26**: 289–351 (1989).
42. N.D.W. Glossop, S. Dubois, W. Tsaw, M. LeBlanc, J. Lymer, R.M. Measures, and R.C. Tennyson, *Composites* **21**(1): 71–80 (1990).
43. T. Liu, M. Wu, Y. Rao, D.A. Jackson, and G.F. Fernando, *Smart Mater. Struct.* **7**(4): 550–556 (1998).
44. C.K.Y. Leung, N. Elvin, N. Olson, T.F. Morse, and Y.F. He, *Eng. Fracture Mech.* **65**(2–3): 133–148 (2000).
45. R.A. Badcock and G.F. Fernando, *Smart Mater. Struct.* **4**(4): 223–230 (1995).
46. F.J. Arregui, I.R. Matias, and M. Lopez-Amo, *Sensors and Actuators A: phys.* **79**(2): 90–96 (2000).
47. A.R. Martin, G.F. Fernando, and K.F. Hale, *Smart Mater. Struct.* **6**(4): 470–476 (1997).
48. D. Inaudi, S. Vurpillot, N. Casanova, and P. Kronenberg, *Smart Mater. Struct.* **7**(2): 199–208 (1998).
49. C.I. Merzbacher, A.D. Kersey, and E.J. Friebelle, *Smart Mater. Struct.* **5**(2): 196–208 (1996).
50. V.M. Murukeshan, P.Y. Chan, O.L. Seng, and A. Asundi, *Smart Mater. Struct.* **8**(5): 544–548 (1999).
51. Y.J. Rao, *Opt. Lasers Eng.* **31**(4): 297–324 (1999).
52. M. Vries, V. Bhatia, T. D'Alberto, V. Arya, and R.O. Claus, *Eng. Struct.* **20**(3): 205–210 (1998).
53. V.E. Saouma, D.Z. Anderson, K. Ostrander, B. Lee, and V. Slowik, *Mater. Struct.* **31**: 259–266 (1998).
54. Y. Okabe, S. Yashiro, T. Kosaka, and N. Takeda, *Smart Mater. Struct.* **9**(6): 832–838 (2000).
55. Y.J. Rao, D.A. Jackson, L. Zhang, and I. Bennion, *Opt. Lett.* **21**: 683–685 (1996).
56. X. Tao, L. Tang, W.C. Du, and C.L. Choy, *Composites Sci. Technol.* **60**(5): 657–669 (2000).
57. V. Bhatia, D.K. Campbell, D. Sherr, and R.O. Claus, *Proc. SPIE* **3042**: 78–88 (1997).
58. Z. Zhang and J.S. Sirkis, *Proc. 12th Inter. Conf. Optical Fiber Sensors (OFS-12)*, Williamsburg, VA, Oct. 1997, pp. 294–297.
59. A. Mita, *2nd Int. Workshop Struct. Health Monitoring*, Stanford University, USA, 1999, pp. 56–67.
60. H. Ohno, Y. Uchiyama, and T. Kurashima, *Proc. SPIE* **3670**: 486–496 (1999).
61. M.D. DeMerchant, A.W. Brown, X. Bao, and T.W. Bremner, *Proc. SPIE* **3670**: 352–358 (1999).
62. E. Udd, D. Nelson, and C. Lawrence, *Proc. 12th Int. Conf. Optical Fiber Sensors (OFS-12)*, Williamsburg, VA, Oct. 1997, pp. 48–51.
63. P. Sivanesan, J.S. Sirkis, V. Venkat, Y.C. Shi, C.J. Reddy, S.N. Sankaran, and H. Singh, *Proc. SPIE* **3670**: 92–103 (1999).
64. J. Smith, A.W. Brown, M.D. DeMerchant, and X. Bao, *Proc. SPIE* **3670**: 366–373 (1999).
65. T. Liu, G.F. Fernando, Z.Y. Zhang, and K.T.V. Grattan, *Sensors and Actuators A: Phys.* **80**(3): 208–215 (2000).
66. S. Kitade, T. Fukuda, K. Osaka, and A. Hamamoto, *US-Japan Workshop Smart Mater. Struct.*, Seattle, WA, Dec. 1996, pp. 283–290.
67. K. Kageyama, I. Kimpara, T. Suzuki, I. Ohsawa, H. Murayama, and K. Ito, *Smart Mater. Struct.* **7**(4): 472–478 (1998).
68. F. Knowles, B. E. Jones, C. M. France, and S. Purdy, *Sensors Actuators A: Phys.* **68**(1–3): 320–323 (1998).
69. K.J. Peters, M. Studer, J. Botsis, A. Iocco, H.G. Limberger, and R.P. Salathe, *Proc. SPIE* **3670**: 195–206 (1999).
70. C. Doyle and G. Fernando, *Smart Mater. Struct.* **7**(4): 543–549 (1998).
71. W.K. Chiu, S.C. Galea, H. Zhang, R. Jones, and Y.C. Lam, *J. Intelligent Mater. Syst. Struct.* **5**(5): 683–693 (1994).
72. H. Zhang, S.C. Galea, W.K. Chiu, and Y.C. Lam, *Smart Mater. Struct.* **2**(4): 208–216 (1993).
73. A.S. Islam and K.C. Craig, *Smart Mater. Struct.* **3**(3): 318–328 (1994).
74. D.K. Shah, W.S. Chan, and S.P. Joshi, *Smart Mater. Struct.* **3**(3): 293–301 (1994).
75. C. Wölfinger, F.J. Arendts, K. Friedrich, and K. Drechsler, *Aerosp. Sci. Technol.* **2**(6): 391–400 (1998).
76. J.F. Campbell, E.G. Vanderheiden, and L.A. Martinez, *J. Composite Mater.* **26**: 334–349 (1992).
77. S.C. Galea, W.K. Chiu, and J.J. Paul, Monitoring Damage in Composites, *J. Intelligent Mater. Syst. Struct.* **4**(3): 330–336 (1993).
78. K. Choi and F.K. Chang, *J. Intelligent Mater. Syst. Struct.* **4**: 864–869 (1993).
79. M. Tracy, Y.S. Roh, and F.K. Chang, *Proc. 3rd ICIM/ECSSM '96*, Lyon, France June 1996, pp. 118–123.
80. J.W. Ayres, F. Lalande, Z. Chaudhry, and C.A. Rogers, *Smart Mater. Struct.* **7**(5): 599–605 (1998).
81. H. Kaczmarek, C. Simon, and C. Delebarre, *Proc. 3rd ICIM/ECSSM '96*, Lyon, France, June 1996, pp. 130–135.
82. E. Moulain, J. Assaad, C. Delebarre, H. Kaczmarek, and D. Balageas, *J. Appl. Phys.* **82**(5): 2049–2055 (1997).
83. R.S.C. Monkhouse, P.W. Wilcox, M.J.S. Lowe, R.P. Dalton, and P. Cawley, *Proc. 4th ESSM 2nd MIMR Conf.*, Harrogate, UK, July 1998, pp. 397–404.
84. M. Veidt, T. Liu, and S. Kitipornchai, *Smart Mater. Struct.* **9**(1): 19–23 (2000).
85. S.R. White, *Int. Composites Expo(SPI/ICE'99)*, Cincinnati, OH, May 1999, SESSION 22-E, pp. 1–6.
86. L. Sandlund, M. Fahlander, T. Cedell, A.E. Clark, J.B. Restorff, and M.W. Fogle, *J. Appl. Phys.* **75**(10): 5656–5658 (1994).
87. M.R. Jolly, J.D. Carlson, B.C. Munoz, and T.A. Bullions, *J. Intelligent Mater. Syst. Struct.* **7**(6): 613–621 (1996).

88. J. Trovillion, J. Kamphaus, R. Quattrone, and J. Berman, *Int. Composites Expo (SPI/ICE'99)*, Cincinnati, OH, May 1999, SESSION 22-D, pp. 1–6.
89. V. Giurgiutiu, Z. Chen, F. Lalande, and C.A. Rogers, *J. Intelligent Mater. Syst. Struct.* **7**(6): 623–634 (1996).
90. K. Schulte and C.H. Baraon, *Composites Sci. Technol.* **36**: 63–76 (1989).
91. D.D.L. Chung, *Mater. Sci. Eng.* **R22**: 57–78 (1998).
92. D.D.L. Chung, *Smart Mater. Struct.* **4**(1): 59–61 (1995).
93. X. Wang and D.D.L. Chung, *Smart Mater. Struct.* **6**(4): 504–508 (1997).
94. P.E. Irving and C. Thiagarajan, *Smart Mater. Struct.* **7**(4): 456–466 (1998).
95. A. Todoroki, *Proc. 4th ESSM 2nd MIMR Conf.*, Harrogate, UK, July 1998, pp. 429–434.
96. J.C. Abry, S. Bochar, A. Chateauminois, M. Salvia, and G. Giraud, *Composite Sci. Technol.* **59**: 925–935 (1999).
97. M. Sugita, H. Yanagida, M. Hiroaki, and N. Muto, *Smart Mater. Struct.* **4**(1A): A52–A57 (1995).
98. G. Bartelds, *J. Intelligent Mater. Syst. Struct.* **9**: 906–910 (1998).
99. T.G. Gerardi, *J. Intelligent Mater. Syst. Struct.* **1**(3): 375–385 (1990).
100. G.A. Hickman, J.J. Gerardi, and Y. Feng, *J. Intelligent Mater. Syst. Struct.* **2**(3): 411–429 (1991).
101. R. Jones, L. Molent, and S. Pitt, *Theor. Appl. Fracture Mech.* **32**(2): 81–100 (1999).
102. U.G. Goranson, *Int. J. Fatigue* **20**(6): 413–431 (1998).
103. E.J. Friebele, C.G. Askins, A.B. Bosse, A.D. Kersey, H.J. Patrick, W.R. Pogue, M.A. Putnam, W.R. Simon, F.A. Tasker, W.S. Vincent, and S.T. Vohra, *Smart Mater. Struct.* **8**: 813–838 (1999).
104. A.E. Aktan, A.J. Helmicki, and V.J. Hunt, *Smart Mater. Struct.* **7**(5): 674–692 (1998).
105. M.L. Wang, G. Heo, and D. Satpathi, *Smart Mater. Struct.* **7**(5): 606–616 (1998).
106. M. Nakamura, S.F. Masri, A.G. Chassiakos, and T.K. Caughey, *Earthquake Eng. Struct. Dynamics* **27**(9): 997–1010 (1998).
107. H. Storoy, J. Saether, and K. Johannessen, *J. Intelligent Mater. Syst. Struct.* **8**: 633–643 (1997).
108. R. Brönnimann, P.M. Nellen, and U. Sennhauser, *Smart Mater. Struct.* **7**(2): 229–236 (1998).
109. P.L. Fuhr, D.R. Huston, P.J. Kajenski, and T.P. Ambrose, *Smart Mater. Struct.* **1**(1): 63–68 (1992).
110. D.J. Pines and P.A. Lovell, *Smart Mater. Struct.* **7**(5): 627–636 (1998).

DRUG DELIVERY SYSTEMS

JOSEPH KOST

SMADAR A. LAPIDOT

Ben-Gurion University of the Negev

Beer Sheva, Israel

INTRODUCTION

The rapid advancement of biomedical research has led to many creative applications for biocompatible polymers. As modern medicine discerns more mechanisms of both physiology and pathophysiology, the approach to healing is to mimic, or if possible, to recreate the physiology of healthy functioning. Thus, the area of responsive drug delivery has evolved. Also called “smart” polymers, for drug delivery, the developments fall in two categories: externally regulated or pulsatile systems (also known as “open-loop” systems) and self-regulated systems (also called “closed-loop”). This article outlines the fundamentals of this research area and gives a detailed account of the most recent advances in both pulsatile and self-regulated drug delivery systems.

DEVELOPMENT OF CONTROLLED DRUG DELIVERY

Control of Drug Concentration Levels Over Time

The overall goal in developing controlled release devices is maintaining the drug in the therapeutic range (zero-order release kinetics) and targeting delivery to specific tissues (lowering systemic exposure and side effects). Polymers have been used in developing all four types of devices, classified by release mechanism: (1) diffusion controlled, both reservoir and monolithic; (2) chemically controlled release, that is, bioerodible carriers; (3) solvent controlled release, where swelling of the matrix is the mechanism that enables the entrapped drug to come out; and (4) externally controlled release (1).

Although newer and more powerful drugs continue to be developed, increasing attention is being given to the methods of administering these active substances. In conventional drug delivery, the drug concentration in the blood rises when the drug is taken, then peaks, and declines. Maintaining drug in the desired therapeutic range by using just a single dose or targeting the drug at a specific area (lowering the systemic drug level) are goals that have been successfully attained by using commercially available controlled release devices (2). However, there are many clinical situations where the approach of a constant drug delivery rate is insufficient, such as the delivery of insulin for patients who have diabetes mellitus, antiarrhythmics for patients who have heart rhythm disorders, gastric acid inhibitors for ulcer control, nitrates for patients who have angina pectoris, as well as selective β -blockade, birth

control, general hormone replacement, immunization, and cancer chemotherapy. Furthermore, studies in the field of chronopharmacology indicate that the onsets of certain diseases exhibit strong circadian temporal dependence. Thus, treatment of these diseases could be optimized by using responsive delivery systems (3), which are, in essence, man-made imitations of healthy function.

Biocompatibility

When designing a controlled delivery device, the effects of the drug must be taken into account and also the potential effects of the device itself on the biological system (4). In other words, both the effects of the implant on the host tissues and the effects of the host on the implant must be considered. These are some of the important potential effects: inflammation and the “foreign body reaction,” immunologic responses, systemic toxicity, blood–surface interactions, thrombosis, device-related infection, and tumorigenesis (4). Many of these effects actually comprise the body’s defense mechanism against injury; placement of a drug delivery device in the body causes injury and therefore, elicits these reactions. However, the degree of perturbation is strongly impacted by the biomaterial that comprises the device.

The first response to be triggered is inflammation. The cellular and molecular mechanisms have been well described, but avoiding them has not yet been achieved. Many of the inflammatory responses are local to the site of implantation and dissipate relatively quickly. Some of the most potent chemical mediators, such as lysosomal proteases and oxygen-derived free radicals also play an important role in the degradation and wear of biomaterials (1).

The products of degradation and wear can cause immune responses and/or nonimmune systemic toxicity. Thus, when testing a delivery device, both the intact device and its degradation products must be thoroughly examined *in vitro* before implantation *in vivo*. An additional phenomenon that can hamper the device’s function is fibrous encapsulation of the biomaterial. These reactions can be very specific to the host, and *in vivo* experiments are not always indicative of the human response. There is a wealth of literature regarding biocompatibility mechanisms and testing into which the interested reader is encouraged to delve (4).

Classification of “Smart” Polymers

“Intelligent” controlled release devices can be classified as open- or closed-loop systems, as shown in Fig. 1. Open-loop control systems (Fig. 1a) are those where information about the controlled variable is not automatically used to adjust the system inputs to compensate for the change in process variables. In the controlled drug delivery field, open-loop systems are known as pulsatile or externally regulated.

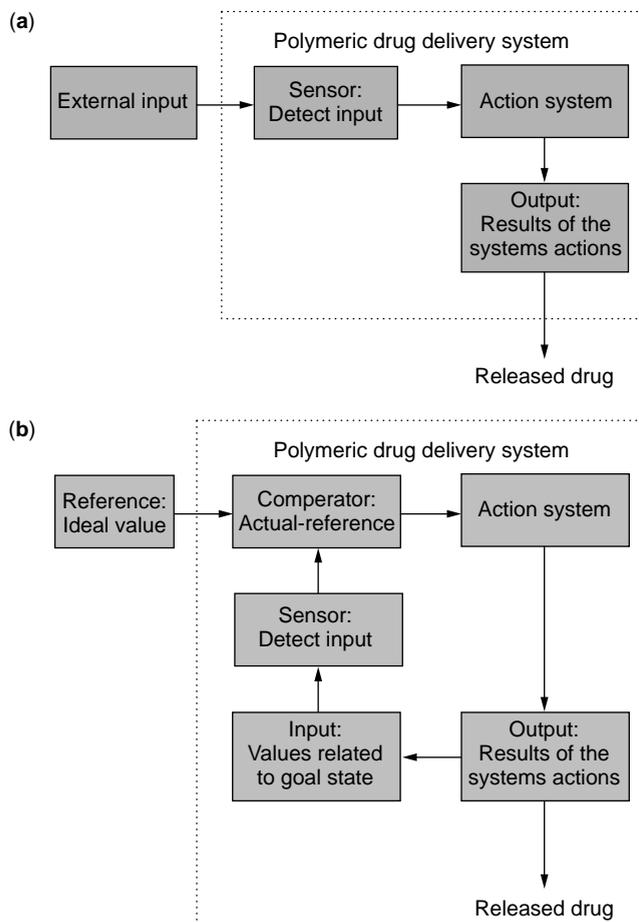


Figure 1. Schematic representation of drug delivery systems and their control mechanisms: (a) open-loop system; (b) closed-loop system.

Externally controlled devices apply external triggers such as magnetic, ultrasonic, thermal, or electric irradiation for pulsatile delivery.

Closed-loop control systems, on the other hand, are defined as systems where the controlled variable is detected, and as a result the system output is adjusted accordingly. Closed-loop systems are known in the controlled drug delivery field as self-regulated. The release rate in self-regulated devices is controlled by feedback information without any external intervention, as shown in Fig. 1b. Self-regulated systems use several approaches for rate control mechanisms (5,6) such as pH-sensitive polymers, enzyme–substrate reactions, pH-sensitive drug solubility, competitive binding, antibody interactions, and metal concentration-dependent hydrolysis.

Many approaches for mimicking the physiological healthy state are undergoing research. The focus of this article is on “smart” polymers; therefore, other important areas, such as using pumps for controlled drug delivery, microencapsulation of living cells, or gene therapy, are not covered. An additional area of significant research that is not covered in this article is site-directed or targeted drug delivery, where the release is constant (as in chemotherapy

for cancer treatment). Here, the fundamental principles and recent advances of *responsive* drug delivery for both pulsatile and self-regulated systems are reviewed.

PULSATILE SYSTEMS

Magnetically Stimulated Systems

Feasibility. Drug molecules and magnetic beads are uniformly distributed within a solid polymeric matrix in magnetically triggered systems. Although drug is released by diffusion when the device is exposed to fluids, a much higher release rate is obtained in the presence of an external oscillating magnetic field. The magnetic system was characterized *in vitro* (7–9). Subsequent *in vivo* (10) studies showed that when polymeric matrices made of ethylenevinyl acetate copolymer (EVAc) that contain insulin and magnetic beads are placed subcutaneously in diabetic rats for two months, glucose levels can be repeatedly and reproducibly decreased on demand by applying an oscillating magnetic field.

Mechanisms. The two principal parameters that control the release rates in these systems are the magnetic field characteristics and the mechanical properties of the polymer matrix. It was found that when the frequency of the applied field was increased from 5 to 11 Hz, the release rate of bovine serum albumin (BSA) from EVAc copolymer matrices rose linearly (7). Saslavski et al. (11) investigated the effect of magnetic field frequency and repeated field application on insulin release from alginate matrices and found that using repeated applications, inverse effects can occur: high frequencies gave a significant release enhancement for the second magnetic field application. Subsequent stimulation resulted in decreased enhancement due to faster depletion at high frequencies.

The mechanical properties of the polymeric matrix also affect the extent of magnetic enhancement (7). For example, the modulus of elasticity of the EVAc copolymer can easily be altered by changing the vinyl acetate content of the copolymer. The release rate enhancement induced by the magnetic field increases as the modulus of elasticity of EVAc decreases. A similar phenomenon was observed for cross-linked alginate matrices: higher release rate enhancement for less rigid matrices (11). Edelman et al. (12) also showed that enhanced release rates observed in response to an electromagnetic field (50 G, 60 Hz) applied for 4 minutes were independent of the duration of the interval between repeated pulses.

Ultrasonically Stimulated Systems

Feasibility. Release rates of substances can be repeatedly modulated at will from a position external to the delivery system by ultrasonic irradiation (13). Both bioerodible and nonerodible polymers were used as drug carrier matrices.

The bioerodible polymers evaluated were polyglycolide, polylactide, poly(bis(*p*-carboxyphenoxy) alkane

anhydrides and their copolymers with sebacic acid. Both the polymer erosion and drug release rates were enhanced when the bioerodible samples were exposed to ultrasound. The system's response to ultrasonic triggering was rapid (within 2 minutes) and reversible. The releasing agents, *p*-nitroaniline, *p*-aminohippurate, bovine serum albumin, and insulin, were tested for integrity following exposure to ultrasonic energy and were found intact.

The enhanced release was also observed in nonerodible systems exposed to ultrasound where the release is diffusion-dependent. Release rates of zinc bovine insulin from EVAc copolymer matrices were 15 times higher when exposed to ultrasound compared to the unexposed periods.

In vivo studies (13) have suggested the feasibility of ultrasound-mediated drug release enhancement. Implants composed of polyanhydride polymers loaded with 10% *p*-aminohippuric acid (PAH) were implanted subcutaneously in the backs of catheterized rats. When exposed to ultrasound, a significant increase in the PAH concentration in urine was detected (400%). Rat's skin histopathology of the ultrasound-treated area after an exposure of 1 hour at 5 W/cm² did not reveal any differences between treated and untreated skin.

Similar phenomena were observed by Miyazaki et al. (14) who evaluated the effect of ultrasound (1 MHz) on the release rates of insulin from ethylene vinyl alcohol copolymer matrices and reservoir type drug delivery systems. When diabetic rats that received implants containing insulin were exposed to ultrasound (1 W/cm² for 30 min), a sharp drop in blood glucose levels was observed after the irradiation, indicating a rapid rate of release of insulin at the implanted site.

During the past 40 years, numerous clinical reports have been published concerning phonophoresis (15), the technique of using ultrasonic irradiation to enhance transdermal drug delivery. Ultrasound nearly completely eliminated the usual lag time for transdermal delivery of drugs. Ultrasound irradiation (1.5 W/cm² continuous wave or 3 W/cm² pulsed wave) for 3–5 minutes increased the transdermal permeation of insulin and mannitol in rats by 5–20-fold within 1–2 hours after ultrasound application.

Miyazaki et al. (16) performed similar studies that evaluated the effect of ultrasound (1 MHz) on indomethacin permeation in rats. Pronounced effects of ultrasound on transdermal absorption for all three ranges of intensities (0.25, 0.5, and 0.75 W/cm²) were observed. Bommannan et al. (17) examined the effects of ultrasound on the transdermal permeation of the electron-dense tracer, lanthanum nitrate, and demonstrated that exposure of the skin to ultrasound can induce considerable and rapid tracer transport through an intercellular route. Prolonged exposure of the skin to high-frequency ultrasound (20 min, 16 MHz), however, resulted in structural alterations of epidermal morphology. Tachibana et al. (18–20) reported using low-frequency ultrasound (48 KHz) to enhance transdermal transport of lidocaine and insulin through hairless mice skin. Low-frequency ultrasound was also used by Mitragotri et al. (21,22) to enhance transport of various low molecular weight drugs, including salicylic

acid and corticosterone, as well as high molecular weight proteins, including insulin, γ -interferon, and erythropoietin, through human skin *in vitro* and *in vivo*.

Mechanisms. It was proposed (13) that cavitation and acoustic streaming are responsible for the augmented degradation and release of bioerodible polymers. In experiments conducted in a degassed buffer where cavitation was minimized, the observed enhancement in degradation and release rates was much smaller. It was also considered that several other parameters (temperature and mixing effects) might be responsible for the augmented release due to ultrasound. However, experiments were conducted which suggested that these parameters were not significant. It has also been demonstrated that the extent of release rate enhancement can be regulated by the intensity, frequency, or duty cycle of the ultrasound.

Miyazaki et al. (14) speculate that the ultrasound caused increased temperatures in their delivery system, which may facilitate diffusion. The increased temperature caused by ultrasound or other forms of irradiation can be used as a trigger to cause collapsing of a hybrid hydrogel that has protein domains, as described by Wang et al. (23). Additional thermostimulated polymers are discussed later in the temperature-sensitive section of self-regulated systems.

Mitragotri et al. (24) evaluated the role played by various ultrasound-related phenomena, including cavitation, thermal effects, generation of convective velocities, and mechanical effects during phonophoresis. The authors' experimental findings suggest that among all the ultrasound-related phenomena evaluated, cavitation plays the dominant role in sonophoresis using therapeutic ultrasound (frequency: 1–3 MHz; intensity; 0–2 W/cm²). Confocal microscopy results indicate that cavitation occurs in the keratinocytes of the stratum corneum upon ultrasound exposure. The authors hypothesized that oscillations of the cavitation bubbles induce disorder in the stratum corneum lipid bilayers, thereby enhancing transdermal transport. The theoretical model developed to describe the effect of ultrasound on transdermal transport predicts that sonophoretic enhancement depends most directly on the passive permeant diffusion coefficient in water, not on the permeant diffusion coefficient through the skin.

Electrically Stimulated Systems

Feasibility. Electrically controlled systems provide drug release by the action of an applied electric field on a rate-limiting membrane and/or directly on the solute and thus control its transport across the membrane. The electrophoretic migration of a charged macromolecule within a hydrated membrane results from the combined response to the electrical forces on the solute and its associated counterions in the adjacent electrolyte solution (25).

Electrically controlled membrane permeability has also been of interest in the field of electrically controlled or enhanced transdermal drug delivery (e.g., iontophoresis, electroporation) (26,27).

Anionic gels as vehicles for electrically modulated drug delivery were studied by Hsu and Block (28). Agarose and

combinations of agarose and anionic polymers (polyacrylic acid, xanthan gum) were evaluated. The authors conclude that the use of carbomer (polyacrylic acid) in conjunction with agarose enables the formulator to achieve zero-order release by electrical field application. Increased anisotropy of a gel system due to the application of electrical current could alter the effectiveness of the drug delivery system.

D'Emanuele and Staniforth (29) proposed a drug delivery device that consists of a polymer reservoir that has a pair of electrodes placed across the rate-limiting membrane. By altering the magnitude of the electric field between the electrodes, the authors proposed to modulate the drug release rates in a controlled and predictable manner. A linear relationship was found between current and propranolol HCL permeability through poly(2-hydroxyethyl methacrylate) (PHEMA) membranes cross-linked with ethylene glycol dimethacrylate (1%v/v). It was found that buffer ionic strength, drug reservoir concentration, and electrode polarity have significant effects on drug permeability (30).

Labhassetwar et al. (31) propose a similar approach for modulating cardiac drug delivery. The authors studied a cardiac drug implant in dogs that can modulate electric current. A cation-exchange membrane was used as an electrically sensitive rate-limiting barrier on the cardiac-contacting surface of the implant. The cardiac implant demonstrated *in vitro* drug release rates that were responsive to current modulation. *In vivo* results in dogs confirmed that electrical modulation resulted in regional coronary enhancement of drug levels and a current-responsive increase in drug concentration.

A different approach for electrochemical controlled release is based on polymers that bind and release bioactive compounds in response to an electric signal (32). The polymer has two redox states, only one of which is suitable for ion binding. Drug ions are bound in one redox state and released from the other. The attached electrodes switch the redox states, and the amount of current passed can control the amount of ions released. Hepel and Fijalek (33) propose to use this method of electrochemical pulse stimulation on a novel composite polypyrrole film for delivering cationic drugs directly to the central nervous system (CNS).

By encapsulating drugs in multicomponent hydrogel microspheres, Kiser et al. (34) propose a synthetic mimic of the secretory granule that can be triggered to release the bioactive agent by various forms of external stimulation. In the report, the external protective lipid membrane was porated by electrical stimulation. Following electroporation, the hydrogel microsphere quickly swells to dissipate the pH gradient. The swelling leads to a burst of drug release. Thus, an off/on irreversible mechanism is described that can be triggered in a controlled fashion.

Mechanisms. Grimshaw (35) reported four different mechanisms for the transport of proteins and neutral solutes across hydrogel membranes: (1) electrically and chemically induced swelling of a membrane to alter the effective pore size and permeability, (2) electrophoretic augmentation of solute flux within a membrane, (3)

electrosmotic augmentation of solute flux within a membrane, and (4) electrostatic partitioning of charged solutes into charged membranes.

Kwon et al. (36) studied the effect of electric current on solute release from cross-linked poly(2-acrylamido-2-methylpropane sulfonic acid-*co-n*-butylmethacrylate). Edrophonium chloride, a positively charged solute, was released in an on-off pattern from a matrix (monolithic) device by an electric field. The mechanism was explained as an ion exchange between a positive solute and the hydroxonium ion, followed by fast release of the charged solute from the hydrogel. The fast release was attributed to electrostatic force, a squeezing effect, and electro-osmosis of the gel. However, the release of neutral solute was controlled by diffusion effected by swelling and deswelling of the gel.

Photostimulated Systems

Feasibility. Photoinduced phase transition of gels was reported by Mamada et al. (37). Copolymer gels of *N*-isopropylacrylamide and the photosensitive molecule bis(4-dimethylamino)phenyl(4-vinylphenyl)methyl leucocyanide showed a discontinuous volume phase transition upon ultraviolet irradiation that was caused by osmotic pressure of cyanide ions created by the ultraviolet irradiation.

Yui et al. (38) proposed photoresponsive degradation of heterogeneous hydrogels comprised of cross-linked hyaluronic acid and lipid microspheres for temporal drug delivery. Visible light induced degradation of cross-linked hyaluronic acid gels by photochemical oxidation using methylene blue as the photosensitizer. [The authors also proposed that hyaluronic acid gels are inflammation-responsive (39)].

By combining technologies developed for targeted drug delivery and external photostimulation of the active agent released, Taillefer et al. (40) propose using polymeric micelles to deliver water-insoluble, photosensitizing anticancer drugs.

Mechanisms. Photoresponsive gels reversibly change their physical or chemical properties upon photoradiation. A photoresponsive polymer consists of a photoreceptor, usually a photochromic chromophore, and a functional part. The optical signal is captured by the photochromic molecules, and then isomerization of the chromophores in the photoreceptor converts it to a chemical signal.

Suzuki and Tanaka (41) reported a phase transition in polymer gels induced by visible light, where the transition mechanism is due only to the direct heating of the network polymer by light.

SELF-REGULATED SYSTEMS

Environmentally Responsive Systems

Polymers that alter their characteristics in response to changes in their environment have been of great recent interest. Several research groups have been developing drug delivery systems based on these responsive polymers

that more closely resemble the normal physiological process. Drug delivery in these devices is regulated by an interaction with the surrounding environment (feedback information) without any external intervention. The most commonly studied polymers that have environmental sensitivity are either pH- or temperature-sensitive. There are also inflammation-sensitive systems and systems that use specific binding interactions, which are discussed in the next section.

Temperature-Sensitive Systems. Temperature-sensitive polymers can be classified into two groups based on the origin of the thermosensitivity in aqueous media. The first is based on polymer–water interactions, especially, specific hydrophobic/hydrophilic balancing effects and the configuration of side groups. The other is based on polymer–polymer interactions in addition to polymer–water interactions. When polymer networks swell in a solvent, there is usually a negligible or small positive enthalpy of mixing or dilution. Although a positive enthalpy change opposes the process, the large gain in entropy drives it. The opposite is often observed in aqueous polymer solutions. This unusual behavior is associated with a phenomenon of polymer phase separation as the temperature is raised to a critical value that is known as the lower critical solution temperature (LCST). *N*-Alkyl acrylamide homopolymers and their copolymers, including acidic or basic comonomers, show this LCST (42,43). Polymers characterized by LCST usually shrink as the temperature is increased through the LCST. Lowering the temperature below the LCST results in swelling of the polymer. Bioactive agents such as drugs, enzymes, and antibodies may be immobilized on or within temperature-sensitive polymers; examples of such uses are discussed below. Responsive drug release patterns regulated by external temperature changes have been recently demonstrated by several groups (42,44–57).

pH-Sensitive Systems. The pH range of fluids in various segments of the gastrointestinal tract may provide environmental stimuli for responsive drug release. Several research groups (58–72) studied polymers that contain weakly acidic or basic groups in the polymeric backbone. The charge density of the polymers depends on the pH and ionic composition of the outer solution (the solution to which the polymer is exposed). Altering the pH of the solution causes swelling or deswelling of the polymer. Thus, drug release from devices made from these polymers display release rates that are pH-dependent. Polyacidic polymers are unswollen at low pH because the acidic groups are protonated and hence un-ionized. Polyacid polymers swell as the pH increases. The opposite holds for polybasic polymers because ionization of the basic groups increases as the pH decreases. Siegel et al. (73) found that the swelling properties of polybasic gels are also influenced by buffer composition (concentration and pKa). A practical consequence proposed is that these gels may not reliably mediate pH-sensitive, swelling-controlled release in oral applications because the levels of buffer acids in the stomach (where swelling and release are expected) generally cannot be controlled. However, the gels may be

useful as mediators of pH-triggered release when precise rate control is of secondary importance.

Annaka and Tanaka (59) reported that more than two phases (swollen and collapsed) are found in gels that consist of copolymers of randomly distributed positively and negatively charged groups. Polymer segments in these gels interact with each other through attractive or repulsive electrostatic interactions and through hydrogen bonding. The combination of these forces seems to result in the existence of several phases, each characterized by a distinct degree of swelling, and abrupt jumps occur between them. The existence of these phases presumably reflects the ability of macromolecular systems to adopt different stable conformations in response to changes in environmental conditions. The largest number of phases was seven in copolymer gels prepared from acrylic acid (the anionic constituent) and methacryl-amido-propyl-trimethyl ammonium chloride (460 mmol/240 mmol). A similar approach was proposed by Bell and Peppas (69); membranes made from grafted poly (methacrylic acid-glycol) copolymer showed pH sensitivity due to complex formation and dissociation. Uncomplexed equilibrium swelling ratios were 40 to 90 times higher than those of complexed states and varied according to copolymer composition and polyethylene glycol graft length.

Giannos et al. (74) proposed temporally controlled drug delivery systems that couple pH oscillators and membrane diffusion properties. By changing the pH of a solution relative to the pKa, a drug may be rendered charged or uncharged. Because only the uncharged form of a drug can permeate across lipophilic membranes, a temporally modulated delivery profile may be obtained by using a pH oscillator in the donor solution.

Heller and Trescony (75) were the first to propose using pH-sensitive bioerodible polymers. In their approach, described in the section on systems that utilize enzymes, an enzyme–substrate reaction produces a pH change that is used to modulate the erosion of a pH-sensitive polymer that contains a dispersed therapeutic agent.

Bioerodible hydrogels that contain azoaromatic moieties were synthesized by Ghandehari et al. (76). Hydrogels that have lower cross-linking density underwent a surface erosion process and degraded at a faster rate. Hydrogels that have higher cross-linking densities degraded at a slower rate by a process in which the degradation front moved inward to the center of the polymer.

Recently, recombinant DNA methods were used to create artificial proteins that undergo reversible gelation in response to changes in pH or temperature (77). The proteins consist of terminal leucine zipper domains that flank a central, flexible, water-soluble polyelectrolyte segment. Formation of coiled-coil aggregates of the terminal domains in near-neutral aqueous solutions triggers formation of a three-dimensional polymer network, where the polyelectrolyte segment retains solvent and prevents precipitation of the chain. Dissociation of the coiled-coil aggregates by elevating pH or temperature causes dissolution of the gel and a return to the viscous behavior that is characteristic of polymer solutions. The authors suggest that these hydrogels have potential in bioengineering

applications that require encapsulation or controlled release of molecules and cellular species.

Inflammation-Responsive Systems. Yui et al. (39) proposed an inflammation-responsive drug delivery system based on biodegradable hydrogels of cross-linked hyaluronic acid. Hyaluronic acid is specifically degraded by hydroxyl radicals that are produced locally at inflammatory sites by phagocytic cells such as leukocytes and macrophages. In their approach, drug-loaded lipid microspheres were dispersed into degradable matrices of cross-linked hyaluronic acid.

Brown et al. (78) developed a biodegradable, biocompatible, inflammation-responsive microsphere system. The gelatin microspheres were synthesized by complex coacervation, a low temperature method that does not denature the encapsulated active agent. Gelatinase and stromelysin are activated in the synovial fluid of an inflamed joint. These enzymes degrade the gelatin microspheres and thus cause release of the bioactive protein, making this delivery system potentially useful for treating osteoarthritis.

An infection-responsive delivery system was developed by Tanihara et al. (79). As in an inflammatory response, inflection responses are characterized by the secretion of specific proteins. By responding to thrombin-like activity in infected wound fluid, the novel system released gentimycin as needed, thus avoiding problematic overexposure to antibiotics.

Systems Using Specific Binding Interactions

All of the following drug delivery systems use a specific binding interaction to manipulate the microenvironment of the device and thus modulate the rate of drug release from the polymer. The basic principles of binding and competitive binding are the underlying mechanism of the function of these systems. Due to the vast amount of literature on the subject of glucose-responsive insulin delivery systems, they are discussed in a separate section.

Systems Using Antibody Interactions. Pitt et al. (80) proposed utilizing hapten-antibody interactions to suppress the enzymatic degradation and permeability of polymeric reservoirs or matrix drug delivery systems. The delivery device consists of naltrexone contained in a polymeric reservoir or dispersed in a polymeric matrix configuration. The device is coated by covalently grafting morphine to the surface. Exposure of the grafted surface to antibodies to morphine results in coating of the surface by the antibodies, a process that can be reversed by exposure to exogenous morphine. Antibodies on the surface or in the pores of the delivery device block or impede the permeability of naltrexone in a reservoir configuration or enzyme-catalyzed surface degradation and the concomitant release of the drug from a matrix device. A similar approach was proposed for responsive release of a contraceptive agent. The β subunit of human chorionic gonadotropin (HCG) is grafted to the surface of a polymer, which is then exposed to antibodies to β -HCG. The appearance of HCG in the circulatory system (indication of pregnancy) causes release of a contraceptive drug. (HCG competes for the polymer-bound

antibodies to HCG and initiates release of the contraceptive drug.)

Pitt et al. (80,81) also proposed a hypothetical reversible antibody system for controlled release of ethinyl estradiol (EE). EE stimulates biosynthesis of sex-hormone-binding globulin (SHBG). High serum levels of EE stimulate the production of SHBG, which increases the concentration of SHBG bound to the polymer surface and reduces the EE release rate. When the EE serum level falls, the SHBG level falls, as does binding of the SHBG to the polymer surface, which produces an automatic increase in the EE release rate.

The reversible binding of antigen to antibody that is the basis for swelling of a hydrogel that could lead to release of a bioactive agent was recently reported (82). Miyata et al. describe the grafting of both antigen and antibody in the polymer network that causes the formation of reversible cross-linking. In the presence of free antigen that competes with the immobilized antigen, swelling ensues (82) and creates an antigen-responsive hydrogel.

Systems Using Chelation. Self-regulated delivery of drugs that function by chelation was also suggested (83). These include certain antibiotics and drugs for treating arthritis, as well as chelators used for treating metal poisoning. The concept is based on the ability of metals to accelerate the hydrolysis of carboxylate or phosphate esters and amides by several orders of magnitude. Attachment of the chelator to a polymer chain by a covalent ester or amide link prevents premature loss by excretion and reduces its toxicity. In the presence of the specific ion, a complex with the bound chelating agent forms, followed by metal-accelerated hydrolysis and subsequent elimination of the chelated metal. Measurement of the rates of hydrolysis of polyvinyl alcohol coupled with quinaldic acid chelator (PVA-QA) in the presence of Co(II), Zn(II), Cu(II), and Ni(II) confirmed that it is possible to retain the susceptibility of the esters to metal-promoted hydrolysis in a polymer environment.

Recently, Goldbart and Kost (84) reported developing a calcium-responsive drug delivery system. Calcium in external media reactivates α -amylase that was immobilized after being reversibly inactivated in a starch matrix. The activated enzyme causes degradation of the matrix, thus releasing an entrapped active agent. These investigators also developed a compartmental mathematical model that describes the release and degradation mechanisms involved (85).

Systems Using Enzymes

In this approach, the mechanism is based on an enzymatic reaction. One possible approach studied is an enzymatic reaction that results in a pH change and a polymer system that can respond to that change.

Urea-Responsive Delivery. Heller et al. (75) were the first to attempt using immobilized enzymes to alter local pH and thus cause changes in polymer erosion rates. The proposed system is based on converting urea to NH_4HCO_3 and NH_4OH by the action of urease. Because this reaction

causes a pH increase, a polymer that is subjected to increased erosion at high pH is required.

The authors suggested a partially esterified copolymer of methyl vinyl ether and maleic anhydride. This polymer displays release rates that are pH-dependent. The polymer dissolves by ionizing the carboxylic acid group. The pH-sensitive polymer that contains dispersed hydrocortisone is surrounded by urease immobilized in a hydrogel that is prepared by cross-linking a mixture of urease and BSA with glutaraldehyde. When urea diffuses into the hydrogel, its interaction with the enzyme leads to a pH increase, therefore, resulting in enhanced erosion of the pH-sensitive polymer and concomitant increases in the release rate of hydrocortisone.

Ishihara et al. (86,87) suggested a nonerodible system based on a similar idea. The system is comprised of a pH-sensitive membrane, produced by copolymerizing 4-carboxy acrylanilide with methacrylate, sandwiched within a membrane that contains urease immobilized in free radically cross-linked *N,N*-methylenebisacrylamide. The permeation of a model substance, (1,4-bis(2-hydroxyethoxy) benzene, varied with the urea concentration in the external solution.

Morphine Triggered Naltrexone Delivery System. Heller and co-workers (88–95) have been developing a naltrexone drug delivery system that would be passive until drug release is initiated by the appearance of morphine external to the device. Naltrexone is a long acting opiate antagonist that blocks opiate-induced euphoria, and thus the intended use of this device is to treat heroin addiction. Activation is based on the reversible inactivation of enzymes achieved by the covalent attachment of hapten close to the active site of the enzyme–hapten conjugate with the hapten antibody. Because the antibodies are large molecules, access of the substrate to the enzyme's active site is sterically inhibited and thus effectively renders the enzyme inactive. Triggering of drug release is initiated by the appearance of morphine (hapten) in the tissue and dissociation of the enzyme–hapten–antibody complex that renders the enzyme active. This approach is being developed by incorporating the naltrexone in a bioerodible polymer. The polymer matrix is then covered by a lipid layer that prevents water entry, and this prevents its degradation and therefore also the release of naltrexone. The system is placed in a dialysis bag. The bag contains lipase (enzyme) that is covalently attached to morphine and reversibly inactivated by antimorphine complexation. Thus, when morphine is present in the tissues that surround the device, morphine diffuses into the dialysis bag, displaces the lipase-morphine conjugate from the antibody, and allows the now activated enzyme to degrade the protective lipid layer. This in turn permits degradation of the polymeric core and subsequent release into the body of the narcotic antagonist, naltrexone.

A key component of this morphine-responsive device is the ability to inactivate an enzyme reversibly and completely and to disassociate the complex rapidly using concentrations as low as 10^{-8} to 10^{-9} M. To achieve this sensitivity, lipase was conjugated with several morphine analogs and complexed with polyclonal antimorphine

antibodies purified by affinity chromatography. *In vivo* studies (91) suggest that the concentration of morphine in a device implanted in a typical heroin-addicted patient is estimated at about 10^{-7} to 10^{-8} M. Recent studies have shown that reaching such sensitivity is possible (89).

Many of the glucose-responsive systems discussed in the next section also use enzymes.

Glucose-Responsive Insulin Delivery

The development of glucose-sensitive insulin-delivery systems has used several approaches, including immobilized glucose oxidase in pH-sensitive polymers, competitive binding, and a polymer–complex system. None of the present modes of treatment, including insulin pumps, fully mimics the physiology of insulin secretion. Therefore the development of a “smart” insulin-delivery system could significantly help patients who have diabetes to control their blood glucose level and thus avoid the various severe complications including eye disease, gangrene of the extremities, cardiovascular disease, and renal failure (96).

Polymer–Complex System. Kitano et al. (97) proposed a glucose-sensitive insulin release system based on a sol–gel transition. A phenylboronic acid (PBA) moiety was incorporated in poly(*N*-vinyl-2-pyrrolidone) by the radical copolymerization of *N*-vinyl-2-pyrrolidone with *m*-acrylamidophenylboronic acid [poly(NVP-*co*-PBA)]. Insulin was incorporated into a polymer gel formed by a complex of poly(vinyl alcohol) with poly(NVP-*co*-PBA). PBA can form reversible covalent complexes with molecules that have diol units, such as glucose or PVA. By adding glucose, PVA in the PVA–boronate complex is replaced by glucose. This leads to a transformation of the system from the gel to the sol state that facilitates the release of insulin from the polymeric complex. The same group (98) modified the approach and suggested glucose-responsive gels based on complexation between polymers that have phenylboronic acid groups and PVA. The introduction of an amino group into phenylborate polymers was effective in increasing the complexation ability and the glucose responsivity at physiological pH.

Shiino et al. (99) attached gluconic acids to insulin. The modified insulin that contains two gluconic acid units per insulin (G-Ins) was bound into a PBA gel column, and the G-Ins release profile in response to varying concentrations of glucose was studied. Concentration of released G-Ins from PBA gel responded to concentration changes of the eluting glucose. These polymeric complexes have been applied as interpenetrating polymer networks to achieve pulsatile insulin release in response to changes in glucose concentration.

Competitive Binding. The basic principle of competitive binding and its application to controlled drug delivery was first presented by Brownlee and Cerami (100) who suggested the preparation of glycosylated insulins that are complementary to the major combining site of carbohydrate binding proteins such as Concanavalin A (Con A). Con A is immobilized on Sepharose™ beads. The glycosylated insulin, which is biologically active, is displaced from the Con A by glucose in response to, and proportional to, the

amount of glucose present that competes for the same binding sites. Kim et al. (101–108) found that the release rate of insulin also depends on the binding affinity of an insulin derivative to Con A and can be influenced by the choice of the saccharide group in glycosylated insulin. By encapsulating the glycosylated insulin-bound Con A by using a suitable polymer that is permeable to both glucose and insulin, the glucose influx and insulin efflux would be controlled by the encapsulation membrane.

It was found (102) that glycosylated insulins are more stable to aggregation than commercial insulin and are also biologically active. The functionality of the intraperitoneally implanted device was tested in pancreatectomized dogs by an intravenous glucose tolerance test (IVGTT). The effect of an administered 500 mg/kg dextrose bolus on blood glucose level was compared with normal and pancreatectomized dogs without an implant. The results of this study indicated that the diabetic dogs that had the implant had normal glucose levels (107). In addition, the blood glucose profile for a period of 2 days demonstrated that a diabetic dog, implanted with the self-regulating insulin delivery system, could maintain acceptable glucose levels (50–180 mg/dL) for the majority of the experiment (40 hours) (104–106). Makino et al. (101) proposed a modification based on hydrophilic nylon microcapsules that contained Con A and succinyl-amidophenyl-glucopyranoside insulin. The thin wall of these microcapsules and large surface area resulted in rapid diffusion of glucose and glycosylated insulin and therefore, a much shorter lag time.

To limit the leakage of Con A (which is toxic) and allow preparation of porous microspheres, Pai et al. (109) cross-linked the Con A by first blocking the sugar binding sites and then reacted it with glutaraldehyde. The porous microspheres demonstrated rapid exchange between succinyl-amidophenyl-glucopyranoside insulin and glucose, and had a short response time.

Kokufata et al. (110) reported a gel system that swells and shrinks in response to specific saccharides. The gel consists of a covalently cross-linked polymer network of *N*-isopropylacrylamide in which the lectin, Con A, is immobilized. Con A displays selective binding affinities for certain saccharides. For example, when the saccharide dextran sulphate is added to the gel, it swells to a volume up to fivefold the original volume. Replacing dextran sulphate with nonionic saccharide α -methyl-D-mannopyranoside brings about collapse of the gel, almost to its native volume. The process is reversible and repeatable.

Taylor et al. (111) proposed a similar approach for delivering insulin. It was shown that a self-regulating delivery device, responsive to glucose, operates *in vitro*. The device comprises a reservoir of insulin and a gel membrane that determines the delivery rates of insulin. The gel consists of a synthetic polysucrose and the lec, Con A. The mechanism is one of displacing the branched polysaccharide from the lec receptors by incoming glucose. The gel loses its high viscosity as a result but reforms upon removal of glucose and thus provides the rate-controlling barrier to the diffusion of insulin or any other antihyperglycemic drugs.

A similar approach was presented by Park et al. (112–114) who synthesized glucose-sensitive membranes based on the interaction between polymer-bound glucose and Con A.

Immobilized Glucose Oxidase in pH-Sensitive Polymers.

Responsive drug delivery systems based on pH-sensitive polymers have been developed along three different approaches: pH-dependent swelling, degradation, and solubility.

pH-Dependent Solubility. Glucose-dependent insulin release was proposed by Langer and co-workers (115–117) based on the fact that insulin solubility is pH-dependent. Insulin was incorporated into ethylene vinyl acetate (EVAc) copolymer matrices in solid form. Thus, the release was governed by its dissolution and diffusion rates. Glucose oxidase was immobilized to Sepharose™ beads which were incorporated along with insulin into EVAc matrices. When glucose entered the matrix, the gluconic acid produced caused a rise in insulin solubility and consequently enhanced release. To establish this mechanism at the physiological pH of 7.4, the insulin was modified by three additional lysine groups so that the resultant isoelectric point was 7.4. *In vitro* and *in vivo* studies demonstrated the response of the system to changes in glucose concentration. In the *in vivo* experiments, a catheter was inserted into the left jugular vein, and polymer matrices that contained insulin and immobilized enzyme were implanted subcutaneously in the lower back of diabetic rats. Serum insulin concentrations were measured for different insulin matrix implants. A 2 M glucose solution was infused, 15 minutes into the experiments, through the catheter. Rats that received trilycine insulin/glucose oxidase matrices showed a 180% rise in serum insulin concentration which peaked at 45 minutes into the experiment. Control rats that received matrices that contained no insulin, or insulin but no glucose oxidase, or diabetic rats without implants showed no change in serum insulin.

pH-Dependent Degradation. Heller et al. (88,89,118) suggested a system in which insulin is immobilized in a pH-sensitive bioerodible polymer prepared from 3,9-bis-(ethylidene 2,4,8,10-tetraoxaspiro(5,5)undecane and *N*-methyl-diethanolamine), which is surrounded by a hydrogel that contains immobilized glucose oxidase. When glucose diffuses into the hydrogel and is oxidized to gluconic acid, the resultant lowered pH triggers enhanced polymer degradation and release of insulin from the polymer in proportion to the concentration of glucose. The response of the pH-sensitive polymers that contained insulin to pH pulses was rapid. Insulin was rapidly released when the pH decreased from 7.4 to 5.0. Insulin release was shut off when the pH increased. The amount of insulin released showed dependence on pH change. However, when the *in vitro* studies were repeated in a physiological buffer, the response of the device was only minimal, even at very low pH pulses. The authors found that the synthesized amine-containing polymer undergoes general acid catalysis and the catalyzing species is not the hydronium ion but rather the specific buffer molecules used. Therefore, further development of this system will require developing a bioerodible polymer that has adequate pH sensitivity and also undergoes specific ion catalysis.

pH-Dependent Swelling. Systems based on pH-sensitive polymers consist of immobilized glucose oxidase in a pH-responsive hydrogel that encloses a saturated insulin solution or is incorporated with insulin (119–129). As glucose diffuses into the hydrogel, glucose oxidase catalyzes its conversion to gluconic acid, thereby lowers the pH in the microenvironment of the hydrogel, and causes swelling. Because insulin should permeate the swelled hydrogel more rapidly, faster delivery of insulin in the presence of glucose is anticipated. As the glucose concentration decreases in response to the released insulin, the hydrogel should contract and decrease the rate of insulin delivery.

Horbett and co-workers (119–126) immobilized glucose oxidase in a cross-linked hydrogel made from *N,N*-dimethylaminoethyl methacrylate (DMA), hydroxyethyl methacrylate (HEMA), and tetraethylene glycol dimethacrylate (TEGDMA). It was previously shown that membranes prepared at -70°C by radiation polymerization retain enzymatic activity (130). To obtain sufficient insulin permeability through the gels, porous HEMA/DMA gels were prepared by polymerization under conditions which induce a separation into two phases during polymerization: one phase is rich in polymer, and the other is rich in solvent plus unreacted monomer. When gelation occurs after phase separation, the areas where the solvent/monomer phase existed become fixed in place as pores in the polymer matrix. The authors used a dilute monomer solution to obtain a porous gel, whose pores were typically 1–10 μm in diameter (106).

The rate of insulin permeation through the membranes was measured in the absence of glucose in a standard transport cell; then glucose was added to one side of the cell to a concentration of 400 mg/dL, and the permeation measurement was continued. The results indicated that the insulin transport rate is enhanced significantly by the addition of glucose. The average permeability after addition of 400 mg/dL glucose was 2.4 to 5.5 times higher than before glucose was added. When insulin permeabilities through the porous gels were measured in a flowing system, where permeabilities were measured as fluid flowed continuously past one side of the membrane, no effect of glucose concentration on insulin permeabilities could be detected. The authors propose that inappropriate design of the membranes used in the experiments is the explanation for their lack of response to glucose concentration (105).

A mathematical model that describes these glucose-responsive hydrogels demonstrates two important points (119,120): (1) Progressive response to glucose concentration over a range of glucose concentrations can be achieved only by using a sufficiently low glucose oxidase loading; otherwise, depletion of oxygen makes the system insensitive to glucose. (2) A significant pH decrease in the membrane and resultant swelling can be achieved only if the amine concentration is sufficiently low that pH changes are not prevented by the buffering of the amines.

The great advantage of reservoir systems is the ease by which they can be designed to produce constant release rate kinetics, but their main disadvantage is leaks that are dangerous because all of the incorporated drug could be released rapidly. To overcome this problem, Goldraich and Traitel (127,128) proposed incorporating the drug (insulin) and the enzyme (glucose oxidase) into the pH-responsive

polymeric matrices. Furthermore, Traitel et al. (129) developed a compartmental math model that describes the pH-responsive swelling of the matrix and can be used to optimize the system further.

Ishihara et al. (86,87,131–134) investigated two approaches to glucose-responsive insulin delivery systems: One approach is similar to that investigated by Horbett et al. (125). The polymers were prepared from 2-hydroxyethyl acrylate (HEA)-*N,N*-dimethylaminoethyl methacrylate (DMA), 4-trimethylsilylstyrene (TMS), by radical polymerization of the corresponding monomers in dimethylformamide (DMF). The mole fractions of HEA, DMA, and TMS in the copolymer were 0.6, 0.2, and 0.2, respectively. Membranes were prepared by solvent casting. Capsules that contained insulin and glucose oxidase were prepared by interfacial precipitation using gelatin as an emulsion stabilizer. The average diameter of the polymer capsules obtained was 1.5 mm (131,134). The water content of HEA–DMA–TMS copolymer membranes increased as the pH of the medium decreased. An especially drastic change was observed in the pH range of 6.15 to 6.3. The permeation of insulin through the copolymer membrane increases in response to pH decreases. The permeation rate of insulin at pH 6.1 was greater than that at pH 6.4 by about 42 times. The permeation of insulin through the copolymer membranes was very low in buffer solution without glucose. Adding 0.2 M glucose to the upstream compartment induced an increase in the permeation rate of insulin. When glucose was removed, the permeation rates of insulin gradually returned to their original levels (131).

Podual et al. recently reported a glucose-sensitive system that works on the same principle but utilizes a different polymer (135). The authors found that poly(diethyl aminoethyl methacrylate-*g*-ethylene glycol) that contained glucose oxidase and catalase resulted in matrices that were reproducibly and reversibly glucose-sensitive.

The other approach proposed by Ishihara et al. (133) is based on a glucose oxidase immobilized membrane and a redox polymer that has a nicotinamide moiety. The device consists of two membranes. One membrane that contains the immobilized glucose oxidase acts as a sensor for glucose and forms hydrogen peroxide by an enzymatic reaction; the other membrane is a redox polymer that has a nicotinamide moiety that controls the permeation of insulin by an oxidation reaction with the hydrogen peroxide formed. The oxidation of the nicotinamide group increases hydrophilicity and therefore should enhance the permeability to water-soluble molecules such as insulin. The results showed relatively small increases in insulin permeability.

Iwata et al. (136,137) pretreated porous poly(vinylidene fluoride) membranes (average pore size of 0.22 μm) by air plasma, and subsequently, acrylamide was graft polymerized on the treated surface. The polyacrylamide was then hydrolyzed to poly(acrylic acid). In the pH range of 5–7, grafted poly(acrylic acid) chains are solvated and dissolved but cannot diffuse into the solution phase because they are grafted to the porous membrane. Thus, they effectively close the membrane pores. In the pH range of 1 to 5, the chains collapse, and the permeability increases. To achieve sensitivity of the system to glucose, glucose oxidase was immobilized onto a poly(2-hydroxyethyl methacrylate) gel.

Ito et al. (138) adopted the approach proposed by Iwata et al. (137) using a porous cellulose membrane that had surface-grafted poly(acrylic acid) as a pH-sensitive membrane. By immobilizing glucose oxidase onto the poly(acrylic acid)-grafted cellulose membrane, it became responsive to glucose concentrations. The permeation coefficient after glucose addition was about 1.7 times that before the addition of glucose. The authors suggest improving the proposed system (sensitivity of insulin permeability to glucose concentrations) by modifying the graft chain density, length, and size, or density of pores.

Siegel and co-workers (139,140) proposed an implantable "mechanochemical" pump that functions by converting changes in blood glucose activity into a mechanical force, generated by the swelling polymer that pumps insulin out of the device.

More recently, Siegel (141) proposed self-regulating oscillatory drug delivery based on a polymeric membrane whose permeability to the substrate of an enzyme-catalyzed reaction is inhibited by the product of that reaction. This negative feedback system can, under certain conditions, lead to oscillations in membrane permeability and in the levels of substrate and product in the device. Any one of these oscillating variables can then be used to drive a cyclic delivery process. The product concentration in the chamber inhibitorily affects the permeability of the membrane to the substrate. That is, increasing product concentration causes decreasing flux of substrate into the device. Siegel proposed several means of controlled drug delivery based on this idea. Drug solubility could be affected by substrate or product concentration, which oscillates. Alternatively, the drug permeability of the membrane can oscillate with time along with the substrate permeability.

CONCLUDING REMARKS

During the last three decades, polymeric controlled drug delivery has become an important area of research and development. In this short time, a number of systems that display constant or decreasing release rates have progressed from the laboratory to the clinic and clinical products. Although these polymeric controlled delivery systems are advantageous compared to the conventional methods of drug administration, they are insensitive to a changing metabolic state. Responsive mechanisms must be provided to control the physiological requirements of specific drugs more closely. The approaches discussed represent attempts conducted during the past two decades to achieve pulsatile release. These drug delivery systems are still in the developmental stage, and much research will have to be conducted for such systems to become practical clinical alternatives. Critical considerations are the biocompatibility and toxicology of these multicomponent polymer-based systems, the response times of these systems to stimuli, the ability to provide practical levels of the desired drug, and addressing necessary formulation issues in dosage or design (e.g., shelf life, sterilization, reproducibility). A key issue in the practical use of pulsatile, externally triggered systems (i.e., magnetic, ultrasound, electrically regulated, and photoresponsive) will be the design of small portable trigger units that the patient can use easily. Ideally, such

systems could be worn by patients, wristwatch-like. They could either be preprogrammed to go on and off at specific times or patients could turn them on when needed. A critical issue in the development of responsive, self-regulated systems such as those that contain enzymes or antibodies are the stability and/or potential leakage and possible immunogenicity of these bioactive agents. The successful development of responsive polymer delivery systems is a significant challenge. Nevertheless, the considerable pharmacological benefit that these systems could provide, particularly in view of ongoing research in biotechnology, pharmacology, and medicine that may provide new insights into the desirability and requirements for pulsatile release, should make this an important and fruitful area for future research.

BIBLIOGRAPHY

1. S.W. Shalaby, A.S. Hoffman, B.D. Ratner, and T.A. Horbett, eds., *Polymers as Biomaterials*. Plenum Press, NY, 1984.
2. R. Langer, Drug delivery and targeting, *Nature* **392**: 5–10 (1998).
3. J. Kost and R. Langer, *Adv. Drug Delivery Rev.* **16**: 19–50 (1991).
4. B.D. Ratner, A.S. Hoffman, F.J. Schoen, and J.E. Lemons, eds., *Biomaterials Science: An Introduction to Materials in Medicine*. Academic Press, NY, 1996.
5. K. Dusek, ed., *Responsive Gels: Volume Transitions*, Vols. I and II. Springer-Verlag, 1993.
6. I.Y. Galaev and B. Mattiasson, *Trends Biotechnol.* **17**(8): 335–40 (1999).
7. J. Kost, R. Noecker, E. Kunica, and R. Langer, *J. Biomed. Mater. Res.* **19**: 935 (1986).
8. E. Edelman, J. Kost, H. Bobeck, and R. Langer, *J. Biomed. Mater. Res.* **67** (1985).
9. D. Hsieh, R. Langer, and J. Folkman, *Proc. Natl. Acad. Sci. USA* **78**: 1863 (1981).
10. J. Kost, J. Wolfrum, and R. Langer, *J. Biomed. Mater. Res.* **21**: 1367 (1987).
11. O. Saslavski, P. Couvrer, and N. Peppas, in *Controlled Release of Bioactive Materials*, J. Heller, F. Harris, H. Lohmann, H. Merkle, and J. Robinson, eds., Controlled Release Society, Basel, 1988, p. 26.
12. E. Edelman, L. Brown, J. Taylor, and R. Langer, *J. Biomed. Mater. Res.* **21**: 229 (1987).
13. J. Kost, K. Leong, and R. Langer, *Proc. Natl. Acad. Sci. USA* **86**: 7663–7666 (1989).
14. S. Miyazaki, C. Yokouchi, and M. Takada, *J. Pharm. Pharmacol.* **40**: 716–717 (1988).
15. J. Kost, in *Electronically Controlled Drug Delivery*, B. Berner and S. Dinh, eds., CRC, Boca Raton, 1998.
16. S. Miyazaki, O. Mizuoka, and M. Takada, *J. Pharm. Pharmacol.* **43**: 115–116 (1990).
17. D. Bommannan, G. Menon, H. Okuyama, P. Elias, and R. Guy, *Pharm. Res.* **9**: 1043–1047 (1992).
18. K. Tachibana, *Pharm. Res.* **9**: 952–954 (1992).
19. K. Tachibana and S. Tachibana, *Anesthesiology* **78**: 1091–1096 (1993).
20. K. Tachibana and S. Tachibana, *J. Pharm. Pharmacol.* **43**: 270–271 (1991).
21. S. Mitragotri, D. Blankschtein, and R. Langer, *Pharm. Res.* **13**: 411–420 (1996).

22. S. Mitragotri, D. Blankschtein, and R. Langer, *Science* **269**: 850–853 (1995).
23. C. Wang, R.J. Stewart, J. Kopecek, *Nature* **397**: 417–420 (1999).
24. S. Mitragotri, D. Edwards, D. Blankschtein, and R. Langer, *J. Pharm. Sci.* **84**: 697–706 (1995).
25. A.J. Grodzinsky and P.E. Grimshaw, in *Pulsed and Self-Regulated Drug Delivery*, J. Kost, ed., CRC, Boca Raton, 1990, pp. 47–64.
26. M. Prausnitz, R. Bose, R. Langer, and J. Weaver, *Proc. Natl. Acad. Sci. USA* **90**: 10504–10508 (1993).
27. D. Rolf, *Pharm. Technol.* **12**: 130–140 (1988).
28. C. Hsu and L. Block, *Pharm. Res.* **13**: 1865–1870 (1996).
29. A. D'Emanuele and J.N. Staniforth, *Pharm. Res.* **8**: 913–918 (1991).
30. A. D'Emanuele and J.N. Staniforth, *Pharm. Res.* **9**: 215–219 (1992).
31. V. Labhassetwar, T. Underwood, S. Schwendemann, R. Levy, *Proc. Natl. Acad. Sci. USA* **92**: 2612–2616 (1995).
32. L.L. Miller, G.A. Smith, A. Chang, and Q. Zhou, *J. Controlled Release* **6**: 293–296 (1987).
33. M. Hepel and Z. Fijalek, in *Polymeric Drugs and Drug Administration*, 1994, pp. 79–97.
34. P.F. Kiser, G. Wilson, D. Needham, *Nature* **394**: 459–462 (1998).
35. P.E. Grimshaw, A.J. Grodzinsky, M.L. Yarmush, and D.M. Yarmush, *Chem. Eng. Sci.* **104**: 827–840 (1989).
36. I.C. Kwon, Y.H. Bae, T.O. and S.W. Kim, *J. Controlled Release* **17**: 149–156 (1991).
37. A. Mamada, T. Tanaka, D. Kugwatchkakun, and M. Irie, *Macromolecules* **23**: 1517–1519 (1990).
38. N. Yui, T. Okano, and Y. Sakurai, *J. Controlled Release* **26**: 141–145 (1993).
39. N. Yui, T. Okano, and Y. Sakurai, *J. Controlled Release* **22**: 105–116 (1992).
40. J. Taillefer, M.-C. Jones, N. Brasseur, J.E. Van Lier, and J.-C. Leroux, *J. Pharm. Sci.* **89**(1): 52–62 (2000).
41. A. Suzuki and T. Tanaka, *Nature* **346**: 345–347 (1990).
42. A.S. Hoffman, *J. Controlled Release* **6**: 297–305 (1987).
43. T. Tanaka, in *Encyclopedia of Polymer Science and Technology*, Vol. 7, H.F. Mark and J.I. Kroschwitz, eds., Wiley, NY, 1985, pp. 514–531.
44. G.H. Chen and A.S. Hoffman, *Nature* **373**: 49–52 (1995).
45. T. Ueda, K. Ishihara, and N. Nakabayashi, *Macromol. Chem. Rapid Commun.* **11**: 345 (1990).
46. D.W. Urry, B. Haynes, H. Zhang, R.D. Harris, and K.U. Prasad, *Proc. Natl. Acad. Sci. USA* **85**: 3407–3411 (1988).
47. M. Yoshida, M. Asano, M. Kumakura, R. Kataki, T. Mashimo, H. Yuasa, and H. Yamanaka, *Drug Design and Delivery* **7**: 159–174 (1991).
48. R. Spohr, N. Reber, A. Wolf, G.M. Alder, V. Ang, C.L. Bashford, C.A. Pasternak, H. Omichi, and M. Yoshida, *J. Controlled Release* **50**(1–3): 1–11 (1998).
49. M. Palasis and H. Gehrke, *J. Controlled Release* **18**: 1–12 (1992).
50. T. Okano, Y.H. Bae, and S.W. Kim, in *Pulsed and Self-Regulated Drug Delivery*, J. Kost, ed., CRC, Boca Raton, 1990, pp. 17–46.
51. Y. Okahata, H. Noguchi, and T. Seki, *Macromolecules* **19**: 493–494 (1986).
52. H. Katano, A. Maruyama, K. Sanui, N. Ogata, T. Okano, and Y. Sakurai, *J. Control. Release* **16**: 215–228 (1991).
53. S. Vakalanka, C. Brazel, and N. Peppas, *J. Biomater. Sci. Polym. Ed.* **8**: 119–129 (1996).
54. A. Serres, M. Baudys, and S. Kim, *Pharm. Res.* **13**: 196–201 (1996).
55. N. Ogata, *J. Controlled Release* **48**: 149–155 (1997).
56. J.E. Chung, T. Yokoyama, and T. Okano, *J. Controlled Release* **65**(1–2): 93–103 (2000).
57. M. Baundys, A. Serres, C. Ramkisoon, and S.W. Kim, *J. Controlled Release* **48**: 304–305 (1997).
58. L. Brannon-Peppas and N.A. Peppas, *J. Controlled Release* **8**: 267–274 (1989).
59. M. Annaka and T. Tanaka, *Nature* **355**: 430–432 (1992).
60. B.A. Firestone and R.A. Siegel, *Polym. Commun.* **29**: 204–208 (1988).
61. L.-C. Dong and A.S. Hoffman, *Proc. Int. Symp. Controlled Bioactive Mater.*, **17**: 325–326 (1990).
62. J.H. Kou, D. Fleisher, and G. Amidon, *J. Controlled Release* **12**: 241–250 (1990).
63. M. Pradny and J. Kopecek, *Makromol. Chem.* **191**: 1887–1897 (1990).
64. R.A. Siegel, M. Falmazian, B.A. Firestone, and B.C. Moxley, *J. Controlled Release* **8**: 179–182 (1988).
65. K. Kono, F. Tabata, and T. Takagishi, *J. Membrane Sci.* **76**: 233–243 (1993).
66. D. Hariharan and N.A. Peppas, *J. Controlled Release* **23**: 123–136 (1993).
67. R.A. Siegel and B.A. Firestone, *Macromolecules* **21**: 3254–3259 (1988).
68. H. Allcock and A. Ambrosio, *Biomaterials* **17**: 2295–2302 (1996).
69. C. Bell and N. Peppas, *Biomaterials* **17**: 1201–1218 (1996).
70. K. Jarvinen, S. Akerman, B. Svarfvar, T. Tarvainen, P. Viinikka, and P. Paronen, *Pharm. Res.* **15**: 802–805 (1998).
71. V. Pillay and R. Fassih, *J. Controlled Release* **59**(2): 229–42 (1999).
72. A.B. Madhan Kumar and K. Panduranga Rao, *Biomaterials* **19**(7–9): 725–732 (1998).
73. R.A. Siegel, I. Johannes, A. Hunt, and B.A. Firestone, *Pharm. Res.* **9**: 76–81 (1992).
74. S. Giannos, S. Dinh, and B. Berner, *J. Pharm. Sci.* **84**: 539–543 (1995).
75. J. Heller and P.V. Trescony, *J. Pharm. Sci.* **68**: 919–921 (1979).
76. H. Ghandehari, P. Kopeckova, and J. Kopecek, *Biomaterials* **18**: 861–872 (1997).
77. W. Petka, J. Harden, K. McGrath, D. Wirtz, and D. Tirrell, *Science* **281**: 389–392 (1998).
78. K.E. Brown, K. Leong, C.H. Huang, R. Dalal, G.D. Green, H.B. Haimes, P.A. Jimenez, and J. Bathon, *Arthritis Rheum.* **41**(12): 2185–2195 (1998).
79. M. Tanihara, Y. Suzuki, Y. Nishimura, K. Suzuki, Y. Kakimaru, and Y. Fukunishi, *J. Pharm. Sci.* **88**(5): 510–514 (1999).
80. C.G. Pitt, Z.-W. Gu, R.W. Hendren, J. Thompson, and M.C. Wani, *J. Controlled Release* **2**: 363–374 (1985).
81. C.G. Pitt, *Pharm. Int.* 88–91 (1986).
82. T. Miyata, N. Asami, and T. Urugami, *Nature* **399**: 766–769 (1999).
83. C.G. Pitt, T.T. Bao, Z.W. Gu, M.C. Wani, and Z.H. Zhu, in *Pulsed and Self-Regulated Drug Delivery*, J. Kost, ed., CRC, Boca Raton, 1985, pp. 117–127.
84. R. Goldbart and J. Kost, *Pharm. Res.* **16**(9): 1483–1486 (1999).

85. R. Goldbart, J. Kost, and S.A. Lapidot, *Pharm. Res.* (submitted).
86. K. Ishihara, N. Muramoto, H. Fuji, and I. Shinohara, *J. Polym. Sci. Polym. Chem. Ed.* **23**: 2841–2850 (1985).
87. K. Ishihara, N. Muramoto, H. Fuji, and I. Shinohara, *J. Polym. Sci. Polym. Lett. Ed.* **23**: 531–535 (1985).
88. J. Heller, R.V. Sparer, and G.M. Zenter, in *Biodegradable Polymers as Drug Delivery Systems*, M. Chasin and R. Langer, eds., Marcel Dekker, NY, 1990, pp. 121–161.
89. J. Heller, in *Controlled Drug Delivery: Challenges and Strategies*, K. Park, ed., American Chemical Society, Washington, DC, 1997, pp. 127–146.
90. K. Roskos, V. Tefft, and J. Heller, *J. Clin. Mater.* **13**: 109–119 (1993).
91. K.V. Roskos, B.K. Fritzing, J.A. Tefft, and J. Heller, *Biomaterials* **16**: 1235–1239 (1995).
92. K.V. Roskos, J.A. Tefft, B.K. Fritzing, and J. Heller, *J. Controlled Release* **19**: 145–160 (1992).
93. J.A. Tefft, K.V. Roskos, and J. Heller, *J. Biomed. Mater. Res.* **26**: 713–724 (1992).
94. G.R. Nakayama, K.V. Roskos, B.K. Fritzing, and J. Heller, *J. Biomed. Mater. Res.* **29**: 1389–1396 (1995).
95. J. Heller, A.C. Chang, G. Rodd, and G.M. Grodsky, *J. Controlled Release* **13**: 295–304 (1990).
96. J. Kost, in *Polymeric Materials Encyclopedia*, J.C. Salamone, ed., CRC, Boca Raton, 1996, Vol. 4, pp. 2825–2829.
97. S. Kitano, Y. Koyama, K. Kataoka, T. Okano, and Y. Sakurai, *J. Controlled Release* **19**: 161–170 (1992).
98. I. Hisamitsu, K.T.O. Kataoka, and Y. Sakurai, *Pharm. Res.* **14**: 289–293 (1997).
99. D. Shiino et al., *J. Controlled Release* **37**: 269–276 (1995).
100. M. Brownlee and A. Cerami, *Science* **26**: 1190–1191 (1979).
101. K. Makino, E.J. Mack, T. Okano, and S.W. Kim, *J. Controlled Release* **12**: 235–239 (1990).
102. S. Sato, S.Y. Jeong, J.C. McRea, and S.W. Kim, *J. Controlled Release* **67**–77 (1984).
103. S. Sato, S.Y. Jeong, J.C. McRea, and S.W. Kim, *Pure Appl. Chem.* **56**: 1323–1328 (1984).
104. L. Seminoff, G. Olsen, J. Zheng, D. Wilson, and S.W. Kim, in *Controlled Release of Bioactive Materials*, 1988, Vol. 15, pp. 161–161.
105. L. Seminoff and S.W. Kim, in *Pulsed and Self-Regulated Drug Delivery*, J. Kost, ed., CRC, Boca Raton, 1990.
106. S.W. Kim, C.M. Pai, L.A. Makino, L.A. Seminoff, D.L. Holmberg, J.M. Gleeson, D.E. Wilson, and E.J. Mack, *J. Controlled Release* **11**: 193–201 (1990).
107. S.Y. Jeong, S.W. Kim, D. Holumberg, and J.C. McRea, *J. Controlled Release* **2**: 143–152 (1985).
108. S.Y. Jeong, S.W. Kim, M.J.D. Eenink, and J. Feijen, *J. Controlled Release* **1**: 57–66 (1984).
109. C.M. Pai et al., *J. Pharm. Sci.* **81**: 532 (1992).
110. E. Kokufata, Y.-Q. Zhang, and T. Tanaka, *Nature* **351**: 302–304 (1991).
111. M. Taylor, S. Tanna, P. Taylor, and G. Adams, *J. Drug Targeting* **3**: 209–216 (1995).
112. A. Obaidat, and Park, *Biomaterials* **11**: 801–806 (1997).
113. A. Obaidat and K. Park, *Pharm Res.* **13**: 998–995 (1996).
114. S. Lee and K. Park, *J. Mol. Recognition* **9**: 549–557 (1996).
115. L. Brown, F. Ghodsian, and R. Langer, *Proc. Int. Symp. Controlled Release Bioactive Mater.* 1988, pp. 116–167.
116. L. Brown, E. Edelman, F. Fishel-Ghodsian, and R. Langer, *J. Pharm. Sci.* **85**: 1341–1345 (1996).
117. F. Fischel-Ghodsian, L. Brown, E. Mathiowitz, D. Brandenburg, and R. Langer, *Proc. Natl. Acad. Sci. USA* **2403**–2466 (1995).
118. J. Heller, *Adv. Drug Delivery Rev.* **10**: 163–204 (1993).
119. G. Albin, T. Horbett, and B. Ratner, in *Pulsed and Self-Regulated Drug Delivery*, J. Kost, ed., CRC, Boca Raton, 1990, pp. 159–185.
120. G. Albin, T.A. Horbett, S.R. Miller, and N.L. Ricker, *J. Controlled Release* **6**: 267–291 (1987).
121. G. Albin, T.A. Horbett, and B.D. Ratner, *J. Controlled Release* **3**: 153–164 (1985).
122. L.A. Klumb and T.A. Horbett, *J. Controlled Release* **18**: 59–80 (1992).
123. L.A. Klumb and T.A. Horbett, *J. Controlled Release* **27**: 95–114 (1993).
124. J. Kost, T.A. Horbett, B.D. Ratner, and M. Singh, *J. Biomed. Mater. Res.* **19**: 1117–1133 (1985).
125. T. Horbett, J. Kost, and B. Ratner, *Am. Chem. Soc. Div. Polym. Chem.* **24**: 34–35 (1983).
126. T. Horbett, J. Kost, and B. Ratner, in *Polymers as Biomaterials*, S. Shakaby, A. Hoffman, T. Horbett, and B. Ratner, eds., Plenum Press, NY, 1984, pp. 193–207.
127. M. Goldraich and J. Kost, *Clin. Mater.* **13**: 135–142 (1993).
128. T. Traitel and J. Kost, *Biomaterials* **21**: 1679–1687 (2000).
129. T. Traitel, J. Kost, and S.A. Lapidot, (submitted).
130. I. Kaetsu, M. Kumakura, and M. Yoshida, *Biotechnol. Bioeng.* **21**: 847–861 (1979).
131. K. Ishihara, in *Controlled Release of Bioactive Molecules*, 1988, Vol. 15, pp. 168–169.
132. K. Ishihara, M. Kobayashi, and I. Shonohara, *Macromol. Chem. Rapid Commun.* **4**: 327–331 (1983).
133. K. Ishihara, M. Kobayashi, N. Ishimaru, and I. Shinohara, *Polym. J.* **16**: 625–631 (1984).
134. K. Ishihara and K. Matsui, *J. Polym. Sci. Polym. Lett. Ed.* **24**: 413–417 (1986).
135. K. Podual, F. Doyle, and N. Peppas, *J. Controlled Release* **67**: 9–17 (2000).
136. H. Iwata and T. Matsuda, *J. Membrane Sci.* **38**: 185–199 (1988).
137. H. Iwata, H. Amemiya, T. Hata, T. Matsuda, H. Takano, and T. Akutsu, *Proc. Int. Symp. Controlled Release Bioactive Mater.* **15**: 170–171 (1988).
138. Y. Ito, M. Casolaro, K. Kono, and Y. Imanishi, *J. Controlled Release* **10**: 195–203 (1989).
139. R. Siegel, in *Pulsed and Self-Regulated Drug Delivery*, J. Kost, ed., CRC, Boca Raton, 1990, pp. 129–157.
140. R.A. Siegel and B.A. Firestone, *J. Controlled Release* **11**: 181–192 (1990).
141. R. Siegel, in *Controlled Release: Challenges and Strategies*, K. Park, ed., American Chemical Society, Washington, DC, 1997.

ELECTRICALLY CONDUCTIVE ADHESIVES FOR ELECTRONIC APPLICATIONS

DAOQIANG LU
C.P. WONG
Institute of Technology
Atlanta, GA

INTRODUCTION

Soldering processes using tin/lead solders (Sn/Pb) are standard interconnection technologies for electronic components on printed circuit boards (PCBs). The most common reflow soldering process for surface mounted technology (SMT) uses tin/lead solder pastes (1–4). The pressure to reduce the industrial use of lead is growing, particularly in Europe (5–7). Lead is a well-known hazard to human health. Even small quantities can damage the brain, nervous system, liver, and kidneys when ingested. When Sn/Pb solders are disposed of in landfills, lead can leach into soils and pollute ground water. As such, the pressure to remove or minimize the use of lead is steadily building. Most European communities, in fact, have proposed a ban on the landfill disposal of electronic products containing leaded printed circuit boards, as well as on the sale of products containing the metal. In United States, consumer electronics were identified as the second largest source of lead (30%) in the municipal solid waste stream after lead-acid batteries (65%), which were already being separated from trash before disposal (5). Thus, usage of tin/lead solder paste in the SMT process is considered environmentally harmful and must be reduced or eliminated. Lead-free and environmentally sound interconnect bonding processes are urgently needed. Among the possibilities are electrically conductive adhesives (ECAs) and lead-free solders (8). Compared to soldering technology, ECA technology can offer numerous advantages such as fewer processing steps which reduces processing cost, lowers the processing temperature which makes the use of heat-sensitive and low cost substrates possible, and fine pitch capability (8).

ELECTRICALLY CONDUCTIVE ADHESIVES (ECAs)

There are two main types of conductive adhesives, anisotropic conductive adhesives (ACAs) and isotropic conductive adhesives (ICAs) (9–15). All electrically conductive adhesives consist of a polymer binder and a conductive filler. ACAs provide unidirectional electrical conductivity in the vertical or Z axis. This directional conductivity is achieved by using a relatively low-volume loading of the conductive filler. The low-volume loading is insufficient for interparticle contact and prevents conductivity in the X, Y plane of the adhesive. The ACA in film or paste form is interposed between the surfaces to be connected. Heat and pressure are simultaneously applied to this

stack-up until the particles bridge the two conductor surfaces on the two adherends. Once electrical continuity is produced, the polymer binder is hardened by a thermally initiated chemical reaction (for thermosets) or by cooling (for thermoplastics). The advantage of this approach is that the material conducts only in the direction in which the force was applied (12). A major disadvantage, however, is that pressure and heat must be simultaneously applied while the polymeric matrix is hardened, otherwise the conductive pathway is lost. ACAs are now being used to connect flat panel displays, tape-automated bonding (TAB), flip chips and fine pitch surface-mounted devices (SMDs). A cross section of an ACA junction is shown in Fig. 1.

The discussion in the following sections is focused mainly on isotropic conductive adhesives (ICAs). ICAs consist of metallic particles (normally in flake form) in a polymer binder. Typical filler loadings are 25–30 volume %. At these loadings, the materials have achieved the percolation threshold and are electrically conductive in all directions after they are cured. The most common ICAs are silver-flake-filled thermosetting epoxies which are typically provided as one-part thixotropic pastes (13–15). They can provide both electrical and mechanical interconnections between components after they are cured (Fig. 2) (12–15). Because the mechanical strength of the joint is provided by the polymer binder, the challenge in a formulation is to use the maximum metal filler loading without sacrificing the required strength. ICAs are often called “polymer solders” (15). ICAs have been used in the electronic packaging industry, primarily as die attach adhesives (16–18). Recently, ICAs have been proposed as an alternative to tin/lead solders in surface-mounted technology (SMT) applications. The benefits and drawbacks of ICAs compared to Sn/Pb solders have been thoroughly discussed in literatures. Briefly, the advantages of ICAs include environmental friendliness (lead-free and no flux), low processing temperature, low thermomechanical fatigue, compatibility with a wide range of surfaces (includes nonsolderable substrates), finer pitch interconnect capability, simple processing, and low cost (9,18–21).

However, conductive adhesive technology is still in its infancy, and concerns and limitations do exist. The main limitations of current commercial ICAs include lower conductivity, unstable contact resistance with nonnoble metal finished components, and poor impact performance. The electrical conductivity of an ICA is lower than that of Sn/Pb solders. Although this conductivity is adequate for most electronics applications, the electrical conductivity of ICAs must still be improved. Contact resistance between an ICA and nonnoble metal (such as Sn/Pb, Sn, and Ni) finished components increases dramatically especially under elevated temperature and humidity aging conditions (18,20–23). In addition, printed circuit board assemblies are subject to significant shocks during assembly, handling, and throughout product life. The packages cannot survive

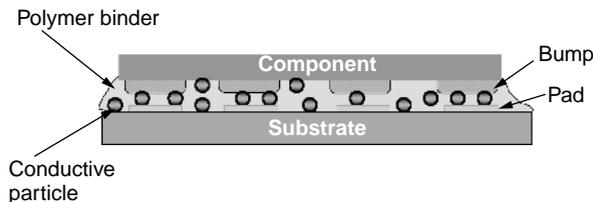


Figure 1. A cross section of an anisotropic conductive adhesive junction.

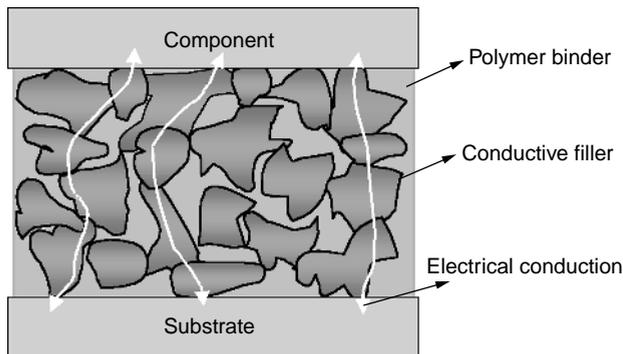


Figure 2. A cross section of an ICA joint.

without desirable impact performance. Most of the current commercial ICAs have poor impact performance. Components assembled using ICAs tend to fall from the substrate when the package experiences a sudden shock (14,21). For conductive adhesive technology to find universal appeal as a solder replacement, new conductive adhesives with desirable overall properties must be developed (22). In the past few years, there has been a great deal of effort to improve the properties of ICAs and to make them smarter materials. Recent smart material designs and formulations have resulted in improved electrical conductivity, contact resistance stability, and impact performance of ICAs. These will be reviewed in the following sections.

IMPROVEMENT OF ELECTRICAL CONDUCTIVITY OF ICAs

The electrical conductivity of ICAs is inferior to that of solders (24). Even though the conductivity of ICAs is adequate for most applications, higher electrical conductivity is still needed. To develop a novel ICA for modern electronic interconnect application, a thorough understanding of the materials is needed.

An ICA is generally composed of a polymer binder and Ag flakes. There is a thin layer of organic lubricant on the Ag flake surface. This lubricant layer plays an important role in the performance of ICAs, including the dispersion of the Ag flakes in the adhesives and the rheology of the adhesive formulations (24–27). This organic layer is a Ag salt formed between the Ag surface and the lubricant which typically is a fatty acid such as stearic acid (27,28). This lubricant layer affects the conductivity of an ICA because it is electrically insulating (27,28). To improve the conductivity, the organic lubricant layer

must be partially or fully removed. Some chemical substances can be used to dissolve the organic lubricant layer (27–29). However, the viscosity of an ICA paste might increase if the lubricant layer is removed. An ideal chemical substance (or lubricant remover) should be latent (does not remove the lubricant layer) at room temperature but be active (capable of removing the lubricant layer) at a temperature slightly below the cure temperature of the polymer. The lubricant remover can be a solid, short-chain acid, a high boiling point ether such as diethylene glycol monobutyl ether or diethylene glycol monoethyl ether acetate, or a low molecular weight polyethylene glycol (27–29). These chemical substances can improve the electrical conductivity of ICAs by removing the lubricant layer on the Ag flake surfaces and providing intimate flake–flake contact (27,28).

In general, an ICA paste has low conductivity before cure, but the conductivity increases dramatically after the ICA is cured. The ICA achieves conductivity during cure, mainly through more intimate contact between Ag flakes caused by cure shrinkage of the polymer binder (30). Cure shrinkage of the polymer binder affects the electrical conductivity of an ICA. An ICA that has higher cure shrinkage generally shows better conductivity (30). Therefore, increasing cure shrinkage of the polymer binder is another method for improving electrical conductivity. For ICAs based on epoxy resins, a small amount of a multifunctional epoxy resin can be added into an ICA formulation to increase cross-linking density, shrinkage, and thus increase conductivity (30).

Electrical conductivity of ICAs can be improved by applying an electrical field before or while curing them. An applied electrical field could improve electrical conductivity significantly. However, the exact reasons for this improvement are not yet clear (31).

Another approach to improving conductivity is to incorporate transient liquid-phase sintering metallic fillers in ICA formulations. The filler used is a mixture of a high melting point metal powder and a low melting point alloy powder. The electrical conduction is established through a plurality of metallurgical connections formed *in situ* from these two powders in a polymer binder. When heated at a specific high temperature, the low melting point alloy powders melt and diffuse rapidly into the high melting point filler and eventually solidify at that temperature. The polymer binder fluxes both the metal powders and the metals to be joined and facilitates transient liquid bonding of the powders to form a stable metallurgical network for electrical conduction and also forms an interpenetrating polymer network that provides adhesion (Fig. 3). High electrical conductivity can be achieved by using this method (32–34). One critical limitation of this technology is that the number of combinations of lower melting point filler and high melting point filler are limited. Only certain combinations of two metallic fillers can dissolve each other and form metallurgical interconnections.

IMPROVEMENT OF CONTACT RESISTANCE STABILITY

Contact resistance between an ICA (generally a Ag-flake-filled epoxy) and nonnoble metal finished components

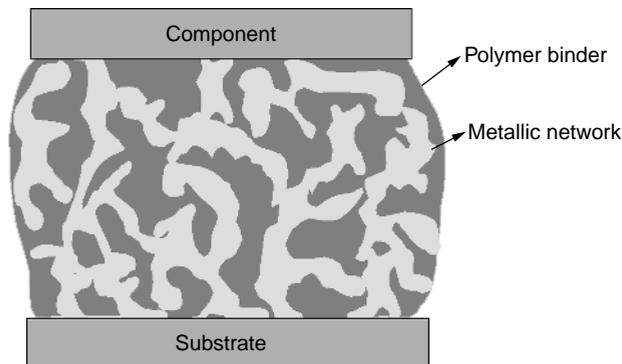


Figure 3. Diagram of transient liquid-phase sintering conductive adhesives.

increases dramatically during elevated temperature and humidity aging, especially at 85°C/85% relative humidity. The National Center of Manufacturing and Science (NCMS) set a criterion for solder replacement conductive adhesives. The criterion is that if the contact resistance shift after a 500-hour 85/85 aging is less than 20%, then the contact resistance is defined as “stable” (21).

Two main mechanisms, simple oxidation and corrosion of the nonnoble metal surfaces, have been proposed in the literature. Simple oxidation of the nonnoble metal surfaces was claimed as the main reason for the increased resistance in most of the literature. Corrosion was claimed as the possible mechanism for resistance shift in only a few papers (18,20,35–37). A recent study strongly indicated that galvanic corrosion rather than simple oxidation of the nonnoble metal at the interface between an ICA and the nonnoble metal was the main mechanism for the shift in contact resistance of ICAs (Fig. 4) (38,39). The nonnoble metal acts as an anode, is reduced by losing electrons, and turns into metal ion ($M - ne^- = M^{n+}$). The noble metal acts as a cathode, and its reaction generally is $2H_2O + O_2 + 4e^- = 4OH^-$. Then M^{n+} combines with OH^- to form a metal hydroxide or metal oxide. After corrosion, a layer of metal hydroxide or metal oxide is formed at the interface. Because this layer is electrically insulating, the contact resistance increases dramatically (38,39). A galvanic corrosion process has several characteristics: (1) happens only under wet conditions, (2) an electrolyte must be present, and (3) oxygen generally accelerates the process. Based on this finding, several methods can be employed to stabilize the shift in contact resistance.

Galvanic corrosion happens only in wet conditions. An electrolyte solution must be formed at the interface before galvanic corrosion can happen. Therefore, one way to prevent galvanic corrosion at the interface between an ICA and a nonnoble metal surface is to lower the moisture pickup of the ICA. ICAs that have low moisture absorption generally showed more stable contact resistance on nonnoble surfaces (40,41). Without electrolyte, the rate of galvanic corrosion is very low. The electrolyte in this case is mainly from the impurities in the polymer binder (generally epoxy resins). Therefore, ICAs formulated with resins of high purity should show better performance.

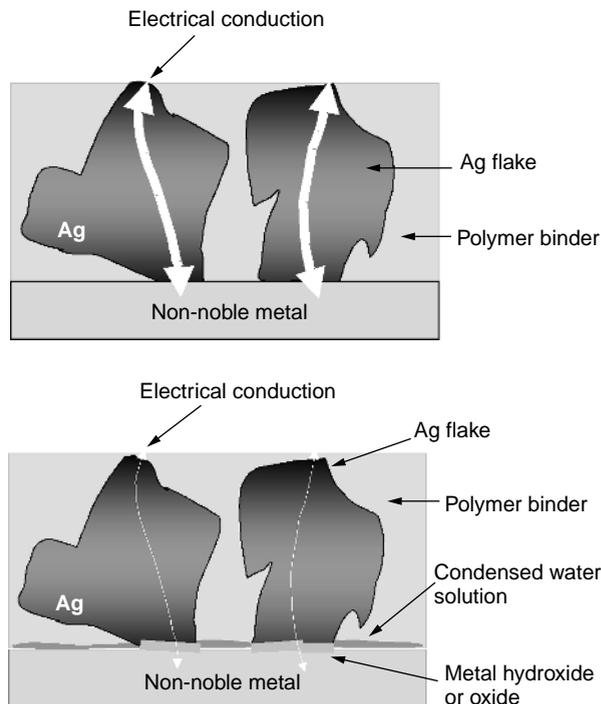


Figure 4. Metal hydroxide or oxide formation after galvanic corrosion.

The second method of preventing galvanic corrosion is to introduce some organic corrosion inhibitors into ICA formulations (39,41). In general, organic corrosion inhibitors act as a barrier layer between the metal and environment by being adsorbed as a film over the metal surfaces (43–46). Some chelating compounds are especially effective in preventing metal corrosion (45). Most organic corrosion inhibitors can react with epoxy resins at certain temperatures. Therefore, if an ICA is epoxy-based, the corrosion inhibitors must not react with the epoxy resin when the epoxy resin is cured. Otherwise, the corrosion inhibitors are consumed by reacting with the epoxy resin and lose their effect. Organic corrosion inhibitors are thoroughly summarized in the literature (44,46). As an example, Fig. 5

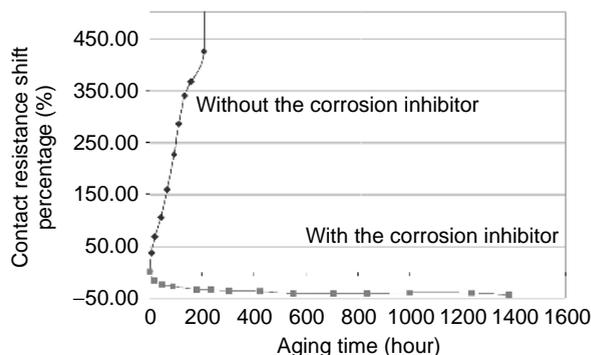


Figure 5. Effect of a corrosion inhibitor on the contact resistance between an ICA and a Sn/Pb surface (aging conditions: 85°C/85% relative humidity) (38).

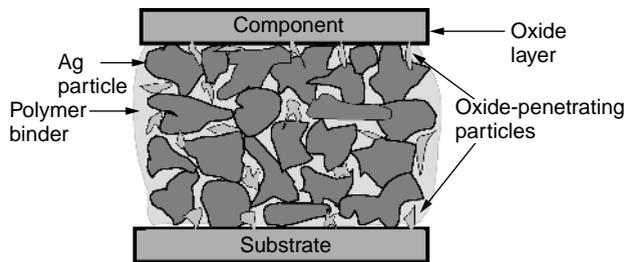


Figure 6. A joint connected with an ICA that contains oxide-penetrating particles and silver powders.

shows the effect of a chelating corrosion inhibitor on contact resistance between an ICA and Sn/Pb surface. It can be seen that this corrosion inhibitor is very effective in stabilizing the contact resistance.

Oxygen can accelerate galvanic corrosion. Therefore, another way to slow down the corrosion process is to incorporate some oxygen scavengers into ICA formulations (43). When an oxygen molecule diffuses through the polymer binder, it reacts with the oxygen scavenger and is consumed. However, when the oxygen scavenger is depleted completely, then oxygen can still diffuse into the interface and accelerate the corrosion process. Therefore, oxygen scavengers can delay galvanic corrosion process only for some time. Similarly, the oxygen scavengers used must not react with the epoxy resin at its cure temperature. Common oxygen scavengers include hydrazine, carbonylhydrazide, hydroquinone, gallic acid, propyl gallate, hydroxylamines and related compounds, dihydroxyacetone, 1,2-dihydro-1,2,4,5-tetrazines, erythorbic acid, and oximes (43,47–50).

It is claimed that transient liquid-phase sintering conductive adhesives mentioned in the previous section have stable contact resistance during elevated temperature and humidity aging (32). The joints formed include metallurgical alloying to the junctions as well as within the adhesive itself. This provides a stable electrical connection during aging.

Another approach for improving contact resistance stability during aging is to incorporate some electrically conductive particles, which have sharp edges, into the ICA formulations. The particle is called an oxide-penetrating filler. Force must be provided to drive the oxide-penetrating particles through the oxide layer and hold them against the adherend materials. This can be accomplished by employing polymer binders that show high shrinkage when cured (Fig. 6) (51). This concept is used in Poly-Solder which has good contact resistance stability with standard surface-mounted devices (SMDs) on both solder-coated and bare circuit boards.

IMPROVEMENT OF IMPACT PERFORMANCE

Impact performance is a critical property of solder replacement ICAs. Effort has been continued in developing ICAs that have better impact strength and will pass the drop test, a standard test used to evaluate the impact strength of ICAs.

Nanosized metal particles were used in ICAs to improve electrical conduction and mechanical strength. Using nanosized particles, agglomerates are formed due to surface tension (52). Another approach is simply to decrease the filler loading to improve impact strength (53). However, such a process reduces the electrical properties of the conductive adhesives. A recent development was reported where conductive adhesives were developed using resins of low modulus, so that this class of conductive adhesives could absorb the impact energy developed during the drop (54). However, the electrical properties of these materials were not mentioned in the paper. Conformal coating of the surface-mounted devices was also used to improve mechanical strength. It was demonstrated that conformal coating could improve the impact strength of conductive adhesives joints (13,55).

More recently, a new class of conductive adhesives which is based on an epoxide-terminated polyurethane (ETPU) has been developed (56,57). This class of conductive adhesives has the properties of polyurethane materials, such as high toughness and good adhesion. The modulus and glass transition temperature of the ICAs can be adjusted by incorporating some epoxy resins such as bisphenol-F epoxy resins. Conductive adhesives based on ETPU showed a broad loss factor ($\tan \delta$) peak with temperature and a high $\tan \delta$ value at room temperature. The $\tan \delta$ value of a material is a good indication of the damping property and impact performance of the material. In general, the higher the $\tan \delta$ value, the better the damping property (impact strength) of the material. As an example, changes in $\tan \delta$ and modulus of an ETPU-based ICA with temperature are shown in Fig. 7. ICAs based on ETPU resins also showed a much higher loss factor in a wide frequency range than ICAs based on bisphenol-F epoxy resins (Fig. 8). This indicated that the ICAs based on ETPU resins should exhibit good damping properties and improved impact performance in different electronic packages. This class of conductive adhesives showed superior impact performance and substantial stable contact resistance with non-noble metal surfaces such as Sn/Pb, Sn, and Cu (56,57).

Again, transient liquid-phase sintering conductive adhesives showed good impact strength due to the

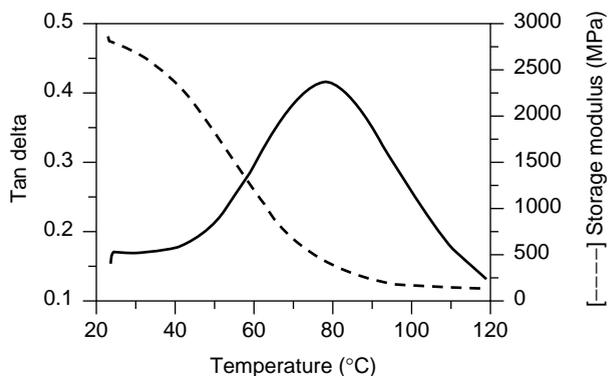


Figure 7. Changes in $\tan \delta$ and modulus with temperature of an ETPU-based conductive adhesive measured by a dynamic mechanical analyzer (DMA) (56).

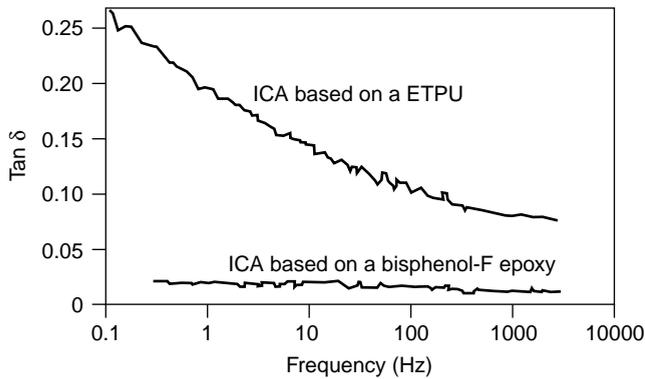


Figure 8. Loss factor versus frequency of two ICA materials.

metallurgical interconnection between the conductive adhesive and the components (32).

HIGH-PERFORMANCE CONDUCTIVE ADHESIVES

A solder replacement conductive adhesive should have the following main properties (22):

1. Bulk resistivity lower than 1 milliohm-cm ($m\Omega\text{-cm}$);
2. Shift of contact resistance (after 500-hour $85^\circ\text{C}/85\%\text{RH}$ aging) less than 20%; and
3. Passes six drops from a 60-inch height.

Based on the review in the previous sections, a high-performance conductive adhesive can be formulated mainly by two approaches which are described as follows.

Conductive Adhesives Based on an Epoxide-Terminated Polyurethane (ETPU)

A hardener (usually a liquid anhydride), which can provide the ICA with the lowest moisture pickup, should be selected first. A conductive adhesive formulation should consist of the following components: the epoxide-terminated polyurethane, an epoxy resin (optional), the hardener, an adhesion promoter, a conductivity promoter, Ag flakes (mixture of large and small particles), and some other additives, for example, corrosion inhibitors and oxygen scavengers. Conductive adhesives based on this formulation showed desirable conductivity, superior impact performance, and substantial contact resistance on nonnoble metal surfaces such as Cu, Sn, and Sn/Pb (56,57). These materials are being tested in surface-mounted applications.

Transient Liquid-Phase Sintering Conductive Adhesives

A conductive adhesive from this category should include a high melting point metal filler, a low melting point metal filler, a polymer binder, and other (optional) additives. The low melting point metal filler must alloy with the high melting point metal filler and the metal surface to be joined at a specific acceptable temperature. The alloy formed should not liquify at the processing or usage temperature of the ICA. Otherwise, the contact resistance and adhesion might be diminished. For the low melting point alloy to wet

the high melting point filler and for the metal surfaces to be joined, the polymer binder must also contain a flux or have a fluxing function to remove the oxide on the high melting point metal filler and the metal surfaces to be joined. Due to the metallurgical connection at the junctions and inside the conductive adhesive, the junction showed stable contact resistance during elevated temperature and humidity aging and desirable impact strength.

BIBLIOGRAPHY

1. H.H. Manko, *Solders and Soldering, Materials, Design, Production, and Analysis for Reliable Bonding*, 2nd ed. McGraw-Hill, New York, 1979.
2. R.J.K. Wassink, *Soldering in Electronics*, Electrochemical Publications, 1984.
3. F.N. Sinnadurai, *Handbook of Microelectronics Packaging and Interconnection Technologies*, Electrochemical Publications, 1985.
4. L.P. Lambert, *Soldering for Electronic Assemblies*, New York, 1987.
5. B. Trumble, Get the lead out! *IEEE Spectrum*, May, pp. 55–60 (1998).
6. B.T. Alpert and A.J. Schoenberg, Conductive adhesives as a soldering alternative. *Electron. Packag. Prod.*, November, pp. 130–132 (1991).
7. O. Deubzer and A. Middendorf, Environmentally friendly materials and processes for SMT-reflow soldering and conductive adhesive joining in comparison. pp. 163–168.
8. R. Cdenhead and D. DeCoursey, *Int. J. Microelectron.* **8**(3), 14 (1985).
9. J.C. Jagt, Reliability of electrically conductive adhesive joints for surface mount applications: A summary of the state of the art. *IEEE Trans. Components, Packag., Manuf. Technol., Part A* **21**(2), 215–225 (1998).
10. A.O. Ogunjimi, O. Boyle, D.C. Whalley, and D.J. Williams, A review of the impact of conductive adhesive technology on interconnection. *J. Electron. Manuf.* **2**, 109–118 (1992).
11. P.G. Harris, Conductive adhesives: A critical review of progress to date. *Soldering Surf. Mount Technol.* **20**, 19–21 (1995).
12. K. Gilileo, Assembly with conductive adhesives. *Soldering Surf. Mount Technol.* **19**, 12–17 (1995).
13. J. Liu, Z. Lai, H. Kristiansen, and C. Khoo, Overview of conductive adhesive joining technology in electronics packaging applications. *Proc. 3rd Int. Conf. Adhes. Joining Coat. Technol. Electron. Manuf.*, Binghamton, NY, 1998, pp. 1–17.
14. S. Corbett and M.J. Dominano, Processing with polymer solders. *Surf. Mount Technol.*, June, pp. 48–52 (1997).
15. J. Bolger and S. Morano, Conductive adhesives: How and where they work. *Adhes. Age*, June, pp. 17–20 (1984).
16. J. Greaves, Evaluation of solder alternatives for surface mount technology. *Proc. Nepcon West Tech. Program, 1993*, pp. 1479–1488.
17. M.A. Gaynes, R.H. Lewis, R.F. Saraf, and J.M. Roldan, Evaluation of contact resistance for isotropic conductive adhesives. *IEEE Trans. Components, Packag., Manuf. Technol., Part B* **18**(2), 299–304 (1995).
18. G. Nguyen, J. Williams, F. Gibson, and T. Winster, Electrical reliability of conductive adhesives for surface mount applications. *Proc. Int. Electron. Packag. Conf., 1993*, pp. 479–486.

19. G. Nguyen, J. Williams, and F. Gibson, Conductive adhesives: Reliable and economical alternatives to solder paste for electrical applications. *Proc. ISHM, 1992*, pp. 510–517.
20. J.C. Jagt, P.J.M. Beric, and G.F.C.M. Lijten, Electrically conductive adhesives: A prospective alternative for SMD soldering? *IEEE Trans. Components, Packag., Manuf. Technol., Part B*. **18**(2), 292–298 (1995).
21. M. Zwolinski, J. Hickman, H. Rubon, and Y. Zaks, Electrically conductive adhesives for surface mount solder replacement. *Proc. 2nd Int. Conf. Adhes. Joining & Coat. Technol. Electron. Manuf.*, Stockholm, Sweden, June 3–5, 1996, pp. 333–340.
22. H. Botter, Factors that influence the electrical contact resistance of isotropic conductive adhesive joints during climate chamber testing. *Proc. 2nd Int. Conf. Adhes. Joining Coat. Technol. Electron. Manuf.*, Stockholm, Sweden, June 3–5, 1996, pp. 30–37.
23. C.P. Wong, D. Lu, S. Vona, and Q.K. Tong, A fundamental study of electrically conductive adhesives. *Proc. 1st IEEE Int. Sympo. Polym. Electron. Packag.*, Norrkoping, Sweden, 1997, pp. 80–85.
24. L. Smith-Vargo, Adhesives that possess a science all their own. *Electron. Packag. Prod.*, August, pp. 48–49 (1986).
25. E.M. Jost and K. McNeilly, Silver flake production and optimization for use in conductive polymers. *Proc. ISHM, 1987*, pp. 548–553.
26. S.M. Pandiri, The behavior of silver flakes in conductive epoxy adhesives. *Adhes. Age*, October, pp. 31–35 (1987).
27. D. Lu, Q.K. Tong, and C.P. Wong, A study of lubricants on silver flakes for microelectronics conductive adhesives. *IEEE Trans. Components, Packag., Manuf. Technol., Part A* **22**(3), 365–371 (1999).
28. D. Lu, Q. Tong, and C.P. Wong, A fundamental study on silver flakes for conductive adhesives. *Proc. Int. Sympo. Exhib. Adv. Packag. Mater., Processes, Properties, Interfaces*, Braselton, GA, 1998, pp. 256–260.
29. A.J. Lovinger, Development of electrical conduction in silver-filled epoxy adhesives. *J. Adhes.* **10**, 1–15 (1979).
30. D. Lu, Q.K. Tong, and C.P. Wong, Conductivity mechanisms of isotropic conductive adhesives (ICAs). *IEEE Trans. Components, Packag., Manuf. Technol., Part C* **22**(3), 223–227 (1999).
31. J. Morris and C. Cook, Recent results of ICA testing. *Proc. 2nd IEEE Int. Symp. Polym. Electron. Packag.*, Gothenburg, Sweden, 1999, pp. 15–26.
32. C. Gallagher, G. Matijasevic, and J.F. Maguire, Transient liquid phase sintering conductive adhesives as solder replacement. *Proc.—Electron. Components Technol. Conf.*, **47**, 554–560 (1997).
33. J.W. Roman and T.W. Eagar, Low stress die attach by low temperature transient liquid phase bonding. *Proc. ISHM, 1992*, pp. 52–57.
34. C. Gallagher, G. Matijasevic, and A. Capote, Transient liquid phase sintering conductive adhesives. U.S. Pat. 5,863,622 (1998).
35. H. Botter, Factors that influence the electrical contact resistance of isotropic conductive adhesive joints during climate chamber testing. *Proc. 2nd Int. Conf. Adhes. Joining Coat. Technol. Electron. Manuf.*, Stockholm, Sweden, June 3–5, 1996, pp. 30–37.
36. K. Gilleo, Evaluating polymer solders for lead free assembly. Part I. *Circuits Assembly*, pp. 50–51 (1994).
37. K. Gilleo, Evaluating polymer solders for lead free assembly. Part II. *Circuits Assembly*, pp. 51–53 (1994).
38. D. Lu, Q.K. Tong, and C.P. Wong, Mechanisms underlying the unstable contact resistance of conductive adhesives. *IEEE Trans. Components, Packag., Manuf. Technol., Part C* **22**(3), 228–232 (1999).
39. Q.K. Tong, G. Fredrickson, R. Kuder, and D. Lu, Conductive adhesives with superior impact resistance and stable contact resistance. *Proc.—Electron. Components Technol. Conf.*, **49**, 347–352 (1999).
40. D. Lu and C.P. Wong, Novel conductive adhesives for surface mount applications. *J. Appl. Polym. Sci.* **74**, 399–406 (1999).
41. D. Lu and C.P. Wong, Novel conductive adhesives with stable contact resistance. *Proc. 4th Int. Sympo. Exhib. Adv. Packag. Mater., Processes, Properties, Interfaces*, Braselton, GA, 1999, pp. 288–294.
42. Q.K. Tong, A. Xiao, and D. Lu, Conductive and resistive materials with electrical stability for use in electronics devices. U.S. Pat. (1999).
43. H. Leidheiser, Jr., Mechanism of corrosion inhibition with special attention to inhibitors in organic coatings. *J. Coat. Technol.* **53**(678), 29–39 (1981).
44. G. TrabANELLI and V. Carassiti, Mechanism and phenomenology of organic inhibitors. In *Advanced Corrosion Science and Technology* (M. G. Fontana and R.W. Staehle, eds.), Vol. 1, pp. 147–229. Plenum, New York, 1970.
45. G. TrabANELLI, Corrosion inhibitors. In *Corrosion Mechanisms* (F. Mansfeld, ed.), pp. 119–164. Dekker, New York, 1987.
46. O.L. Riggs, Jr., Theoretical aspects of corrosion inhibitors and inhibition. In (C.C. Nathan, ed.), pp. 2–27. NACE, Houston, TX, 1973.
47. P.A. Reardon, New oxygen scavengers and their chemistry under hydrothermal conditions. *Corrosion'86*, Pap. No. 175. NACE, Houston, TX, 1986.
48. M.G. Noack, Oxygen scavengers. *Corrosion'89*, Pap. No. 436. NACE, Houston, TX, 1989.
49. P.A. Reardon and W.E. Bernahl, New insight into oxygen corrosion control. *Corrosion'87*, Pap. No. 438. NACE, Houston, TX, 1987.
50. S. Romaine, Effectiveness of a new volatile oxygen scavenger. *Proc. Am. Power Conf.*, **48**, 1066–1073 (1986).
51. D. Durand, D. Vieau, A.L. Chu, and T.S. Weiu, Electrically conductive cement containing agglomerates, flake and powder metal fillers. U.S. Pat. 5,180,523 (1989).
52. S. Kotthaus, R. Haug, H. Schafer, and O.D. Hennemann, *Proc. 1st IEEE Int. Symp. Polym. Electron. Packag.*, Norrkoping, Sweden, 1997, pp. 64–69.
53. S. Macathy, *Proc. Surf. Mount Int.*, San Jose, CA, August 27–31, 1995, pp. 562–567.
54. S.A. Vona and Q.K. Tong, Surface mount conductive adhesives with superior impact resistance. *Proc. 4th Int. Symp. Exhib. Adv. Packag. Mater., Processes, Properties, Interfaces*, Braselton, GA, 1998, pp. 261–267.
55. J. Liu and B. Weman, Modification of processes and design rules to achieve high reliable conductive adhesive joints for surface mount technology. *Proc. 2nd Int. Symp. Electron. Packag. Technol.*, 1996, pp. 313–319.
56. D. Lu and C.P. Wong, High performance electrically conductive adhesives. *IEEE Trans. Components, Packag., Manuf., Part C* **22**(4), 324–330 (1999).
57. D. Lu and C.P. Wong, Electrically conductive adhesives having superior impact performance. U.S. Pat. (pending)

ELECTROCERAMICS

HARRY TULLER

Massachusetts Institute of Technology
Cambridge, MA

INTRODUCTION

Ceramics have traditionally been admired for their mechanical, chemical, and thermal stability; however, their unique electrical, optical, and magnetic properties have become of increasing importance in many key technologies including communications, energy conversion and storage, electronics, and automation. Such materials are now classified as *electroceramics* as distinguished from other functional ceramics such as advanced structural ceramics. It is the combination of the traditional ruggedness of conventional ceramics and the exceptional functionality of select electroceramic materials that make electroceramics particularly attractive as *smart materials* and as components of *smart systems*.

ELECTROCERAMICS AND SMART SYSTEMS

The term *smart* suggests emulation of living systems that sense their environments, evaluate these inputs in relation to knowledge stored in memory, and respond accordingly, thereby operating by using a built-in *feedback* mechanism. A very general schematic of the feedback loop is shown in Fig. 1.

From a technological standpoint, the feedback system's functions are commonly subdivided into the categories of (1) sensor, (2) microprocessor and memory, and (3) actuator. The *brain* of the feedback network operates nearly exclusively in the electrical domain, so the sensor normally functions as a transducer that converts various physical and chemical inputs or stimuli into electrical outputs, and the actuators convert electrical inputs from the microprocessor into the appropriate physical and chemical outputs. These feedback system features are illustrated in Fig. 2. Both sensor and actuator often require auxiliary inputs to enable them to operate properly. This will become obvious

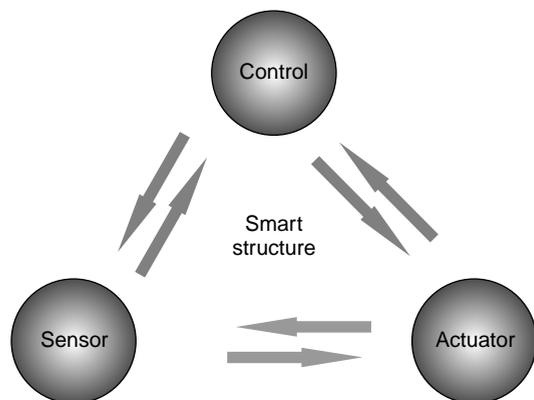


Figure 1. Smart structures: Incorporation of sensing and actuating functions in a control system. A feedback loop can go through a control circuit or directly between sensor and actuator.

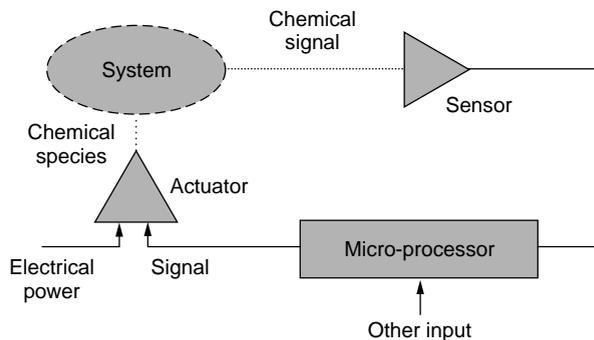


Figure 2. A schematic of a feedback system including sensor, microprocessor, and actuator that monitors and controls the chemical state of the system. Both sensor and actuator may require auxiliary inputs to enable them to operate properly.

as specific examples of sensors and actuators are discussed. Pursuing biological counterparts further, we are aware that some stimuli bypass the brain. Some responses are triggered automatically, for example, when one pulls one's hand away from a hot object. Corresponding automatic responses occur in materials systems. Newnham subdivides smart materials into two categories (1). *Passively smart materials* are defined as those that respond automatically to external change without assistance, whereas *actively smart materials* use a feedback loop that enables them to recognize the change and initiate an appropriate response through an actuator, as illustrated in Fig. 1. Obviously, *passively smart materials* potentially provide much simpler and thereby less costly alternatives, whereas *actively smart materials*, though more complex, via programming provide considerably more flexibility in their response. Examples of both types of smart materials are described in following sections.

Based on this definition of *smart*, any feedback system, independent of size, complexity, power requirements, cost, etc., would qualify. However, as additional constraints of speed, size, power, harsh environmental operation, and cost are imposed on such systems, then the types of materials, their modes of fabrication, transduction efficiency, and stability begin to play important roles. *Electroc ceramic* materials excel for many of these reasons.

Enormous strides have been made in microprocessor speed and memory density during the past few decades, but comparable developments in sensors and actuators have, in many cases, lagged considerably. Many such devices remain bulky, consume substantial power, do not readily interface with electronics, and respond sluggishly. The effective integration of functional *electroc ceramic* materials into microelectronic and microelectromechanical (MEMS) technologies remains a key challenge for this field. The ability to harvest the benefits of miniaturization, low power, integration, and the ability to fabricate large arrays, will, however, pay major dividends in performance, versatility, speed, and reduced costs.

To appreciate the versatility of *electroc ceramics* and their potential use in smart systems, we summarize some of the key properties most commonly associated with sensing and actuating functions. For example, piezoelectrics such as $\text{Pb}(\text{Zr}_{1-x}\text{Ti}_x)\text{O}_3$ (PZT) are used to detect mechanical

Table 1. Energy Transduction Types, Materials, and Applications

Energy Transduction	Materials	Applications
Electrical \Rightarrow mechanical	Pb(Zr _{1-x} Ti _x)O ₃ (PZT)	Positioner, optical tuner
Mechanical \Rightarrow electrical		Igniter
Electrical \Rightarrow chemical	Zr _{1-x} Y _x O _{2-x/2} (YSZ),	Oxygen pump
Chemical \Rightarrow electrical	SnO ₂	Gas sensor
Electrical \Rightarrow optical	LiNbO ₃	Electro-optic modulator
Optical \Rightarrow electrical	GaN	Light detector
Electrical \Rightarrow magnetic	AB ₂ O ₄ spinel: A,B = Mn, Zn	Write head
Magnetic \Rightarrow electrical	La _{1-x} Ca _x MnO ₃	Read head
Electrical \Rightarrow thermal	SiC	Resistive heater
Thermal \Rightarrow electrical	AB ₂ O ₄ spinel: A,B = Mn, Ni, Fe, Co, Cu	NTCR thermistor
	BaTiO ₃ perovskite	PTCR thermistor

displacement electrically and inversely to induce mechanical displacements by applying electric fields. Solid electrolytes, such as Zr_{1-x}Y_xO_{2-x/2} (YSZ), are used potentiometrically to monitor the chemical state (oxygen activity) of an environment or to modify that state by electrochemically pumping oxygen into that environment. These and other examples that relate to transduction between electrical, optical, magnetic, and thermal forms of energy are summarized in Table 1. Typical materials and applications are included in this table.

The performance of *electroc ceramic* materials and devices depends on the complex interplay among processing, chemistry, and structure at many levels. Because *electroc ceramics* are commonly polycrystalline, their microstructure (grain size and distribution, porosity, phase distribution) and their crystal and defect structures often play crucial roles in determining their properties. Likewise, domain structures dominate the properties of ferroelectric and ferromagnetic *electroc ceramic* materials. These factors are discussed when addressing some of the specific applications of these materials relevant to the design and development of smart materials.

Articles in the professional literature deal with the processing, characterization, structure, properties, modeling, and performance of *electroc ceramics*. Recent advances in these areas are described in the *Journal of Electroceramics* and other general or more specialized journals such as the *Journal of Applied Physics*, *Journal of the American Ceramic Society*, *Ferroelectrics*, and *Solid State Ionics*.

The numbers and types of applications for smart materials and systems are limited only by the imagination, so we selected five that (1) are representative from the standpoint of dealing, in one form or another, with each of the six major energy types (electrical, optical, magnetic, chemical, mechanical, and thermal); (2) illustrate popular applications; and (3) illustrate the use of passively and actively smart materials. For each of these examples, we summarize the underlying physical principles, discuss figures of merit, and consider materials of choice. When appropriate, we also list opportunities for further development.

ELECTROMECHANICAL ACTUATORS

Applications

Vibrational Damping. There are many applications where vibrational reduction and/or control is required. The

military has an interest in reducing acoustic emissions from submarines, commercial airlines are concerned with the vibrations in passenger cabins, and automobile manufacturers, deal with the vibrations experienced by passengers riding over rough terrain. Two recent examples where such concepts have been applied to the world of sports are examined: the ski/snowboard, and the bicycle shock absorber. Figure 3 shows the piezoelectric stack used in ski and snowboard vibrational control. The heart of the vibrational damper is a piezoelectric element connected to an electric circuit that is tuned to dissipate energy at frequencies deemed undesirable, as illustrated in Fig. 3.

A more sophisticated design is the smart shock absorber or suspension system. Smart suspensions are already installed in many cars. For simplicity, we examine a shock absorber designed for bicycles (Fig. 4). The magnetic GMR sensor sends data relating to the rate of motion to the control electronics that in turn instructs the piezoelectric actuator to manipulate a valve in the hydraulic shock absorber that controls damping.

Precision Positioning. In the microelectronics industry, piezoelectric motors or stacks enable mask positioning in

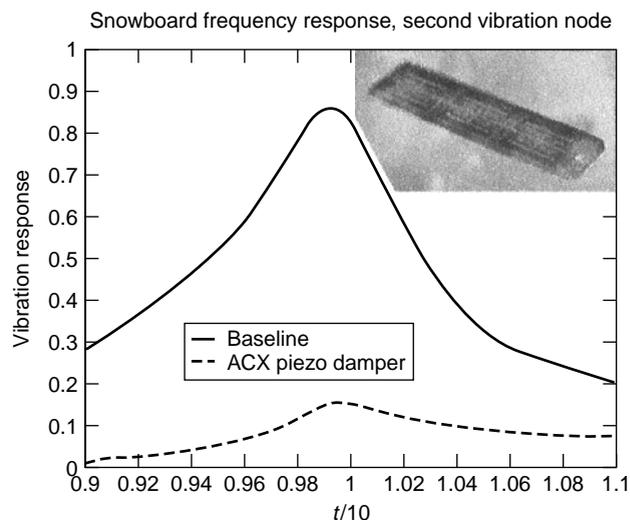


Figure 3. Vibrational damper in skis and snowboards. An electrical circuit is tuned to damp out vibrations in the undesired frequency band (ACX designs).

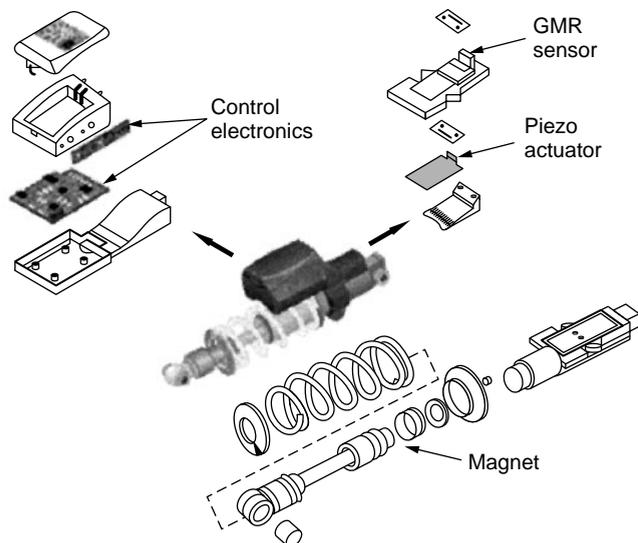


Figure 4. Design for a smart shock absorber. The rate of motion is monitored by the magnetic GMR sensor, and the piezoelectric actuator manipulates a valve that controls damping (Design by K2 SmartShock™).

the submicron range. Similar positioning systems are used in electron microscopy and atomic force microscopy (AFM) to control stage movements of 10–100 nm. In AFM, for example, the position of the microcantilever beam relative to the substrate is detected optically by reflecting a laser beam off the microcantilever beam and measuring its movement by using position-sensitive photodetectors. Thus, the feedback loop involves an optical sensor, computer, and piezoelectric actuator.

Camera Components. Consumer applications include camera lens controllers and video camera stabilizers. Auto-focus camera lenses now contain built-in piezoelectric ultrasonic motors. These motors operate in a feedback mode and reposition the optical components in the lens to achieve fast and accurate autofocusing (for more details see section on smart optical systems).

Some manufacturers use piezoelectric sensors to sense video camera vibrations. Two piezoelectric velocity sensors respond to the camera's movement along its vertical and horizontal axes. The sensors control motors that realign a

prism that bends the light to compensate for vibrations. The operation of this system is illustrated in Fig. 5.

ACTUATOR MATERIALS

General

Mechanical actuators are encountered in many forms and sizes. Although electromagnetic motors and hydraulic jacks serve many applications, there are limitations to their use in smaller scale applications. A comparison of conventional and piezoelectric actuation methods is shown in Table 2. Piezoelectrics excel in bandwidth and power density.

The development of precision electromechanical actuators (EMA) has grown in recent years. The world market for medical ultrasonic probes alone stands at \$250 million/year (2). The major advantages of these actuators are high precision (on the order of 10 nm), rapid response time (10 μ s), and high generative force with low driving power (see Table 2). These properties make EMAs attractive for optical systems, precision machinery, and small, high-power motors (3).

Another category of piezoelectric applications uses resonant strain obtained by an applied ac field. Ultrasonic motors, pumps and surgical knives are just a few examples of high-frequency, high-accuracy applications that use resonating piezoelectric materials. A list of some general types of applications that use piezoelectric and electrostrictive actuators is given in Table 3. The table also lists some of the major research institutes that are developing those applications.

BASIC RELATIONSHIPS AND PHENOMENA

The Piezoelectric Family of Materials

The piezoelectric effect relates electric polarization to applied stress. Polarization requires a nonsymmetrical structure; therefore all crystallographic point groups that have a center of symmetry cannot support piezoelectricity. Furthermore, although the cubic class 432 lacks a center of symmetry, it cannot accommodate piezoelectricity. Therefore, only 20 of the 32 crystallographic point groups exhibit piezoelectricity. Of these groups, 10

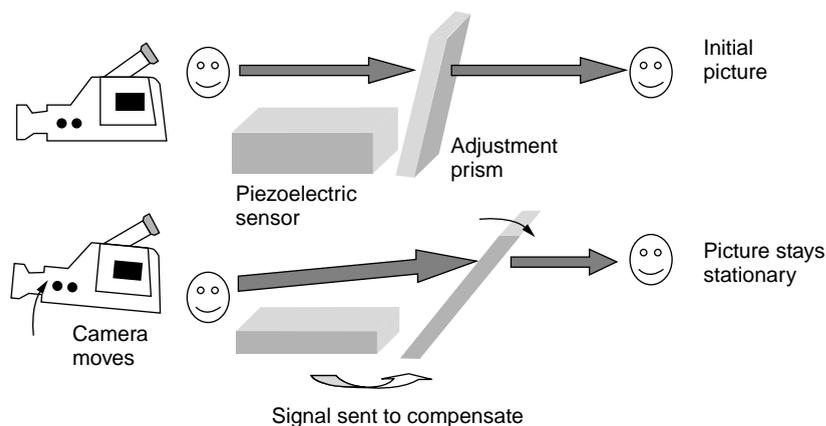


Figure 5. Piezoelectrics sense camera movement and adjust the prism to stabilize the image in a video camera.

Table 2. Comparison of Conventional and Piezoelectric Actuation

Actuation Method		Stress (Mpa)	Strain (%)	Efficiency (%)	Bandwidth (Hz)	Work (J/cm ²)	Power (W/cm ³)
Conventional	Electromagnetic	0.02	50	90	20	0.005	0.1
	Hydraulic	20	50	80	4	5	20
	Pneumatic	0.7	50	90	20	0.175	3.5
Piezoelectric	Polycrystal	35	0.2	50	5000	0.035	175
	Shape-memory	200	1.0	3	3	10	30
	Single-crystal	300	1.7	90	5800	2.55	15000

have a unique crystallographic axis and therefore, can have an electric dipole even at zero applied field (electrical and mechanical). These materials are defined as pyroelectric and show a change of polarization when temperature changes.

Pyroelectric materials include a subclass of materials in which an applied external field can change the direction of polarization. These materials are referred to as ferroelectric. In some cases, although a polar unit cell exists in the material, the overall behavior is nonpolar. Materials that have polar subcells organized so that adjacent subcells have opposite polarization are called antiferroelectric. The relationships between the different symmetry groups and polarization properties are shown in Fig. 6.

The Piezoelectric Effect

The piezoelectric modulus d is a third-rank tensor that relates the polarization P_i to the applied stress σ_{jk} :

$$P_i = d_{ijk} \sigma_{jk}. \quad (1)$$

The units of the piezoelectric modulus are [C/N] or [m/V].

The converse piezoelectric effect relates changes in crystal shape and size to an applied electric field. It can be shown that the coefficient that relates strain ε_{ij} to an applied electric field E_k is identical to the direct piezoelectric

modulus, so that

$$\varepsilon_{jk} = d_{ijk} E_i. \quad (2)$$

For a detailed discussion of the tensor behavior of the piezoelectric coefficient and a proof of the equality of the direct and converse effects see (5, Chaps. VII and X).

Piezoelectric Figures of Merit

The piezoelectric coefficient is a very important parameter of piezoelectric actuators. There are, however, other important parameters of these materials to consider.

Polarization is not an easy property to measure. In many sensors, the measured value is the voltage, or the electric field E , that is induced by an applied stress σ . The relationship between the field and the stress is given by the piezoelectric voltage constant g :

$$E_i = g_{ijk} \sigma_{jk}. \quad (3)$$

Taking into account the relationship between the electric field and polarization, one can easily show that the two piezoelectric constants are related by

$$g = d/\varepsilon \cdot \varepsilon_0. \quad (4)$$

This is a simplified form of the tensor relationship that holds for homogeneous materials or along a principal axis.

Table 3. Applications of Piezoelectric and Electrostrictive Actuators^a

Device	Research Institute	Application	Driving Control	Strain
High-sensitive interferometry	Sophia Univ.	Optical	Servo	Rigid
Deformable mirror	Tokyo Inst. Tech., Sophia Univ.	Optical	Servo	Rigid
Microscope stage	NBS (USA)	Optical	Servo	Rigid
Ultraprecision guide mechanism	Hitachi	Mechanical	Servo	Rigid
Cutting-error compensation actuator	Tokyo Inst. Tech., Sophia Univ., Hitachi	Mechanical	Servo	Rigid
Swing CCD image sensor	Toshiba	Optical	On/off	Rigid
Microangle adjusting device	Sophia Univ., Toshiba	Optical	On/off	Rigid
Piezoelectric relay	PEPI (USA), Omron	Mechanical	On/off	Rigid
Ink jet	NEC, etc.	Mechanical	On/off	Rigid
Ultrasonic surgical knife	Tokyo Inst. Tech.	Energy transfer	ac	Resonating
Piezoelectric pump	Phys. Internation. (USA) Misuzu-Erie	Energy transfer	ac	Resonating
Ultrasonic motor	IBM (USA), Shinsei Kogyo, Hitachi Maxell, Matsushita	Optical/mechanical	ac	Resonating

^aRef. 4.

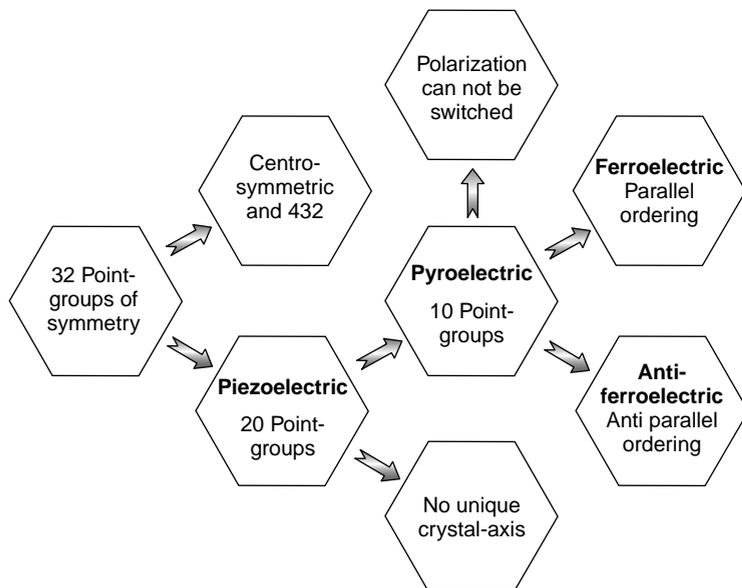


Figure 6. Crystallographic symmetry groups and their relationship to polarized materials.

Both the strain and the voltage piezoelectric constants give the magnitude of the effect; however, they do not tell us how efficient the system is. Three parameters describe the transfer between electrical and mechanical energies:

1. Efficiency: $\eta = (\text{output energy})/(\text{consumed energy})$
2. Energy transmission coefficient: $\lambda_{\max} = [(\text{output energy})/(\text{input energy})]_{\max}$
3. Electromechanical coupling factor: $k^2 = (\text{stored energy})/(\text{input energy})$

Although the differences between these parameters are small, they are important. They are not, however, independent. It can be shown that the electromechanical coupling coefficient is related to the stiffness c or the compliance s and to the piezoelectric and dielectric constants (6):

$$k^2 = d^2 c / \varepsilon \cdot \varepsilon_0 = d^2 / s \cdot \varepsilon \cdot \varepsilon_0 \quad (5)$$

From this, the relationship between the coupling coefficient and the energy transmission coefficient can be calculated as

$$\lambda_{\max} = \left[\left(\frac{1}{k} \right) - \sqrt{\left(\frac{1}{k^2} \right) - 1} \right]^2 \quad (6)$$

By taking the consumed energy as the difference between input and output energies, the relationship between the

efficiency and the energy transmission can be written as

$$\eta = \frac{\lambda_{\max}}{1 - \lambda_{\max}} \quad (7)$$

Piezoelectric materials have a characteristic electromechanical resonant frequency. The material response close to that frequency is highly nonlinear, and the parameters described before are insufficient. The resonance behavior is very important for applications such as sensors and ultrasonic motors. The frequency width $\Delta\omega$ of the resonance is used to describe the quality of the resonance in the material around the resonant frequency ω_0 :

$$Q_m = \omega_0 / 2\Delta\omega \quad (8)$$

A full description of resonators and acoustic wave sensors is beyond the scope of this article. Comprehensive coverage of the topic can be found in (7). Table 4 summarizes the figures of merit.

Ferroelectric Hysteresis

Ferroelectric materials have a remanent polarization at a zero applied field. The energy needed to rotate domains and to move domain walls creates a hysteric loop, as shown in Fig. 7. Similar behavior is seen in the strain-field curve. This results in a permanent strain after removing the electric field that can be used for shape-memory

Table 4. Figures of Merit for Piezoelectric Materials^a

Figure of merit	Parameter (units)	Important Relationships
Piezoelectric strain constant	d (C/N)	$\varepsilon_{jk} = d_{ijk} E_i$; $P_i = d_{ijk} \sigma_{jk}$
Piezoelectric voltage constant	g (V · m/N)	$E_i = g_{ijk} \sigma_{jk}$; $g = d / \varepsilon \cdot \varepsilon_0$
Electromechanical coupling coefficient	k^2 (-)	$k^2 = d^2 c / \varepsilon \cdot \varepsilon_0 = d^2 / s \cdot \varepsilon_0$
Mechanical quality factor	Q_m (-)	$Q_m = \omega_0 / 2\Delta$

^aSymbol definitions can be found in the text.

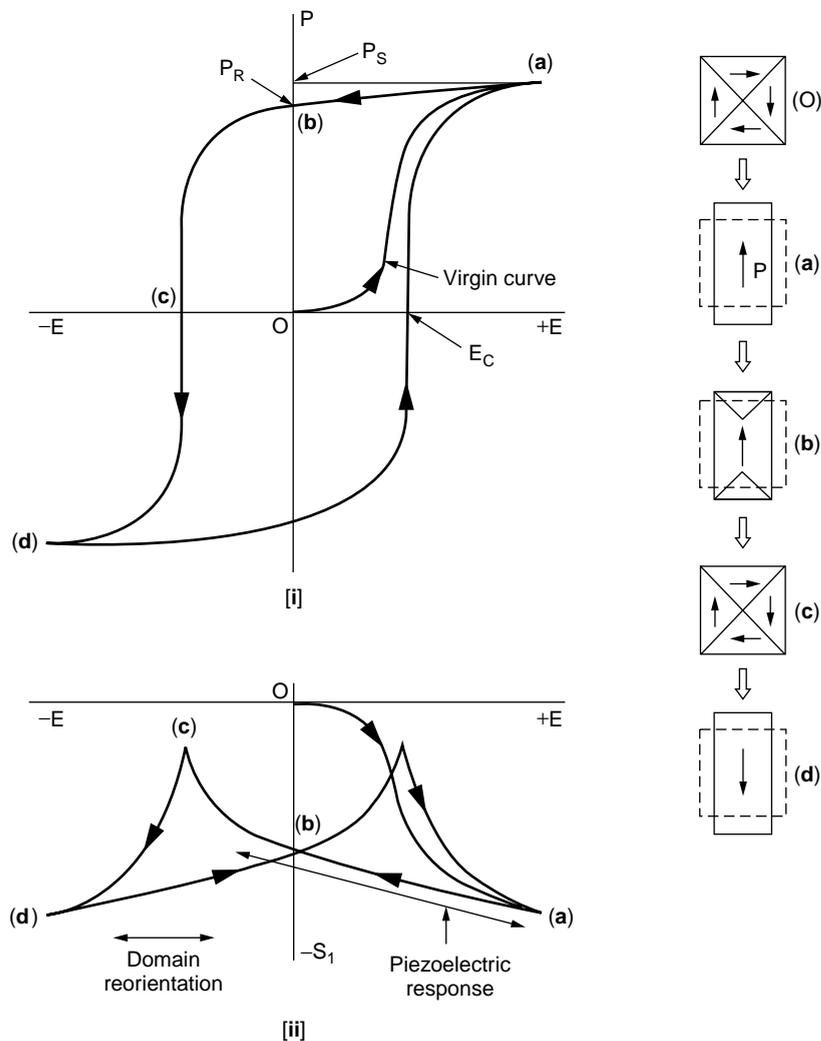


Figure 7. [i] Polarization curve and [ii] strain curve, in a ferroelectric material. The domain images on the right side are schematic representations of the polarization state in the material at different points along the curves (Based on 8,9).

applications. The drawback of these systems is the relatively high loss due to hysteresis. In high-frequency applications, this leads to heating, and special care must be taken to limit temperature excursions to less than the Curie temperature.

Two important parameters are related to the hysteretic loop; the remanent polarization and the coercive field. The remanent polarization is the polarization level in the absence of an external electric field (P_R in Fig. 7). This is the effective amount of charge that the material can hold after the electric field is removed and therefore, is the amount of “memory” that the material can have. The remanent polarization is accompanied by a remanent strain. This is important in actuators because it can effectively latch a strained state without the need to keep a constant voltage on it and therefore can save energy. The coercive field (E_C in Fig. 7) is the field required to remove the remanent polarization. The higher that field, the more stable the latching effect. This can be important for preventing spontaneous depoling; however, a high coercive field requires a higher operating voltage and is more limited in high-frequency applications.

Actuation by Phase Transition

At times, electric fields can be used to induce phase transitions in ferroelectric materials. Application of an appropriate field to the rhombohedral phase causes rotation of the polarization vector toward the direction of the applied field. Under sufficient rotation, the induced structural change induces a phase transition illustrated in Fig. 8.

At times, the new phase is metastable so that it remains after the electric field is removed (see Fig. 8). In such cases, a latching, or memory effect is observed, and the material remains strained with respect to its initial state. An electric field in the opposite direction must be applied to return it to the unstrained state. This is similar to the memory effect described in the previous section, although the remanent field and strain are usually much higher.

ELECTROSTRICTIVE BEHAVIOR IN MATERIALS

Electrostriction is a second-order effect of an electromechanical interaction. This is not to say that it is always

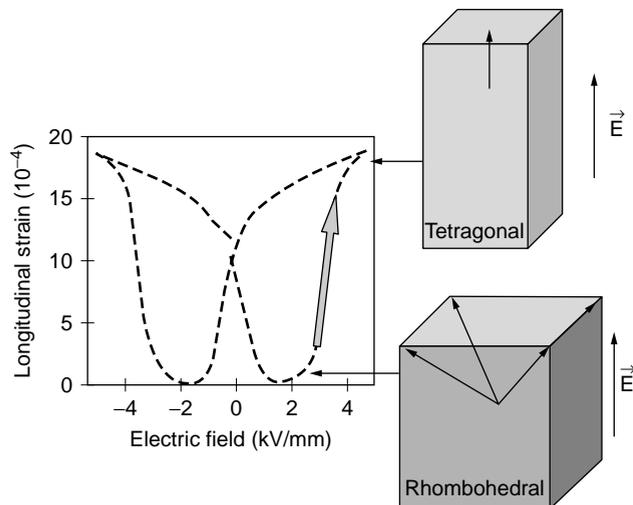


Figure 8. Field-induced phase transition from the rhombohedral to the tetragonal phase. Data is for PLZHT (10).

negligible compared to first-order piezoelectricity. When looking at a simplified rigid sphere model, the piezoelectric effect is the result of the difference in the spring constants of the different ions. The electrostrictive effect is represented by divergence from a perfect harmonic oscillator (4), as illustrated in Fig. 9. Because it is a second-order effect, the electrostriction is not limited to systems that have no center of symmetry. Furthermore, electrostriction is not limited to specific crystallographic systems.

The combined electrostrictive and piezoelectric converse effect is given by

$$\varepsilon_{jk} = d_{ijk} E_i + q_{iljk} E_i E_l. \quad (9)$$

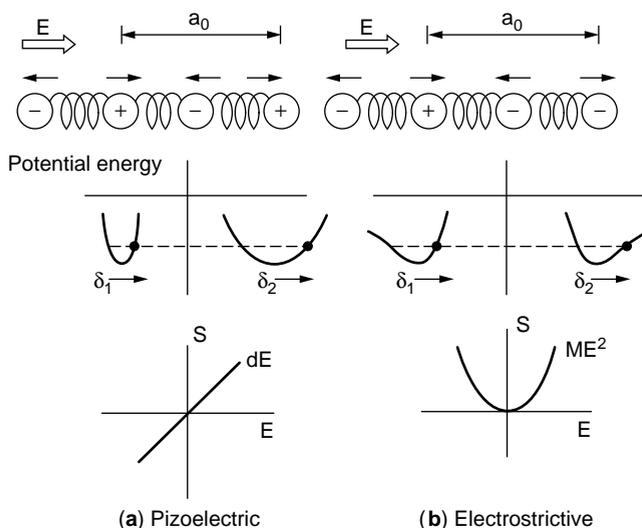


Figure 9. Rigid ion model and the corresponding strain-field behavior of (a) a non centrosymmetric, piezoelectric material, and (b) a centrosymmetric, electrostrictive crystal (8).

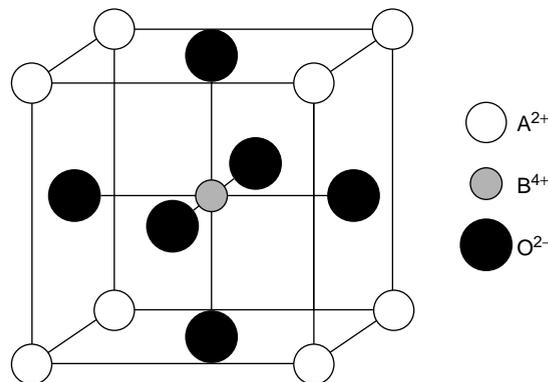


Figure 10. A unit cell of the perovskite structure. The B ion is octahedrally coordinated by the oxygen ions.

It is important to note that, even if one begins with a centrosymmetric structure, the application of an electric field can result in breaking that symmetry and consequently the appearance of a piezoelectric element (5).

MATERIALS SYSTEMS

Perovskite Structure

Most oxygen-octahedral structures that have ferroelectric properties and high Curie temperatures are based on corner linking of the oxygen octahedra (11). Of these, the simplest and most common structure is the perovskite family shown in Fig. 10. The figure describes the high symmetry cubic prototype ($m\bar{3}m$) for this family. At lower temperatures, phase transitions to ferroelectric / antiferroelectric phases occur due to lattice distortion and oxygen octahedral tilting. This results in a loss of symmetry, as shown in Fig. 11.

PZT-Based Materials

Lead zirconate titanate ($\text{PbZr}_x\text{Ti}_{1-x}\text{O}_3$) solid solutions are some of the most commonly used piezoelectric materials. A morphotropic phase boundary between a tetragonal and a rhombohedral structure occurs around $x = 0.52$. Materials whose compositions are close to that boundary exhibit a high piezoelectric coefficient and higher coupling coefficient. A second morphotropic phase boundary exists between ferroelectric and antiferroelectric phases around $x = 0.92$. This region is particularly useful for shape-memory actuators.

Barium-Titanate-Based Materials

Barium Titanate (BaTiO_3) is another very common ferroelectric material. For many years, the main interest in this materials system was for its excellent dielectric properties. BaTiO_3 has a low piezoelectric coefficient at room temperature; however, by using a solid solution of BaTiO_3 that contains other materials, such as BaZrO_3 , the rhombohedral phase can be stabilized, and a much higher piezoelectric response can be achieved (12).

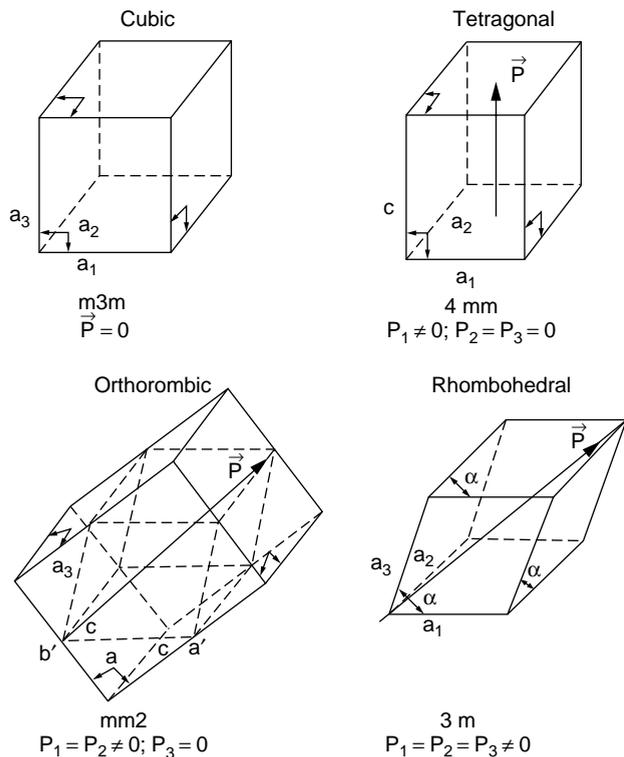


Figure 11. The four key structural types exhibited by compounds that have perovskite structures. Note the orientations of the polarization vector in the tetragonal, orthorhombic, and rhombohedral distortions of the cubic structure.

Single-Crystal Materials

The most widely used single-crystal piezoelectric materials are quartz, lithium niobate (LiNbO_3), and lithium tantalate (LiTaO_3). Although the piezoelectric coefficients of these materials are much smaller than that of PZT, they hold an important position in resonance applications. The quality factor of single crystalline materials can be three orders of magnitude higher than that of polycrystals. This increases their response close to the resonant frequency.

Quartz oscillators are used for clock source and frequency stabilizers in many household appliances. LiNbO_3 and LiTaO_3 have very high electromechanical coupling coefficients for surface acoustic waves (SAW) and are often used in SAW devices.

Relaxor-Based Ferroelectric Single Crystals

In recent years, a new family of ferroelectrics has received great attention for use in electromechanical actuation. Relaxor-based ferroelectrics have shown great promise as actuator materials due to their high strain. Strain levels up to 1.7% have been reported for PZN-PT (13) compared to 0.15% for PZT.

Table 5 summarizes the key properties of the piezoelectric materials discussed before.

THERMAL SYSTEMS

Applications

The most pervasive form of energy that has traditionally interested humankind in measuring and controlling is thermal. Presently, this ranges from monitoring and controlling the temperatures of our homes, offices, and automobiles, the foods that we prepare, and various key industrial processes (chemical and drug syntheses, semiconductor fabrication, metal forming, etc.) on which we depend.

Self-Regulating Heaters. Electroceramics offer passively smart systems that regulate temperature without temperature controllers. Certain types of electroceramic materials, fabricated as honeycombs, are used as heating elements that have built-in self-regulation of temperature. Heat is obtained by forcing air through hundreds of holes in each element (see Fig. 12). The elements, largely based on barium titanate, have a positive temperature coefficient (PTC) of electrical resistance. As the temperature of the ceramic increases, its resistance increases rapidly, thereby reducing the current flow through it. Therefore, the temperature of the PTC heater is self-regulating and does not exceed a given *set* temperature that is fixed by the composition of the ceramic. Some typical applications include portable fan heaters, clothes dryers, and fuel evaporators.

PTC ceramics are also used for overcurrent and overtemperature protection of devices such as motors, transformers, and electronic circuits. When a fault occurs, the element heats up and the resulting increase in resistance prevents damage to the device by limiting the current. In this manner, the *thermistor* acts like a smart fuse;

Table 5. Properties of Piezoelectric Materials^a

Material	T_c ($^{\circ}\text{C}$)	ϵ_r	k_{33}	d_{33} (pC/N)	Q_m
PZT 4 (hard)	325	1200	0.70	285	500
PZT 5H (soft)	195	2800	0.75	593	65
BaTiO_3	130	1900	0.49	190	500
Quartz, single-crystal		4.6	0.1 (k_{11})	2.3 (d_{11})	$>10^6$
LiNbO_3 , single-crystal	1210	29	0.17	6	
BaTiO_3 (modified)	80	1500	—	480	—
PZN-8%PT	175	5000	0.94	2500	70

^aBased on (6,14,15).

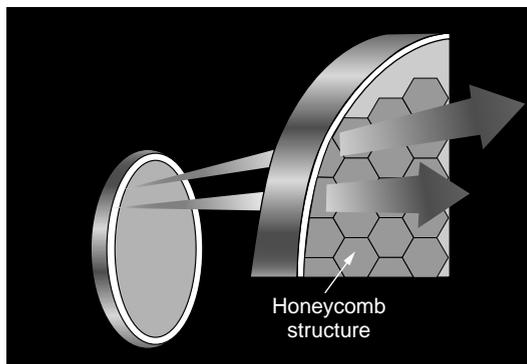


Figure 12. A self-regulated heater element fabricated as a honeycomb. This provides a convenient means for heating air forced through the structure (16).

once the circuit fault is corrected, it returns to its normal protective function.

Miniaturized Flow Sensor. Miniaturization provides the means to integrate multiple functions into a single package that enables designing and fabricating smarter sensors. An example that uses elements both for generating thermal energy and detecting its flow and direction is the micromachined airflow sensor illustrated in Fig. 13 (17). The mass airflow sensor consists of thermally isolated silicon nitride microbridges suspended over a cavity etched into a silicon wafer. Thin film sensing elements (thermally sensitive resistors) are positioned on both sides of centrally located heating elements that maintain the bridge at a well-defined temperature above ambient ($\sim 150^\circ\text{C}$). The thermal isolation limits the temperature rise on the surrounding chip to less than 1°C . Airflow directed laterally across the chip cools the upstream resistor and heats the downstream resistor. The output of a Wheatstone bridge then becomes proportional to the airflow, and the polarity of the voltage depends on the direction of flow. One such device, manufactured by Honeywell (18), is illustrated in Fig. 14. The sensitivity depends, in part, on the temperature coefficient of resistivity of the resistors and the thermal mass of the elements. These devices provide sensitivity to low flows (0.1 sccm), consume little power (30 mW), and have rapid response times. Typical applications include medical respirators, leak detection, gas metering and analysis, process control, and air conditioning systems.

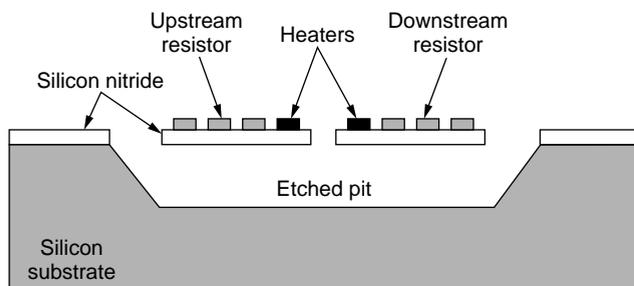


Figure 13. A schematic of a micromachined flow sensor (17).

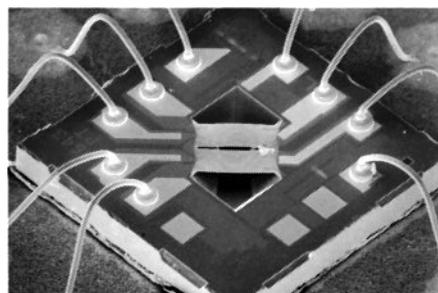


Figure 14. A micromachined flow sensor manufactured by Honeywell similar in design to this schematic in Fig. 13 (18).

Operating Principles

Resistive. All matter exhibits temperature-dependent resistivity, but the nature and magnitude of this sensitivity varies widely among different materials. The resistivity of metals increases as temperature increases and thus the designation *positive temperature coefficient of resistance* (PTCR). This relatively weak variation with temperature results from thermally induced scattering of carriers and is reflected in a nearly linear increase of resistivity with temperature. On the other hand, the resistivity of semiconductors and insulators, in general, decreases with temperature and thus the designation *negative coefficient of resistance* (NTCR). The decrease in resistivity is exponential and is most commonly tied to the thermal activation of charge carriers across an energy gap. However, among *electroceramic* materials, many are characterized by mobilities (small polaron) that themselves are thermally activated. This becomes relevant in our subsequent discussions of NTCR thermistors.

A limited number of materials exhibit phase change induced anomalous and strong temperature-dependent changes in resistivity across relatively narrow temperature limits. Examples of the various types of temperature-dependent resistivities are illustrated in Fig. 15, and the

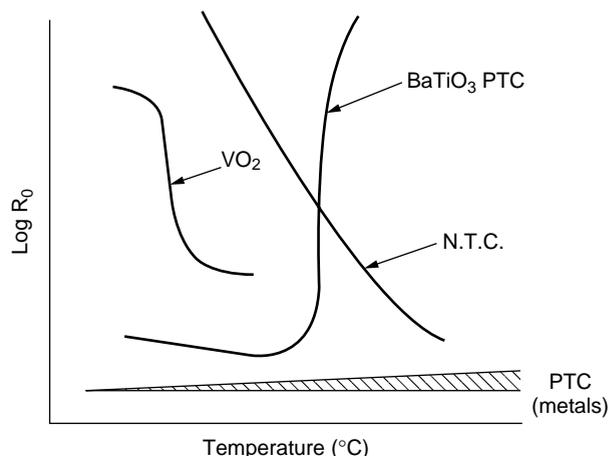


Figure 15. Characteristic resistance-temperature dependencies for various sensor materials (20).

Table 6. Sensitivity, Stability, and Operating Range of Sensor Materials^a

Sensor Material	Resistance Change (1/R)dR/dT (%/°C)	Stability	Temperature Range (°C)
Metals	+0.1–0.5	Fair to good	100–1000
Silicon	+0.1	Fair to good	150
BaTiO ₃ (PTC)	+10 to 100	Good at T_C	200+
NTC	–6.0 to –1.0	Good to excellent	300

^aRef. 20.

relative sensitivity, stability, and temperature range of usage are summarized in Table 6.

Devices designed to detect temperature changes by monitoring corresponding changes in resistance are called *thermistors*. The resistivity ρ of NTCR thermistor materials takes the form

$$\rho = \rho_0 \exp(E/kT), \quad (10)$$

where E is the effective activation energy that takes into account both carrier generation and mobility contributions. This is conventionally rewritten in terms of resistance R as

$$R = A \exp(B/T), \quad (11)$$

where the material constant B , or beta factor, is simply E/k and A is a constant related to the device dimensions and materials resistivity. Typical values of B fall around 3500 K, which corresponds to an activation energy of ~ 0.3 eV. The temperature coefficient of resistance (TCR) is often designated by the symbol $\sigma = (1/R) (dR/dT)$ and is related to B by $\alpha = -B/T^2$.

PTCR thermistors, in contrast to the majority of NTCR thermistors, which are grain controlled devices, depend instead on the resistance of the boundaries between the grains. N-type barium titanate BaTiO₃, prepared using appropriate additives and annealed in oxidizing atmospheres, develops grain boundary regions depleted in electrons. This results in creating space charge barriers (see Fig. 16) that block the flow of electrons between grains and a corresponding increase in the overall resistance of the material. For a small bias, the resistance is given by (21)

$$R \approx (A/T) \exp(\phi_B/kT). \quad (12)$$

The barrier height ϕ_B is related to N_D the donor density, d the space charge width, and ϵ the dielectric constant by

$$\phi_B = qN_D d^2 / 2\epsilon. \quad (13)$$

For materials such as BaTiO₃, which become ferroelectric at reduced temperatures and comply with the Curie–Weiss law,

$$\epsilon = C/T - T_C. \quad (14)$$

Combining Eqs. (12)–(14), one obtains the following dependence of R on T :

$$R \approx (A/T) \exp\{(qN_D d^2 / 2Ck)[(T - T_C)/T]\}. \quad (15)$$

As the temperature approaches the Curie temperature T_C from above, the dielectric constant increases and leads to a rapid drop in resistance. This correlation between R (or equivalently the resistivity ρ) and ϵ is clearly illustrated in Fig. 17 which shows that the resistivity decreases by nearly eight orders of magnitude as the effective dielectric constant increases from ~ 2000 to $\sim 30,000$. This resultant strong PTCR effect in the vicinity of the Curie temperature is used in a number of applications for temperature measurement and control.

Materials Systems

NTCR thermistors are traditionally fabricated from complex oxide spinels of the general formula AB₂O₄. The A and B sites are, respectively, tetrahedrally and octahedrally coordinated by oxygen. Transport reportedly occurs largely between neighboring B-site cations of different valence. The temperature dependence of resistivity is largely tied to the thermally activated mobility of the carriers, given that these systems already possess high carrier densities. Because the spinel structure accommodates large variations in composition, these allow great flexibility in tailoring both the magnitude of resistivity and its temperature coefficient of resistance (TCR), as illustrated in Fig. 18. Furthermore, thermistors fabricated from this family of materials can be reproduced with high precision and reproducibility (20).

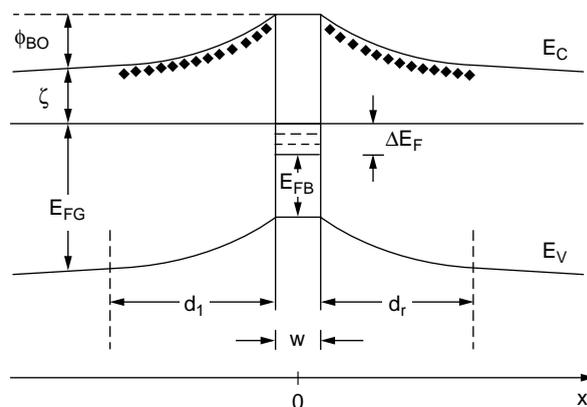


Figure 16. Schematic of the energy band diagram of the interfacial region between two n-type semiconducting grains. Interfacial traps induce depletion of electrons and band bending in the adjacent grains. Depletion widths are given by d , grain boundary width $w \ll d$, ϕ_{B0} the barrier height, and $\xi = E_C - E_{FB}$, the energy difference between the conduction band edge and the Fermi energy in the bulk (21).

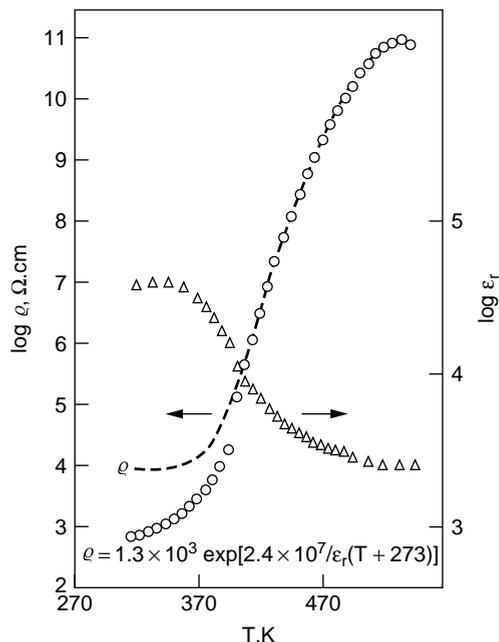


Figure 17. Temperature dependence of resistivity and dielectric constant of a PTCR thermistor (22).

PTCR thermistors are largely based on the perovskite barium titanate (see earlier section for more details). As discussed before, a key parameter is the Curie temperature which defines the onset temperature of the PTCR effect. This corresponds to the cubic to tetragonal phase transition

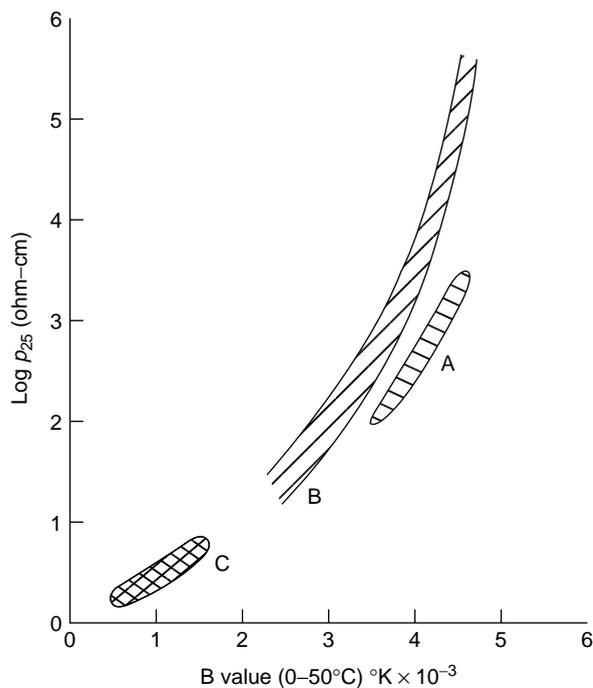


Figure 18. Relationship between room temperature resistivity and beta value for selected NTCR thermistor materials classes: (A) Li-doped $(\text{Mn},\text{Ni},\text{Co})\text{O}$; (B) complex spinels AB_2O_4 : $(\text{Ni},\text{Mn})_3\text{O}_4$, $(\text{Ni},\text{Mn},\text{Fe},\text{Co})_3\text{O}_4$; (C) hematite $(\text{Fe},\text{Ti})_2\text{O}_3$ (19).

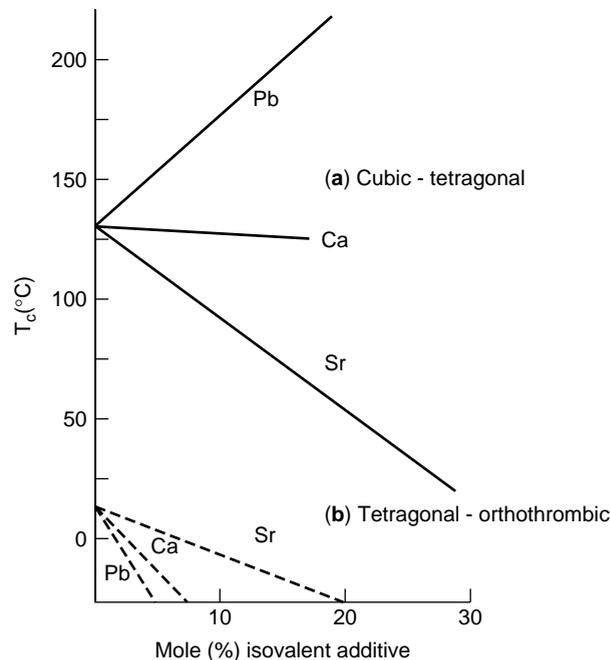


Figure 19. Curie temperature control of BaTiO_3 -based PTCR materials that contain isovalent additives: (a) cubic-tetragonal phase transformation; (b) tetragonal-orthorhombic phase transformation (20).

that can be controlled by many hundreds of degrees Celsius by forming solid solutions using Pb (raises T_C) or Sr (lowers T_C), as illustrated in Fig. 19. Other additives are added to control the level of resistivity of the grains, the height of the space charge barrier, and to aid in densifying the ceramic body. These are summarized in Table 7.

SMART CHEMICAL SYSTEMS

Applications

Auto Exhaust Sensor. Smart chemical systems are of growing importance, given society's increasing sensitivity to and concern for changes in the environment. Sensors are needed to detect and monitor an increasing number of chemical species. Gases such as NO_x , SO_2 , and CO_2 , which contribute to smog, acid rain, and the greenhouse effect, are of particular concern. An important example of such a smart system is represented by the auto exhaust sensor that, in conjunction with the fuel injector, maintains the air-to-fuel (A/F) ratio within narrow limits. The sensor, based on the oxygen ion conductor calcia or yttria stabilized zirconia, is designed to measure the oxygen partial pressure of the exhaust gas across more than fifteen orders of magnitude that represent conditions ranging from fuel-rich to fuel-lean operation. By maintaining A/F close to the so-called stoichiometric ratio, the *three-way catalyst*, placed in the exhaust stream, can simultaneously reduce NO_x , CO, and hydrocarbon emissions with high efficiency. Figure 20 illustrates the position of the zirconia sensor within the auto exhaust manifold, the feedback loop to the electronic control unit (ECU), and the fuel injector.

Table 7. Chemistry Considerations for PTC Sensor Manufacture: Additions Impacting Barium Titanate Semiconductor Properties^a

Addition Type	Addition Function	Typical Substitution Level for Ba or Ti	Addition Elements Used
Isovalent	Solid-solution substitution for Ba or Ti in the titanate	0–0.20	Sr, Pb, Ca, Sn
Altrivalent	Semiconduction or resistivity control	0–0.008	Y, La, Sb, Ce, Nb
Boundary Sintering	Barrier layer control	0–0.002	Mn, Fe, V, Cu
Poisons	Liquid-phase sintering	0.005–0.020	Si, Ti, Ge
	Impurities that degrade PTC quality	0.0002–0.0050	Na, K, Al, P, Mg, transition metals

^aRef. 20.

Smart Gas Sensors Demonstrating Selectivity. The performance materials depend on their chemical makeup. In the semiconductors, such as silicon (Si), even a parts per million (ppm) variation in its chemical makeup has a marked impact on properties. Traditionally, in chemicals production, samples extracted from the process line are chemically analyzed off-line. This procedure is slow, awkward, and often costly. Increasingly, manufacturers are looking for sensors that can operate *in situ* and provide continuous feedback regarding the chemical state of the process. A smart system also provides means for modifying the

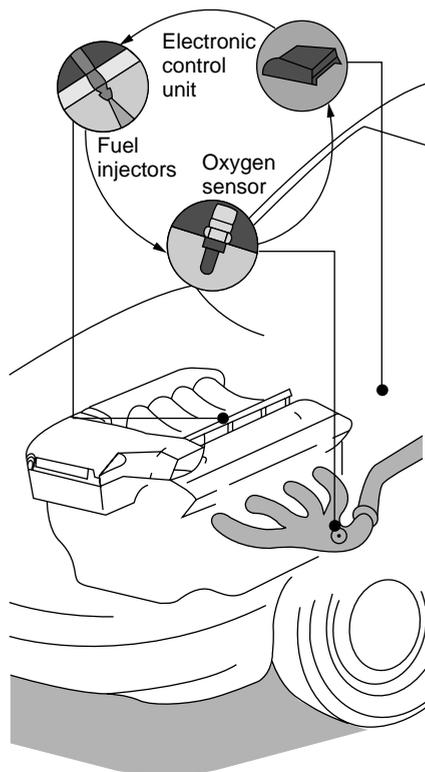


Figure 20. The feedback loop between the oxygen exhaust sensor, electronic control unit, and fuel injector used to maintain the air-to-fuel ratio within prescribed limits (22).

process to maintain it within prescribed limits. For example, this might be achieved by controlling the flow rate of certain gases or liquids into the system via an electromechanical flow valve or in principle via an electrochemical pump (24). Similar schemes are under development for drug delivery to patients in which the blood chemistry is continuously monitored and drugs administered appropriately. Gases may also represent health or safety hazards. In contrast to the auto exhaust sensor, which is designed to monitor only oxygen, gas sensors for many of the applications discussed earlier require additional levels of intelligence to monitor a number of gases simultaneously and distinguish among potentially interfering gases.

As in the flow sensor, integration of electroceramics into microelectromechanical systems (MEMS) provides important advantages. Figure 21 shows an image of a 2×2 array of NIST developed microhotplate-based sensors (25,26). Each microhotplate consists of a low thermal mass, resistively heated membrane thermally isolated from the remainder of the silicon wafer by thin tethers. A cross section of the device (Fig. 22) illustrates its multilayer nature that includes a polysilicon heater, a metal thermometer/heat distribution plate, and

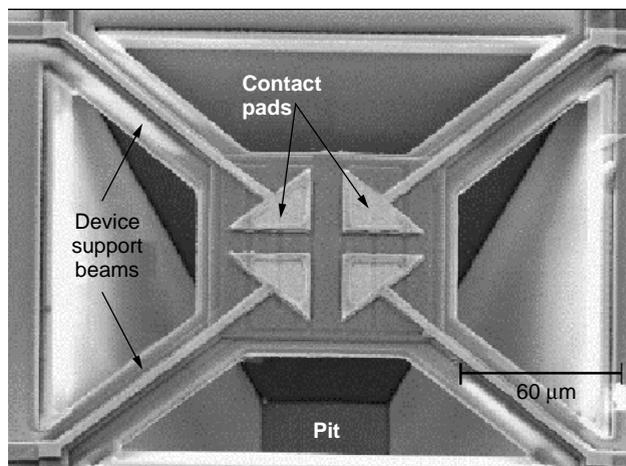


Figure 21. A 2×2 array of microhotplate gas sensor platforms developed at NIST (25).

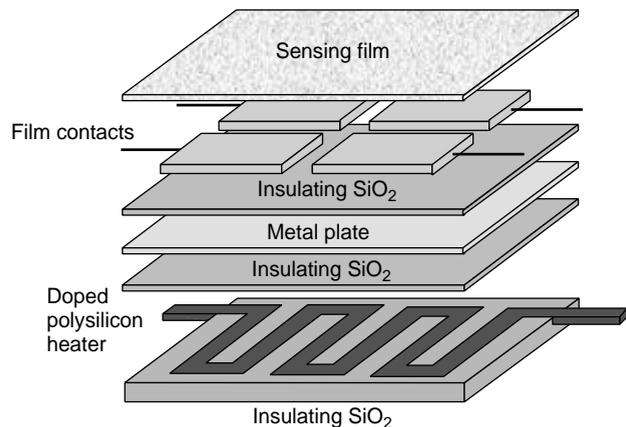


Figure 22. A cross section of the microhotplate sensor in Fig. 21 that show heater, thermometer, contact pads, and sensing film (26).

sensing film electrical contacts separated by insulating SiO_2 layers. Electroceramic films such as SnO_2 or ZnO are deposited onto the top of each device, and gases are sensed by monitoring the change in the resistance of these films. Key features of the device include low power consumption (50 mW for continuous 500°C operation), short turn-off and turn-on times ($\sim\text{ms}$), and wide operating temperature range (typically $20\text{--}500^\circ\text{C}$). The short heating and cooling times enable programmed temperature cycling that leads to improved gas selectivity, as illustrated in Figure 23. These results represent the response of a number of gases, including acetone, formaldehyde, ethanol, and methanol, to a program that consists of ten 100 ms pulses at each of eight temperatures ranging from $20\text{--}370^\circ\text{C}$ in 50°C increments. These gases are normally difficult to distinguish by using conductometric sensors such as SnO_2 when operated at a fixed temperature. By cycling the temperature, characteristic thermal fingerprints

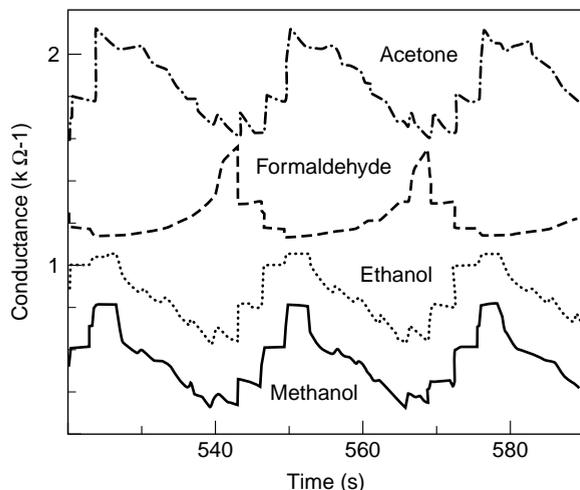


Figure 23. Conductance-time plots showing the response of SnO_2 films to period changes in sensor temperature. Note that each gas exhibits a distinctive thermal “fingerprint” (25).

of the different gases are obtained that are more easily distinguished. By using arrays of sensors, each optimized for different gases, the types of gases and their relative concentrations can be determined much more accurately (28). Because these hotplates are fabricated onto silicon wafers, the option exists to integrate the necessary electronics directly onto the same chip. Similar microhotplate-based sensors are now being micromachined from the more refractory SiC enabling such sensors to operate *in situ* at high temperature and/or in chemically corrosive environments (29,30).

Operating Principles

Electrolytic. Solids that have largely ionic bonding and large band gaps exhibit predominantly ionic conductivity. When these solids are highly defective and the crystalline structure provides high diffusivity pathways between equivalent sites, they may exhibit ionic conductivities comparable to those of liquids (31). Such solids are typically called *solid electrolytes* or *fast ion conductors*. When a concentration gradient of a chemical species contained in the electrolyte is imposed across it, a voltage, known as the Nernst emf is induced across it. For an oxygen ion conducting electrolyte such as stabilized zirconia mentioned before, the voltage takes the form

$$E_N = (kT/4q) \ln(P\text{O}_2''/P\text{O}_2'), \quad (16)$$

where $P\text{O}_2''$ and $P\text{O}_2'$ are the oxygen partial pressures on the two sides of the electrolyte. Examining a schematic of an oxygen sensor used in automotive exhaust systems (see Fig. 24), if $P\text{O}_2'$ is set to a reference gas such as air, then the oxygen partial pressure in the exhaust gas $P\text{O}_2''$ can be determined from the induced output voltage (E in the figure) and the sensor temperature T . Such sensors are called *potentiometric* sensors (32).

By applying a voltage across an electrolyte, it is possible to pump chemical species electrochemically from one chamber to the other thereby modifying its chemical

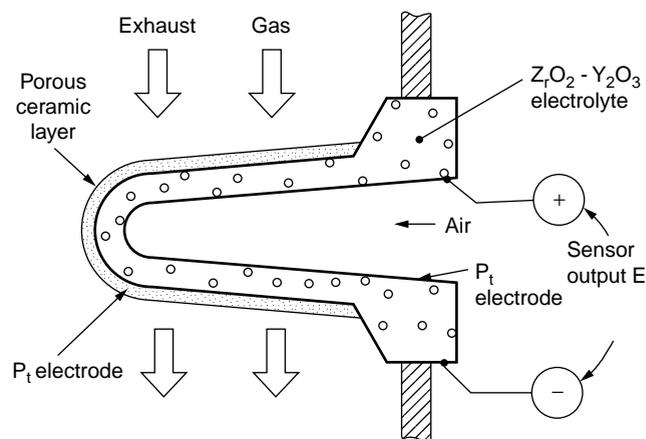


Figure 24. Schematic of a zirconia-based potentiometric auto-exhaust sensor. The sensor is shown inserted into the wall of the exhaust manifold.

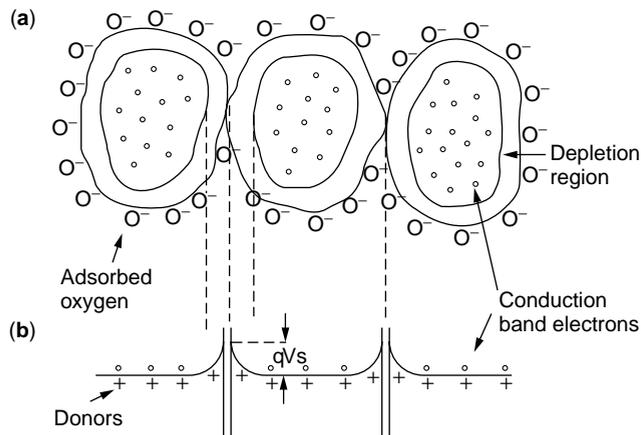


Figure 25. (a) Illustration of semiconductor grains in contact with oxygen ions adsorbed at the surfaces of the grains thereby forming depleted regions near the surfaces. (b) An energy band diagram, as in Fig. 16 that shows barriers of energy qVS between grains (34).

composition. Such *amperometric* devices may be used as sensors or as actuators (24,33).

Conductometric. Compound semiconductors exposed to different gases often exhibit changes in their resistance or equivalently their conductance. At lower temperatures, this effect is limited to gas molecules adsorbing and desorbing from the surface of the semiconductor. Such molecules may form energy states at the surface of the semiconductor and trap free carriers thereby forming depletion layers and space charge barriers at the interface (34), as illustrated in Fig. 25. In the most common electroceramic gas sensor materials, such as SnO_2 , these barriers block the flow of charge carriers between adjacent grains in porous polycrystalline ceramics. Adsorbed oxygen increases the barrier height, and reducing gases decrease it. Such sensors exhibit high sensitivity, but this detection mechanism does

not readily distinguish between different reducing gases. Improved selectivity is achieved by adding catalysts or incorporation into arrays of the type discussed earlier.

At sufficiently high temperatures, the bulk of the semiconductor is reduced or oxidized, depending both on temperature and gas composition. The defects formed during the redox process act as donors or acceptors and thereby modify the conductance of the semiconductor. In general, the conductivity σ follows a well-defined power law dependence on the gas partial pressure (20):

$$\log \sigma = \log (ne\mu) = \frac{1}{r} \log P_{\text{O}_2} + C(T), \quad (17)$$

where the factor r depends on the defect regime (negative for electrons, positive for holes) and $C(T)$ is a temperature-dependent constant. Sensors of this type require means for compensating for cross sensitivity to temperature [the $C(T)$ term] and for achieving reduced response times (20).

Materials Systems

Solid Electrolytes. In the most common form of potentiometric sensor, the chemical species to be sensed is also the mobile ion in the solid electrolyte. For example, the oxygen ion conductor, yttria stabilized ZrO_2 , may be used to monitor oxygen partial pressure. Likewise, one would use a sodium ion conductor, such as sodium beta alumina, to monitor the sodium concentration (or activity) in a solution or melt that contains sodium. In Table 8, we list a number of popular solid electrolytes and their properties.

A number of factors need be considered in selecting a particular solid electrolyte for sensor and/or actuator application. Considering the materials listed in Table 7, phase stability is an obvious criterion. A number of the materials go through phase transitions from relatively insulating to conductive states. One must consider the operating range required before selecting a candidate electrolyte. For potentiometric (open circuit) measurements, the magnitude of conductivity is not as critical as it is in amperometric applications where a given current must flow through the

Table 8. Mobile Ions and Electrical Properties of Solid Electrolytes

Material	Mobile Ion	Properties	Remarks
$\text{Zr}_{0.85} \text{Y}_{0.15} \text{O}_{2-x}$ (35)	O^{2-}	$\sigma(1000^\circ\text{C}) = 0.12 \text{ S/cm}$ $E = 0.8 \text{ eV}$	Material of choice in auto exhaust sensors
$\text{Ce}_{0.95} \text{Y}_{0.05} \text{O}_{2-x}$ (36)	O^{2-}	$\sigma(1000^\circ\text{C}) = 0.15 \text{ S/cm}$ $E = 0.76 \text{ eV}$	Semiconducting at low PO_2
$\alpha\text{-AgI}$ (37)	Ag^+	$\sigma(1000^\circ\text{C}) = 1.6 \text{ S/cm}$ $E = 0.1 \text{ eV}$	Phase transition at 146°C
$\text{RbAg}_4 \text{I}_5$ (37)	Ag^+	$\sigma(1000^\circ\text{C}) = 0.3 \text{ S/cm}$ $E = 0.09 \text{ eV}$	
$\alpha\text{-CuI}$ (38)	Cu^+	$\sigma(450^\circ\text{C}) = 10^{-1} \text{ S/cm}$ $E = 0.15 \text{ eV}$	Phase transition at 407°C
$\text{Na } \beta \text{ alumina}$ (39)	Na^+	$\sigma(100^\circ\text{C}) = 10^{-4} \text{ S/cm}$ $E = 0.48 \text{ eV}$	Stoichiometry varies between $\text{Al}_2\text{O}_3/\text{Na}_2\text{O} = 5.3 - 8.5$
$60\text{Li}_2 \text{S} - 40\text{SiS}_2$ (40)	Li^+	$\sigma(25^\circ\text{C}) = 5 \times 10^{-4} \text{ S/cm}$ $E = 0.25 \text{ eV}$	Amorphous
$\beta \text{ PbF}_2$ (41)	F^-	$\sigma(100^\circ\text{C}) = 10^{-4} \text{ S/cm}$ $E = 0.48 \text{ eV}$	

Table 9. SnO₂-Based Gas Sensors for Emission Detection

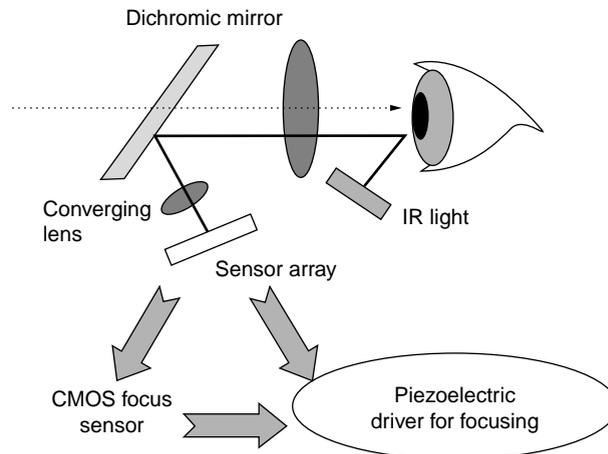
Materials	Gas	Temperature (°C)
SnO ₂ + ThO ₂ + SiO ₂	CO	190–210
SnO ₂ + Sb	CO	440–460
SnO ₂ + Bi ₂ O ₃	CO	430–500
SnO ₂ + Pd+Cu	CO	100–200
SnO ₂ + Sb ₂ O ₃	H ₂	180–440
SnO ₂ + Pd	H ₂	230–260
SnO ₂ + In	H ₂	330–380
SnO ₂ + Bi ₂ O ₃	NO _x	200–430
SnO ₂ + Cd	NO _x	250–400
SnO ₂ + V	NO _x	290–310
SnO ₂ + Pd	NO _x	200–300
SnO ₂ + Bi ₂ O ₃	CH ₄	480–520
SnO ₂ + Bi	CH ₄	430–480
SnO ₂ + La ₂ O ₃	C ₂ H ₅ OH	310–330
SnO ₂ + Pd	C ₂ H ₅ OH	380–450

device. In general, higher conductivities are advantageous even for potentiometric devices because they enable the use of low input impedance voltmeters. Secondly, the electronic conductivity must typically be less than about one part per thousand. Electronic conductivity provides short circuiting pathways for the atoms through the electrolyte, even under open circuit conditions. Such permeation may result in polarizing the electrodes and generating overpotentials that disturb the Nernst potential. Because electronic carriers are typically created under sufficiently oxidizing or reducing conditions, a key criterion is the “electrolytic domain width” that specifies the range of chemical activities for which the electrolyte remains predominantly ionic. These and related issues are discussed in (42).

Semiconductors. Many semiconducting oxides have been examined for their response to various gases, but the greatest emphasis by far has been on several of the binary oxides, including SnO₂, ZnO, and TiO₂.

It is likely that fabricating them is typically much simpler and more reproducible than the more complex oxides. Furthermore, a large band gap is desirable (≥ 3 eV) to insure minimal electron–hole generation even at operating temperatures of 250–450°C. Finally, outside of its surface adsorption/desorption characteristics, the material should remain inert to minimize aging effects. For this reason, SnO₂ has been the material of choice for surface controlled semiconducting gas sensors. Its versatility is illustrated in Table 9 (34).

These results demonstrate that selectivity to specific gases can be enhanced by the adding various dopants and/or catalysts.

**Figure 26.** An example of a Canon EOS SLR camera focusing system that shows the use of photovoltaic and pyroelectric materials.

For high temperature measurements, bulk conductive properties are impacted. Here, materials that equilibrate rapidly with the atmosphere (i.e., have high oxygen diffusivities) and exhibit well-defined dependence on the atmosphere are desired. These and other key figures of merit are summarized in Table 10.

Materials under investigation for such purposes include SrTiO₃, TiO₂, Nb₂O₅, and CeO₂ (20,34,43).

OPTICAL MATERIALS

Applications

Light sensors are used in many applications: pyroelectric materials in automatic door openers and automatic light switches, photovoltaic and photoconductive materials in cameras, and electrochromic materials for smart windows.

Camera Focusing System. A good example of integrating smart materials and systems is the EOS camera focusing system by Canon. The system contains three major components; two sensors and one actuator. The first sensor is an IR sensor array that follows the eye movement and analyzes the direction in which the user is looking. The second sensor is a CMOS sensor array to measure distance to the object. The data from these two sensors controls a piezoelectric motor that moves the optical lenses and changes the focal length according to the object the user is looking at. Figure 26 shows a schematic diagram of the system. Another smart component in the system is the dichromic window. It allows the passage of visible

Table 10. Materials Requirements for Resistive Exhaust Gas Oxygen Sensors

1. High electronic conductivity strongly dependent on oxygen partial pressure
2. Small temperature coefficient of resistance
3. High mobility of oxygen-related lattice defects (e.g., oxygen vacancies)
4. High porosity and small particle size (for ceramics)
5. High catalytic activity and good oxygen-transfer characteristics
6. Wide oxygen partial pressure and temperature range of stability

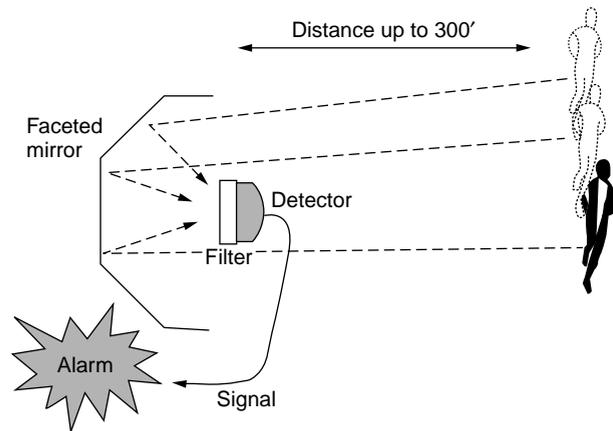


Figure 27. Pyroelectric motion detector and intruder alarm. A simpler design can be used to open doors automatically.

light but reflects IR. This allows it to minimize the number of optical components in the system and simplify the design.

Pyroelectric Motion Detector. Pyroelectric motion detectors are very common in intruder alarm systems (Fig. 27) and in automatic door openers and light dimmers. These devices are designed to sense the blackbody radiation emitted from the human body ($T \approx 300$ K, $\lambda \approx 10$ μm). A short wavelength filter that has a cutoff of about 5 μm is used to prevent the device from responding to changes in background lighting levels (14).

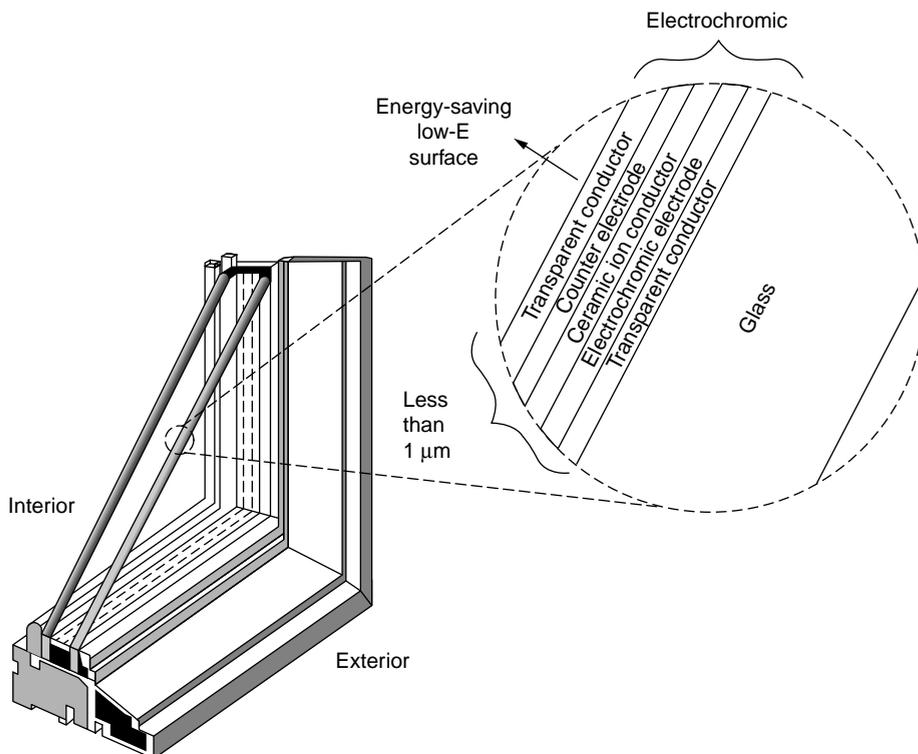


Figure 28. Cross section of a smart electrochromic window illustrating the five-layer structure (46).

Electrochromic Windows

Smart electrochromic windows are used to conserve energy by manipulating their optical reflectivity in response to changes in the ambient illumination. Thus, on a hot summer day, the reflectivity edge is shifted to longer wavelengths to reduce the amount of visible light allowed into the building. Alternatively, in the evening, the reflectivity edge is shifted to short wavelengths to allow the heat from the building to radiate outward. The system is based on the ability to insert and reversibly remove protons and alkali ions such as Li, electrochemically, from certain oxides such as WO_3 . For example, when Li atoms are inserted into WO_3 , the ions enter the lattice interstitially, and the electrons enter directly into the conduction band. An increasing free electron density increases the plasma energy E_P ($\propto n^{1/2}$) and thus shifts the reflectivity edge to shorter wavelengths (44). The electrical conductivity can be shifted in this manner from as little as 10^{-6} to as much as 5×10^4 S/cm (45).

The cross section of such a window is shown in Fig. 28. Aside from the outer transparent conducting electrodes, the key layers include the solid electrolyte middle layer (see earlier section), the electrochromic electrode (e.g., WO_3), and the counterelectrode that itself may be electrochromic (e.g., LiCoO_2) but of a complementary nature (i.e., it is transparent when loaded with Li and becomes more reflective as Li is removed). In reality, this structure represents a thin film battery whose optical properties depend on its state of charge. The overall smart system requires an optical sensor to monitor the light intensity, a microprocessor to evaluate this data in the light of other data, for example,

the time of day, the date, and temperature, and the actuator that is an electrochemical pump whose state of charge determines its reflectivity. Figures of merit include rate of reflectivity shift, ratio of maximum-to-minimum reflectivity, number of reproducible cycles, life, and cost. Another key application would be a smart automobile rear view mirror whose reflectance is conditioned on the amount of glare that reaches the mirror from other traffic.

Basic Relationships and Phenomena

Index of Refraction, Birefringence, and the Electro-optic Effect. The index of refraction n is defined as the ratio between the speed of light in vacuum and that in a given material. The index of refraction is related to the electric and magnetic permittivities by

$$n = \frac{c_0}{c} \Leftrightarrow n = \sqrt{\frac{\varepsilon}{\varepsilon_0} \cdot \frac{\mu_0}{\mu}} \approx \sqrt{\frac{\varepsilon}{\varepsilon_0}}. \quad (18)$$

The last approximation is justified for nonmagnetic materials and for the range of visible light where the magnetic permittivity is nearly constant.

In nonisotropic materials, these values differ in different crystallographic directions, and one has to look at the dielectric tensor. It is convenient to work in the major electro-optic axis (where the dielectric constant is diagonalized), as shown in Fig. 29. Materials that have two polarization vectors of light propagating at different speeds are called birefringent.

The dielectric constant (and therefore the index of refraction) is related to polarization in the material by the Clausius–Mossotti equation (47):

$$(\varepsilon - 1)(\varepsilon + 2)^{-1} = \frac{4\pi}{3V}\alpha, \quad (19)$$

where V is the unit-cell volume and α is the polarizability. This implies that changing the polarization can change the index of refraction. Therefore, piezoelectric materials

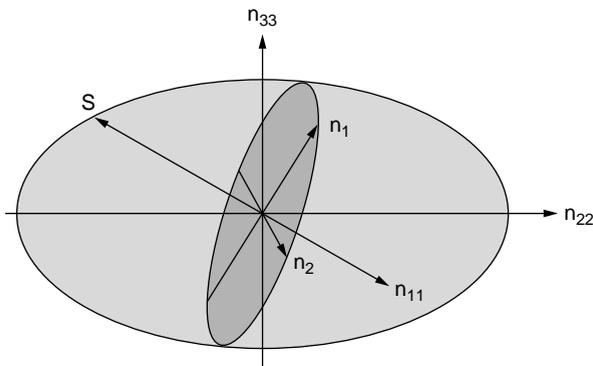


Figure 29. The index ellipsoid for n_{11} , n_{22} , and n_{33} as major axis indexes. For light propagating in the S direction, the field can be considered one that has two polarizations—one has index n_1 , and the other has index n_2 . The ellipse perpendicular to S defines the polarization plane.

usually also show electro-optic behavior, a change in the optical index of refraction due to an applied electric field. From the Clausius–Mossotti equation, one can derive the change in index as a function of strain and polarization (48):

$$\frac{\Delta n_i}{n_i} = \frac{(n_i^2 - 1)(n_i^2 + 2)}{6n_i^2} \left(-\frac{\Delta V}{V} + \frac{\Delta \alpha_i}{\alpha_i} \right). \quad (20)$$

In our previous discussion of electromechanical actuation, we described three types of a ctuation: piezoelectric, electrostrictive, and phase transition. The equivalent electro-optic behaviors in piezoelectric materials are (1) quadratic birefringence (Kerr effect), (2) linear birefringence (Pockels effect), and (3) memory scattering. The different behaviors for these effects and a typical setup to generate them are shown in Fig. 30.

Photoconductive and Photovoltaic Materials. The electron–hole pair production rate R in semiconductors depends on the flux of light ϕ , the absorption efficiency η , and the illuminated volume (limited by either the thickness or the characteristic absorption depth) (50):

$$R = \frac{\eta\phi}{V}. \quad (21)$$

These excess charge carriers are free to move in the material until they recombine and therefore contribute to a change in conductivity:

$$\Delta\sigma = e(\mu_e + \mu_h)\Delta n = \frac{e(\mu_e + \mu_h)\eta\tau}{V}\phi, \quad (22)$$

where μ is the mobility for holes and electrons and τ is the excess-carrier recombination lifetime.

The rate of photogeneration G under optical intensity I can be written as

$$G = s(N_D - N_D^+)I, \quad (23)$$

where s is the photoionization cross section, N_D is the density of dopants (donors or acceptors), and the $+$ superscript represents ionized dopants. The recombination rate R is

$$R = \gamma_R n N_D^+ \quad (24)$$

where γ_R is the recombination constant that is inversely proportional to the recombination lifetime. And by comparing Eqs. (15) and (16) at equilibrium, one can calculate the density of carriers n :

$$n = \frac{s}{\gamma_R} \frac{N_D - N_D^+}{N_D^+} I. \quad (25)$$

The photovoltaic current density \mathcal{J}^{ph} is given by (51):

$$j_i^{\text{ph}} = \beta_{ijk} E_j E_k^*, \quad (26)$$

where E_i is the electric field component of the light wave.

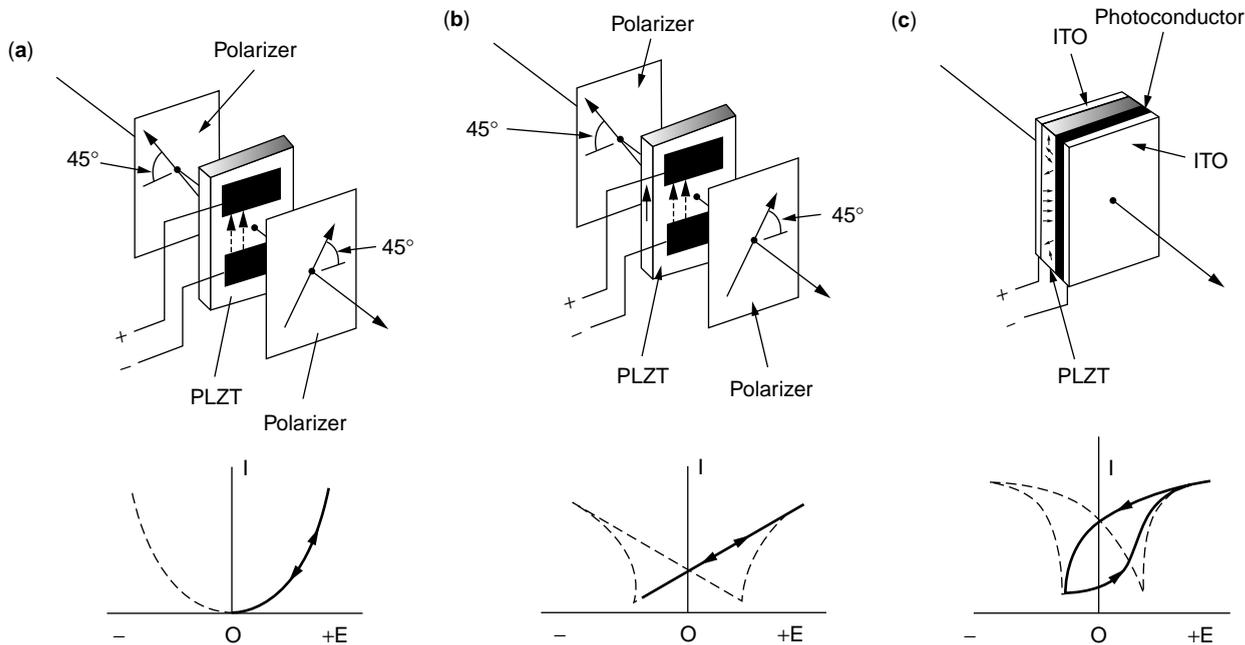


Figure 30. Operational configurations and typical light output response for (a) quadratic birefringence, (b) linear birefringence, and (c) memory scattering (49).

Pyroelectric Materials. Pyroelectric materials are a subgroup of piezoelectric materials (see Fig. 6). Pyroelectric materials experience a change in polarization due to a change in temperature, which makes them excellent candidates for IR detectors. Photoconductivity in pyroelectric crystals can be accompanied by current flow without an external field due to the anisotropy, as shown in Fig. 31.

The pyroelectric coefficient is defined as the change in spontaneous polarization P_s as a function of temperature T :

$$p = -\frac{\partial P_s}{\partial T}. \quad (27)$$

In general, polarization increases as temperature decreases, resulting in a positive pyroelectric coefficient.

The responsivity r_i of the material is defined as the current density per unit power of the radiating light. For a

sample of area A and radiation power W ,

$$r_i = \frac{1}{WA} \frac{\partial q}{\partial t} \propto \frac{p}{\rho \cdot c_p}. \quad (28)$$

In this equation, ρ is the density and c_p the heat capacity. Materials that have a high heat capacity take longer to change their temperature and are therefore slower sensors. Because response time is important, the ratio between the pyroelectric coefficient and the heat capacity is used as a figure of merit. A list of modified figures of merit for different applications is given in Table 11.

Pyroelectric materials are used mainly for detecting infrared radiation. The pyroelectric effect requires a change in temperature to produce a change in polarization. This means that only changes in incident radiation can be measured. When response from a stationary object is required, a radiation chopper is used. Such choppers can be rotating

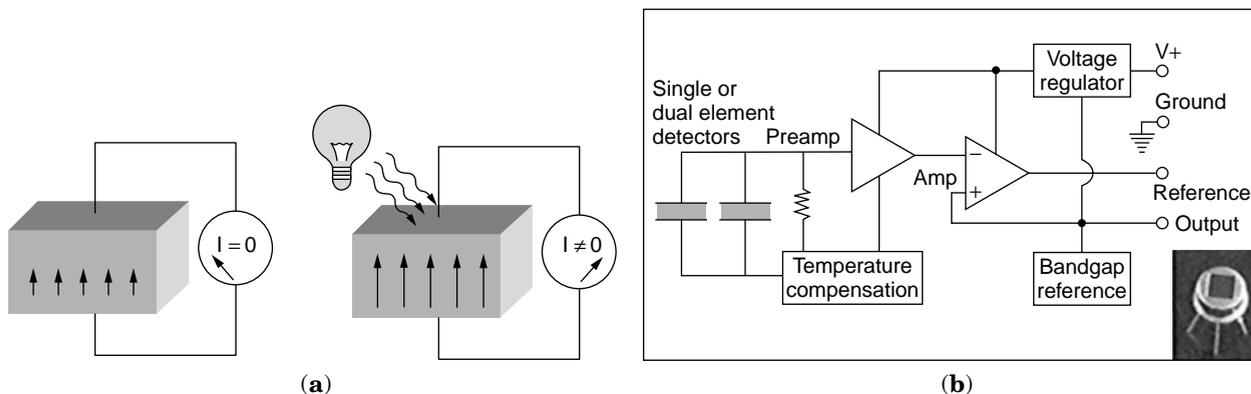


Figure 31. Pyroelectric effect. (a) The change in polarization due to heating of the surface causes current to flow in the circuit. (b) Example of a commercial pyroelectric sensor (442-3 IR-EYE™ by Acroname Inc.).

Table 11. Figures of Merit for Pyroelectric Materials^a

Application	Figure of Merit	Symbols
Low impedance amplifier	$\frac{p}{c_p}$	p : Pyroelectric coefficient c_p : Specific heat
High impedance amplifier	$\frac{p}{\epsilon c_p}$	ϵ : Relative permittivity α : Thermal diffusivity
Thermal imaging device	$\frac{p}{\alpha \epsilon c_p}$	

^aRef. 52.**Table 12. Properties of Pyroelectric Materials^a**

Material	p (μ C/m ² K)	ϵ	$\tan \delta$	C_P (MJ/m ³ K)	T_C (°C)	$\frac{P}{\epsilon C_P}$ (m ² C ⁻¹)
TGS, single crystal	280	38	1×10^{-2}	2.3	49	0.36
DTGS, single crystal	550	43	2×10^{-2}	2.4	61	0.60
LiTaO ₃ , single crystal	230	47	5×10^{-3}	3.2	665	0.17
(Sr _{0.5} Ba _{0.5})Nb ₂ O ₆ , single crystal	550	400	3×10^{-3}	2.3	121	0.07
PbZrO ₃ , modified	3.8	290	3×10^{-3}	2.5	230	0.0005
PbTiO ₃ , modified	4.2	220	5×10^{-3}	2.5	494	0.0008
(Pb _{0.94} La _{0.06})(Zr _{0.8} Ti _{0.2})O ₃	760	1000	3×10^{-3}	2.6	230	0.03

^aBased on (6,14,54).

discs, although new piezoelectric bimorphs have been developed (53).

Another important design consideration derives from the fact that pyroelectric materials are a subgroup of piezoelectric materials. Any external stress or electric field can cause an electric signal that will interfere with the measured response. Furthermore, any small changes in environment temperature will further compromise the measurements. To compensate for these effects, a second pyroelectric element, identical to the measurement unit, is set up with a light-blocking layer (see Fig. 31). Any environmental changes create the same voltage on both elements, and therefore any voltage difference between them is due to the incident light.

Materials Systems

Some key pyroelectric materials and their properties are listed in Table 12. The temperature change in the sample is larger the smaller the specific heat under constant heating, and the voltage generated by a given induced charge is larger for a smaller dielectric constant. Other factors to consider include materials stability, reproducibility, and ease of fabrication. TGS has a high figure of merit, but it is a fragile and water-soluble crystal. Lithium tantalate is readily grown as a single crystal, is insensitive to moisture, and stable in vacuum to temperatures required to enable outgassing (14).

CONCLUSIONS

Electroceramics, as has been shown, exhibit a wide range of properties that make them ideally suited to serve as transducers in the form of sensors and/or actuators. In some cases, as in the PTCR self-regulated heater, the device is inherently smart and requires no microprocessor

to respond appropriately to a stimulus. In the vast majority of cases, a microprocessor is needed for the appropriate feedback. Here, the start of a trend toward integrating electroceramics into microelectronic and microelectromechanical systems (MEMS) will hasten the development of self-contained miniaturized smart systems. The microhotplate gas sensor array represents a graphic example. Last, we have emphasized transduction to and from electrical forms of energy because this is what most signal handling systems need to operate. This is beginning to change, particularly as all optical systems develop. Here again, given the electro-optic, piezo-optic, thermo-optic, magneto-optic effects that are exhibited by electroceramics, this group of materials should make important contributions to smart optical systems.

ACKNOWLEDGMENTS

The authors acknowledge financial support for research related to the topic of this article by the National Science Foundation under Contract No. DMR-9701699, the Army Research Office under Contract No. DAAH04-95-1-0104, and Nanovations Corporation via the MIT Microphotonics Laboratory.

BIBLIOGRAPHY

1. R.E. Newnham, Smart, *MRS Bull.* **18**: 24–26 (1993).
2. Y. Yamashita, *Jpn. J. Appl. Phys. Pt I* **37**: 6a (1998).
3. K. Uchino, *Piezoelectric Actuators and Ultrasonic Motors*. Kluwer Academic, Norwood, MA, 1997.
4. K. Uchino, *Ceram. Bull.* **65**: 4 (1986).
5. J.E. Nye, *Physical Properties of Crystals*, Oxford, 1985.
6. K. Uchino, *Ferroelectric Devices*, Marcel Dekker, NY, 2000.
7. D.S. Ballantine et al., *Acoustic Wave Sensors*, Academic Press, 1997.

8. S. Takahashi, in *Ferroelectric Ceramics*, N. Setter and E.L. Colla, eds., Birkhauser Velgar, Basel, 1993, pp. 349–362.
9. G.H. Haertling, in *Ceramic Materials for Electronics*, R.C. Buchanan, ed., Marcel Dekker, NY, 1991.
10. C. Heremans and H.L. Tuller, *J. Appl. Phys.* **87**(3): 1458–1486 (2000).
11. L.E. Cross, in *Ferroelectric Ceramics*, N. Setter, E.L. Colla, eds., Monte Verita, 1993.
12. A. Avrahami and H.L. Tuller, unpublished.
14. A.J. Moulson and J.M. Herbert, *Electroceramics Materials, Properties, Applications*. Chapman and Hall, 1990.
15. Morgan Matroc, Inc. *Guide to Modern Piezoelectric Ceramics*.
16. <http://www.pelonis.com/oemtpc.html#Technology>
17. G.T.A. Kovaks, *Micromachined Transducers Sourcebook*, McGraw-Hill, Boston, 1998, p. 599.
18. AWM Series Airflow Sensors, Honeywell Micro Switch Sensing and Control, Freeport, IL.
19. http://content.honeywell.com/sensing/prodinfo/massairflow/catalog/c15_64.pdf
20. D.C. Hill and H.L. Tuller, in *Ceramic Materials for Electronics*, 2nd ed., R.C. Buchanan, ed., Marcel Dekker, NY, 1991, pp. 249–347.
21. G.E. Pike and C.H. Seager, *J. Appl. Phys.* **50**: 3414–3422 (1979).
22. M. Kuwabara, in *Additivies and Interfaces in Electronic Ceramics*, M.F. Yan and A.H. Heuer, eds., Advances in Ceramics, Vol. 7, American Ceramic Society, Columbus, OH, 1983, p. 128.
23. <http://www.ramac.com/bosch.htm>
24. H.L. Tuller and M.W. Barsoum, *Proc. 1985 Int. Conf. Solid State Sensors Actuators—Transducers '85*, IEEE, Piscataway, NJ, 1985, pp. 256–259.
25. <http://www.cstl.nist.gov/div836/836.04/SensorProj>
26. J. Suehle, R.E. Cavicchi, M. Gaitan, and S. Semancik, *IEEE Electron Device Lett.* **14**: 118–120 (1993).
27. R.E. Cavicchi, J.S. Suehle, P. Chaparala, K.G. Kreider, M. Gaitan, and S. Semancik, *Proc. 1994 Solid State Sensor Actuator Workshop*, Hilton Head, S.C., 1994, pp. 53–56.
28. R.E. Cavicchi, J.S. Suehle, K.G. Kreider, M. Gaitan, and P. Chaparala, *IEEE Electron Device Lett.* **16**: 286–288 (1995).
29. H.L. Tuller and R. Mlcak, *J. Electroceramics* **4**: 415–425 (2000).
30. R. Mlcak, Boston MicroSystems, Inc. Woburn, MA, private communication.
31. H.L. Tuller and P.K. Moon, *Mat. Sci. Eng. B* **1**: 171–191 (1988).
32. N. Yamazoe and N. Miura, *J. Electroceramics* **2**: 243–244 (1998).
33. R.E. Hetrick, W.A. Fate, and W.C. Vassell, *Soc. Automot. Eng. (SAE)* paper 810, 433, 1981.
34. C.C. Wang, S.A. Akbar, and M.J. Madou, *J. Electroceramics* **2**: 273–282 (1998).
35. R.M. Dell and A. Hooper, in *Solid Electrolytes*, P. Hagenmuller and W. Van Gool, eds., Academic Press, NY, 1978, pp. 291–312.
36. H.L. Tuller and A.S. Nowick, *J. Electrochem. Soc.* **122**: 255–259 (1975).
37. J.S. Kasper, in *Solid Electrolytes*, P. Hagenmuller and W. Van Gool, eds., Academic Press, NY, 1978, pp. 217–235.
38. T. Matsui and J.B. Wagner, Jr., in *Solid Electrolytes*, P. Hagenmuller and W. Van Gool, eds., Academic Press, NY, 1978, pp. 237–252.
39. R. Collongues, J. Thery, and J.P. Boilot, in *Solid Electrolytes*, P. Hagenmuller and W. Van Gool, eds., Academic Press, NY, 1978, pp. 253–276.
40. F.A. Fusco and H.L. Tuller, in *Superionic Solids and Solid Electrolytes—Recent Trends*, A.L. Laskar and S. Chandra, eds., Academic Press, San Diego, CA 1989, pp. 43–110.
41. J.-M. Reau and J. Portier, in *Solid Electrolytes*, P. Hagenmuller and W. Van Gool, eds., Academic Press, NY, 1978, pp. 313–333.
42. H.L. Tuller, Mixed Conduction in Nonstoichiometric Oxides, in *Non-Stoichiometric Oxides*, O.T. Sorensen, ed., Academic Press, NY, 1981, pp. 271–335.
43. W. Menesklou, H.-J. Schreiner, K.H. Härdtl, and E. Ivers-Tiffée, *Sensors and Actuators B* **59**: 184–189 (1999).
44. R.B. Goldner, F.O. Arntz, K. Dickson, M.A. Goldner, T.E. Hass, T.Y. Liu, S. Slaven, G. Wei, K.K. Wong, and P. Zerigian, *Solid State Ionics* **70/71**: 613–618 (1994).
46. S.E. Park and T.R. Shrout, *J. Appl. Phys.* **82**: 4 (1997).
47. N.W. Ashcroft, N.D. Mermin, *Solid State Physics*, Saunders College, 1976.
48. P.D. Townsend, *Rep. Prog. Phys.* 501–558 (1987).
49. G.H. Haertling, in *Ceramic Materials for Electronics*, R.C. Buchanan, ed., Marcel Dekker, NY, 1991, pp. 129–202.
50. B.E.A. Saleh, and M.C. Teich, *Fundamentals of Photonics*. Wiley, 1991.
51. P. Gunter, J.P. Huignard, in P. Gunter, ed., *Photorefractive Materials and their Applications*, Springer-Verlag, 1988.
52. K. Uchino, *Ferroelectric Devices*, Marcel Dekker, NY, 2000.
53. K. Shibata et al., *Jpn. J. Appl. Phys.* **24**(Suppl. 24–3): 181 (1985).
54. W. Wersing, in *Ferroelectric Ceramics*, N. Setter and E.L. Colla, eds., Birkhauser Velgar, Basel, 1993, pp. 312–319. *Ame. Ceram. Soc. Bull.* **75**(4): (1996).

ELECTROCHROMIC SOL-GEL COATINGS

L.C. KLEIN
Rutgers—The State University of New Jersey
Piscataway, NJ

DEFINITION OF ELECTROCHROMISM

Certain inorganic compounds, especially oxides of polyvalent metals, exhibit coloration that depends on the oxidation state of their cations. This property leads to electrochromism, which is a reversible and visible change in transmittance and/or reflectance. The oxidation–reduction reactions are electrochemically induced, using low voltages of the order of ± 1 V dc.

An electrochromic (EC) device is a multilayer construction in which one of the layers shows electrochromic properties (Fig. 1) (1–5). An EC device operates on the principle of a galvanic cell. The best known electrochromic material is tungsten trioxide (WO_3), which forms deep blue alkali and hydrogen tungsten bronzes (M_xWO_3) on reduction. The reaction is expressed by the following

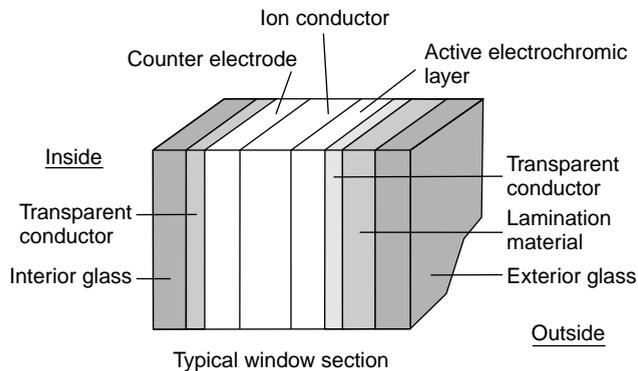
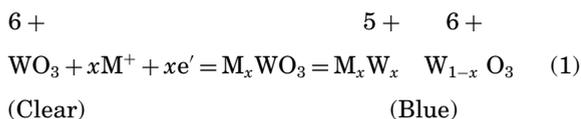


Figure 1. Schematic of an electrochromic (EC) device, substrate (indium–tin oxide coated commercial float glass), electrode (tungsten oxide), solid electrolyte (lithium-containing inorganic gel), counterelectrode (vanadium oxide), and substrate.

equation:



where M is hydrogen or an alkali and $0 < x < 1$. Typically, a low-voltage EC device that has an electrochromic cathode is colored in the charged state and bleached in the discharged state.

A major application of electrochromic devices is so-called “smart windows.” These electronically controlled windows, which can lighten or darken depending on charge insertion/extraction, are designed to adjust to the amount of sunlight, the time of the day, or the season of the year. Control of transmittance through windows is a factor in energy usage in a building, as well as the creation of a pleasant environment in interior spaces. Using them on buildings reduces the heat lost through architectural windows, if they are colored on bright, summer days and bleached on cloudy, winter days.

EC windows that fulfill these expectations have been demonstrated using thin film technologies other than sol-gel processing (6,7). Substitution of a sol-gel layer for any one of the layers—the EC electrode, the counterelectrode or the electrolyte—is a breakthrough in making affordable EC windows (8–10).

Principle of the Galvanic Cell

Recall for a moment, the basic galvanic cell. The cell consists of an anode, a cathode, and an electrolyte. In an EC device, the layers are referred to as electrode, counterelectrode, and electrolyte.

Requirements for the Electrodes. The performance of an EC device is shown schematically in Fig. 2. Assuming that the EC device is operating with a lithium ion ($\text{M} = \text{Li}^+$), when charging, the Li is oxidized at the cathode (electrode)

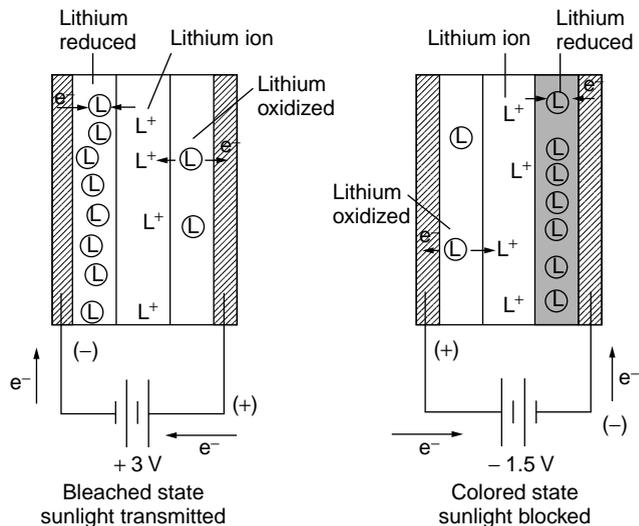


Figure 2. Schematic operation of an EC device, comparing the colored state (charged) and bleached state (discharged).

and reduced at the anode (counterelectrode)



In all cases, the cathode is an intercalation compound that has a layered or framework structure (11). In a reversible cell, which is rechargeable, both the anode and cathode are intercalation compounds (12,13). In addition to tungsten oxides, other common electrode materials are vanadium oxides, niobium oxides, and titania-doped cerium oxides (14–16).

The requirements for the electrodes are

- high coloration efficiency
- fast switching between colored and bleached states
- high diffusivity for lithium; high capacity for lithium
- compatibility with electrolyte

Requirements for Solid Electrolytes. The electrolyte, which has to transport ions back and forth between the cathode and anode, is a critical feature in the galvanic cell. (Because we have worked on solid electrolytes for about 15 years, the electrolyte receives more attention in this chapter.) The electrolyte has to have

- a wide electrochemical potential window
- stability during charge/overcharge and discharge/overdischarge.

Most commercial electrolytes are polymers and “gel” electrolytes (17,18). Some problems with commercial electrolytes are that aqueous electrolytes may generate H_2 and nonaqueous electrolytes have conductivities that are too low.

Solid electrolytes are attractive because there are effectively no moving parts, meaning no mobile protons. While a solid electrolyte is functioning as a separator and membrane, it is mechanically rigid and holds the components of

the electrochromic device fixed. But most solid electrolytes do not have high enough ionic conductivity at ambient temperature (19). One way around this problem is thin-film technology (20), which is, in part, the motivation for using sol-gel processing.

Sol-Gel Processing

A gel is a material that has solid-like properties, which means that a gel has a fixed shape. On the one hand, organic “gels”, which are frequently used in batteries, are rigid, but only in the sense that they “set” when shear stress is absent. Inorganic gels, on the other hand, that result from the sol-gel process are irreversible gels [see Brinker and Scherer (21) for general background]. These gels are based on an oxide network. These gels are rigid structures, and they are also inert in contact with organics and solvents.

A natural advantage of the sol-gel process is the fact that a film of oxide can be placed on a substrate directly from solution (22–24). Other methods of depositing films of oxides are sputtering or chemical vapor deposition (CVD). These methods may succeed in some cases, but the sol-gel process is simpler than either one.

Dipping and spinning are common processes for preparing sol-gel thin films. Among the available techniques, dip coating is more widely used. Its advantages include the following:

- deposits uniform thin films on large areas
- allows coating any size or shape
- allows control of film microstructure
- delivers multicomponent films without changing the deposition equipment
- provides flexibility in manufacturing.

Another important consequence of the sol-gel process is the fact that the film is less than a micron thick. The film is rigid, but it is thin enough to resist cracking, so that dimensional changes in the electrodes can be accommodated even when the electrolyte is in place.

Sol-Gel Processing of the Electrochromic Layer. Sol-gel thin films used in EC devices have exceptional requirements:

- significant spectral switching over the visible and infrared regions of the solar spectrum
- acceptable chemical and physical durability during a long exposure time [~ 30 year lifetime of architectural windows (25)].

There are three basic approaches to preparing transition-metal oxides (26–31). One of the routes is to prepare a solution using alkoxides, metal organic compounds with alkoxy groups. The available alkoxides are only slightly soluble in common organic solvents and tend to be very reactive with water. The second route is to use salts such as oxychlorides as the network-forming oxide precursor. The third route is to use colloidal sols (32).

A sol is a colloidal suspension of small particles (1–500 nm) in a liquid. Oxide sols are obtained after hydrolysis–condensation reactions of the precursors. Hydrolysis occurs when metal cations (M) are solvated by water molecules. Condensation occurs as soon as hydroxyl groups are present in the coordination sphere of M. These reactions lead to M–O–M bonds. When the number of links is high enough, the sol becomes a gel. The gel point occurs when the solid phase becomes continuous. After this point, the gel continues to age, primarily causing changes in structure. This is followed by drying, when evaporation of the liquid can further shrink the gel structure.

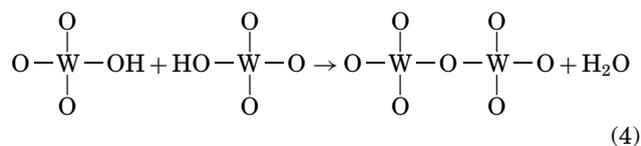
Tungsten Oxide Gels. There are several ways to obtain WO_3 thin films (14), and the least expensive precursor is tungstic acid. The reasons for choosing tungstic acid sols are

- their low cost
- their aqueous nature
- their ability to be recycled
- the presence of W–O–W bonds in the sols
- their amorphous nature and nondirectional morphology.

A drawback of tungstic acid sols is that they gel in 30 minutes.

A solution of aqueous sodium tungstate, $\text{Na}_2\text{WO}_4 \cdot 2\text{H}_2\text{O}$, is made by mixing tungstic acid powder, sodium hydroxide, and water. The solution is poured through a proton exchange resin, which is conditioned before use by washing with 2N HCl. Excess acid is removed from the column by washing with distilled water.

The sol is collected in a beaker that contains a defined amount of solvent. The sol, which is clear yellow, contains polymeric species. The mechanism of polymerization is approximately



Solvents can be used to delay gelling of the sols, but they must be easy to remove during drying and low temperature curing. A general trend is that more homogeneous films are obtained by using solvents that have higher surface tension. Wetting is also better for sols that contain aprotic solvents, rather than protic solvents. For a given solvent, less solvent results in better wetting. In addition, the higher the boiling point of the solvent, the more slowly the solvent evaporates, so drying is slower. Slow drying (hours vs. minutes) can be useful in preventing cracking of the films, although cracking is not a major issue for film thicknesses from 200–400 nm. The net result in one study was that more uniform films were obtained by using acetone at a slow dipping/withdrawal speed (~ 2.5 cm/min) (32).

As-prepared WO_3 films that are poorly crystallized have higher coloring efficiency than crystallized WO_3 (27,33). Crystallization of WO_3 thin films occurs after heating at

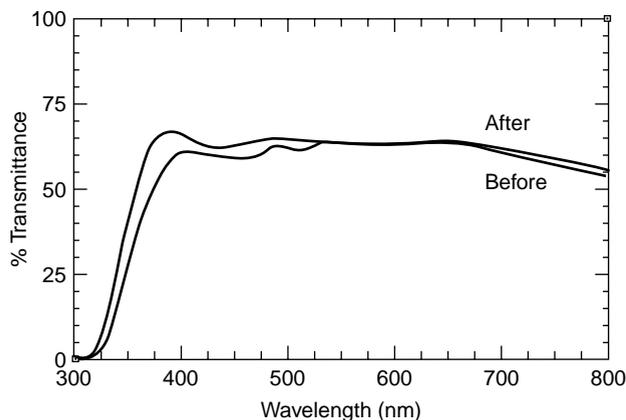


Figure 3. Percent light transmittance across the visible range (350–800 nm) for WO_3 films deposited on glass substrates with an ITO coating, measured by using a Perkin-Elmer Lambda 9 Spectrometer. The spectra were collected before and after 10 cycles.

400°C, so films should not crystallize when heat treated for one hour at only 90°C.

Comparisons between uncoated glass and coated glass showed that the transmittance in the visible range is lower by about 10–15% for coated samples. Nevertheless, the transmittance of the coated glass is acceptable for most applications. A typical transmittance curve before and after 10 lithiation cycles is shown in Fig. 3. The WO_3 sol was deposited on commercial float glass that was coated with indium tin oxide (ITO).

Vanadium Oxide Gels. Again, there are several ways to obtain V_2O_5 thin films (34–39). These include nitric acid dissolution of vanadate salts, proton exchange, hydrolysis of vanadium alkoxides, and hydration of amorphous vanadia. The least expensive precursor is vanadia powder that is melted at 1000°C ($T_{\text{melt}} = 690^\circ\text{C}$) and quenched in water to produce amorphous vanadia. This glassy material is soluble in water and yields a golden brown, colloidal sol that is suitable for dip coating. In many respects, the behavior of vanadia sols parallels that of the tungstic acid sols, so that thin, transparent coatings can be produced.

Sol-Gel Processing of the Solid Electrolyte. Sol-gel processing has been used for more than 15 years to prepare solutions that contain oxide components desirable as fast ion conductors (FIC) (19). The sol-gel process has been used to duplicate conventional crystalline fast ion conductors, as well as fast ion conducting gels and glasses. Glasses are desirable because they

- are isotropic
- have few restrictions on composition
- have more channels for fast ion conduction than crystalline compounds.

Ionic transport in solids is an activated process. In most fast ion conductors, the activation energy is indicative of the energy of motion rather than the energy to form charge

carriers. The desired features in a fast ion conductor are

- large number of mobile ions
- large number of sites for the ions
- relatively low activation energy
- wide compositional range
- ease of fabrication in complex shapes and thin films.

Classically, high ionic conductivity requires high alkali concentration. The system originally investigated for the solid electrolyte was lithium silicates (40). In the $\text{Li}_2\text{O}-\text{SiO}_2$ gels, the second component was introduced typically as lithium nitrate. The nitrate salt dissolves in water and then crystallizes in the dried gel as nitrate nanocrystals (41). Alternatively, lithium hydroxide was used to create very high pH conditions, increasing the chance of a reaction between LiOH and the silica network to give $\text{Si}-\text{O}-\text{Li}$ bonds (42). The alkoxide-based gels, or so-called polymerized gels, showed values of ionic conductivity that are stable and in the range needed for operating a lithium battery at around 250°C. Lithium nitrate gels had a characteristic behavior described by a composite mechanism (43,44).

Thermodynamic calculations were performed to find out which oxide could improve the stability of the gel. Using the standard heats of formation, it is easy to see that no oxide addition makes lithium silicates stable compared to lithium metal. At best, the addition improves the kinetic stability at low temperatures. Alumina is one candidate for increasing the kinetic stability of lithium silicate gels.

In the past, small alumina substitutions were made in lithium silicate gels to suppress crystallization (45). When the $\text{Al}_2\text{O}_3/\text{SiO}_2$ ratio was varied for a constant $(\text{LiCl})_2$ mole%, the role of Al_2O_3 was identified in charge compensation for nonbridging oxygens. The ionic conductivity increased as the LiCl content and the $\text{Al}_2\text{O}_3/\text{SiO}_2$ ratio increased (46). Consequently, silicate gels were prepared using different amounts of alumina.

Lithium aluminosilicate gels were prepared by using an alkoxide and salts (46). Tetraethylorthosilicate (TEOS, $\text{SiO}(\text{CH}_2\text{CH}_3)_4$) was used as the precursor for the network-forming oxide, silica. TEOS was added to ethanol/water mixtures that contained lithium nitrate and aluminum nitrate hydrate. For most samples studied, the Li_2O content was 15 mol%, which corresponds to a reasonable concentration for ionic conductivity.

Conductivity can be measured under dc or ac conditions, but dc conductivity measurements provide little information about these materials other than discriminating long-range ion migration from localized motions. ac conductivity measurements, on the other hand, provide information about possible conduction mechanisms (47). They allow discriminating electrode capacitance, grain boundary capacitance, and electronic contributions.

Ionic conductivity measurements were carried out by the ac complex impedance method using a Solartron 1250 frequency response analyzer and 1186 electrochemical interface, which were programmed by a Hewlett-Packard 9816 desktop computer for data collection and analysis. Experiments were performed on pressed pellets that had

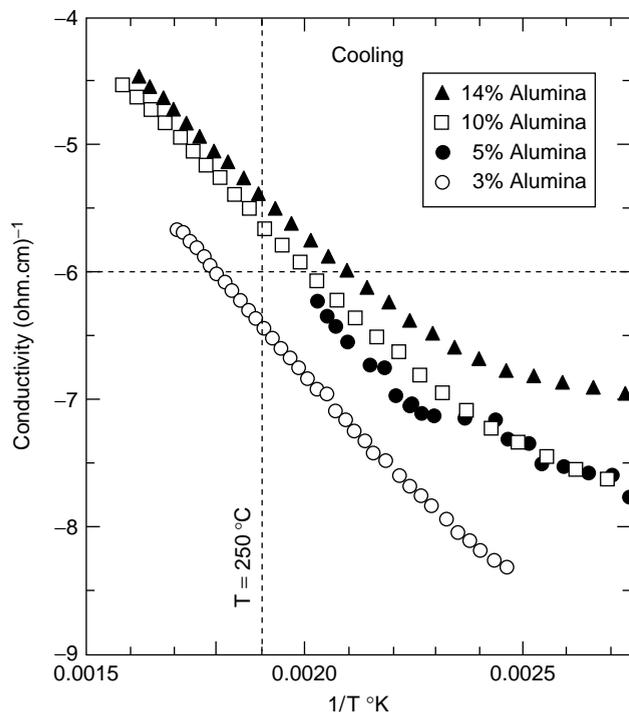


Figure 4. Ionic conductivity of alumina-containing lithium silicate bulk gels, plotted as an Arrhenius plot of log (conductivity) vs. reciprocal temperature.

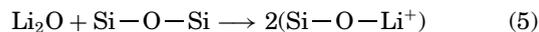
both sides coated by a layer of platinum paste to provide contacts with platinum electrodes.

Thermal analysis shows that weight loss occurs at 160°C for adsorbed species and at 210°C for organics. Heat treatment of the gels to 250°C for 1 hour removes volatiles. Ionic conductivity was measured on gels heated from 100°C to around 360°C at a heating rate of 10°C/min in the frequency range of 10 Hz to 65 kHz. Measurements were repeated and reproduced to eliminate any concern about a contribution to the measurements from protons. A summary of measured conductivities is shown in Fig. 4. Ideally, the solid electrolyte has only mobile lithium ions as charge carriers (48). As seen in Fig. 4, in bulk materials, the conductivities are too low at room temperature for use in an electrochromic device.

For measurements on thin films, the surface of a microelectrode (Microsensor Systems SAW 302), that consisted of an interdigitated array of gold electrodes deposited on a quartz crystal substrate, was coated with the solution (49). A sample containing 10 mol% Al₂O₃ was selected because measurements on the bulk showed good ionic conductivity for this composition. The values of ionic conductivity measured on thin films matched those for bulk ionic conductivity.

In lithium aluminosilicate compositions, conductivity is due to the conventional hopping mechanism, taking two interdependent factors into account. First, when alkali ions are incorporated into the network, the excess oxygen is accommodated by the network by breaking bridging oxygens to create nonbridging oxygens (NBO). According to the

following relationship:



the Li⁺ ion is available to charge compensate for the nonbridging Si-O bond. Secondly, borrowing from results found for (LiCl)₂-SiO₂-Al₂O₃ in ⁷Li, ²⁷Al, and ²⁹Si NMR experiments (46), the silicon ions and most of the aluminum ions are four coordinated with oxygen ions, and there is some octahedrally coordinated aluminum. Because one oxygen ion shares its two charges with two tetrahedra, oxygen ions can not compensate for the charge deficit provided by Al³⁺. The additional charge needed is provided by one lithium ion located near the tetrahedron. In a sense, there are two types of lithium ions associated with the network: the first category compensates for [AlO₄] units, and the second compensates for NBO in the silica network.

Assuming that the only species that can move is the lithium ion, the changes in conductivity are related to changes in mobility. Further, it is observed that the conductivity increases when the Al₂O₃ content increases, indicating a link between mobility and Al₂O₃ content. The higher the Al₂O₃ mol% in the gels, the more [AlO₄] units should replace SiO₄ units, and the more Li ions compensate for these [AlO₄] units. This means that the mobility of Li⁺ ions compensating for [AlO₄] units is different from that of Li⁺ compensating for NBO of [SiO₄]⁻. This difference is undoubtedly due to differences in electrostatic interactions between oxygen and lithium ions. When a lithium ion acts as a charge compensator in the silica network, the lithium ion is very close to the oxygen ion, and the electrostatic interaction is very strong. When Li plays the part of compensating for [AlO₄] units, the interaction is weaker, and Li⁺ is more able to move under a current flow, meaning a lower activation energy.

Stable and reproducible conductivities have been demonstrated in thin films, but the conductivity values are still too low. Further advances in electrochromic devices depend on finding a solution to the electrolyte problem. Some recent studies of organically modified electrolytes are leading in the right direction (50,51).

CONCLUSIONS

In further pursuing the development of completely sol-gel EC devices, physical properties need to be considered, in addition to the ionic conductivity. Of course, the electrolyte has to have

- high electronic resistivity,
- high ionic conductivity
- no defects.

In addition, all of the layers require

- a film thickness less than about 1000 nm (generally less than 500 nm)
- high transparency across the visible range

- good adhesion between adjacent layers
- thermodynamic stability between adjacent layers across a wide electrochemical window.

Overall, an EC device should operate under the approximate conditions of +2.5 volts (100 s)/-1.5 volts (100 s) for more than 10,000 cycles at 15 mC/cm².

For practical EC windows, there has been significant recent progress using a variety of technologies (52,53). Sol-gel processing, of course, is only one route under investigation. Interest continues in sol-gel processing because precursors for all of the layers that lead to the appropriate oxide gels are readily available. The resulting gels are inorganic and rigid. In particular, the sol-gel electrolyte operates while playing the role of a mechanical divider between the electrodes. In addition, sol-gel processing is a logical batch process for large area coatings. Several studies show that sol-gel coatings have the required transmittance across the visible light range. Finally, the layers have "no moving parts."

ACKNOWLEDGMENT

This work was supported in the past by the Assistant Secretary for Conservation and Renewable Energy, Office of Transportation Technologies, Electric and Hybrid Propulsion Division of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098, Subcontract No. 4593610 with the Lawrence Berkeley Laboratory. We acknowledge the financial support of the N.J. Commission on Science and Technology Innovation Partnership Program and the many useful discussions with G. Amatucci, L. Laby, G. Lous, A. Wojcik, M. Greenblatt, A. Salkind, and J. Van Dine.

BIBLIOGRAPHY

1. F.G.K. Baucke, *Mater. Sci. Eng.* **B10**, 285 (1991).
2. C.B. Greenberg, *Thin Solid Films* **251**, 81 (1994).
3. C.B. Greenberg, *J. Electrochem. Soc.* **140**(11) (1993).
4. C.M. Lampert and C.G. Granqvist, *SPIE IS4* (1990).
5. I.D. Watkins, G.E. Tulloch, T. Maine, L. Spiccia, D.R. McFarlane, and J.L. Woolfrey, *Ceram. Trans.* **81**, 223 (1998).
6. J.-G. Zhang, C.E. Tracy, D.K. Benson, and S.K. Deb, *J. Mater. Res.* **8**, 2649 (1993).
7. A. Agrawal, J.P. Cronin, and R. Zang, *Sol-Gel Opt. II, SPIE 1758*, 330 (1992).
8. J.E. Van Dine, V.D. Parkhe, L.C. Klein, and F.A. Trumbore, U.S. Pat. 5,404,244 (1995).
9. J.E. Van Dine, V.D. Parkhe, L.C. Klein, and F.A. Trumbore, U.S. Pat. 5,659,417 (1997).
10. J.E. Van Dine, V.D. Parkhe, L.C. Klein, and F.A. Trumbore, U.S. Pat. 5,699,192 (1997).
11. J.M. Tarascon, *J. Electrochem. Soc.* **134**, 1345 (1987).
12. P. Barboux, J.M. Tarascon, and F.K. Shokoohi, *J. Solid State Chem.* **94**, 185 (1991).
13. S. Colson, S.-P. Szu, L.C. Klein, and J.M. Tarascon, *Solid State Ionics* **46**, 283 (1991).
14. J. Livage, *MRS Symp. Proc.* **293**, 261 (1993).
15. A. Pawlicka, C. Avellaneda, M. Schmitt, S. Heusing, and M.A. Aegerter, *Ceram. Trans.* **81**, 229 (1998).
16. M.A. Macedo, L.H. Dall'Antonia, and M.A. Aegerter, *Sol-Gel Opt. II, SPIE 1758*, 320 (1992).
17. B.G. Dixon, J.R. Deans, and L.B. Jenney, U.S. Pat. 4,834,772 (1989).
18. C.F. Mason, U.S. Pat. 5,016,991 (1991).
19. J.P. Boilot and P. Colomban, in *Sol-Gel Technology for Thin Films, Fibers, Preforms, Electronics and Specialty Shapes* (L.C. Klein, ed.), p. 303. Noyes Data Corp., Park Ridge, NJ, 1988.
20. B. Wang, M. Greenblatt, J. Yan, and Y. Wu, *J. Sol-Gel Sci. Technol.* **2**, 323 (1994).
21. C.J. Brinker and G.W. Scherer, *Sol-Gel Science: The Physics and Chemistry of Sol-Gel Processing*, Academic Press, Boston, MA, 1990.
22. C.J. Brinker, A.J. Hurd, G.C. Frye, P.R. Schunk, and C.S. Ashley, in *Chemical Processing of Advanced Materials* (L.L. Hench and J.K. West, eds.) p. 395. Wiley New York, 1992.
23. I.M. Thomas, in *Sol-Gel Optics* (L.C. Klein, ed.), p. 141. Kluwer, Boston, MA, 1994.
24. L.C. Klein, in *Handbook of Nanophase Materials* (A.N. Goldstein, ed.), p. 43. Dekker, New York, 1996.
25. C.E. Tracy, J.-G. Zhang, D.K. Benson, A.W. Czanderna, and S.K. Deb, *Electrochim. Acta* **44**, 3195 (1999).
26. A. Chemseddine, M. Henry, and J. Livage, *Rev. Chim. Miner.* **21**, 487 (1984).
27. N.R. Lynam, F.H. Moser, and B.P. Hichura, *SPIE 823*, 130 (1987).
28. J.P. Cronin, D.J. Tarico, J.C.L. Tonazzi, A. Agrawal, and S.R. Kennedy, *Sol. Energy Mater. Sol. Cells* **29**, 371 (1993).
29. B. Munro, S. Kramer, P. Zapp, and H. Krug, *J. Sol-Gel Sci. Technol.* **13**, 673 (1998).
30. B. Pecquenard, H. Lecacheux, S. Castro-Garcia, and J. Livage, *J. Sol-Gel Sci. Technol.* **13**, 923 (1998).
31. J. Livage and G. Guzman, *Solid State Ionics* **84**, 205 (1996).
32. L.C. Klein, L. Laby, and G. Lous, in *Sol-Gel Processing of Advanced Ceramics* (F.D. Gnanam, ed.), p. 205. Oxford/IBH, New Delhi, 1996.
33. J.-D. Guo, K.P. Reis, and M.S. Whittingham, *Solid State Ionics* **53-56**, 305 (1992).
34. J. Livage and J. Lemerle, *Annu. Rev. Mater. Sci.* **12**, 103 (1982).
35. J.K. Bailey, G.A. Pozarnsky, and M.L. Mecartney, *J. Mater. Res.* **7**, 2530 (1992).
36. S. Mege, D. Routhier, F. Ansart, and J.M. Savariault, *Mol. Cryst. Liq. Cryst.* **311**, 95 (1998).
37. S. Mege, M. Verelst, P. Lecante, E. Perez, F. Amsart, and J.M. Savariault, *J. Non-Cryst. Solids* **238**, 37 (1998).
38. R.D. Cussler, M.G. Kulkarni, and E.L. Cussler, *J. Membr. Sci.* **127**, 153 (1997).
39. N. Ozer, *Thin Solid Films* **305**, 80 (1997).
40. L.C. Klein, S.F. Ho, S.-P. Szu and M. Greenblatt, in *Applications of Analytical Techniques to the Characterization of Materials* (D. L. Perry, ed.), p. 101. Plenum, New York, 1991.
41. S.-P. Szu, L.C. Klein, and M. Greenblatt, *J. Non-Cryst. Solids* **121**, 119 (1990).
42. S.-P. Szu, M. Greenblatt, and L.C. Klein, *J. Non-Cryst. Solids* **124**, 91 (1990).
43. S.-P. Szu, M. Greenblatt, and L.C. Klein, *Solid State Ionics* **46**, 291 (1991).

44. L.C. Klein, S.-P. Szu, A.M. Buckley, and M. Greenblatt, in *Nucleation and Crystallization in Glasses and Liquids* (M. Weinberg, ed.), p. 167. Am. Ceram. Soc., Westerville, OH, 1993.
45. H. de Lambilly and L.C. Klein, *J. Non-Cryst. Solids*, **102**, 269 (1988).
46. B. Wang, S.-P. Szu, M. Greenblatt, and L.C. Klein, *Chem. Mater.* **4**, 191 (1992).
47. J. Farcy, R. Messina, and J. Perichon, *J. Electrochem. Soc.* **137**, 1337 (1990).
48. E. Mouchon, L.C. Klein, V. Picard, and M. Greenblatt, *MRS Symp. Proc.* **346**, 189 (1994).
49. L. Laby, L.C. Klein, J. Yan, and M. Greenblatt, *Solid State Ionics* **81**, 217 (1995).
50. K. Dahmouche, M. Atik, N.C. Mello, T.J. Bonagamba, H. Panepucci, M.A. Aegerter, and P. Judeinstein, *J. Sol-Gel Sci. Technol.* **8**, 711 (1997).
51. P. Judeinstein, J. Titman, M. Stamm, and H. Schmidt, *Chem. Mater.* **6**, 127 (1994).
52. C.G. Granqvist, *Electrochim. Acta* **44**, 3005 (1999).
53. C.G. Granqvist, A. Azens, A. Hjelm, L. Kullman, G.A. Niklasson, D. Ronnow, M. Stromme Mattsson, M. Veszelei, and G. Vainars, *Sol. Energy* **63**, 199 (1998).

ELECTORRHEOLOGICAL FLUIDS

TIAN HAO

Rutgers—The State University of New Jersey
Piscataway, NJ

INTRODUCTION

An electrorheological (ER) fluid is a colloidal system in which one kind of particulate material is dispersed in a nonconductive liquid base. When an electric field of several kilovolts per millimeter is applied, ER fluids reversibly change from a liquid state to a solid crystallized state within a millisecond. What the working mechanism of the ER effect is and how to design a high-performance ER fluid for industrial uses have been studied for many years; however, there is no clear answer yet. Because the ER effect is induced by an external electric field, it is believed that polarization of both the liquid medium and the solid particles plays an important role. The dielectric tool is thus widely used to characterize the ER fluids, and many efforts have been spent on revealing how the dielectric properties of the ER fluids correlate physically with the ER effect. The purpose of this article is to draw a logical relationship between the dielectric properties of the materials used for making an ER fluid and the ER mechanism. To achieve this goal, the dielectric investigation that was done to characterize an ER fluid and the related ER models that were developed to describe ER phenomena are reviewed following. Attention is paid to the ER models proposed on the basis of the dielectric studies; other ER models are only briefly described.

The ER phenomenon was first reported in 1947 by Winslow (1) in an insulating oil that contained starch or flour particulate material. The first example of using a dielectric tool to investigate the ER phenomena was carried

out by Klass (2) in 1967. It was found that the dielectric constant of the ER suspension increases linearly with the particle volume fraction, and an electric-double-layer (EDL) mechanism was proposed to explain the observed experimental results. The interaction among the EDLs, as well as the interaction between the EDL and the particles, becomes much stronger once an electric field is applied. The EDLs can be polarized and then distorted; thus the electrostatic force between particles can dramatically increase, generating the observed ER effect. A deeper dielectric investigation of the way the dielectric constant of the whole suspension changes as the water content in the suspension and the particle volume fraction change was carried out by Ujima (3). At that time, it was believed that the "water" (4) and the EDL (2) are crucial to the ER effect because all ER fluids were active only if the dispersed particles adsorb a trace amount of water or surfactants.

After the innovation of anhydrous ER fluids (5,6) in 1985, the mechanism of the ER effect was substantially modified because neither the "water" nor the EDL mechanism can explain the considerably strong ER effect in ER fluids without any amount of water. The dielectric properties of polymer-based water-free ER fluids were studied by Block (7), who proposed a flow modified polarization (FMP) theory to explain the ER effect. According to the FMP theory, there should be resonance between the applied electric field and the applied shear field. It was found that the dielectric constants and losses of the ER suspensions peak when the shear rate is 4π times the frequency of the electric field. An extra surface charge is generated once a particle is polarized in an electric field, which in turn drives the particles to rearrange in the suspensions. It was thought that this is a possible reason that the ER suspensions can be fibrillated to form a bridge spanning between two electrodes. However, the physical picture of the way the dielectric properties of the particulate material can change ER performance is not clear in Block's theory.

The influence of the conductivities and dielectric constants of both the liquid medium and particulate material on the ER effect was addressed by Anderson (8) and Carino et al. (9). It was found that the ER effect becomes stronger once the dielectric constant ratio of the particle-to-liquid increases. This issue was further addressed theoretically by Davis (10–12), and he found that the ER effect is controlled by the conductivity ratio of the particle-to-liquid in a low-frequency electric field and by the dielectric constant ratio in a high-frequency field, which is the well-known polarization model. It was believed that the dielectric constant mismatch between the suspended particles and the medium is essential. However, the ER effect of a barium titanate suspension (the dielectric constant of BaTiO_3 is around 2000, depending on its crystallization state), inactive in a dc field (13) and active after adsorbing a small amount of water (14) or being stimulated by an ac field (9), suggests that the polarization model does not adequately reveal the physical essence of the ER effect. The polarization model has much room for improvement.

Because of the obvious shortcomings of the polarization model, another model called the conduction model was introduced by Atten (15,16) and was well developed by other researchers (17–21). Good agreement between the

theoretically predicted parameters (the current density, the yield stress, etc.) of the conduction model and the experimentally measured values determined either by them or by Inoue (22) was achieved. However, as stated in Khusid's paper (23), the conduction model can be used only when the suspension microstructure has been fully formed. A major shortcoming of the conduction model is that it cannot explain the response time of ER fluids that vary from 0.1 ms to 1 s because only particle interaction is considered in the conduction model regardless of the microstructural change before and after an electric field is applied. The dynamic process that occurs in ER fluids in an electric field is not taken into account. A contrary experimental result was obtained in the magnesium hydroxide/poly (methylphenylsiloxane) system (24): A negative ER effect was observed in this system in a dc field, although a positive ER effect was expected according to the conduction model, strongly suggesting that the conduction model too, does not reflect the physical essentials of ER effect. The limitations of both the polarization and the conduction models stem, we think, from the fact that both are static and do not involve dynamic processes that occur in ER fluids.

Besides the polarization and conduction models, many other models were also proposed to unveil the ER mechanism. An excellent theory was proposed by Khusid (23): Both the electric-field-induced particles aggregative process and the interfacial polarization process were considered. Almost all of the currently observed ER phenomena can be explained, especially those that cannot be understood by the polarization and the conduction models. However, some discrepancies in the experimental results still exist, which arise from the assumptions on which the theory is based. Two presumptions were used to derive the theory: The dispersed particles and the suspending medium have no intrinsic dielectric dispersion, and the variation of the applied electric field is very slow compared with the rate of polarization; these are not always valid in ER fluids. A so-called layered model proposed by See and Saito (25) gives derivations similar to those from the conduction model. The first-principle method was used for calculating how ER properties (the yield stress and the modules) change as the frequency and water content change under the assumption that the Debye relaxation process is dominant (26,27). It was found that the theoretical predictions and the experimental results agree very well. However, because only few solid materials exhibit Debye-type relaxation behavior, good agreement between the theoretical predictions and the experimental results on the transient ER response was also achieved by Tang and Conrad (18) without using the Debye-type relaxation assumption. Whether the Debye-type relaxation process is important in ER fluids thus remains a question.

The dielectric studies of anhydrous ER fluids provide deep insight into the ER mechanism. The intrinsic properties of dispersed particles, it was thought, contribute to the ER effect of anhydrous ER fluids (6,28–30); mobile charges (ions or electrons) on the surfaces of particles that can move freely on the surfaces but cannot move off are essential. The ER effect should somewhat correlate with the dielectric dispersion in a low-frequency field; however, the ER effect does not correlate directly with charge concentration. The

polarization rate, as well as its magnitude, was stressed (7): Too high a rate of polarization and too low a rate of polarization were experimentally found unfavorable to a strong ER effect. An ER fluid might deliver a strong ER effect if it meets the following empirical requirements: Its dielectric relaxation frequency should be between 100 and 10^5 Hz, and the difference between the dielectric constants below and above the relaxation frequency must be large (31–33). Those experimental results, we think, imply that the ER effect may correlate with particle dielectric loss because charges moving within the particles in an electric field would be one of origins of dielectric loss. The influence of the polarization rate on the ER effect was theoretically addressed (34), and good agreement with experimental results was obtained. However, why the polarization rate is important is not clear. The explanation rests on interfacial polarization, but there is no clear physical picture of the ER mechanism from employing the interfacial polarization framework.

This article attempts to provide a clear physical picture of the ER response mechanism through dielectric analysis of ER phenomena. We begin with a general description of the dielectric properties of heterogeneous systems such as colloidal suspension systems, which is introduced in the following section. In the subsequent section, experimental evidence of the way dielectric parameters change the ER effect are discussed, and an empirical dielectric criterion for designing high-performance ER fluids is proposed. A theoretical treatment on the dielectric criterion is described in the following section IV, and an ER mechanism is proposed in the section thereafter. A yield stress equation based on the ER mechanism is theoretically derived in the next section, and comparisons with currently available experimental results are also made. Final conclusions of this article are discussed in the last Section.

DIELECTRIC PROPERTIES OF HETEROGENEOUS SYSTEMS

Four kinds of polarization are generated in a heterogeneous system under an electric field: the electronic, atomic, Debye, and interfacial polarizations. They are schematically illustrated in Fig. 1 on the basis of the relative relaxation times (35). The former three polarizations occur in a homogeneous phase, whereas interfacial polarization occurs only in heterogeneous systems. Electronic and atomic polarization are induced by the displacement of an electron and atom, separately; Debye polarization is due to dipole orientation, and interfacial polarization is due to heterogeneous interfaces. Debye polarization usually takes place in liquids and in a few solids because dipole orientation becomes possible only if the dipole is not firmly fixed in bulk materials. Electronic and atomic polarizations appear in a high-frequency field; usually they are regarded as fast polarizations. Debye and interfacial polarizations are rather slow processes compared with the electronic and the atomic polarizations, and are usually called slow polarizations. Another type of polarization is ionic displacement polarization that takes place only if the particulate material is ionic. If the ER effect is controlled by polarization, then the kind of polarization that is responsible for the ER

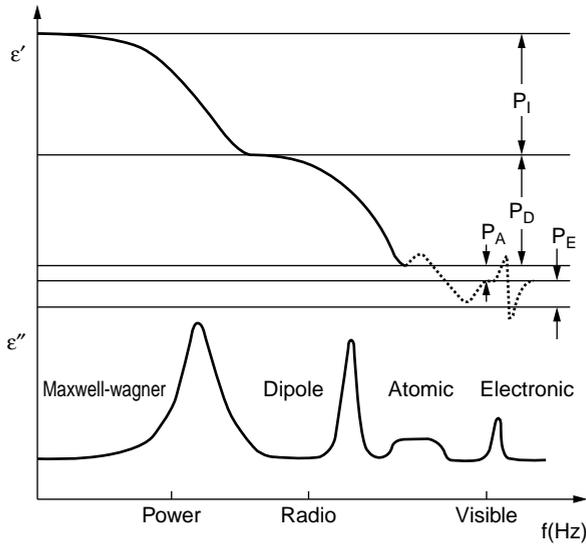


Figure 1. Dielectric constant and dielectric dispersion of a heterogeneous system vs. frequency [reproduced from (49)].

effect would be important in understanding the ER effect. This article attempts to clarify this issue and addresses it in the next section.

Debye polarization can be described by the Debye equation, and interfacial polarization can be described by the Wagner–Maxwell equation. If the particle conductivity σ_p is much larger than that of the medium, a simplified Maxwell–Wagner equation can be expressed by (36)

$$\varepsilon = \varepsilon_\infty \left[1 + \frac{K}{(1 + \omega^2 t^2)} \right]; \quad (1)$$

in this equation,

$$\varepsilon_\infty = \varepsilon_{sm} \left[1 + \frac{3\Phi(\varepsilon_{sp} - \varepsilon_{sm})}{2\varepsilon_{sm} + \varepsilon_{sp}} \right], \quad (2)$$

$$K = \frac{9\Phi\varepsilon_{sm}}{2\varepsilon_{sm} + \varepsilon_{sp}}, \quad (3)$$

$$t = \frac{\varepsilon_0(2\varepsilon_{sm} + \varepsilon_{sp})}{\sigma_p}, \quad (4)$$

where ε_0 , ε_{sm} , and ε_{sp} are the static dielectric constants of the vacuum, the medium, and the particles; Φ is the particle volume fraction, ω is the frequency of the alternating electric field, ε_∞ is the infinite frequency dielectric constant of the suspension, and t is the relaxation time of polarization. The dielectric loss tangent is given by

$$\tan \delta = K\omega t / (1 + K + \omega^2 t^2) \quad (5)$$

The Wagner theory may not hold for concentrated dispersions because it was derived for diluted systems whose particle volume fraction is less than 10%. Hanai et al. (37)

proposed a new theory applicable to concentrated dispersions and derived an equation similar to Bruggeman's (37,38). There is no obvious difference between Bruggeman's and Wagner's equations until the volume fraction of dispersed particles exceeds 40% (37,38). So, applying the Wagner equation for concentrated ER suspensions whose particle volume fraction is as high as 40% is acceptable.

EXPERIMENTAL FACTS

The experimental facts presented here demonstrate (1) the kind of polarization that is responsible for the ER effect, (2) whether the Wagner–Maxwell equation describes the dielectric properties of ER fluids, and (3) whether the dielectric loss is also important in the ER effect. A criterion is extracted from these experimental facts for designing high-performance ER fluids. We will follow this sequence to show the experimental evidence.

As indicated in the introduction, it is most likely that the ER effect is controlled by slow polarization (Debye and interfacial polarization) rather than fast polarization (electronic and atomic polarization) because the response time of the ER effect occurs on a millisecond scale. Dielectric dispersions in many good ER fluids are usually observed in low-frequency fields (31–33), further confirming that the ER effect may have something to do with slow polarization. As an example, the dielectric loss tangents of two kinds of ER suspensions consisting of aluminosilicate powder dispersed in polydimethylsiloxane (PDMS) oil (powder treated at 550 and 750°C for 8 hours and named AS1 and AS2, respectively) are plotted against frequency in Fig. 2. These two ER suspensions give a very strong ER effect, and the dielectric dispersions appear in low-frequency fields of less than 10^3 Hz, indicating that dielectric dispersions arise from slow polarization, which in turn contributes to the ER effect. Experimental results obtained by Filisko (29,40) also suggested

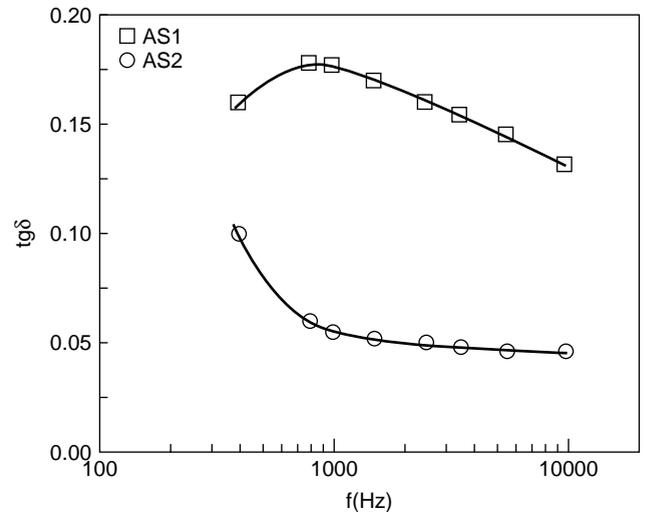


Figure 2. Dielectric loss tangents of the AS1 and AS2 suspensions vs. frequency [reproduced from (49)].

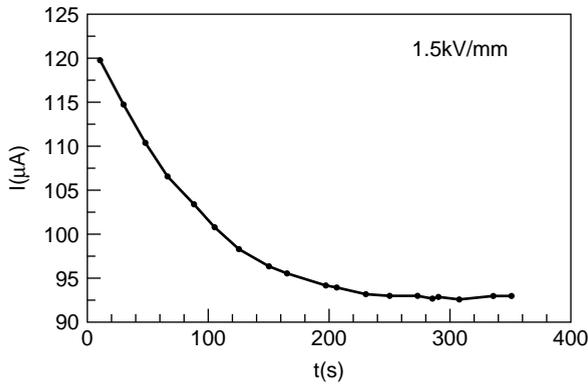


Figure 3. dc current passing through AS2 vs. time [reproduced from (49)].

that the ER effect might result from low-field dielectric dispersion. Other stronger evidence is that the dc currents that pass through ER fluids are not stable, exponentially decrease with time, and then level off; this was found both in a polymeric ER fluid (oxidized polyacrylonitrile/PDMS) and an inorganic ER fluid (aluminosilicate/PDMS) (39). A typical example of dc current passing through the AS2 suspension in an electric field (1.5 kV/mm) versus time is shown in Fig. 3. This is the so-called dielectric adsorption phenomenon that arises from heterogeneity and slow polarization, further indicating that slow polarization rather than fast polarization is very distinct in the ER system. All of these experimental facts demonstrate that slow polarization in ER systems is likely to be the most important process.

If slow polarization is responsible for the ER effect, another question arises: which slow polarization, Debye or interfacial polarization, controls the ER effect or jointly control the ER effect? As we know, Debye polarization is generated due to dipole orientation in an electric field. The dipole of most solid materials is almost unable to reorient because solidification usually fixes the molecule with such rigidity in the solid lattice that there is little or no orientation of dipoles, even in an extremely strong electric field. If the ER effect stems from polarization of solid particulate material, Debye polarization can almost be ruled out. This can be further confirmed by measuring the ER response time that should be determined by the relaxation time of polarization that controls the ER effect. The relaxation times of Debye and interfacial polarization are governed by different factors; Debye polarization is governed by the potential barrier between two equilibrium positions between which the dipole would orient (41), and interfacial polarization is governed by the dielectric constants and conductivities of the two components (35,36). The relaxation time of the Debye t_d , can be expressed as (41)

$$t_d = (\pi/2\omega_a) A e^{H/kT}, \quad (6)$$

where $\pi/2\omega_a$ is the average time required by an excited molecule to turn from one equilibrium direction to the

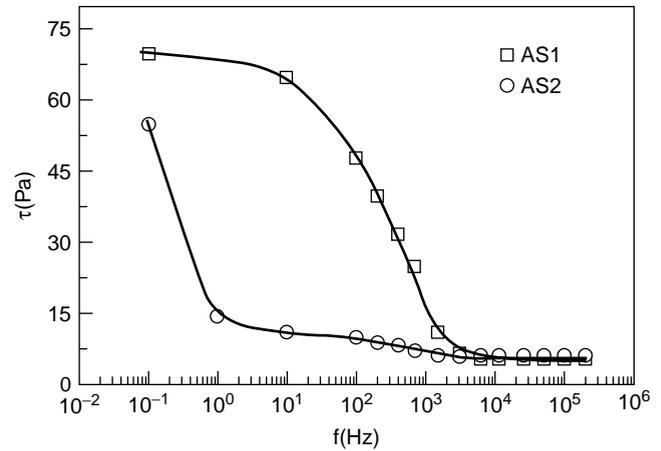


Figure 4. The shear stresses of AS1 and AS2 suspensions vs. frequency (0.3 kV/mm) [reproduced from (49)].

other, A is a factor that varies only slowly with temperature T , H is the energy barrier that separates the two equilibrium positions, and k is the Boltzmann constant. The relaxation time t_i for interfacial polarization in one system where σ_p (conductivity of the particle) \gg σ_m (conductivity of the medium) can be expressed as (35,36)

$$t_i = \varepsilon_0(2\varepsilon_m + \varepsilon_p)/\sigma_p. \quad (7)$$

Clearly, t_d depends on the material state (liquid or solid), the molecular interaction and the environment, whereas t_i depends on the dielectric properties of the two components, especially on particle conductivity. The response times of AS1 and AS2 were determined by measuring the shear stress dependence on field frequency (see Fig. 4). The response times of these two suspensions are very different. The response time of the suspension whose particle conductivity is 6.0×10^{-7} S/m (AS1) is 0.6 ms, and that of the suspension whose particle conductivity is 8.4×10^{-10} S/m (AS2) is 0.22 s. If Debye polarization controls the ER effect, these two ER fluids should give, almost the same response time. According to Eq. (6), because the solid particulate material in these two suspensions is almost same, and the energy barrier H should not be very different. The quite different response times can be reasonably understood only from Eq. (7), which indicates that interfacial polarization probably governs the ER effect.

The publications previously cited, support this conclusion. The data presented in (42) showed that the experimentally measured response time is inversely proportional to the particle conductivity and agrees well with Eq. (7) Also, Wen et al. (27) and Conrad and Chemn (43) found that even the same particulate material displays different response time after adsorbing different amounts of water; the larger amount of water leads to a shorter response time. These are also consistent with Eq (7). Because oil is hydrophobic, the added water can adsorb just on the particle surface and enhance particle conductivity considerably. Whittle et al. (44) also found the same inversely proportional relationship between particle conductivity and ER response time.

The dielectric could provide a direct way to distinguish Debye and interfacial polarization in ER fluids. Differentiating Eq. (5) with respect to frequency ω , one can easily obtain the value of $tg\delta_{\max}$,

$$tg\delta_{\max} = \frac{K}{2\sqrt{1+K}} = \frac{9\phi\epsilon_{sm}}{2\sqrt{(9\phi\epsilon_{sm} + 2\epsilon_{sm} + \epsilon_{sp})(2\epsilon_{sm} + \epsilon_{sp})}} \quad (8)$$

Assuming that both ϵ_{sm} and ϵ_{sp} depend on temperature T and differentiating Eq (8) with respect to T ,

$$\begin{aligned} \frac{dtg\delta_{\max}}{dT} &= \frac{(2+K)\frac{dK}{dT}}{4(1+K)^{3/2}} = \frac{(2+K)}{4(1+K)^{3/2}} \\ &\times \frac{9\phi\left(\epsilon_{sp}\frac{d\epsilon_{sm}}{dT} - \epsilon_{sm}\frac{d\epsilon_{sp}}{dT}\right)}{(2\epsilon_{sm} + \epsilon_{sp})^2} \end{aligned} \quad (9)$$

The sign of $(dtg\delta_{\max}/dT)$ obviously depends on the sign of $(\epsilon_{sp}\frac{d\epsilon_{sm}}{dT} - \epsilon_{sm}\frac{d\epsilon_{sp}}{dT})$ because the rest of the equation is always positive. According to Frohlich’s theory (41), the static dielectric constant of a polar liquid decreases as temperature increases. Usually, the most commonly used dispersing medium for ER fluids is PDMS oil, a polar liquid; thus $d\epsilon_{sm}/dT$ would be less than zero. If $d\epsilon_{sp}/dT > 0$, then, the term $(\epsilon_{sp}\frac{d\epsilon_{sm}}{dT} - \epsilon_{sm}\frac{d\epsilon_{sp}}{dT})$ would be less than zero, and thus, $dtg\delta_{\max}/dT < 0$. If $\frac{d\epsilon_{sp}}{dT} < 0$, the situation is somewhat complicated, and one needs to know all of the parameters involved in the term $(\epsilon_{sp}\frac{d\epsilon_{sm}}{dT} - \epsilon_{sm}\frac{d\epsilon_{sp}}{dT})$. These feature the dielectric characteristics of an ER suspension in dominant Maxwell–Wagner polarization.

It is well known that the maximum value of the dielectric loss tangent, $tg\delta_{d\max}$ of Debye polarization can be

expressed as (41)

$$tg\delta_{d\max} = \frac{\epsilon_{sp} - \epsilon_{\infty p}}{2\sqrt{\epsilon_{sp}\epsilon_{\infty p}}} \quad (10)$$

where $\epsilon_{\infty p}$ is the dielectric constant at very high frequency. Differentiating Eq. (10) with respect to T and assuming that only ϵ_{sp} is temperature dependent,

$$\frac{dtg\delta_{d\max}}{dT} = \frac{\left(1 + \frac{\epsilon_{\infty p}}{\epsilon_{sp}}\right)\frac{d\epsilon_{sp}}{dT}}{4\sqrt{\epsilon_{\infty p}\epsilon_{sp}}} \quad (11)$$

Comparing Eqs. (9) and (11), if $d\epsilon_{sp}/dT > 0$, the observed $(dtg\delta_{\max}/dT)$ determined by Maxwell–Wagner polarization gives a sign opposite to that determined by Debye polarization. Then, it would be easier to distinguish the polarization type. If $d\epsilon_{sp}/dT < 0$, one has to know the exact value of $d\epsilon_{sp}/dT$.

A suspension consisting of Molecular Sieve 3A (MS3A) powder dispersed in PDMS oil was used to address this issue by direct dielectric means. The dielectric properties of the PDMS oil were measured at different temperatures, as shown in Fig. 5. Because PDMS oil is a good insulator, it almost does not have any dielectric loss (the dielectric loss in the measurement frequency range is almost zero). Its dielectric constant remains unchanged in the measured frequency range, and thus, its low-frequency dielectric constant can be regarded as a static dielectric constant. Figure 6 gives the temperature dependence of the dielectric constant of PDMS oil at 20 Hz. The dielectric constant obviously decreases as temperature decreases, and the slope of the linear regression line is $-2.4 \times 10^{-3}(1/^{\circ}\text{C})$, which can be reasonably regarded as the value of $d\epsilon_{sm}/dT$ in Eq. (9). The dielectric constant and the dielectric loss of pure MS3A

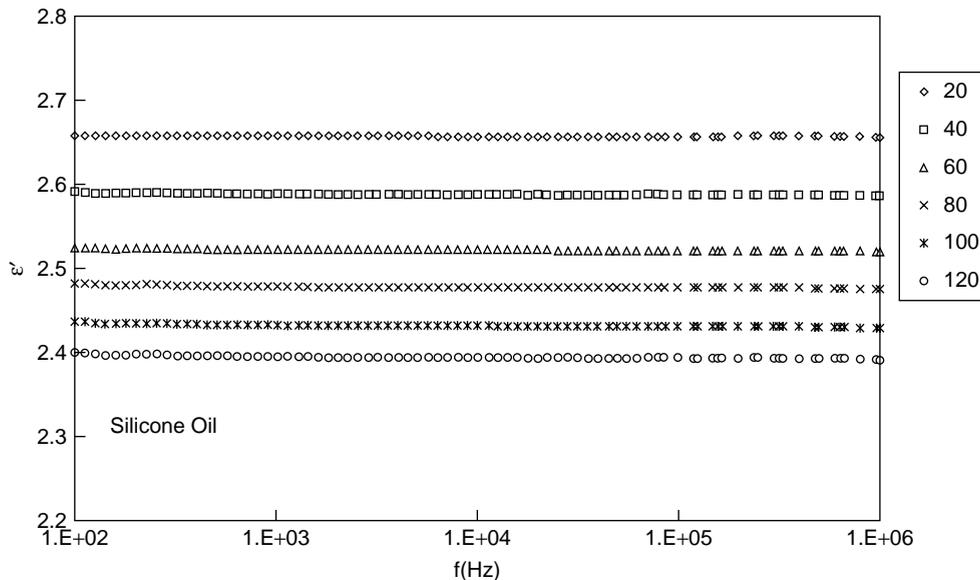


Figure 5. Dielectric constant of poly(dimethylsiloxane) oil vs. frequency at temperatures from (20–120°C).

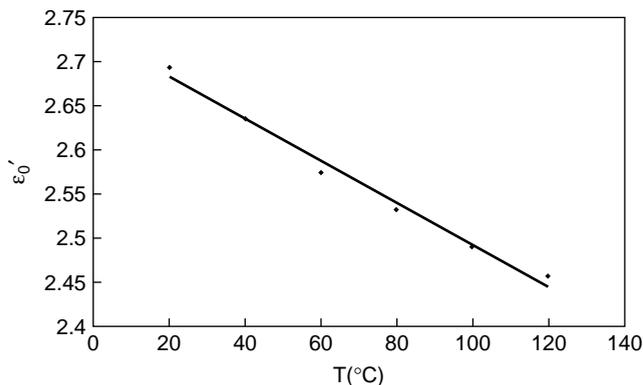


Figure 6. Static dielectric constant of poly(dimethylsiloxane) oil vs. temperature.

versus frequency measured at 20°C are shown in Fig. 7. The static dielectric constant of pure MS3A was calculated by using the Cole–Cole plot and is shown in Fig. 8. The static dielectric constant of pure MS3A increases as temperature increases, that is, $(d\varepsilon_{sp}/dT) = 0.40 > 0$. Thus, the term $(\varepsilon_{sp} \frac{d\varepsilon_{sm}}{dT} - \varepsilon_{sm} \frac{d\varepsilon_{sp}}{dT})$ in Eq. (9) can be determined and is equal to -1.21 , which means that $(dtg\delta_{max}/dT)$ is less than zero. According to the criteria given before, it would be easier to distinguish the polarization type in the MS3A suspension— $(dtg\delta_{max}/dT)$ would be less than zero in dominant Maxwell–Wagner polarization, and it would be larger than zero in dominant Debye polarization. This result indicates that Maxwell–Wagner polarization is dominant in the particle-type MS3A ER suspension. The direct dielectric measurement of the MS3A suspension further confirms the conclusion. Figure 9 shows the dielectric loss tangent of the MS3A suspension as a function of frequency obtained at temperatures from 20 to 120°C. The maxima of the dielectric loss tangent obviously decrease as temperature increases, in accordance with the theoretical prediction.

If mainly interfacial polarization occurs in ER fluids, the Wagner–Maxwell equation should be able to describe the dielectric properties of ER fluids. Filisko (40) and

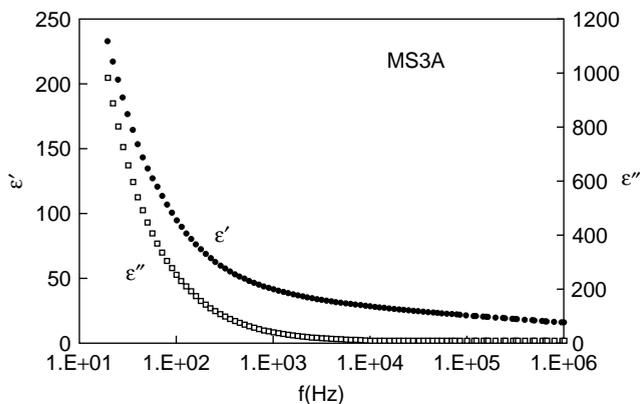


Figure 7. The dielectric constant and the dielectric loss of pure MS3A vs. frequency at 20°C.

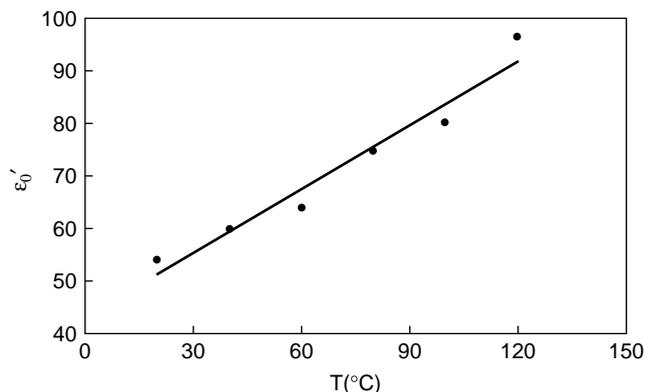


Figure 8. The static dielectric constant of pure MS3A vs. temperature.

Weiss and Carlson (45) measured the dielectric constants and dielectric losses of their ER fluids and found that the Wagner–Maxwell equation is a suitable model for describing the dielectric properties. The Wagner–Maxwell equation should also be able to explain the frequency dependence of the yield stress of an ER fluid. The yield stress of the oxidized polyacrylonitrile (OP) /PDMS

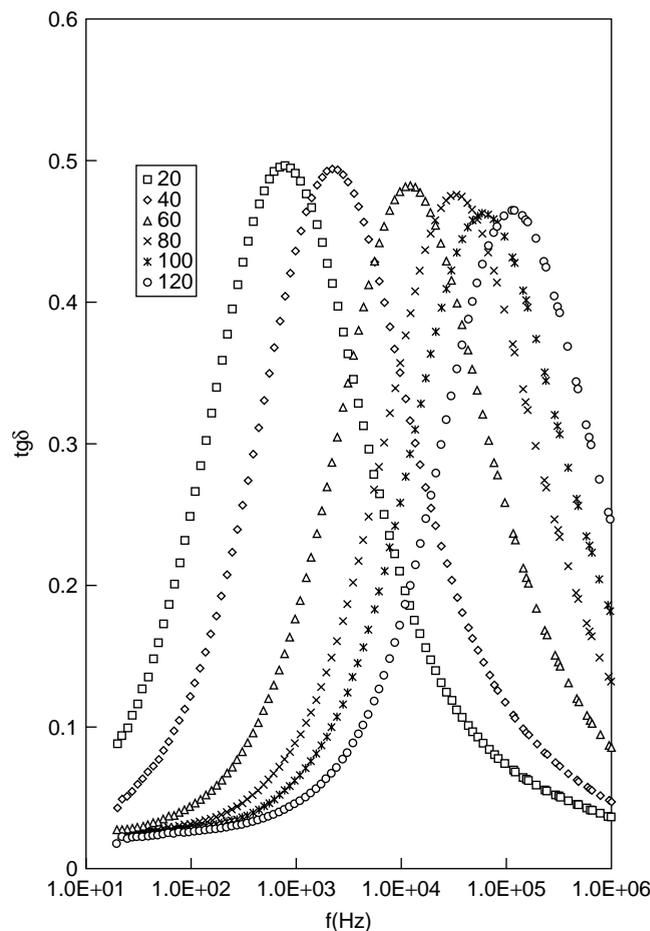


Figure 9. The dielectric loss tangent of the MS3A suspension as a function of frequency obtained at temperatures from 20–120°C.

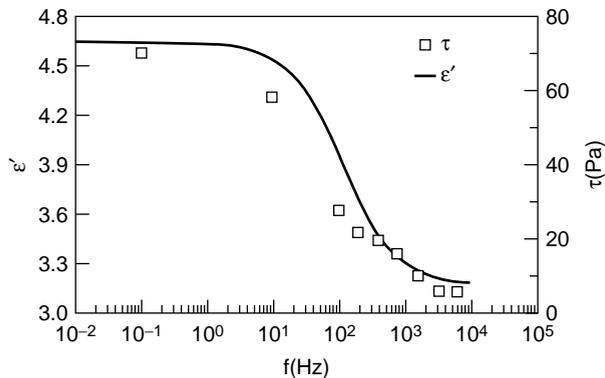


Figure 10. The frequency dependence of the yield stress of the OP suspension obtained at 300 V/mm. The solid line is the Wagner model's prediction [reproduced from (56)].

oil suspension versus frequency is shown in Fig. 10. The yield stress was obtained at an oscillatory field strength of 300 V/mm and a frequency range from 10^{-1} to 10^4 Hz, and the dielectric constant data were calculated by using the Wagner–Maxwell equation. The frequency dependences of the yield stress and the dielectric constant coincided, indicating that the Wagner–Maxwell polarization controls the dielectric properties of ER suspensions and then the ER effect.

As the experimental facts indicated, the importance of the Wagner–Maxwell polarization in ER fluids is obvious. However, it is still not clear how interfacial polarization controls the ER effect or how the material dielectric constant and dielectric loss correlate with the ER effect. From the Wagner–Maxwell Eq. (4), one can see that interfacial polarization depends greatly on the conductivity of the solid particulate material. A material, whose conductivity can be changed easily and thus whose ER performance would be different due to the different interfacial polarization contribution, would be very useful for evaluating whether the dielectric constant or the dielectric loss are important in the ER effect. Oxidized polyacrylonitrile (OP) and aluminosilicate (AS) were selected as candidates for this purpose because the conductivity of OP can be easily varied by changing the carbonating temperature and that of the AS can also be easily controlled by changing the

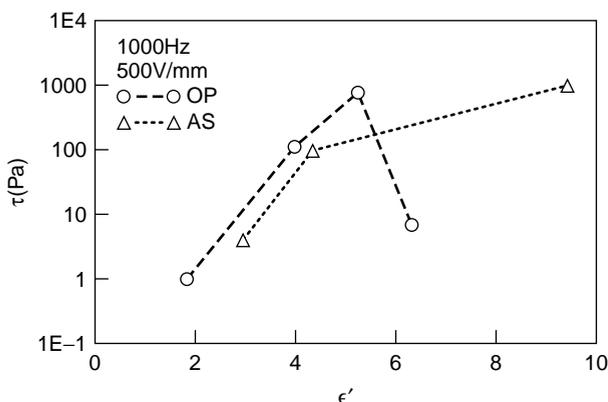


Figure 11. The shear stress of OP and AS suspensions versus the dispersed particle constant [reproduced from (52)].

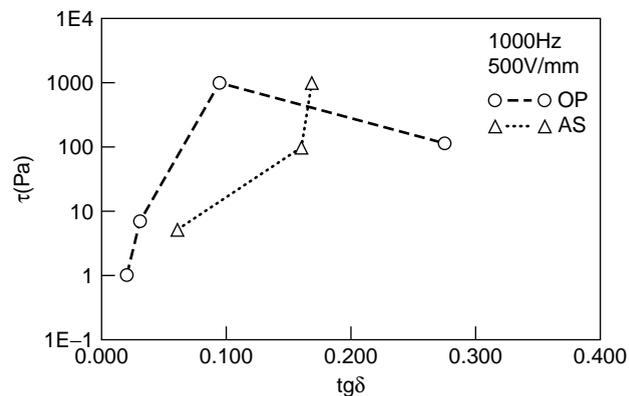


Figure 12. The shear stress of OP and AS suspensions versus the dispersed particle dielectric loss tangent [reproduced from (52)].

heat treatment temperature. The heat processing is controlled merely to change the material's conductivity rather than the material itself. The shear stresses versus the particulate dielectric constants and dielectric loss tangents of both of these materials are shown in Figs. 11, 12. The shear stresses do not change monotonically with the dielectric constant or dielectric loss tangent even for the same series of samples. From these two figures, one can see that the material that has a larger dielectric constant and a large dielectric loss would give a stronger ER effect. The ER effect would be weak if the material has only a large dielectric constant but a small dielectric loss. This means that the dielectric loss is also very important. By analyzing the experimental results presented in Figs. 11 and 12, one can conclude that the dielectric loss tangent of the solid particulate material should be around 0.10 at 1000 Hz, which is the precondition for the ER effect. Furthermore, the larger the dielectric constant, then the stronger the ER effect. These results are little surprising because one cannot imagine what the function of the dielectric loss in the ER effect is. However, taking a look at the dielectric loss tangent data of solid materials that it is already found, present a good ER effect (see Table 1), one has to come to the same conclusion, indicating that the empirical criterion proposed before is reasonable.

Table 1. Dielectric Constant ϵ and Dielectric Loss Tangent ($tg\delta$) of Currently Used ER Solid Materials at 20°C

Material	ϵ	$tg\delta$	Frequency (Hz)
Silica glass	9.7	0.24	1000
Sandy soil	3.42	0.08	1000
Titanium dioxide	200	0.35	1000
Phenol-formaldehyde	7	0.11	1000
Cellulose nitrate	8.4	0.1	100
Methyl cellulose	7.6	0.128	100
Polyvinyl alcohol acetate			
(25°C)	7.8	0.05	1000
(85°C)	100	1.3	1000
Polyethyl methacrylate			
(25°C)	2.75	0.042	1000
(80°C)	3.36	0.11	1000
Oxidized polyacrylonitrile	5.5	0.085	1000

THEORETICAL TREATMENT

What is the physical basis of the empirical criterion described before? This section will deal with this issue theoretically. One typical feature of ER fluids is that ER particulates can form a fibrillated structure in an electric field, whereas non-ER particulates cannot. Because the order in ER fluids increases dramatically after an electric field is applied, the entropy changes in the ER systems should obviously be less than zero.

Consider a general ER system. Assuming that the volume of the ER fluid is always constant and that the temperature T is the only parameter besides the electric field E that varies, the increase dU of internal energy U per unit volume of ER suspension, after an electric field is applied, can be expressed as follows, according to the first thermodynamic law and electromagnetic theory (41):

$$dU = dQ + \frac{E}{4\pi}dD, \quad (12)$$

where dQ is the influx of heat per unit volume, D is the electric displacement, and $\frac{E}{4\pi}dD$ represents the influx of energy into the ER fluid. For ER fluids, it is already known that, both theoretically and experimentally, there is a critical electric field at which the liquid–solid (disorder–order) transition takes place (23). The critical electric field is usually less than 200 V/mm, and a typical value is around 45 V/mm for many ER systems. When the applied electric field is slightly larger than the critical field, one may assume that the static dielectric constant ε_s of an ER suspension is independent of E because the applied electric field is not very strong. Even in a very strong field, the static dielectric constant of many materials would not substantially change (46). Then $D = \varepsilon_s E$; thus,

$$dD = d(\varepsilon_s E) = \varepsilon_s dE + E d\varepsilon_s = \varepsilon_s dE + E \frac{\partial \varepsilon_s}{\partial T} dT, \quad (13)$$

because ε_s depends just on T . It will be useful to take T and E^2 as independent variables. Thus Eq. (12) can be rewritten as

$$dU = dQ + \frac{\varepsilon_s}{8\pi}d(E^2) + \frac{E^2}{4\pi} \frac{\partial \varepsilon_s}{\partial T} dT = \frac{\partial U}{\partial(E^2)}d(E^2) + \frac{\partial U}{\partial T}dT. \quad (14)$$

One also can think that entropy S is a function of T and E^2 , as $dS = dQ/T$, the second thermodynamic law for a reversible process, so that

$$dS = \frac{\partial S}{\partial T}dT + \frac{\partial S}{\partial(E^2)}d(E^2). \quad (15)$$

Inserting dQ from Eqs. (14) and (15),

$$dS = \frac{1}{T} \left(\frac{\partial U}{\partial T} - \frac{E^2}{4\pi} \frac{\partial \varepsilon_s}{\partial T} \right) dT + \frac{1}{T} \left(\frac{\partial U}{\partial(E^2)} - \frac{\varepsilon_s}{8\pi} \right) d(E^2). \quad (16)$$

The condition that dS is a total differential requires that

$$\frac{\partial}{\partial(E^2)} \left[\frac{1}{T} \left(\frac{\partial U}{\partial T} - \frac{E^2}{4\pi} \frac{\partial \varepsilon_s}{\partial T} \right) \right] = \frac{\partial}{\partial T} \left[\frac{1}{T} \left(\frac{\partial U}{\partial(E^2)} - \frac{\varepsilon_s}{8\pi} \right) \right]. \quad (17)$$

Carrying out the differentiation,

$$\frac{\partial U}{\partial(E^2)} = \frac{1}{8\pi} \left(\varepsilon_s + T \frac{\partial \varepsilon_s}{\partial T} \right). \quad (18)$$

Comparing Eqs. (14) and (15) and inserting $\partial U/\partial(E^2)$ from Eq. (18),

$$\frac{\partial S}{\partial(E^2)} = \frac{1}{8\pi} \frac{\partial \varepsilon_s}{\partial T}. \quad (19)$$

Integrating,

$$S = S_0(T) + \frac{\partial \varepsilon_s}{\partial T} \frac{E^2}{8\pi}, \quad (20)$$

where $S_0(T)$ is the entropy in the absence of a field. Thus

$$\Delta S = \frac{\partial \varepsilon_s}{\partial T} \frac{E^2}{8\pi}. \quad (21)$$

For an ER fluid, $\Delta S < 0$; this obviously requires that $\partial \varepsilon_s/\partial T < 0$. The static dielectric constant ε_s of an ER fluid is the dielectric constant at $\omega = 0$. Thus, it can be expressed from Eq. (1) as

$$\varepsilon_s = \varepsilon_\infty(1 + K). \quad (22)$$

Assuming that both ε_{sm} and ε_{sp} are functions of T ,

$$\begin{aligned} \frac{d\varepsilon_s}{dT} = & \left[1 + 3\Phi + \frac{54\Phi^2\varepsilon_{sm}(\varepsilon_{sp}^2 - 1.5\varepsilon_{sm}\varepsilon_{sp} - \varepsilon_{sm}^2)}{(2\varepsilon_{sm} + \varepsilon_{sp})^3} \right] \frac{d\varepsilon_{sm}}{dT} \\ & - \frac{27\varepsilon_{sm}^2\Phi^2(\varepsilon_{sp} - 4\varepsilon_{sm})}{(2\varepsilon_{sm} + \varepsilon_{sp})^3} \frac{d\varepsilon_{sp}}{dT}. \end{aligned} \quad (23)$$

For polar liquids, $d\varepsilon_{sm}/dT$ is always less than zero. So, if $\varepsilon_{sp} > 4\varepsilon_{sm}$ and $d\varepsilon_{sp}/dT > 0$, $d\varepsilon_s/dT$ would be less than zero under any conditions; if $\varepsilon_{sp} > 4\varepsilon_{sm}$, $d\varepsilon_{sp}/dT < 0$, the situation becomes complicated, and the ER effect is still theoretically possible if

$$\begin{aligned} & \left(\frac{d\varepsilon_{sp}}{dT} \right) / \left(\frac{d\varepsilon_{sm}}{dT} \right) \\ & < \frac{(1 + 3\Phi)(2\varepsilon_{sm} + \varepsilon_{sp})^3 + 54\Phi^2\varepsilon_{sm}(\varepsilon_{sp}^2 - 1.5\varepsilon_{sm}\varepsilon_{sp} - \varepsilon_{sm}^2)}{27\Phi^2\varepsilon_{sm}^2(\varepsilon_{sp} - 4\varepsilon_{sm})}. \end{aligned}$$

However, the possibility is low because the entropy decrease is not so large (the particles contribute positively) and the ER solidification process would need a much larger decrease in entropy to form a body-centered tetragonal crystalline structure. So, only the weak ER effect or even

no ER effect would be generated. If the value of $(\frac{d\varepsilon_{sp}}{dT})/(\frac{d\varepsilon_{sm}}{dT})$ is larger than the value of

$$\frac{(1 + 3\Phi)(2\varepsilon_{sm} + \varepsilon_{sp})^3 + 54\Phi^2\varepsilon_m(\varepsilon_{sp}^2 - 1.5\varepsilon_{sm}\varepsilon_{sp} - \varepsilon_{sm}^2)}{27\Phi^2\varepsilon_{sm}^2(\varepsilon_{sp} - 4\varepsilon_{sm})},$$

$\frac{d\varepsilon_s}{dT}$ would be larger than zero, a negative ER effect, i.e., a decrease in the viscosity of the fluid as the applied electric field increases, would be observed. If $\varepsilon_{sp} < 4\varepsilon_{sm}$, $d\varepsilon_{sp}/dT$ is unlikely to be less than zero because a small dielectric constant is generally due to polar group orientation, as well as electronic and atomic polarization. For a polar material that has a comparatively small dielectric constant, its $d\varepsilon_{sp}/dT$, likely of the same order as $d\varepsilon_{sm}/dT$, would be larger than zero (41). So the ER effect might also be generated but would still be very weak. Here, we just concentrate on the positive ER situation, $d\varepsilon_{sp}/dT > 0$.

According to the Clausius–Mossotti formula, the static dielectric constant ε of a pure material can be expressed as (47)

$$\frac{\varepsilon - 1}{\varepsilon + 2} = \frac{4}{3}\pi N(\alpha_e + \alpha_a + \alpha_i + \alpha_d), \quad (24)$$

where α_e , α_a , α_i , and α_d are the polarizabilities of the electron, atom, ion displacement, and dipole rotation; and N is the number of molecules per unit volume. For a solid material, dipole polarization is negligible because solidification usually fixes the molecule with such rigidity in the lattice that little or no orientation of the dipole occurs in an electric field (41). Thus for ER particles,

$$\frac{\varepsilon_{sp} - 1}{\varepsilon_{sp} + 2} = \frac{4}{3}\pi N(\alpha_e + \alpha_a + \alpha_i). \quad (25)$$

In a very high frequency region, Eq. (25) leads to

$$\frac{\varepsilon_{\infty p} - 1}{\varepsilon_{\infty p} + 2} = \frac{4}{3}\pi N(\alpha_e + \alpha_a). \quad (26)$$

Differentiating Eq. (25) with respect to T ,

$$\frac{3}{(\varepsilon_{sp} + 2)^2} \frac{d\varepsilon_{sp}}{dT} = \frac{4}{3}\pi(\alpha_e + \alpha_a + \alpha_i) \frac{dN}{dT} + \frac{4}{3}\pi N \frac{d\alpha_i}{dT}. \quad (27)$$

Because $\frac{1}{N} \frac{dN}{dT} = -3\beta$, where β is the material's linear coefficient of expansion, Eq. (27) can be rewritten as

$$\frac{d\varepsilon_{sp}}{dT} = \left[-3 \frac{\varepsilon_{sp} - 1}{\varepsilon_{sp} + 2} \beta + \frac{4\pi N}{3} \frac{d\alpha_i}{dT} \right] \frac{(\varepsilon_{sp} + 2)^2}{3}, \quad (28)$$

where $\alpha_i = \frac{2q^2}{k}$, $k = \frac{2M(n-1)q^2r_0^{n-1}}{3r^{n+2}}$ [see (47)], q is the ion charge, k is the elastic bonding coefficient between two ions of opposite sign of charge, M and n are constants, and r_0 and r are the equilibrium distance and the distance between the two ions. Assuming that the α_i and k are function of temperature due to the fact that r would change

with temperature,

$$\frac{d\alpha_i}{dT} = -\frac{\alpha_i}{k} \frac{dk}{dT}, \quad (29)$$

$$\frac{dk}{dT} = -(n+2)k \frac{1}{r} \frac{dr}{dT}. \quad (30)$$

By definition, $\frac{1}{r} \frac{dr}{dT} = \beta$. From Eq. (26),

$$\frac{d\alpha_i}{dT} = \alpha_i(n+2)\beta = \frac{3}{4\pi N} \left(\frac{\varepsilon_{sp} - 1}{\varepsilon_{sp} + 2} - \frac{\varepsilon_{\infty p} - 1}{\varepsilon_{\infty p} + 2} \right) (n+2)\beta. \quad (31)$$

Inserting Eq. (31) into Eq. (28) and assuming $\frac{\varepsilon_{sp} - 1}{\varepsilon_{sp} + 2} = p_s$ and $\frac{\varepsilon_{\infty p} - 1}{\varepsilon_{\infty p} + 2} = p_\infty$,

$$\frac{d\varepsilon_{sp}}{dT} = [(n-1)p_s - (n+2)p_\infty] \frac{(\varepsilon_{sp} + 2)^2 \beta}{3}. \quad (32)$$

Only when $\frac{p_s}{p_\infty} > \frac{n+2}{n-1}$, would $d\varepsilon_{sp}/dT$ be larger than zero. This requirement leads to

$$\frac{\varepsilon_{sp} - \varepsilon_{\infty p}}{\varepsilon_{sp}\varepsilon_{\infty p}} > \frac{1}{n} \approx 0.1 \quad (33)$$

because for most materials, $n \approx 10$ [see (47)]. Equation (33) is the dielectric requirement for an ER particulate material.

The conductivities of most ER particulate materials are comparatively low, and one would assume that the particle dielectric loss results just from ion displacement polarization. The dielectric loss due to ion displacement polarization can be calculated from the absorption current. Assuming that the initial conductivity of an ER particle material is σ_0 the absorption current can be expressed as $\sigma_0 e^{-t/\theta}$, t is time, θ is a time constant, $\theta = \frac{\varepsilon_{sp} + 2}{\varepsilon_{\infty p} + 2} \tau$, τ is the relaxation time, thus under an oscillatory electric field $E = E_0 e^{i\bar{\omega}t}$, the induced current density d_i due to ion displacement polarization can be expressed as (47)

$$\begin{aligned} d_i &= \sigma_0 \int_{-\infty}^{t_1} \frac{dE}{dt} e^{-\frac{t-t}{\theta}} dt = i\bar{\omega} \sigma_0 E_0 \int_{-\infty}^{t_1} e^{i\bar{\omega}t - \frac{t-t}{\theta}} dt \\ &= \frac{i\bar{\omega} \theta \sigma_0}{1 + i\bar{\omega} \theta} E_0 e^{-\frac{t_1}{\theta}} e^{\frac{1+i\bar{\omega}\theta}{\theta}}, \\ &= \frac{\bar{\omega}^2 \theta^2 \sigma_0}{1 + \bar{\omega}^2 \theta^2} E_0 e^{i\bar{\omega}t_1} + i \frac{\bar{\omega} \theta \sigma_0}{1 + \bar{\omega}^2 \theta^2} E_0 e^{i\bar{\omega}t_1}, \end{aligned} \quad (34)$$

where t_1 is a constant. The first term of this equation has the same phase as the applied electric field, can be called the ohmic component d_{io} , and would contribute to the dielectric loss; the second term has a $(\pi/2)$ phase difference from the electric field, can be called the capacitive component d_{ic} , and contributes just to polarization. Because atomic and electronic polarizations also contribute to the capacitive current density and the general relationship between the induced current density d and the dielectric

constant ε is $d = \frac{\bar{\omega}\varepsilon E_0}{4\pi}$, the current density due to these two kinds of polarization d_{ae} can be expressed as $d_{ae} = \frac{\bar{\omega}\varepsilon_{\infty p} E_0}{4\pi}$. Then, the total capacitive current density d_c can be written as

$$d_c = d_{ic} + d_{ae} = \bar{\omega} \left(\frac{\varepsilon_{\infty p}}{4\pi} + \frac{\theta\sigma_0}{1 + \bar{\omega}^2\theta^2} \right) E_0. \quad (35)$$

According to the definition, the dielectric loss tangent $tg\delta$ is given by

$$tg\delta = \frac{d_{io}}{d_c} = \frac{\bar{\omega} \theta^2 \sigma_0}{\frac{\varepsilon_{\infty p}}{4\pi} (1 + \bar{\omega}^2 \theta^2) + \theta \sigma_0}. \quad (36)$$

The total dielectric constant ε is given by

$$\varepsilon = \frac{d_c}{\bar{\omega} E_0} 4\pi = \varepsilon_{\infty p} + \frac{4\pi\theta\sigma_0}{1 + \bar{\omega}^2\theta^2}. \quad (37)$$

In a dc field ($\omega = 0$),

$$\varepsilon_{sp} = \varepsilon_{\infty p} + 4\pi\theta\sigma_0. \quad (38)$$

Inserting Eq. (38) into Eq. (36),

$$tg\delta = \frac{\bar{\omega} \theta (\varepsilon_{sp} - \varepsilon_{\infty p})}{\varepsilon_{sp} + \varepsilon_{\infty p} \bar{\omega}^2 \theta^2}. \quad (39)$$

Equation (39) is a general expression for a solid material of very low conductivity and marked ionic displacement polarization. Differentiating Eq. (39) with respect to ω , $tg\delta$ has a maximum value,

$$tg\delta_{\max} = \frac{\varepsilon_{sp} - \varepsilon_{\infty p}}{2\sqrt{\varepsilon_{sp}\varepsilon_{\infty p}}}. \quad (40)$$

Comparing this with Eq. (33),

$$tg\delta_{\max} > \frac{\varepsilon_{sp} - \varepsilon_{\infty p}}{\varepsilon_{sp}\varepsilon_{\infty p}} > 0.1. \quad (41)$$

For many ER particle materials, $\sqrt{\varepsilon_{sp}\varepsilon_{\infty p}}$ is always larger than 2. Equation (34) indicates that the suspended particles can become ordered in an electric field only when the maximum value of the dielectric loss tangent is larger than 0.1, which matches exactly the empirical criterion obtained experimentally.

In conclusion, from the basic fact that the entropy of an ER fluid is greatly reduced after an electric field is applied, one may theoretically find that the maximum value of the dielectric loss tangent of the dispersed particles should be larger than 0.1, which agrees well with experimental results. The negative ER effect becomes possible only if $d\varepsilon_{sp}/dT < 0$.

THE MECHANISM OF THE ER EFFECT

It is found experimentally and theoretically that both the large dielectric loss and the large dielectric constant of dispersed particles are very important for the ER effect.

Interfacial polarization is also crucial. What is the relationship between interfacial polarization and the dielectric constant and loss of dispersed particles? Why can ER particulates form a fibrillated structure?

As shown in Fig.1, the total polarization P of a heterogeneous material can be expressed as

$$P = P_i + P_d + P_a + P_e. \quad (42)$$

Correspondingly, the total dielectric constant ε can be expressed as

$$\begin{aligned} \varepsilon &= \varepsilon_s + \varepsilon_i + \varepsilon_d + \varepsilon_a + \varepsilon_e, \\ &= \varepsilon_s + \varepsilon_i + \varepsilon_d + \varepsilon_{\infty}, \end{aligned} \quad (43)$$

where ε_i , ε_d , ε_a , and ε_p are artificially regarded as the dielectric constant induced by the interfacial, Debye, atomic, and electronic polarizations, respectively; ε_s is the static dielectric constant; and $\varepsilon_{\infty} = \varepsilon_a + \varepsilon_e$, is the high-frequency dielectric constant. As stated before, the dipole orientation contribution of solid particulate materials can be negligible; thus, Eq. (43) can be written as

$$\varepsilon = \varepsilon_s + \varepsilon_i + \varepsilon_{\infty}. \quad (44)$$

The dielectric constant and the dielectric loss are not independent, and the dielectric loss originally results from slow polarization, interfacial polarization in this case. A large dielectric loss tangent for an ER suspension would mean that the ratio of ε_i to ε is large. Obviously, the large dielectric loss of dispersed particles would result in a large dielectric loss for the whole suspension and then high interfacial polarization. So, interfacial polarization stems from the dielectric loss of dispersed particles. No obvious interfacial polarization is expected if the particulate material does not have a marked dielectric loss.

One typical feature of ER fluids is that ER particles can fibrillate between two electrodes; however, non-ER particles cannot do so. In the primary stage, ER and non-ER particles have almost same microstructures—the particles randomly distribute in the medium or stochastically form some clusters. After they are exposed to an external electric field, why do they behave quite differently? A major difference between ER and non-ER particles is that the dielectric loss tangents of many ER particulate materials are comparatively higher, around 0.10 at 1000 Hz., that is, ER suspensions usually have large interfacial polarizations. Why does the large interfacial polarization produce a difference between ER and non-ER particles? Because the interfacial polarization is originally associated with bounded surface charges, one has to presume that large amount of surface charge can make ER particles turn in the direction of an external electric field. The non-ER particles, however, cannot turn due to the lack of surface charge. Although they still can be polarized, the total interparticle force would be canceled out owing to the diversity of particle dipole vectors (see Fig. 13).

Generally, surface charge is associated with the material's dielectric loss and can be produced by the following mechanisms (46,48): (1) Debye polarization from dipole

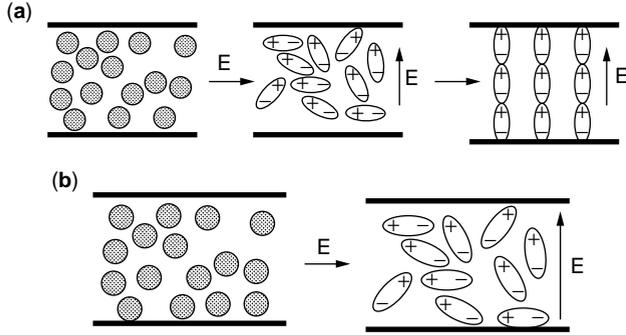


Figure 13. Schematic illustration of ER and non-ER particle behavior before and after an external electric field is applied. (a) ER particle; (b) non-ER particle [reproduced from (49)].

orientation; (2) interfacial polarization (Maxwell–Wagner polarization) from the displacement of charge carriers across a microscopic distance; and (3) leakage conduction from the formation of space charges. The two former effects do not involve any transfer of charge carriers between the dielectric and the electrodes and would contribute to the dielectric absorption; the latter would scale the space charge movement between the dielectric and the electrodes. So the net surface charge of a dielectric would be determined by the former two mechanisms. Because Debye orientation can be overlooked in ER systems, the net surface charge can be produced only by interfacial polarization and can be calculated from dc absorption current measurement. Hao and Xu (40) experimentally proved that even in a very weak electric field $E_{cr} = 12\text{V/mm}$, a particle can turn to align in the direction of the applied electric field if large interfacial polarization can be generated. ER particles turning in an electric field were also detected by using X-ray diffraction (50).

Obviously, after ER particles turn in the direction of an electric field, the interparticle force would be determined mainly by how much the particles are polarized, that is the value of particle dielectric constant. So, the mechanism of the ER effect contains two steps: first, the particles turn in the direction of the applied electric field; second, the particles bind together due to polarization. The first step is controlled by the dielectric loss of the dispersed particles, and the second step is controlled by the dielectric constant.

THE YIELD STRESS EQUATION

According to the ER mechanism proposed, the entropy change in an ER system obviously includes two parts: the particle configurational entropy, which represents the entropy change from the randomly distributed particle state to the body-centered tetragonal (bct) crystalline state; the other is the entropy change from the very weak interparticle force state to the exceptionally strong interparticle force state. The former would contribute to the particle arrangement, whereas the latter would contribute to the ER effect and is originally induced by interfacial polarization. In the first part, interfacial polarization could be presumed thought not to take place, and the interparticle

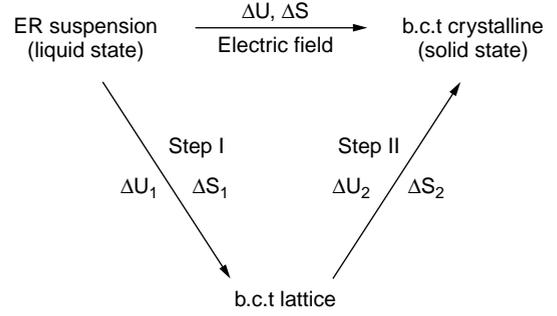


Figure 14. Schematic of the hypothesized two-step process during the solidification transition of an ER suspension in an external electric field [reproduced from (55)].

force could be negligible because it is extremely weak in this state; in the second part, interfacial polarization takes place, making the already well-arranged particles become strongly correlated. This hypothesized process is schematically illustrated in Fig. 14. To determine the yield stress of an ER suspension, one has to know the internal energy and entropy changes of the second step, ΔS_2 and ΔU_2 , respectively. Obviously,

$$\Delta S_2 = \Delta S - \Delta S_1, \quad (45)$$

$$\Delta U_2 = \Delta U - \Delta U_1. \quad (46)$$

Integrating Eq. (18),

$$U = U_0(T) + \frac{E^2}{8\pi} \left(\varepsilon_s + T \frac{\partial \varepsilon_s}{\partial T} \right), \quad (47)$$

where $U_0(T)$ represents the internal energy of an ER suspension in the absence of an electric field. Thus, the internal energy change ΔU due to the applied electric field is

$$\Delta U = \frac{E^2}{8\pi} \left(\varepsilon_s + T \frac{\partial \varepsilon_s}{\partial T} \right). \quad (48)$$

Substituting Eqs. (22), (23) into Eq. (48),

$$\begin{aligned} U = U_0 + \frac{E^2}{8\pi} & \left[(1 + 3\Phi) \left(\varepsilon_{sm} + T \frac{d\varepsilon_{sm}}{dT} \right) \right. \\ & + \frac{27\Phi^2 \varepsilon_{sm}^2 (\varepsilon_{sp} - \varepsilon_{sm})}{(2\varepsilon_{sm} + \varepsilon_{sp})^2} \\ & + \frac{54\Phi^2 \varepsilon_{sm} T (\varepsilon_{sp}^2 - 1.5\varepsilon_{sm}\varepsilon_{sp} - \varepsilon_{sm}^2)}{(2\varepsilon_{sm} + \varepsilon_{sp})^3} \frac{d\varepsilon_{sm}}{dT} \\ & \left. - \frac{27\Phi^2 \varepsilon_{sm}^2 T (\varepsilon_{sp} - 4\varepsilon_{sm})}{(2\varepsilon_{sm} + \varepsilon_{sp})^3} \frac{d\varepsilon_{sp}}{dT} \right]. \quad (49) \end{aligned}$$

For step 1, we assumed that interfacial polarization does not take place, which is likely only under the condition $\varepsilon_{sm} = \varepsilon_{sp}$. Then, the static dielectric constant of this assumed system, according to Eq. (22), is $\varepsilon_{sm}(1 + 3\Phi)$. Thus, the influx of electric energy into the assumed system for

arraying disordered particles to form the very loose bct. lattice can be expressed, analogous to Eq. (47), by

$$\Delta U_1 = (1 + 3\Phi) \left(\varepsilon_{sm} + T \frac{d\varepsilon_{sm}}{dT} \right) \frac{E^2}{8\pi}, \quad (50)$$

and the entropy change

$$\Delta S_1 = (1 + 3\Phi) \frac{E^2}{8\pi} \frac{\partial \varepsilon_{sm}}{\partial T}. \quad (51)$$

Therefore, according to Eqs. (45) and (46),

$$\begin{aligned} \Delta U_2 = & \frac{E^2}{8\pi} \left[\frac{27\Phi^2 \varepsilon_{sm}^2 (\varepsilon_{sp} - \varepsilon_{sm})}{(2\varepsilon_{sm} + \varepsilon_{sp})^2} \right. \\ & + \frac{54\Phi^2 \varepsilon_{sm} T (\varepsilon_{sp}^2 - 1.5\varepsilon_{sm}\varepsilon_{sp} - \varepsilon_{sm}^2)}{(2\varepsilon_{sm} + \varepsilon_{sp})^3} \frac{d\varepsilon_{sm}}{dT} \\ & \left. - \frac{27\Phi^2 \varepsilon_{sm}^2 T (\varepsilon_{sp} - 4\varepsilon_{sm})}{(2\varepsilon_{sm} + \varepsilon_{sp})^3} \frac{d\varepsilon_{sp}}{dT} \right], \quad (52) \end{aligned}$$

$$\begin{aligned} \Delta S_2 = & \frac{E^2}{8\pi} \left[\frac{54\Phi^2 \varepsilon_{sm} T (\varepsilon_{sp}^2 - 1.5\varepsilon_{sm}\varepsilon_{sp} - \varepsilon_{sm}^2)}{(2\varepsilon_{sm} + \varepsilon_{sp})^3} \frac{d\varepsilon_{sm}}{dT} \right. \\ & \left. - \frac{27\Phi^2 \varepsilon_{sm}^2 T (\varepsilon_{sp} - 4\varepsilon_{sm})}{(2\varepsilon_{sm} + \varepsilon_{sp})^3} \frac{d\varepsilon_{sp}}{dT} \right]. \quad (53) \end{aligned}$$

ΔU_2 should be less than zero because the interparticle force in the ER crystalline lattice is attractive. Thus, Eq. (52) should be expressed as

$$\begin{aligned} \Delta U_2 = & -\frac{E^2}{8\pi} \left[-\frac{27\Phi^2 \varepsilon_{sm}^2 (\varepsilon_{sp} - \varepsilon_{sm})}{(2\varepsilon_{sm} + \varepsilon_{sp})^2} \right. \\ & - \frac{54\Phi^2 \varepsilon_{sm} T (\varepsilon_{sp}^2 - 1.5\varepsilon_{sm}\varepsilon_{sp} - \varepsilon_{sm}^2)}{(2\varepsilon_{sm} + \varepsilon_{sp})^3} \frac{d\varepsilon_{sm}}{dT} \\ & \left. + \frac{27\Phi^2 \varepsilon_{sm}^2 T (\varepsilon_{sp} - 4\varepsilon_{sm})}{(2\varepsilon_{sm} + \varepsilon_{sp})^3} \frac{d\varepsilon_{sp}}{dT} \right]. \quad (54) \end{aligned}$$

ΔS_2 should also be less than zero; as in step 2, entropy obviously decreases substantially and requires $\varepsilon_{sp} > 4\varepsilon_{sm}$ when $\frac{d\varepsilon_{sp}}{dT} > 0$ and $\varepsilon_{sp} < 4\varepsilon_{sm}$ when $\frac{d\varepsilon_{sp}}{dT} < 0$. These are the preconditions of Eq. (54).

Equation (54) gives the internal energy change per unit volume of the whole suspension. For an ER suspension of volume V , particle volume fraction ϕ , and particle diameter r , the number of particles N in this volume is

$$N = \frac{V\phi}{\frac{4}{3}\pi r^3} = \frac{3V\phi}{4\pi r^3}. \quad (55)$$

Then, the interparticle energy U_{ip} is

$$U_{ip} = \frac{\Delta U_2 V}{N/2} = \frac{8\pi r^3 \Delta U_2}{3\phi}. \quad (56)$$

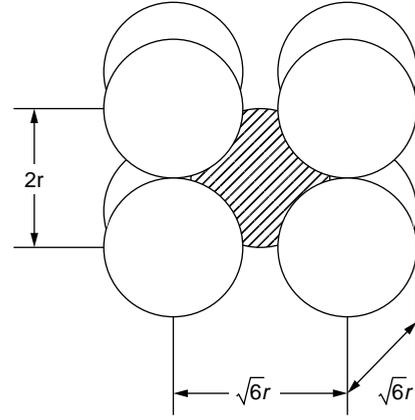


Figure 15. Schematic of the unit cell of a bct lattice formed by ER particles in an electric field. The radius of the particle is r [reproduced from (55)].

The free volume per particle V_{fp} in an off-state electric field is

$$V_{fp} = \frac{V - V\phi}{N} = \frac{4\pi r^3 (1 - \phi)}{3\phi}. \quad (57)$$

It is known that ER particles in an electric field form fibrillation chains of a bct. lattice. The unit cell of a bct. lattice is schematically illustrated in Fig. 15. One center particle, which belongs to a different chain class, would have eight nearest neighbor particles at the interparticle distance $2r$; besides, one center particle, which belongs to the same chain class, also has another six nearest neighbor center particles also at the interparticle distance $2r$. So, the total energy U_{it} for one particle in the unit cell

$$U_{it} = (8 + 6) \times U_{ip} = \frac{112\pi r^3 \Delta U_2}{3\phi}. \quad (58)$$

Because a particle cannot freely leave the cell of the lattice, the free volume per particle in an electric field, V_{fE} , would become

$$V_{fE} = V_{fp} - \left(\frac{2r \times \sqrt{6}r \times \sqrt{6}r}{2} \right) = \frac{4\pi r^3 (1 - \Phi) - 18\Phi r^3}{3\Phi}. \quad (59)$$

Note that there are two equivalent particles in one unit cell. Assuming that the yield stress τ_y is the force that can drive one particle to move within the free volume per particle (51),

$$\begin{aligned} \tau_y = & \frac{|U_{it}|}{V_{fE}} = \frac{112\pi}{4\pi(1 - \Phi) - 18\Phi} |\Delta U_2| \\ = & \frac{14E^2}{4\pi(1 - \Phi) - 18\Phi} \left[-\frac{27\Phi^2 \varepsilon_{sm}^2 (\varepsilon_{sp} - \varepsilon_{sm})}{(2\varepsilon_{sm} + \varepsilon_{sp})^2} \right. \\ & \left. - \frac{54\Phi^2 \varepsilon_{sm} T (\varepsilon_{sp}^2 - 1.5\varepsilon_{sm}\varepsilon_{sp} - \varepsilon_{sm}^2)}{(2\varepsilon_{sm} + \varepsilon_{sp})^3} \frac{d\varepsilon_{sm}}{dT} \right. \\ & \left. + \frac{27\Phi^2 \varepsilon_{sm}^2 T (\varepsilon_{sp} - 4\varepsilon_{sm})}{(2\varepsilon_{sm} + \varepsilon_{sp})^3} \frac{d\varepsilon_{sp}}{dT} \right] \end{aligned}$$

$$+ \frac{27\Phi^2 \varepsilon_{sm}^2 T(\varepsilon_{sp} - 4\varepsilon_{sm})}{(2\varepsilon_{sm} + \varepsilon_{sp})^3} \frac{d\varepsilon_{sp}}{dT} \Big]. \quad (60)$$

Equation (60) presents a correlation between the yield stress of an ER suspension and the physical parameters of the suspension materials. One can easily obtain the yield stress value if all of the parameters are already known. Equation (60) can be further developed to a more meaningful form. According to the Clausius–Mossotti equation, the liquid media used in ER fluids are usually nonpolar, and then,

$$\frac{d\varepsilon_{sm}}{dT} = -\frac{(\varepsilon_{sm} - 1)(\varepsilon_{sm} + 2)}{3} \beta_m, \quad (61)$$

where β_m is the linear expansion coefficient of the liquid. Assuming $\xi = \varepsilon_{sp}/\varepsilon_{sm}$ and substituting Eqs. (32) and (61) in Eq. (60),

$$\begin{aligned} \tau_y = & \frac{126\Phi^2 E^2}{4\pi(1-\Phi) - 18\Phi} - \left[\frac{3\varepsilon_{sm}(\xi^2 + \xi - 2)}{(2+\xi)^3} \right. \\ & + \frac{2\beta_m T(\xi^2 - 1.5\xi - 1)(\varepsilon_{sm}^2 + \varepsilon_{sm} - 2)}{(2+\xi)^3} \\ & \left. + \frac{T\beta_p p(\xi - 4)(\varepsilon_{sm}\xi + 2)^2}{(2+\xi)^3} \right]. \quad (62) \end{aligned}$$

where $p = [(n-1)p_s - (n+2)p_\infty]$. Once the parameters are known, Eq. (62) gives the yield stress value of the ER fluid. First, let us take a look at how the particle-to-oil dielectric constant ratio ξ changes the yield stress of an ER fluid on the basis of Eq. (62). Assuming that $d\varepsilon_{sp}/dT$ is 0.40 and $\beta_p p = 3.82 \times 10^{-4}$ (both are typical values for ER systems based on the experimental data), one can get the numerical relationship between the yield stress and the particle-to-oil dielectric constant ratio of the silicone oil ER system, which is shown in Fig. 16. Figure 16 clearly shows that the yield stress dramatically increases as ξ increases at a low particle-to-oil dielectric constant ratio; however, the yield stress gradually levels off after ξ is higher than 60. The sharp increase in yield stress takes place in the range of $\xi < 50$, indicating that materials whose static dielectric constant is around 150 would potentially display the best ER effect and materials that have a larger static dielectric constant would not always display a good ER effect if other properties are not beneficial to the ER effect.

Note that this conclusion is derived under the assumption that the parameter p has a positive value. If p is negative, the larger particle-to-oil dielectric constant ratio would not generate a larger yield stress. According to the definition and assuming that $\varepsilon_{sp} - \varepsilon_{\infty p} = A$,

$$p = \frac{3(n - \varepsilon_{\infty p})A - 3\varepsilon_{\infty p}(\varepsilon_{\infty p} + 1) + 6}{(\varepsilon_{\infty p} + 2)^2 + (\varepsilon_{\infty p} + 2)A}. \quad (63)$$

If $\varepsilon_{\infty p}$ is not so large (less than 10 for most small dielectric constant materials), $3(n - \varepsilon_{\infty p})$ is always larger than $(\varepsilon_{\infty p} + 2)$, and thus p would increase as A increases, that is, the ER effect would increase approximately with the difference between the dielectric constants below and above

the relaxation frequency; this was already found experimentally by Ikazaki and Kawai (31,32). For a solid material whose leak conductivity is σ , initial conductivity is σ_0 , and relaxation time related constant is θ , it is known that (47)

$$\varepsilon_{sp} - \varepsilon_{\infty p} = 4\pi\theta\sigma_0 \quad (64)$$

and

$$tg\delta = \frac{\sigma(1 + \bar{\omega}^2\theta^2) + \bar{\omega}^2\theta^2\sigma_0}{\frac{\varepsilon_{\infty p}\bar{\omega}}{4\pi}(1 + \bar{\omega}^2\theta^2) + \bar{\omega}\theta\sigma_0}. \quad (65)$$

Thus,

$$A = 4\pi\theta\sigma_0 = \frac{4\pi(1 + \bar{\omega}^2\theta^2) \left(\frac{\varepsilon_{\infty p}\bar{\omega} tg\delta}{4\pi} - \sigma \right)}{\bar{\omega}^2\theta - \bar{\omega} tg\delta}, \quad (66)$$

where $\bar{\omega}$ is the field frequency and $tg\delta$ is the dielectric loss tangent of the solid material. Equation (66) indicates that A obviously increases as $tg\delta$ increases, provided that $\bar{\omega}\theta > tg\delta$ (this always holds because A should have a positive value). Differentiating Eq. (66) with respect to σ , one would find a maximum value for A at

$$\sigma = \frac{\left(\frac{\varepsilon_{\infty p}\bar{\omega}}{4\pi} \frac{dtg\delta}{d\sigma} - 1 \right) \bar{\omega}^2\theta + \bar{\omega} tg\delta}{\bar{\omega} \frac{dtg\delta}{d\sigma}}, \quad (67)$$

because the second derivative of A with respect to σ , $\frac{d^2 A}{d\sigma^2}$, is negative. Considering that $\bar{\omega}\theta > tg\delta$ and $\frac{\varepsilon_{\infty p}\bar{\omega}}{4\pi} \frac{dtg\delta}{d\sigma} \gg 1$, $\sigma \approx \frac{\varepsilon_{\infty p}\bar{\omega}^2\theta}{4\pi}$. For solid materials, one would assume that Debye polarization cannot occur and the dielectric adsorption phenomenon stems just from ion polarization; at $\bar{\omega} = \frac{1}{\theta} \sqrt{\frac{\varepsilon_{sp}}{\varepsilon_{\infty p}}}$, ion polarization-induced $tg\delta$ would reach at a maximum value (47). Then, parameter A also

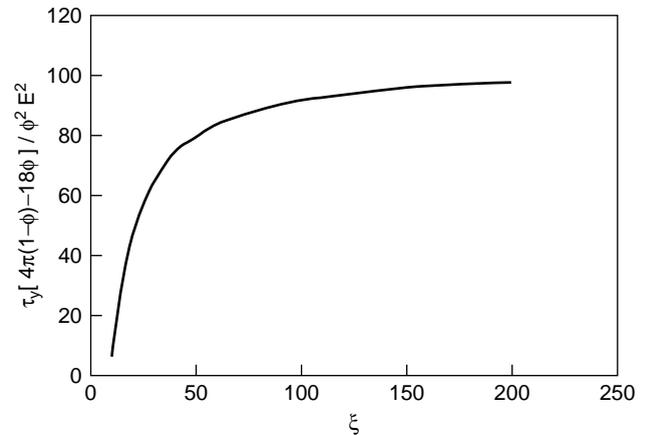


Figure 16. The numerical relationship between the normalized yield stress, $\tau_y[4\pi(1-\Phi) - 18\Phi]/\Phi^2 E^2$, and the particle-to-oil dielectric constant ratio ξ at room temperature for the silicon oil ER system based on Eq. (62) [reproduced from (55)].

reaches a maximum value. Thus, $\sigma \approx \frac{\epsilon_{sp}}{4\pi\theta}$, which indicates that the yield stress for different dispersed solids would peak at different particle conductivities. In the poly(acenequinones)/silicone oil system, the yield stress, it was found, peaks at a particle conductivity around 10^{-5} S/m (7). However, in the oxidized polyacrylonitrile/silicone oil system, the maximum value of the yield stress occurs near 10^{-7} S/m (52). According to the dielectric data presented in each paper, the optimal conductivity ($\sigma \approx \frac{\epsilon_{sp}}{4\pi\theta}$) is crudely estimated at 0.22×10^{-5} S/m for the former system and 0.88×10^{-7} S/m for the latter system, in good agreement with experimental results.

The temperature dependence of the yield stress can also be qualitatively analyzed using Eq. (62). Because σ and $tg\delta$ are more sensitive to temperature than the dielectric constant, one would still center on the temperature dependence of the parameter p . The conductivity of most solid dielectric materials would exponentially increase with temperature. Thus, the conductivity, rather than other parameters, would be a main variable and surely makes a big contribution to the yield stress as temperature varies [see Eq. (62)]. The yield stress would also go through a maximum value at the temperature where the conductivity reaches its optimal value. It was found experimentally that the yield stress first increases and then decreases with temperature (53,54). Accordingly, the yield stress would decrease with temperature if the conductivity of the solid particles were already larger than the optimal value, and the yield stress would increase with temperature if the conductivity were lower than the optimal value; this was also experimentally found before (54).

The relaxation time constant θ would also influence parameter A , and finally p substantially. From either Eqs. (64) or (66), one would find that p increases as θ increases, that is, the ER effect would be stronger if dielectric relaxation is slower. However, too slow a relaxation time (then slow response time) would make the ER fluids become useless. Generally, a ER response time around 1 millisecond is favorable, thus requiring that the relaxation time is of the same timescale, that is, a dielectric relaxation frequency around 10^3 Hz. Block thought that the polarization rate would be important in the ER response process and that polarization too fast or too slow is not good for an ER effect (7). Ikazaki and Kawai found experimentally that the ER fluids whose relaxation frequencies are within the range of 100 – 10^5 Hz would exhibit a large ER effect (31,32), supporting our derivation.

CONCLUSION

The dielectric properties of ER fluids are thoroughly addressed in this article. Among the polarizations (electronic, atomic, Debye, and interfacial) that occur in ER fluids in an electric field, it is found experimentally that interfacial polarization contributes to the ER effect and that the Wagner–Maxwell equation, which deals with interfacial polarization in heterogeneous systems can describe the dielectric phenomena observed in ER fluids. The dielectric loss of a dispersed particulate material, it is found, plays an important role in ER systems, and the dielectric constant

becomes dominant only when the dielectric loss tangent is larger than 0.10 at 1000 Hz. This empirical criterion for selecting an ER solid has a firm physical basis—it is the precondition for the entropy decrease commonly observed in ER systems. Theoretically, one also can come to the same conclusion if we assume that the entropies of the ER systems would greatly decrease, as ER fluids change from a randomly distributed suspension state to a body-centered-tetragonal crystal state after an external electric field is applied. The mechanism of the ER effect consists of two steps. The first is the particle turning process, which is controlled by the particle dielectric loss; the second is the particle's strong interaction, which is controlled by the dielectric constant. The entropy change in ER systems obviously includes two parts: One is the particle configurational entropy, which represents the entropy change from the randomly distributed particle state to the bct. lattice state; the other is the entropy change from the very weak interparticle force state to the exceptionally strong interparticle force state. The former contributes to particle rearrangement, and the latter contributes to the ER effect, which, it is thought, is originally induced by interfacial polarization. On the basis of this physical picture of the ER mechanism, a general form of a yield stress equation is derived from the internal energy and entropy changes of an ER fluid in an external static electric field. The yield stress equation can give very good predictions in accordance with the previous results obtained so far.

The yield stress equation involves a very important parameter p , which gives an appropriate expression for the dielectric loss tangent criterion put forward before and distinguishes our equation from others. When p is positive, the yield stress increases as the particle-to-medium dielectric ratio ξ increases; this is shown by the polarization model and other models. However, the yield stress gradually levels off when $\xi > 60$. The sharp increase in the yield stress takes place in the range of $\xi < 50$, indicating that materials that have a static dielectric constant around 150 would potentially display the best ER effect (in the PDMS oil system) and materials that have a larger static dielectric constant would not always display a good ER effect if other properties are not good for the ER effect. When p is negative, the suspension would give a weak or even no ER effect. To make the value of p as large as possible, very slow relaxation time and suitable particle conductivity are essential. To obtain a good ER fluid, from Eq. (62), one should use a liquid medium and a solid material that have high dielectric constants (a large dielectric constant ratio of the solid to the liquid). In addition, the solid must have a large dielectric loss tangent (a maximum value at least larger than 0.1) to keep the parameter p positive.

BIBLIOGRAPHY

1. U.S. Pat. 2,417,850, 1947, W.M. Winslow.
2. D.L. Klass and T.W. Martinek, *J. Appl. Phys.* **38**(1): 67(1967).
3. H. Ujima, *Jpn. J. Appl. Phys.* **11**(3): 319 (1972).
4. J.E. Stangroom, *Phys. Technol.*, **14**: 290 (1983).
5. H. Block and J.P. Kelly, *Proc. IEE Colloq.* **14**: 1 (1985).
6. F.E. Filisko and L.H. Radzilowski, *J. Rheol.* **34**(4): 539 (1990).

7. H. Block, J.P. Kelly, A. Qin, and T. Watson, *Langmuir* **6**: 6 (1990).
8. R.A. Anderson, *Langmuir* **10**: 2917 (1994).
9. T. Garino, D. Adolf, and B. Hance, in *Proc. Int. Conf. ER Fluids*, R. Tao, ed. World Scientific, Singapore, 1992, p. 167.
10. L.C. Davis, *J. Appl. Phys.* **72**(4): 1334 (1992).
11. L.C. Davis, *J. Appl. Phys.* **73**(2): 680 (1993).
12. L.C. Davis, *Appl. Phys. Lett.* **60**(3): 319 (1992).
13. Y. Otsubo and K. Watanabe, *J. Soc. Rheol. Jpn.* **18**: 111 (1990).
14. C.F. Zukoski, *Annu. Rev. Mater. Sci.* **23**, 45 (1993).
15. P. Atten, J.-N. Foulc, and N. Felici, *Int. J. Mod. Phys. B* **8**: 2731 (1994).
16. J.-N. Foulc, P. Atten, and N. Felici, *J. Electrostatics* **33**: 103 (1994).
17. X. Tang, C.Wu, and H. Conrad, *J. Rheol.* **39**(5): 1059 (1995).
18. X. Tang and H. Conrad *J. Appl. Phys.* **80**(9): 5240 (1996).
19. C. Wu and H. Conrad, *J. Phys. D: Appl. Phys.* **29**: 3147 (1996).
20. C. Wu and H. Conrad, *J. Appl. Phys.* **81**(12): 8057 (1997).
21. C. Wu and H. Conrad, *J. Appl. Phys.* **81**(1): 383 (1997).
22. A. Inoue, in *Proc. Int. Conf. ER Fluids*, J.D. Carlson, A.F. Sprecher, and H. Conrad, eds., Technomic, Lancaster-Basel, 1990, p. 176.
23. B. Khusid, and A. Acrivos, *Phys. Rev. E* **52**: 1669 (1995).
24. J. Trlica, O. Quadrat, P. Bradna, V. Pavlinek, and P. Saha, *J. Rheol.* **40**(5): 943 (1996).
25. H. See and T. Saito, *Rheol. Acta* **35**: 233 (1996).
26. H. Ma, W. Wen, W.Y. Tam, and P. Sheng, *Phys. Rev. Lett.* **77**(12): 2499 (1996).
27. W. Wen, H. Ma, W.Y. Tam, and P. Sheng, *Phys. Rev. E* **55**(2): R1294 (1997).
28. U. Treasurer, L.H. Radzilowski, and F.E. Filisko, *J. Rheol.* **35**(6): 1051 (1991).
29. F.E. Filisko, in *Progress in Electrorheology*, K.O. Havelka and F.E. Filisko, eds., Plenum Press, NY, 1994, p. 3.
30. A.W. Schubring and F.E. Filisko, in *Progress: in Electrorheology*, K.O. Havelka and F.E. Filisko, eds., Plenum Press, NY, 1994, p. 215.
31. A. Kawai, K. Uchida, K. Kamiya, A. Gotoh, S. Yoda, K. Urabe, and F. Ikazaki, *Int. J. Modern Phys. B* **10**: 2849 (1996).
32. F. Ikazaki, A. Kawai, T. Kawakami, K. Edamura, K. Sakurai, H. Anzai, and Y. Asako, *J. Appl. Phys. D* **31**: 336 (1998).
33. F. Ikazaki, A. Kawai, T. Kawakami, M. Konishi, and Y. Asako, *Proc. Int. Conf. Electrorheological Magnetorheological Fluids*, K. Koyama and M. Nakano, eds., World Scientific, Singapore, 1998, p. 205.
34. Z.Y. Qiu, H. Zhang, Y. Tang, L.W. Zhou, C. Wei, S.H. Zhang, and E.V. Korobko, *Proc. Int. Conf. Electrorheological Magnetorheological Fluids*, K. Koyama and M. Nakano, eds., World Scientific: Singapore, 1998, p. 197.
35. S.O. Morgan, *Trans. Am. Electrochem. Soc.* **65**: 109 (1934).
36. R.W. Sillars, *J.I.E.E.* **80**: 378 (1937).
37. T. Hanai, N. Koizumi, and R. Gotoh, *Proc. Symp. Rheol. Emulsion.*, P. Sherman ed., Pergamon, Oxford, 1963, p. 91.
38. D.A.G. Bruggeman, *Ann. Phys.* **24**: 636 (1935).
39. T. Hao and Y. Xu, *J. Colloid Interfacial Sci.* **181**: 581 (1996).
40. F.E. Filisko, *Proc. Int. Conf. ER Fluids*, R. Tao, ed., World Scientific, Singapore, 1992, p. 116.
41. H. Frohlich, *Theory of Dielectrics*. Clarendon Press, Oxford, 1958, p. 80.
42. K.D. Weiss, D.A. Nixon, J.D. Carlson, and A.J. Margida, in *Progress in Electrorheology*, K.O. Havelka and F.E. Filisko, eds., Plenum Press, NY, 1994, p. 207.
43. H. Conrad and Y. Chemn, in *Progress in Electrorheology*, K.O. Havelka and F.E. Filisko, eds., Plenum Press, NY, 1994, p. 55.
44. M. Whittle, W.A. Bullough, D.J. Peel, and R. Firoozian, *Phys. Rev. E* **49**(6): 5249 (1994).
45. K.D. Weiss and J.D. Carlson, *Proc. 3rd Int. Conf. Electrorheological Fluids*, R. Tao, ed., World Scientific, Singapore, 1992, p. 264.
46. C.P. Smyth, *Dielectric Behavior and Structure*. McGraw-Hill, NY, Toronto, London, 1955, p. 201.
47. G.I. Skanavi, *Dielectric Physics*, translated by Y. Chen. High Educational Press, Beijing, 1958, Chaps II and IV.
48. B. Gross, *J. Chem. Phys.* **17**: 866 (1949).
49. T. Hao, A. Kawai, and F. Ikazaki, *Langmuir* **14**: 1256 (1998).
50. W. Wen and K. Lu, *Appl. Phys. Lett.* **68**(8): 1046 (1996).
51. This is a common way to express the yield stress based on the interparticle force. For example, see *J. Rheol.* **41**(3): 769 (1997).
52. T. Hao, Z. Xu, and Y. Xu, *J. Colloid Interfacial Sci.* **190**: 334 (1997).
53. H. Conrad, Y. Li, and Y. Chen, *J. Rheol.* **39**(5): 1041 (1995).
54. T. Hao, H. Yu, and Y. Xu, *J. Colloid Interfacial Sci.* **184**: 542 (1996).
55. T. Hao, A. Kawai, and F. Ikazaki, *Langmuir* **16**: 3058 (2000).
56. T. Hao, *J. Colloid Interfacial Sci.* **206**: 240 (1998).

ELECTORRHEOLOGICAL MATERIALS

FRANK FILISKO
University of Michigan
Ann Arbor, MI

INTRODUCTION

Electrorheological (ER) materials are materials whose rheological properties, flow and deformation behavior in response to a stress, are strong functions of the electric field strength imposed upon them. The materials are typically fluids in the absence of an electric field (although they may be pastes, gels, or elastomers) but under constant shear stress at high enough fields, the materials can solidify into viscoelastic solids. In their solid state, various properties of the solid such as shear and tensile strengths and damping capacity, internal friction, and the ability to adsorb energy under impact are also strong functions of the electric field. Further, all physical and mechanical changes induced by the applied field are virtually instantaneously reversible upon removal of the field; such a material can almost instantaneously be solidified and liquefied by applying and removing the electric field. In the liquid state during flow, these materials exhibit resistivity to flow (or apparent viscosity) which can be increased by hundreds or thousands of times by applying an electric field. The materials can be compounded so that viscosities are near that of water under zero fields but approach infinity at very low shear rates under the

influence of fields of the order of a few thousand volts/mm. In the solid state, these materials can have shear strengths of the order of 5,000–10,000 Nt/m² (1–2 lb/in²) in fields around 5000 volts/mm. In brief, these are materials whose mechanical properties and physical state are determined at any instant by the electric field to which they are exposed.

ER materials are typically dispersions of fine hygroscopic particles in a hydrophobic nonelectrically conducting dispersion medium (1). Particle sizes in the range of 0.1–10 μm are common, although particles much larger have demonstrated ER effectiveness, and certain macromolecules in solution exhibit the effect. For example, materials that work well as the dispersed phase include such diverse materials as corn starch, various clays, silica gel, talcum powder, and various polymers. The fluid phase also may consist of a very wide range of liquids or greases which have the common properties of high electrical resistivity (so that high fields may be applied over the fluids without significant currents) and hydrophobicity. Liquids such as kerosene, mineral oil, toluene and silicone oil work well as do many other fluids. With few very significant exceptions, the vast majority of systems also require that significant amounts of water (10–30%) or other activators be adsorbed onto the particulate phase. This requirement severely limited the potential use of these materials. Dry particulate systems will be discussed later.

Although it is not necessary for an appropriate dispersion to demonstrate ER activity, various other types of additives, called activators, have been reported and are commonly incorporated into the mixtures, including various surfactants, to enhance the effect and to increase the stability of the dispersions. How they work is unclear, but as will be discussed later, they most certainly affect the particulate surface, the dispersing liquid, and the water on the particles.

BACKGROUND

The phenomenon which ultimately became known as electrorheology was first observed in the late 1800s by Duff (2) and others, but it was not until the work of Winslow (3,4) in the 1940s, 1950s, and 1960s that the engineering potential and application of these materials began being fully recognized. The most immediate and obvious applications included torque transmission and damping or vibration control. Upon attempting to use these materials, however, it was soon realized that a seemingly insurmountable obstacle prevented their widespread use; the dispersed phase required significant amounts of water to be adsorbed onto the particles (1,4). Work proceeded to resolve this problem by replacing the water with other substances such as glycerol (5) and silanol (6), but the ER effect was substantially reduced. In effect, it began to be accepted that adsorbed water was necessary. The problem was resolved in the late 1980s (1,7) resulting in a tremendous increase in activity within this field over the next decade and a considerable increase in understanding the physics and chemistry of ER suspensions.

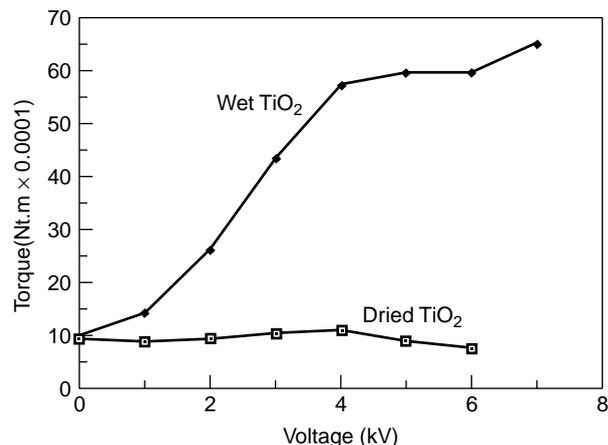


Figure 1. Rheological data for titanium dioxide/paraffin oil (5 g/20 cm³) based ER fluids. The wet TiO₂ is as-received and the dried powder was maintained at 160°C for 5 hours under a liquid N₂ trapped vacuum.

One of the first models was proposed by Winslow (4) and follows from simple observations that particles in an ER fluid align between the electrodes under an electric field in static conditions. He hypothesized that under shear, these chains would become distorted and break but would reform again very rapidly. This could account for the increased stresses but does not address fundamental questions concerning the mechanism of interactions between particles, although a polarization mechanism is mentioned. Klass and Martinek (8) question this model because ER materials show activity in high-frequency ac fields and these chains could not re-form at such speeds. Brooks et al. (9) reported a timescale for fibrillation of around 20 s which is much greater than the submillisecond responses reported (8,10). Other notable discrepancies arise from the fact that chaining is a trivial consequence of polarization of the particles, and therefore it would seem straightforward to increase yield strengths by using particles of higher polarizabilities. However, as illustrated in Fig. 1, for an ERM that contains TiO₂ particulates and has a permittivity of 200, when dried, chaining still occurs, but the fluid loses its ER activity.

Another discrepancy arises from the fact that the strength of particle interactions due to polarization effects in chaining is related to the permittivity difference between the particles and liquid phase (11). This would suggest that metal particles (if appropriately insulated) would result in the strongest interactions between particles. Although this probably occurs, it is not reflected in stronger ER activity. Although there are some reports that metal particle systems are ER active, the strengths are quite low. This and other information suggest that mechanisms other than particle bulk polarizabilities must be involved in a major way in ER activity. An extension of this idea proposes that the particles that interact coulombically flow as clusters, but in static situations, will bridge the electrodes (12,13). Neither of these addresses the basis for particulate interactions on a molecular level.

The particles by themselves and/or in conjunction with the dispersing media must interact with the electric field for the particles to align, provide yield stress, and hold the clusters together. The particles and liquids can interact independently with the field by virtue of their inherent electrical and dielectric properties, and/or the components can act cooperatively by virtue of the electrical double layer (14,15) which develops around colloidal particles in a dispersing liquid, and/or by virtue of interfacial polarization which develops due to mobile charges at the interface of the two materials (16,17). The latter situations are most commonly considered related to ER activity, but it is not clear to what extent these two are interrelated or in fact may be part of the same mechanism. Part of the confusion comes from the fact that although the basis for interfacial polarization is fairly well understood, theories related to electrical double layers [which are well developed for suspensions in electrolytic fluids (18)], are poorly understood (19) for a nonconducting dispersing medium of which ER fluids are an example.

Klass and Martinek (8,10) were the first to involve electrical double layers in their explanation of ER activity. They proposed that the diffuse portion of the double layers would become polarized under the influence of the electric field and that the electrostatic interactions of these distorted double layers require additional energy during flow, especially in concentrated suspensions where the layers overlap. This energy is required due to repulsion of the double layers, so that the particles cannot simply move in a streamline but must have a transverse component that gives rise to the additional dissipation of energy. They do not explicitly discuss the function of the adsorbed water even though without it, there would be no ER effect in these systems; yet, double layers would still exist (19). An interesting observation, based upon the relative permittivities of the systems of particles they used and the relative ER effectiveness, is that the bulk dielectric properties of the dispersed particles does not play an important role. Interfacial and surface properties of the particles are much more important in ER activity. This finding is also supported by others (8,10,20).

Schul'man and Deinega (21) focus more on orientation of the particles and structures that may form in the electric field. They invoke electrical double layers and associate them with a surface-conducting layer on the particles (i.e., water) in a nonconducting fluid where ion exchange with the fluid is presumably negligible. In this case, the mobile charges responsible for the Maxwell-Wagner-Sillars interfacial polarization also involve this water layer. The charge carriers can move along this conductive film under the influence of the electric field and give rise to an MWS polarization. The moisture here serves an essential function. Ion extension into the surrounding medium, the dispersed double layer, may extend to various degrees, depending on among other factors, the degree of conductivity of this medium. In reality, we may speculate that both mechanisms are probably involved in the ER phenomenon. What is certain, however, is that if either of them are correct, then the surface charge conductivity introduced via the water certainly has a dramatic effect on the character of the double layer. This must actually

be the case because the bulk conductivities of the systems increase many orders of magnitude for wet versus dry particles (7), thereby suggesting significant ion transfer to the medium when water is present versus without it.

Uejima (20) presented dielectric measurements that provided the most direct support for these mechanisms. In these studies, he followed loss factor and dielectric constant versus frequency for ER materials composed of cellulose particles and various amounts of adsorbed water. Specifically, he was observing the MWS interfacial dispersion that shifts to higher frequencies as the amount of water on the particles is increased. This is reasonable for this type of polarization (16,22) but has a number of other implications. The first is that the charge carriers involved in this dispersion are characterized by a relaxational spectrum whose characteristic times, temperatures, and presumably distribution depend strongly on the amount and type of water present (23). Whether the MWS dispersion disappears as all the water is removed is an interesting mechanistic question because in these inherently heterogeneous systems, a MWS dispersion should still exist (16,21,23), but charge carriers may be of a different type.

Deneiga and Vinogradov (18) who also made dielectric and rheological measurements, characterized this water layer further by suggesting that upon increasing temperature and field, there is a corresponding rise in ER activity and in the permittivity of the dispersions. However, these quantities peak at some point, and beyond this peak, the bulk electrical conductivity of the system begins to increase dramatically. They suggest that a breakdown of the hydrate layer occurs from both temperature and field and results in lowering of the activation barrier for flow of carriers between particles. A very important point implied here is that the bulk conductivity may not be related to ER activity and the preferred situation is to contain charges on the particles by an infinite activation barrier between particles, if this is possible. This speculation is further supported by the work of Deinega and Vinogradov (18) on, the relationships among ER activity, adsorbed water, and bulk conductivity.

Using various modifications and extensions, virtually all investigations continued to refine the basic concepts of the electrical double layer extending into the liquid phase, a conductive surface layer of water (or other surfactant) on the particles that gives rise to lateral mobility of ions, which are responsible for the classical Maxwell-Wagner-Sillars interfacial polarization. All of them imply the presence of a conductive layer on the particles, most commonly ions in the water, but none explained why the ER effect disappears when the water is removed, even though the double layer and the MWS interfacial polarization still presumably remain.

A major advancement occurred based on reports of particulate systems that produce ER active materials without the need for adsorbed water or any water (7,24). This is critical in resolving the model because either the same mechanism is operating both with and without water or, less likely, that different mechanisms are operating. The implication here is very important because it suggests that the mechanism responsible for ER activity can be an intrinsic

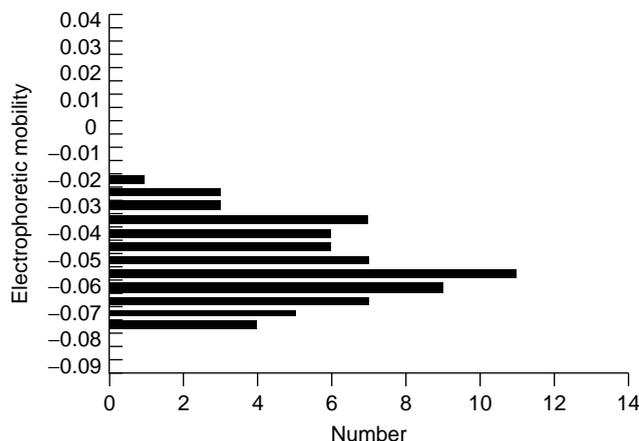


Figure 2. Electrophoretic mobility for an ER material illustrating one that has a net negative EPM. The best ER fluids have curves whose maximum is around zero.

characteristic of the chemistry and physics of materials, not due solely to extrinsic factors such as water.

The models proposed for this activity are similar to those previously discussed but are modified in that the electrical double layer is probably less dominant and the mobile charge carriers are not a consequence of an adsorbed electrolyte. In an article by Block and Kelly (25), more emphasis is put on the particle polarization which is really identical to MWS interfacial polarization. Partially, support of this deemphasis away from the double layer is a consequence of electrophoretic mobility (EPM) measurements on the actual fluids, which indicate that the materials can be very active electrorheologically yet show no significant EPM.

Further, it appears that systems that show significant EPM are less active electrorheologically, (Fig. 2). A suggestion here is that those mechanisms responsible for EPM, fixed surface charges and a diffuse ion layer, are different from those responsible for ER activity and in fact oppose each other to some extent. Permanent fixed surface charges cause attraction or drift of the particles toward one electrode and create an oil or particle-free layer adjacent to the other electrode, thereby giving an apparent viscosity decrease. This also opposes formation of particle-mediated shear transfer between the electrodes which is necessary for ER activity. The mechanism responsible for ER activity, consistent with most others, is the presence of mobile charges (ions or electrons) associated with the particles that can move somewhat freely within the particles but cannot move off of them, that is, a low activation barrier for migration within a particle but an infinite barrier for motion away from the particle.

The explanation for the activity of these dry systems is based essentially on the presence of mobile charge carriers intrinsic to the molecular character or chemistry of the particles. This local mobility of the carriers on the particles is high, but mobility between particles should be very low. In the anhydrous materials of Block and Kelly (1,25), the carriers are presumably electrons because the particles

are semiconductors; in aluminosilicate systems, the charge carriers are ions that are intrinsic to the chemistry of the particles and are located on the surfaces (26). Surface here includes the walls of the extensive interparticulate network of channels and cavities inherent in the morphology of the particles, which can constitute more than 97% of the total surface area (27). An important distinction between the two systems is that the bulk currents in the semiconductor systems are very high (1,25), presumably because electrons can more easily jump or tunnel between the particles and all are available to the outside surface of the particles by standard conduction mechanisms. However, in the zeolite systems, ions on the outer surfaces have a much greater activation barrier to overcome to jump particles, but more important, most are contained within the internal structure of the particles and cannot migrate to the outside surface; yet, they are mobile within the internal labyrinth of channels and cavities afforded by the tremendous porosity inherent in the morphology of the materials. Apparent Maxwell-Wagner-Sillars interfacial dispersions are observed in all dried zeolite systems, as illustrated in Fig. 3. They all occur at very low temperatures.

Anhydrous polyelectrolyte systems are presumed to be ER active by virtue of locally mobile ions that can move within the environment of the chain coils or along a chain (28) but are not free to move easily between chains. It has been shown that ER activity is primarily a function of the pK of various polyelectrolytes (29).

MATERIALS

Although the number of materials that can produce ER active suspensions is almost infinite, it is well understood that for most, the phenomenon is a consequence of an extrinsic component, most commonly adsorbed water (or some other electrolyte) that contains various surfactants added, and has nothing or little to do with the chemistry of the particles. There are, of course properties of the materials that are beneficial in producing better ER materials such as particle porosity, high surface areas, and high affinity for water, but the ER mechanisms are not related to the particle chemistry. Such "wet" or extrinsic systems, most of which were summarized by Block and Kelly (1), were known for many years, and it was as well realized that the water severely limited the potential application of ER technology. Some of the reasons are listed here.

1. Thermal runaway currents, although small, but at the high voltages required cause i^2R heating which drives off some water, which, in turn, increases the current which increases i^2R heating which drives off more water which increases the current, etc., until virtually all the water is off, and the fluid no longer works.
2. Relatively high currents and therefore high powers needed.
3. Limited operating temperature range due to freezing and boiling of the water.
4. Electrolysis

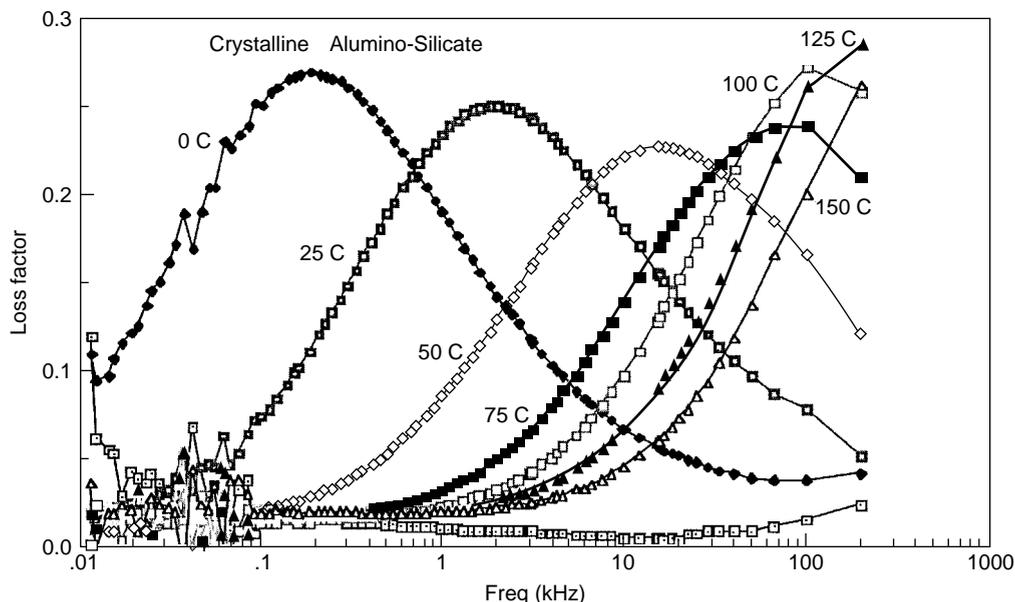


Figure 3. Dielectric dispersion due presumably to MWS interfacial polarization for 4A zeolite.

5. Corrosion of devices containing fluids
6. Instability of fluids with time and operation
7. Irreproducibility of different batches
8. Solid mat formation upon settling due to interparticulate hydrate bond formation.

Other important considerations certainly exist, which will emerge as applications are developed, such as cost, environmental acceptability, raw material availability, settling in low shear applications, plating of particles on one or both electrodes, sealing, effect on pumps, breakdown of particles upon shear of polymeric particles, adsorption of water, and contamination. The items listed before, however, are those associated directly with the presence of adsorbed water.

Because of the discovery of water free or "intrinsic" systems, a number of immediate improvements were realized; many had to do specifically with the water. Some of these are listed here:

1. Thermal runaway eliminated, because no water or other adsorbent is required.
2. Low currents. Currents 10^3 to 10^6 lower as a result of dryness; currents are in the range of microamps/cm², or nanoamps/cm² instead of milliamps/cm² for aluminosilicate particulate systems. Drastic reductions in current do not occur in semiconductor systems.
3. Expanded temperature range. Zeolite-based fluids operate from -60 to 350°C when dispersed in silicone oil.
4. Electrolysis eliminated. Electrolysis does not occur because water is not present and is not needed.
5. Corrosion eliminated. Corrosion does not occur because water is not present.

6. Instability improved. A major cause of instability is loss of water due to operation or heating.
7. Irreproducibility substantially improved. Variability of water is a major source of difficulty in formulating water-based fluids.
8. Solid mat formation and settling substantially reduced.

An important additional consequence of these discoveries implies that mechanisms responsible for ER activity can be associated with the basic chemistry and physics of the particles. Thus, once these mechanisms are understood, materials can be synthesized specifically to optimize these mechanisms and improve ER properties in an intelligent manner.

It is realized that many continue to work on "wet" systems, mostly because it is relatively easy to improve properties to a limited extent. However, we consider that this is a very limited and interim approach because prior attempts to use this method to improve properties substantially in the 1950s and 1960s resulted in virtually complete failure. However it was also realized then that the water made the materials impractical for the most part.

Extrinsic ER (Wet) Systems

Extrinsic ER fluids are suspensions that require adding some substance other than the particles and matrix liquid to make the ER fluid function. This specifically refers to the particulate phase because most solids in suspension do not themselves result in ER active suspensions. Substances that are added to make a suspension ER active are sometimes called activators and may include many additives such as surfactants which are added to stabilize a suspension against settling. The most well known and effective is water, although many others have been reported (1).

Although virtually any particle can be made ER active by adsorbing sufficient water onto it, the actual function of the water is still not known. Theories proposed include that water bridges form that tie the particles together. Another proposes that the high dielectric constant (30) creates a stronger dipolar interaction between particles. Another suggests that water modifies the electrical double layer, and another that it increases the current. Whatever the reason for the effect of adsorbed water, it undoubtedly is the most effective activator. Yet, as explained previously, adsorbed water severely limits the commercial value of the phenomenon.

Intrinsic ER(Dry)Systems

Electrorheological fluids that operate without the need for adsorbed water on the particulate phase can be classified as follows:

1. ionic conductors
2. semiconductors
3. polyelectrolytes
4. solutions

Ionic Conductors. The main particulate systems in this category are alumino silicates, or zeolites. The particles are highly porous and contain numerous cavities and interconnecting channels such that around 97% of the total surface area of the particles is contained within the particles, that is, the walls of the cavities and channels (26). The dimensions of the cavities and channels can be varied by synthetic methods and by varying the aluminum/silicon ratio. Cationic charge carriers arise from the requirement for stoichiometry when some tetravalent Si atoms are replaced by trivalent Al without disrupting the crystal structure. Thus, an Al at the center of a tetrahedron that has oxygens at the vertices can bind to only three of these oxygens, leaving one unbonded and a net negative charge in the structure. This negative charge is balanced by introducing of cations into the system. These cations, however, cannot fit into the closely packed crystal structure and therefore must reside on the surfaces of the cavities and channels. Thus, the cations are present as a consequence of the chemistry of zeolites (not the presence of an electrolyte such as water), and they are mobile because they are on surfaces that are primarily internal and not confined within the crystal structure. Common uses of zeolites are as molecular sieves because they can synthetically control the channel dimensions, and as ion exchange materials due to the presence of unbonded cations that readily exchange with other cations in an aqueous suspension.

The intrinsic ER activity of these materials is associated with the presence of these cations which presumably can move locally under the influence of an electric field. Such materials are susceptible to modification partly by varying the Si/Al ratio, by incorporating atoms other than aluminum, and by varying the types of cations. These materials have been available commercially for years as molecular sieves.

Semiconductors. Most of the work in this area has been performed by Block and associates (24,25) using various polyacene quinone radicals (PAQR) and recently polyaniline. PAQRs are not available commercially but can be prepared by the method described by Pohl (12). The mechanism of activity for these materials is associated with the electronic charge carriers that can move locally under the influence of an electric field. A characteristic of these materials, associated with electron-mediated ER activity, is relatively high bulk current due to the relative ease with which electrons may jump or tunnel between particles.

Although there are many types of semiconductors that can be used to make ER materials, many are ineffective or only a few are effective when dried. The reasons for this may be related to the size of the energy band gaps, the charge mobility, and/or charge concentration, although I know of no studies reported in this regard.

Notable materials that fall into this category are the commercially available "carbonaceous" ER fluids of Bridgestone. The particulates in these fluids are presumably synthesized by the controlled pyrolysis of polymer(s). The most commonly used, although it is unknown what is used in the Bridgestone fluid, is polyacrylonitrile.

Photoconductors represent an interesting group of semiconductors that can be used to make ER materials. In this instance, many fluids that are inactive or weakly active can show much enhanced ER activity when exposed to the correct frequency of light (30,31). Phenothiazine demonstrates this rather dramatically, even when dried. Photoelectrorheological materials (PHERM) represent the clearest proof of the relationship between charge mobility and ER activity because exposing such materials to light produces a tremendous increase in the number of free electronic charge carriers.

Polyelectrolytes. Although many types of polyelectrolytes have been used in ER materials (poly lithium methacrylate is the most well documented), most require adsorbed water to function, presumably to dissociate the cations from the macroions. Treasurer (29) evaluated various polyelectrolytes that are commercially available as ion exchange resins and reported that many function with greatly reduced amounts of water. These were all dried at 120°C under vacuum for 4 days but retain between 0.1 and 2% water. No correlation was found between the ER activity and residual water; however, a strong correlation exists between the dissociation constant pK and ER activity. Another interesting observation was that some systems were ER active at 23°C and 100°C, some were inactive at both temperatures, and some were inactive or weak at 23°C but showed much enhanced activity at 100°C. Materials that have high and low pK s demonstrated activity at both temperatures. Materials that have intermediate pK s showed partial or no activity, and materials that were acidified, that is, contained no cations, showed activity at 100°C but not at 23°C. A common feature of these latter materials was that they all contained quaternary ammoniums, but no reason is given for the relationship to ER activity.

The mechanism for ER activity in these dried materials is presumably due to cations which, even in the absence of an electrolyte, can move locally within the confines of a chain coil in the presence of an electric field but cannot move outside the coil into the surrounding liquid (28). Direct evidence for this has not been obtained, but dielectric dispersions associated with interfacial polarization have been detected and are presumably due to these locally mobile ions.

Systems that contain polymers as the dispersed phase are very popular because the particles are soft and therefore reduce abrasion, because of the relatively low density which will aid the settling problem, and because of the vast body of knowledge on latex suspensions. Additionally, polymers represent an almost infinite range of systems that can be chemically customized for ER materials, once the basic chemical mechanisms for ER activity in "intrinsic" systems are better understood.

Two commercially available ER fluids fall into this category. One is from Nippon Shokubai Co., Ltd., which produces an ER fluid based on sulfonated polystyrene. The particles are prepared in a unique way, so that the sulfonation appears at the surface of the particles. Another commercially available material is made by Bayer-Silicone. This is not strictly a polyelectrolyte but is more correctly referred to as a polymeric electrolyte. The particles are block copolymers of a polyurethane and polyethylene oxide. Solid polyethylene oxide has an interesting capability of dissolving or ionizing small amounts of salts. Presumably, incorporating it into the polyurethane gives this material the capacity of producing ions which, as charge carriers, are considered important in ER activity. Because the material is a good ER fluid, it supports the charge mobility hypothesis.

Solutions. Two of the most common are solutions of poly- γ -benzyl-L-glutamate (PBLG) in various solvents (31) and poly(hexyl isocyanate) (PHIC) in various solvents (32). Difficulties encountered with the PBLG systems include achieving high concentrations before gelling occurs and the better solvents are polar thus resulting in high currents. Nonetheless, these solutions showed very significant increases in viscosity upon applying a field. Further, the effectiveness increased significantly with temperature, the limit is the boiling point of the solvent used.

PHIC systems, on the other hand, are soluble at much greater concentrations and in nonpolar solvents. The ER activities are also significantly higher.

The discovery of ER active solutions represents another very significant advance in the field of ER. One reason is that it would resolve the problem of settling which has remained a major concern in some device designs. These solutions, however, will have unique disadvantages such as the greater toxicity and aggressiveness of the solvents, limited upper operating temperatures due to solvents and thermal degradation of the polymers, and generally higher costs. What is most important, however, is that this discovery emphasizes the enormous versatility in the compositions of ER active materials, as well as the complexity in attempting to ascribe the behavior to a single mechanism.

One ER fluid that may be put into this category is not a suspension of solid particles but a mixture of a high and low viscosity liquid. Both phases are siloxane backbone polymers, but the high viscosity material contains liquid crystallizable side chains that are responsible for its ER activity. Similar to the PHICs in solution, the LC side chains are induced to form nematic structures by the electric field that is reportedly responsible for its ER activity (33–35).

MECHANICAL (RHEOLOGICAL) PROPERTIES OF ER MATERIALS

The rheological behavior of ER materials under the influence of an electric field is commonly characterized by observing their properties during steady-state flow (8,10,21,36). Under these conditions, the flow properties of ERM can be adequately described as Bingham bodies.

ER materials are usually fluids when subjected to unidirectional shearing and under zero field conditions. However, when shearing conditions are maintained constant, the shear stress increases with increasing applied electric field strength. It is commonly reported that the shear stress dependence is proportional to the field squared (4,8,10,21), but many other types of behavior are observed (25,37). According to idealized Bingham behavior (Fig. 4), ER materials are fluids under zero field but are solids under a nonzero field up to a certain critical shear stress (S_c) and liquids at shear stresses above S_c . Although adequate in steady-flow situations where transient or "start up" effects are neglected or unimportant, this model is not applicable when the transient behavior is important or under dynamic loading (i.e., rapid or impact stresses or in damping applications). In these situations, the Bingham model completely overlooks the properties of the materials at stresses less than S_c (38). A more complete description of the behavior of ER materials is illustrated by a plot of stress versus strain, as in Fig. 5.

Under these deformation conditions, ER materials can be described as viscoelastic solids below a certain critical yield stress or yield strain and as viscous liquids at stresses at or above S_c and strains greater than the yield strain. If characterized in this manner, ER materials can be described in terms of their overall rheology as viscoelastic perfectly plastic materials in which S_c and the yield

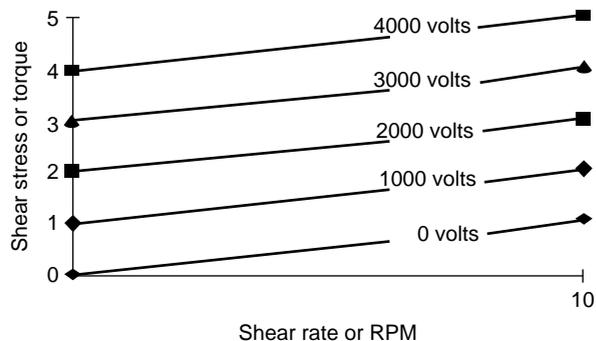


Figure 4. Bingham body illustration of rheological behavior of an ER material. (Oversimplified because nothing is in preyield).

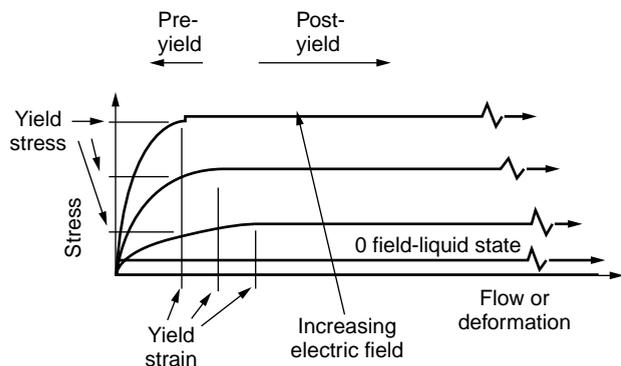


Figure 5. Illustration of more correct rheological behavior of an ER fluid.

strain are strong functions of the electric field strength. The yield stress S_c is highly variable and depends strongly on numerous factors, including the ER material composition. Furthermore, the energy dissipation mechanisms in the preyield region are different from those mechanisms present in steady flow.

In the simplest case, the rheological behavior of ER materials in the preyield region can be characterized by a modulus and a yield stress (S_c), in contrast to the postyield region (i.e., liquid state) in which the material is characterized by an apparent viscosity (η_a). Furthermore, the behavior of the ER material is linear viscoelastic when deformation is restricted to the preyield state and the ER material is characterized by time constants and damping factors that are complex functions of the field strength.

ERM in Steady-State Flow (Postyield Behavior)

In steady-state flow at a shear rate of " $\dot{\gamma}$," ER materials are characterized by an apparent viscosity " η_a " which is defined as

$$\eta_a = [S_c(E) + S_o(\dot{\gamma})]/\dot{\gamma} \quad (1)$$

where S_c is a function of the electric field strength (E) and S_o is a function only of the shear rate and temperature and is material specific.

Two common criteria for evaluating ER materials in flow include (1) the magnitude to which the viscosity can be increased and (2) by what factor it can be increased. The second criterion is more important because it indicates how effective an electric field strength is on the rheology of the material. Regarding the latter point, we can define a fluid effectiveness factor K as

$$K = \eta_a(E)/\eta_o = [(S_o + S_c)/\dot{\gamma}]/(S_o/\dot{\gamma}) = 1 + S_c/S_o \quad (2)$$

where S_c is essentially a constant at a certain field strength but S_o increases continuously with increasing shear rate. Therefore, this suggests that K decreases toward one as the shear rate increases. This is an important first-order relationship for understanding the characteristics of ER materials under flow because it implies that to make a more

effective fluid requires making S_c as large as possible while keeping S_o as small as possible. The second parameter, S_o , is a function of the ER material composition and flow conditions. Thus, ER materials of low or high viscosity can be made by varying the solid concentration or the viscosity of the dispersing liquid. However, though the maximum shear stresses can be increased by making the zero-field materials thicker, the K factor can become small, so that the field-induced change in S_c becomes insignificant. Furthermore, it has been reported that S_o is a much stronger function of concentration than S_c (39), and it appears that S_c is a linear function of concentration (18).

ERM in Oscillatory Shearing (Three Rheological Regions)

The response of ER materials to dynamic loadings can be discussed in terms of three distinct rheological regions: preyield, yield, and postyield regions. In the previous section, discussion was limited to conditions of steady-state flow in which the transient effects of the preyield region were not considered. The preyield region can be effectively studied when oscillatory stresses are applied to the ER material such as may occur in vibration damping, as first reported in detail by Gamota and Filisko (40,41). Under these straining conditions, the amplitude of the shear stress response is a strong function of the applied field strength. However, a limiting shear stress value exists beyond which the shear stress response no longer follows the shape of the shear strain function, but becomes "cutoff" or truncated (40,42). The value of the shear stress at the onset of truncation is a function of the field strength and is also related to S_c . Rheologically, the appearance of the cutoff is an indication that the material is beginning to flow. During an oscillatory shear strain, the ER material may deform as a linear viscoelastic solid over part of the deformation cycle and as a liquid over the other part.

A representative series of stress responses for the ER material when subjected to a sinusoidal shear strain is presented in Fig. 6. Curve a is the applied sinusoidal shear strain of frequency 15 Hz and amplitude 0.25. Curve b is the shear stress response when the material is subjected to a zero-strength electric field. The shear stress response appears sinusoidal, and the phase angle between the applied shear strain and shear stress

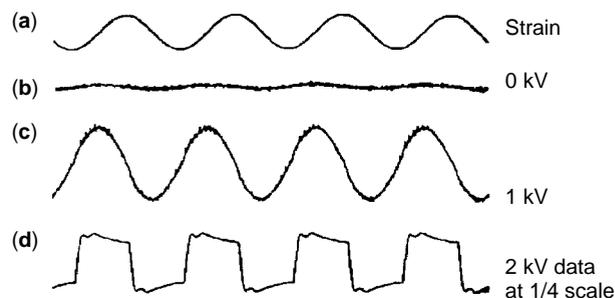


Figure 6. Shear stress response to a constant strain amplitude at various electric fields. Curve a is the strain, b the shear stress where $E = 0$, c the shear stress where $E = 1$ kV/mm, d the shear stress where $E = 2.5$ kV/mm. (D is one-quarter scale.)

response is 90° , suggesting that the material is deforming as a viscous body. Subjecting the ER material to an electric field whose strength is 1.0 kV/mm yields Curve c; the shear stress response amplitude increases, and the phase angle decreases. Thus, when the ER material is subjected to a nonzero electric field, the material behaves as a viscoelastic material. If the strength of the electric field is increased to 2.5 kV/mm , the shear stress response deviates from sinusoidal behavior (Curve d). A nonsinusoidal response suggests that the material is behaving as a nonlinear viscoelastic material. In addition, as the material is subjected to a 2.5-kV/mm field, the fundamental harmonic of the shear stress response increases, and the phase angle between the fundamental harmonic of the shear stress response and the applied shear strain decreases. Thus, as the strength of the applied electric field is increased from 0.0 to 2.5 kV/mm , the ER material transforms from a viscous to a linear viscoelastic to a nonlinear viscoelastic body. Moreover, the energy storing and energy dissipating properties of the ER material are strong functions of the applied electric field strength.

The linear viscoelastic parameters, shear storage modulus (G') and shear loss modulus (G''), are strong functions of the applied electric field, strain amplitude, strain frequency, and material composition (9,18,41). In addition, it was shown (18,41) that the shear storage modulus is a stronger increasing function with increasing electric field strength compared to the shear loss modulus. Furthermore, it is of particular interest to note that ER materials become greater energy storing bodies as they simultaneously become greater energy dissipating bodies.

A second technique for observing the effect of the electric field under cyclic loadings is to observe shear stress—shear strain loops (hysteresis loops) for these materials at a constant strain frequency and amplitude, while varying the strength of the electric field. A sequence of hysteresis loops generated when the ER material is deforming as a linear viscoelastic body is shown in Fig. 7.

As the strength of the electric field increases, both the area within the hysteresis loops and the angle that the major axis of the hysteresis loop makes with the abscissa increase. The hysteresis loops are elliptical which is indicative of a linear viscoelastic response. The viscous component (energy dissipated) is determined by the area within the loop, and the elastic component (stored energy) is determined by the major axis inclination.

The existence of a deformation transition limits the applicability of linear viscoelastic mathematics for quantifying the energy storing and energy dissipating properties of an ER material. The amount of energy dissipated by the ER material during one deformation cycle, irrespective of a linear or nonlinear viscoelastic response, can be obtained by generating a hysteresis loop. The energy dissipated by an ER material is found by calculating the area within the loop. The recorded hysteresis loop for an ER material subjected to a strain of moderate frequency, moderate amplitude, and zero-strength electric field is elliptical, and the major axis of the hysteresis loop is parallel to the abscissa; this response is indicative of a viscous material (Fig. 8a). When the ER material is subjected to a field strength of 1.0 kV/mm , the area within

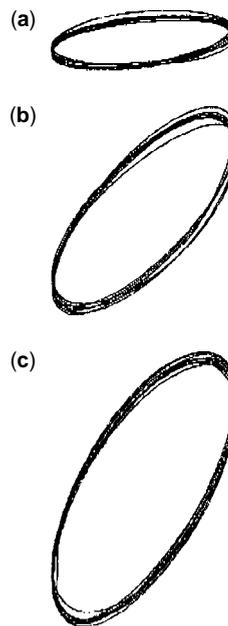


Figure 7. Hysteresis loops for an ER material when subjected to a strain of amplitude of 0.001 radian at 300 Hz . Loop 'a' is under a zero strength field, loop b: $E = 1.0 \text{ kV/mm}$, and loop c: $E = 2.0 \text{ kV/mm}$. 'a' is a hysteresis loop for viscous behavior; b and c represent hysteresis loops indicating viscoelastic behavior where the viscous component of $b < c$ and the elastic component of $b < c$.

the hysteresis loop increases, and the angle between the major axis of the hysteresis loop and the abscissa increases (Fig. 8b). The increased area within the hysteresis loop suggests that the ER material dissipates more energy and the increased angle is related to the energy storing properties of the ER material. Continuing to increase the strength of the electric field yields hysteresis loops that encompass greater areas, suggesting that the material dissipates more energy (Figs. 8c and 8d). However, the loops are no longer elliptical, and thus the ER

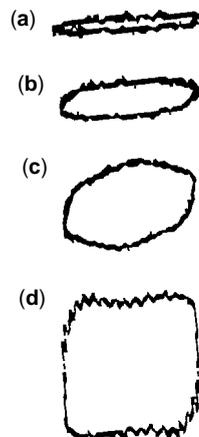


Figure 8. Actual hysteresis loops recorded for an ER material under various electric fields. a: $E = 0$, b: $E = 1 \text{ kV/mm}$, c: $E = 2 \text{ kV/mm}$, d: $E = 3 \text{ kV/mm}$.

material is behaving as a nonlinear viscoelastic material. Moreover, the energy storing properties of the ER material cannot be calculated when nonelliptical hysteresis loops are recorded. Thus, if the amplitude of the imposed impulse is large enough, nonelliptical hysteresis loops are recorded which indicates a nonlinear viscoelastic response. This occurrence suggests that the liquid or flow regime has been encountered.

MECHANICAL MODELS

Presently, the two most serious impediments to the intelligent and innovative design of various types of devices that employ ER materials are the lack of a good mechanical rheological model and corresponding mathematical equations that adequately simulate, at least qualitatively, the behavior of ER materials and the lack of quality engineering design data. The latter point is relative because, as mentioned previously, the composition of materials is virtually infinite and, they will ultimately be customized to optimize desirable and minimize undesirable properties for specific purposes. However, there is sufficient data in the literature to allow estimating reasonable levels of properties and inserting them into appropriate equations. However, an adequate rheological model is presently nonexistent.

In one case, a model by Bullough and Foxon (43) is proposed to simulate only damping and is then only linear viscoelastic so it can be treated mathematically. It is unrealistic in regard to three obvious deletions: first, it incorporates no flow element; second, it incorporates no coulombic damping term; and third, it is completely recoverable. It was not, however, intended to simulate the behavior of ER materials. The only other model by Shul'man (44) suffers from the same problems as that of Bullough and Foxon, but again it was intended only to simulate electric field control of damping.

A current model proposed by Gamota and Filisko (40) shown in Fig. 9 correlates qualitatively with the overall observed behavior of ER materials (i.e., preyield, yield, postyield regions) and has been tested quantitatively for experimentally observed preyield behavior (41).

Element "1" is a dashpot which is essentially field independent but governs the slope of the shear stress versus

shear rate data, that is, the slope of the lines in Fig. 4. Element "2" is a coulombic friction element which is strongly field dependent and controls the magnitude of the ER response under steady shear, that is, the yield stresses or S_c . Elements "3 and 4" combined in parallel form a Voigt element. Both elements are strongly field dependent and are responsible for the transient response and damping behavior under impact or vibration, that is, both the angle of inclination and the areas within the ellipses in Fig. 7. Under low amplitude, it may be only this portion of the model which is deforming. Element "5" is a spring which is strongly field dependent and governs the dynamic response in conjunction with the Voigt element.

Notice that under constant strain rate, elements "3,4, and 5" rapidly reach an equilibrium extension and contribute nothing in steady-state flow. Element "1" simply determines the slope of the stress – strain rate data which is essentially field independent, and element "2" which is field dependent determines the yield stress. However under dynamic loading, at low amplitudes, the block (friction element) may not move, and then only elements "3,4, and 5" determine the material characteristics. In more complex loading situations, all elements may contribute to various degrees to the rheological behavior of the electrorheological materials.

THEORIES OF ER

All current mathematical theories of ER evolve from the assumption that dielectric particles form bridges between electrodes that have a high potential between them. Particles in an electric field are induced by the field to become polarized, that is, become electrically positive on one side and negative on the opposite in perfect analogy to certain particles/objects that can be magnetized, either permanently or temporarily, when exposed to a magnetic field and as such acquire induced north and south magnetic poles. We are all aware that magnets stick together, north pole to south pole, etc., and many can easily be stuck together in progression to form a chain of magnets. In analogy, then, the induced electrical dipolar particles can stick together, positive to negative, etc., to form strings of particles, chain, or columns. Thus, the theories attempt to model mathematically the forces between these electrically induced dipolar particles (11,45).

An extension of this is that at high particle concentrations that are characteristic of ER fluids, typically in the range of 30% solids, particles in an electric field do not form just simple single file chains of particles but aggregates of many particles or columns that can be many hundreds of microns in cross section and therefore can have hundreds of particles across the width of a column (46). This is routinely observed and must be true due to free energy considerations. Simulations that involve monodisperse spheres conclude that the particles in these columns eventually organize into a body-centered tetragonal symmetry which, it is stated, may be the ground state for particle packing (46). This, however, is more of an academic exercise because particles in real ER fluids are neither spheres nor monodisperse and therefore cannot pack in any regular way.

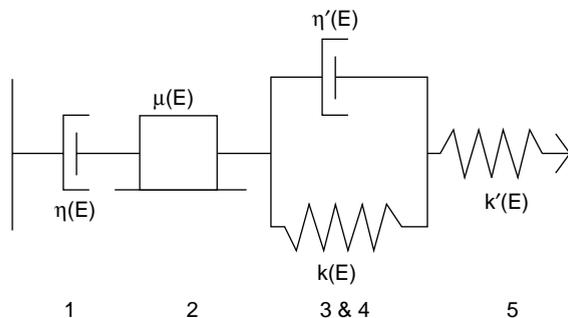


Figure 9. Mechanical model simulating behavior of an ER material.

The third part of the calculations involve determining the force or more specifically the increase in shear stresses due to the electric-field-enhanced interactions between the particles in these columns. It is at this point that some assumptions must be made that are not necessarily rigorous or straightforward. All mathematical theories assume that the increases in shear stresses are due to the columns that break or fail in shear. This has apparently been observed many times but only for very low concentrations of particles where the chains or bridges are essentially composed of single file or a few widths of particles. Until recently, it had not been observed for the very thick columns involved in dispersions of normal concentration. The second assumption implicit in the first is that the columns of particles adhere to the electrodes without failure. Neither of these assumptions has been observed for real ER fluids under flow, but all mathematical theories necessarily assume them. In fact, a significant number of studies conclude that slippage of the particulate structures at the electrodes occurs, not for single chains of particles, but for columns or more highly organized structures that develop (47–49).

Polarization Mechanisms: The Basis of the ER Phenomenon

There are two extremes in the response of a material to a static or dc electric field. One is to establish a steady flow of charge or current between the electrodes, and the other is to cause local separation or segregation of charges leading to asymmetry of charge distribution or polarization which may be most generally meant to describe the response of any material to an electric field where bulk charge movement or a current does not take place. All real materials in an electrical sense lie somewhere between both extremes, and both currents and polarization may have to be considered where applicable. The situation becomes enormously more complicated because of the basic composition of ER materials, that is, a high concentration of particles in a nonconducting oil or liquid. Whereas the liquid phase must be nonconducting or else very large currents will result, the solid phases can be insulators, semiconductors, insulated metals, or insulating materials treated with various activators that typically increase the bulk conductivity of the fluid. Conduction occurs by the net flow of charge which may be accomplished by electrons or ions. Mechanisms of polarization in solids include electronic, ionic, and orientational. A fourth type, interfacial polarization, is a composite of conduction and dielectric mechanisms and of specific value in ER fluids.

Electronic polarization is due to displacement of electron clouds around atoms (16), whereas ionic polarization in solids is due to displacement of positive and negative atoms in ionic crystals. Electronic polarization occurs for all materials but is considered small and not of importance in ER. Ionic polarization can be very large in certain materials, specifically TiO_2 and BaTiO_3 , two materials that have been considered extensively in the development of ER because of their high permittivity. The latter are involved only in materials that contain some ionic bonding, but such materials, exceptions as mentioned, are not common in ER.

A third mechanism involves only solids whose molecular structures have asymmetric charge centers, that is, they contain permanent dipoles. As opposed to electronic and ionic polarization that involve distortion of electron clouds, orientation polarization involves electric-field-induced alignment of these dipoles that are normally random. All three of these mechanisms can occur in homogeneous materials and involve distortion at a molecular or atomic level which recovers after the field is removed, except for electrets.

A fourth type which occurs only in heterogeneous materials or dispersions of particles (17,19,22) involves the presence of a significant number of charge carriers and relatively long range movement of these charge carriers, electrons and/or ions across dimensions of a particle's size. These dispersions involve particles in a nonconducting liquid that have significantly greater permittivities and/or conductivities than the liquid phase. The basis of the phenomenon is that at the interface established between these materials, to satisfy requirements for continuity of displacement and current density across the interface, the charge carriers within the more conductive (particulate) phase must "pile up" at the appropriate interface, thus creating a macroscopic (particle) dipole. At the interface, the carriers experience an activation barrier that hinders their easy movement into the next phase. If infinite, this barrier would prevent any of these charge carriers from crossing it, however, there is always a finite probability that a carrier can penetrate (tunnel) through the barrier, and the result is transport of charge between the particles that results in a flow of charge between the electrodes and thus a current. The mechanism of charge transport between particles via a nonconducting liquid is very complex and not very well understood. However, it is a characteristic of interfacial polarization that a net current can hypothetically occur and does. All fluids, that is, the matrix phase, also have measurable currents as a result of impurity ions or very commonly in ER due to trace amounts of dissolved water. ER fluids that contain semiconducting particles have much greater currents than those that contain ionic carriers due to the very small size of the electrons that allow them to jump more easily between the particles or penetrate the barrier between the particles than the much larger ions.

It is generally felt the interfacial polarization is the polarization mechanism of most importance in ER and as a consequence, particle interactions can result from electrostatic interactions between particles due to the permittivity differences (dielectric properties) and due to polarization of particles as a result of the migration of unbound charge carriers to the interface.

An important distinguishing characteristic of these different polarization mechanisms involves their characteristic times or how fast the "dipole" can respond to an instantaneous field impulse. As illustrated in Fig. 10, electronic and ionic mechanisms respond in timescales of 10^{-16} and 10^{-12} /s, respectively, whereas orientation mechanisms can respond in timescales that range from very slow to about 10^{-6} s. For instance, electrets are very slow in recovering to a random state. This wide range of many orders of magnitude depends on temperature, polar group, and

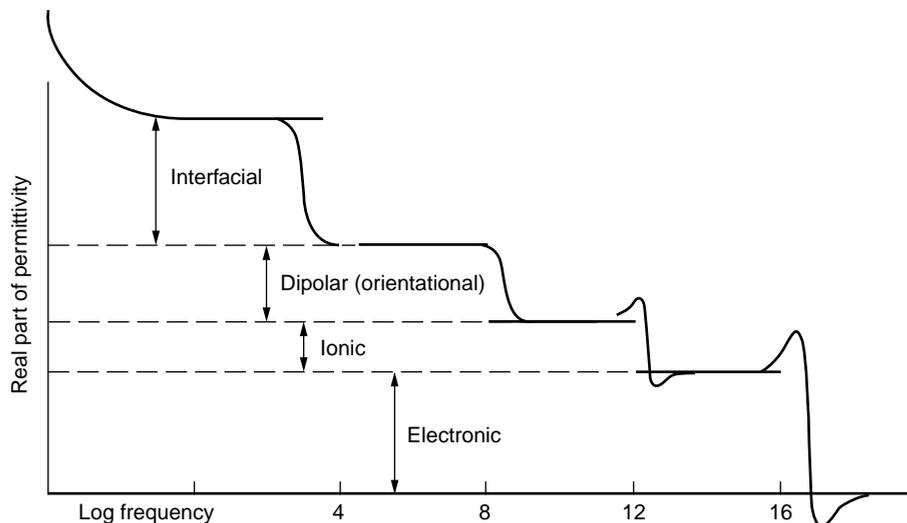


Figure 10. Dielectric constant vs. log. frequency of imposed electric field. The relative positions of the various dispersions are indicated.

physical state of the solid because it involves physical rotation of the polar group under the constraints imposed upon it by its environment, such as crystal structure. Interfacial mechanisms are also relatively slow, and times range from many seconds to around 10^{-4} seconds. Whereas orientational polarization is related to the characteristic speed of dipolar rotation, interfacial polarization involves the drift mobility of charge carries through/on the solid phase. Thus, the mobility of electrons on metals is much greater than those on semiconductors which is typically greater than the mobility of ions on ionic conductors, and the characteristic times go in the reverse direction. The difference of mobilities is reflected in the resistivities of these materials.

Now, as opposed to electronic and ionic polarization which are resonant mechanisms, orientational and interfacial polarizations are relaxational; an identifying characteristic is a strong temperature dependence of the dielectric dispersion (see Fig. 3).

To repeat, electronic, ionic, and orientational polarizations involve electric-field-induced systematic disturbances on atomic or molecular dimensional scales and can occur in all materials, including single-phase materials (although only certain materials have ionic or orientational mechanisms). These determine the classic dielectric parameters of these materials or permittivity (complex permittivity to be correct). The interfacial mechanism, however, occurs only in heterogeneous materials or dispersions such as ER fluids and responds to an electric field in a way characterized by the permittivities and conductivities of both phases. Thus, it is not appropriate to describe an ER fluid (or any suspension) as strictly a dielectric or a conductor. The relative importance of the mechanisms that characterize the way they respond to an electric field reduces to the existence and magnitude of each mechanism and its characteristic time. In a high-frequency imposed field, conductive and possibly orientational (dielectric) mechanisms are prevented from operating, and only ionic and electronic dielectric mechanisms (such as in TiO_2 and BaTiO_3) will be operative. However, at lower frequencies or in a constant dc field, all appropriate mechanisms

operate and thus produce presumably the strongest force of interaction between the particles. A “space charge polarizability” can be defined in analogy to electronic, ionic, and orientational polarizabilities (16) so that this mechanism can be discussed along with the other three true dielectric mechanisms. The current state of thinking, then, is that conduction and dielectric mechanisms are important in the ER phenomenon and both are involved in the Maxwell–Wagner–Sillars (MWS) interfacial effect. Only dielectric mechanisms are operable at higher frequencies, whereas both conduction and dielectric mechanisms are operable at low frequencies and dc, thus giving an overall greater net polarizability as the sum of all possible mechanisms (48).

There are numerous other consequences of conduction and possible mechanisms, including conduction through the particle, conduction around the surface of the particle, conduction through the matrix fluid, situations involving the electrical double layer, polarization layers developing on the outside of the particles due to mobile charge carriers in the oil, greater conduction and different mechanisms at the point of contact of the particles due to the highly intensified fields, field-enhanced dissociation of the matrix phase, and mechanisms of charge transfer through non-conducting liquids. The resulting nonlinear conduction in which the current increases with field at a greater than linear rate has not been extensively studied (51–53) and is still poorly understood. Nonetheless, it is extremely important in electrorheology because it is the major barrier to the use of ER fluids in many applications at elevated temperatures, such as automotive shock absorbers, and is important in understanding the basic mechanisms involved in ER via the conduction theory. More in-depth discussions of conduction mechanisms in nonaqueous media are available from Morrison (54) and Anderson (55).

The situation however is even more complex because of numerous dramatic exceptions to these models. Nonetheless, these mechanisms must somehow be involved in a major way in the response of these materials to a field. Mathematically, the situation is impossible; very severe assumptions are required for any type of solution due

among other things to the high concentrations of particles (and mutual interactions), charge distributions on the particles, irregular particle shapes and sizes, and impurities.

Description of Alternative Model

It has been shown recently (47) that, upon application of an electric field, ER fluids under shear rapidly organize or regiment themselves into numerous tightly packed lamellar formations that are more or less parallel and periodic in relationship to each other. Tightly packed means that the particles are crowded together as closely as possible, if not even compressed together under the influence of the E field, consistent with the irregular size and shape of the particles. While under zero shear conditions, the particles agglomerate into columns under the influence of the E field, and under shear, the particles pack into continuous lamellar structures (i.e., walls) more or less equally spaced and of similar thickness. The relative arrangements and characteristics of these structures depend in a way not yet well determined on the magnitude of the electric field, the shear rate, the particle concentration, the time of shear, and the shear profile.

A major consequence of the tight packing of the particles into these structures is that it results in cooperative strengthening of the structure to failure under shear in the long dimension of the structure because the area of the shear plane increases with lamellar length and thickness and the lack of an easy path for slip planes to develop within the structures due to the irregular packing and irregular size and shapes of the particles.

Thus, it is hypothesized that the lamellar structures remain intact under flow and the most probable path of slippage under flow is at the interface between the particle structures and the electrodes. The structures themselves do not break or shear. Because polarization forces are responsible for holding the structures together, if the structures do not fail, then the forces are irrelevant to postyield behavior, as long as they are strong enough to maintain the structure. Postyield behavior is determined primarily by the mechanics of slippage between the ends of the structures and the adjacent electrode. Further, adjacent lamellar structures may adhere to opposite electrodes, and an additional mechanism of energy dissipation is provided when the matrix fluid is sheared between the lamellar structures. In this case, the slip planes would not be parallel to the electrodes but perpendicular to them and would more reflect the rheological properties of the matrix fluid sheared between them. Because shear between adjacent structures would depend on the gap between these structures as well as on the total number (or particle concentration), if the gaps were large due either to low particle concentration or to structure consolidation, the effect of the lateral shear between the lamellae would be reduced. In any case, it is suspected that shear between the ends of the lamellae and the electrodes is the more dominant source of energy dissipation in ER fluids under field and the source of the yield stress.

APPLICATIONS

Published applications involving the ER effect are numerous and diverse. Many are ingenious, and many are unrealistic in the sense that inventors assign unrealistic or uncharacteristic properties to the fluids. However, the more realistic of the applications are those that recognize the unique characteristics of the fluids and the phenomenon and seek to exploit those characteristics to improve the performance of routine functions, more efficiently, more reliably, in a smaller space, improved operation, better control, less precise tolerances, or to perform functions that cannot be performed any other way. The more unique of these characteristics, as mentioned, are that ER materials change their rheological properties rapidly and reversibly in response to a very low power electric field (less than milliwatts) and can be directly connected to computer control. This technology allows a reduction in the complexity of devices that enhances reliability and subsequently reduces size, weight, and cost.

The more important areas and those which have received the most attention and effort may be categorized as (1) those that involve an electrically variable coupling in "trapped fluid" devices such as clutches, fluid brakes, shock absorbers, or vibration control or isolation devices in general; and (2) those that involve a smart working medium in fluid power circuits where the fluid is pumped at elevated pressures through valves that control the pressure and volumetric flow rate to various actuator devices. The latter covers a wide range of uses from micromachines to heavy machinery, aircraft flight controls, and robotic type devices.

Damping Devices

Damping devices employing ER fluids can broadly be categorized as flow mode, shear mode, or mixed mode. Flow mode devices employ a pseudo-Poiseuille or pressure flow of ER materials, and the electrified flow boundaries are stationary with respect to the flow. Shear mode devices employ a pseudo-Couette flow of ER material, and one electrified flow boundary moves with respect to the flow. Mixed mode devices employ a combination of both shearing and channel flow. Fig. 11 presents three dampers employing the flow mode, mixed mode, and shear mode in ER dampers. In these devices, control of the electric field across the ER flow modulates the force that opposes the motion of the plunger.

Every damping device is simply a method for converting mechanical energy into heat energy. In a flow mode device, ER control is afforded only by the small amount of material in the orifice or valve. The rest of the fluid is not under ER control. This imposes severe requirements of shear rate and heating on the fluid as it passes through the valve. Although simple in nature and many devices incorporate this mode of control, it does not make optimum use of the properties of ER fluids. A shear mode damping device, however, exposes a very large amount of the fluid to shear and thereby makes optimum use of ER control. Such devices are considerably more complex than flow

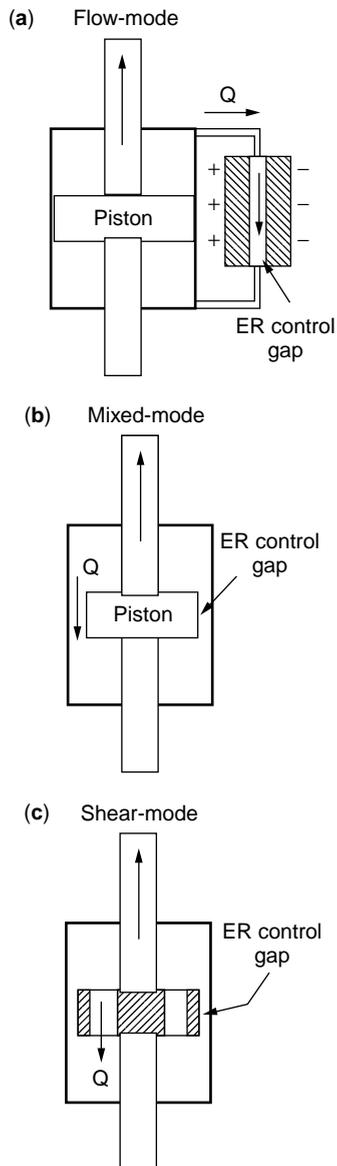


Figure 11. Illustration of three operating modes of an ER damper.

mode devices and involve a series of concentric cylinders that may displace longitudinally or rotate or stacks of discs that have alternate polarity and are fixed so that the fluid is sheared between alternate discs. The latter devices also require a method to transform linear to rotary motion. Although impractical, these devices use the ER phenomenon optimally by exposing the bulk of the fluid to shear flow and electric field control. Mixed mode devices are designed to incorporate both shear and flow modes of control.

Actual devices may be categorized according to the severity of the requirements that they impose on the ER fluid. Vibration isolation devices are meant to minimize or eliminate the transfer of sporadic or systematic vibrations to objects that are adversely affected by them. Examples include isolating vibration inherent in an internal combustion engine from the vehicle for passenger comfort

and safety, isolating vibration inherent in a helicopter from sensitive electronic equipment as well as weapons, isolating vibrations from an earthquake to prevent them from destroying a large structure such as a building or bridge, or protecting sensitive equipment on a ship, for instance, from vibration damage due to the impact of a bomb or torpedo. The most well known of such devices are used for automotive engine mounts. Such devices essentially use an ER fluid to control the flow rate of the active or ER fluid between two chambers, separated by an ER valve which is just a device across which a field can be established. Because of additional considerations, including simplicity in design, packaging, and because such devices are not generally exposed to very severe service, the most common are flow mode devices.

Hydraulic damping devices also transform mechanical energy into heat energy, but as opposed to isolation devices, are involved with much greater displacements or oscillations, have more severe requirements, and correspondingly expose the ER fluid to more severe conditions of shear rates and temperatures and place more severe requirements on the fluids. The most research specific application involves automobile suspension dampers or shock absorbers that have considerably greater flow rates and shear rates and expose the fluids to considerably higher temperatures. Such devices would be designed primarily as shear mode ER dampers; however, within the constraints of packaging and cost, the actual devices are either primarily flow mode and mixed mode devices. Again there are numerous ingenious devices in the patent literature, and considerable research is conducted on these devices primarily in Japan; however, at present no such devices have been commercialized. Although it is generally felt that the "strengths" of ER fluids are more than adequate for this application, a major reason, although not the only one, is the lack of an acceptable ER fluid. Of the currently available ER fluids, all show unacceptable levels of current at the operating temperatures of automobile shock absorbers. Cooling of the devices is unacceptable and impractical, and therefore, fluids that have lower current at high temperatures must be developed.

Torque Transmission Devices

One of the more important areas of application of ER fluids include transferring and controlling torque transfer from a power source or engine to any number of devices (Fig. 12). This is currently performed by torque converters and friction clutches in automobiles and various other types of magnetic or centrifugal clutches for less severe applications such as automobile air conditioners. Such devices present the potential of very simple devices that have direct computer control. These devices present unique problems in that in the uncoupled state or under zero field, fluids must have very low viscosity to minimize drag. However, upon increasing the electric field, the torque transferred is continuously increased thus controlling the rate of acceleration until final speed is attained where lockup or solidification of the fluid occurs. At this point, the device is locked up and results in no viscous losses in the device for the most efficient operation. The maximum amount of

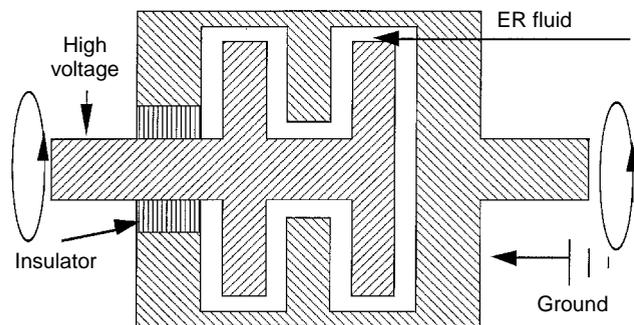


Figure 12. Illustration of an ER clutch that has two plates or rotors.

torque that can be transferred is determined by the area of the shearing surface within the device as well as the apparent yield strength of the fluid. Thus, complex designs that range from series of concentric cylinders to stacks of parallel discs have been patented and are potentially capable of transferring sufficient torque for an automobile. However, at present, the sizes of the devices needed to reach the necessary torque levels do not differ substantially from current devices because of the current levels of yield strengths of available ER fluids. In effect the advantages of using ER clutches to replace conventional ones is not sufficiently cost-effective, primarily because of the fluids, to initiate the use of such devices.

Control of Hydraulic Circuits

One of the more intriguing uses of ER technology and that which potentially can have the most impact in terms of new technology is control in hydraulic systems or circuits. In analogy to electronic devices where a voltage is the driving force for electrons through wires and transistors or electron tubes are gates or valves for controlling the flow of electrons, in hydraulic devices, pressure is the driving force for fluid flow through tubes, and magnetic solenoids are gates or valves to control the flow. Despite the fact that transistors are limited in the currents and voltages that they can handle, their size and speed and low power requirements have resulted in the electronic revolution in solid-state devices, including integrated circuits and of course computers. Hydraulic circuits containing ER fluids, as opposed to large bulky solenoids that have high power requirements, can be controlled with ER valves that consist of little more than electrodes on either side of a hose that carries a fluid. Compared to solenoids, ER valves could modulate flows continuously from full open to full shut, would be much smaller, much faster, require much less power, and could be controlled directly by a computer. The primary control parameters in hydraulic circuits are pressure and volumetric flow rate. In this regard, ER controlled hydraulic circuits are poor compared to solenoid controlled circuits, and a primary limiting factor is the strengths, yield stresses, of the ER fluids. If an ER valve in a hydraulic channel circuit could be imagined as in Fig. 13, at a given flow rate, an increased pressure drop could be controlled by increasing the lengths of the plates. However, there must

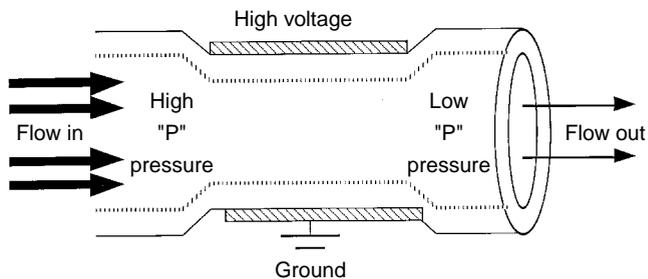


Figure 13. Illustration of an ER valve in which the electrodes on either side of a tube through which an ER active fluid is flowing can be used to regulate the pressure drop across the tube.

be a corresponding drop in volumetric flow rate, when other dimensions are maintained. The volumetric flow rate could be increased by increasing the cross-sectional area of the channel; however, at some point, a practical limit on the length and width of the channel is reached. Long channel lengths are attained in compact spaces by using some ingenuity in the design of devices such as spiraling channels or stacks of plates in which the fluid flows back and forth through progressive layers of plates.

BIBLIOGRAPHY

1. H. Block and J.P. Kelly, Electrorheology. *J. Phys. D.* **21**, 1661–1677 (1988).
2. A.W. Duff, *Phys. Rev.*, **4**, 23 (1896).
3. W.M. Winslow, Methods and means for transmitting electrical impulses into mechanical force. U.S. Pat. 2,417,850, (1947).
4. W.M. Winslow, Induced fibrillation of suspensions. *J. Appl. Phys.* **20**, 1137–1140 (1949).
5. W.M. Winslow, Field response force transmitting compositions. U.S. Pat. 3,047,507 (1962).
6. D.L. Klass, and T.W. Martinek, Preparation of silica for use in fluid responsive compositions. U.S. Pat. 3,250,726 (1966).
7. F.E. Filisko and L.H. Radzilowski, Intrinsic mechanism for activity of alumino-silicate based electrorheological fluids. *J. Rheol.* **34**(4), 539–552 (1990).
8. D.L. Klass and T.W. Martinek, Electroviscous fluids. I. Rheological properties. *J. Appl. Phys.* **38** (1) 67–74 (1967).
9. D. Brooks, J. Goodwin, C. Hjelm, L. Marshall, and C. Zukoski, Viscoelastic studies on an electrorheological fluid. *Colloids Surf.* **18**, 293 (1986).
10. D.L. Klass and T.W. Martinek, Electroviscous fluids. II: Electrical properties. *J. Appl. Phys.* **38** (1), 75–80 (1967).
11. A.P. Gast and C.F. Zukoski, Electrorheological fluids as colloidal suspensions. *Adv. Colloid Interface Sci.* **30**, 153 (1989).
12. H.A. Pohl, The motion and precipitation of suspensions in divergent electric fields. *J. Appl. Phys.* **22** (7), 869–871 (1951).
13. A. Voet, Dielectrics and rheology of non-aqueous dispersions. *J. Phys. Colloid Chem.* **51**, 1037–1063 (1947).
14. J.T. Davies and E.K. Rideal, *Interfacial Phenomena*, Chapter 2. Academic Press, New York, (1961).
15. J. Lyklema, Interfacial chemistry of disperse systems. *J. Mater. Ed.* **7** (2), 211 (1985).
16. A.R. Von Hippel, *Dielectric and Waves*, pp. 228–234. Wiley, New York, 1954.

17. S.S. Dukhin and V.N. Shilov, *Dielectric Phenomena and the Double Layer in Disperse Systems and Polyelectrolyte* (translated from Russian by D. Lederman), Keter Publishing House, Jerusalem, 1974.
18. Y.F. Deinega and G.V. Vinogradov, Electric fields in the rheology of disperse systems. *Rheol. Acta* **23**, 636–651 (1984).
19. A. Kitahara, Nonaqueous systems. In *Electrical Phenomena at Interfaces: Fundamentals, Measurements and Applications* (A. Kitahara and A. Watanabe, eds.) pp. 119–143. Dekker, New York, 1984.
20. H. Uejima, Dielectric mechanism and rheological properties of electro-fluids, *Jpn. J. Appl. Phys.* **11**(3), 319–326 (1972).
21. Z.P. Schul'man, Y.F. Deinega, R.G. Gorodkin, and A.D. Matsepuro, Some aspects of electrorheology. *Prog. Heat Mass Transfer* **4**, 109–125 (1971).
22. J.C. Maxwell, *Electricity and Magnetism*, Vol. 1. Clarendon Press, Oxford, UK, 1892.
23. P. Hedvig, *Dielectric Spectroscopy of polymers*, pp. 282–296. Wiley, New York, 1977.
24. H. Block, and J.P. Kelly, *Proc. Inst. Electr. Eng. Colloq.* **14**(1) (1985).
25. H. Block, J.P. Kelly, A. Qin, and T. Watson, Materials and mechanisms in electrorheology. *Langmuir* **6**(1); 6–14 (1990).
26. S.W. Breck, *Zeolite Molecular Sieves*, Wiley, New York, 1974.
27. R.M. Barrier, *Br. Chem. Eng.* **5**(1) (1959).
28. F. Oosawa, *Polyelectrolytes*, Dekker, New York, 1971.
29. U. Treasurer, L.H. Radzilowski, and F.E. Filisko, Polyelectrolytes as inclusions in water free electrorheological materials: Chemical characteristics. *J. Rheol. (N.Y.)* **35**(4) (1991).
30. H.E. Clark, Electroviscous recording. U.S. Pat. 3,270,637, (1966).
31. F.E. Filisko, Materials aspects of ER fluids. In *Electrorheological Fluids: A Research Needs Assessment*, Report DOE/ER/30172. Department of Energy, Washington, DC, (1993).
32. I.-K. Yang and A.D. Shine, Electrorheology of Poly(n-hexyl isocyanate) solutions. *Soc. Rheol. Meet.*, Rochester, NY, October 20–24, 1991 (1991).
33. A. Inoue and S. Maniwa, *J. Appl. Polym. Sci.*, **55**, 113 (1995).
34. A. Inoue and S. Maniwa, *J. Appl. Polym. Sci.*, **59**, 797 (1996).
35. A. Inoue and S. Maniwa, *J. Appl. Polym. Sci.*, **64**, 1313 (1997).
36. H. Conrad, M. Fisher, and A.F. Sprecher, Characterization of the structure of a model electrorheological fluid employing stereology. *Electrorheol. Fluids, Proc. 2nd Int. Conf. ER Fluids*, Raleigh, NC, 1989 (1990).
37. N. Sugimoto, Winslow effect in ion exchange-resin dispersions. *Bull. JSME* **20**, 1476 (1977).
38. A.F. Sprecher, J.D. Carlson, and H. Conrad, Electrorheology at small strains and strain rates of suspensions of silica particles in silicone oil. *Mater. Sci. Eng.* **95**, 187–197 (1987).
39. H.L. Frisch and R. Simha, The viscosity of colloidal suspensions and macromolecular solutions. In *Rheology: Theory and Applications* (F.R. Eirich, ed.), Vol. 1, Chapter 14. Academic Press, New York, 1956.
40. D.R. Gamota and F.E. Filisko, Dynamic mechanical studies of electrorheological materials: Moderate frequencies. *J. Rheol.* **35**(3), 399–426 (1991).
41. D.R. Gamota and F.E. Filisko, High frequency dynamic mechanical study of an aluminosilicate electrorheological material. *J. Rheol.* **35**(7), 1411–1426 (1991).
42. R. Stanway, J. Sproston, and R. Firoozian, Identification of damping law on an electrorheological fluid: A sequential fitting approach. *J. Dyn. Sys., Meas., Control* **111**, 91–95 (1989).
43. W.A. Bullough and M.B. Foxon, A proportionate Coulomb and viscously damped isolation system. *J. Sound Vibr.* **56**(1), 35–44 (1978).
44. Z.P. Schul'man, B.M. Khusid, E.V. Korobkov, and E.P. Khizhinsky, Damping of mechanical-system oscillations by a non-Newtonian fluid with electric-field dependent parameters. *J. Non-Newtonian Fluid Mech.* **25**, 329–346 (1987).
45. D.J. Klingenberg and C.F. Zukoski, Studies on the steady shear behavior of ER suspensions. *Langmuir* **6**, 15–24 (1990).
46. R. Tao, Electric field induced phase transition in ER fluids. *Phys. Rev. E* **47**, 423 (1993).
47. S. Henley and F.E. Filisko, Flow profiles for electrorheological suspensions: An alternate model for ER activity. *J. Rheol. (N.Y.)* **43**, 5 (1999).
48. T.C. Jordan, M.T. Shaw, and T.C.B. McLeish, Viscoelastic response of ER fluids. 2. Field strength and strain dependence. *J. Rheol. (N.Y.)* **36**; 441 (1992).
49. D.J. Klingenberg, F. van Swol, and C.F. Zukoski, Small shear rate response of ER suspensions. 1. Simulation in the point dipole limit. *J. Chem. Phys.* **94**, 6160–6169 (1991).
50. C.W. Wu and H. Conrad, Dielectric and conduction effects in ohmic ER fluids. *J. Phys. D* **30**, 2634–2642 (1997).
51. P.J. Rankin and D.J. Klingenberg, The electrorheology of barium titanate suspensions. *J. Rheol. (N.Y.)* **42**(3), 639–656 (1998).
52. P. Atten, J.-N. Foulc, and N. Felici, A conduction model of the ER effect. *Int. J. Mod. Phys.* **B8**, 2731 (1994).
53. J.-N. Foulc, P. Atten, and N. Felici, Macroscopic model of interaction between particles in ER fluids. *J. Electrostat.* **33**, 103 (1994).
54. I.D. Morrison, Electrical charges in nonaqueous media. *Colloids Surf. A* **71**, 1–37 (1993).
55. R.A. Anderson, in *Electrorheological Fluids: Mechanisms, Properties, Structure, Technology, and Applications* (R. Tao, ed.). World Scientific, Singapore, 1992.

GENERAL REFERENCES

- W.A. Bullough, ed., *Electrorheological Fluids, Magnetorheological Suspensions and Associated Applications*, World Scientific, Singapore, 1996.
- Electrorheological Fluids: A Research Needs Assessment*, DOE/ER/30172. Department of Energy, Washington, DC, 1993.
- F.E. Filisko, ed., *Progress in Electrorheology*, Plenum, New York, 1995.
- M. Nakano and K. Koyama, eds., *Electrorheological Fluids, Magnetorheological Suspensions and their Applications*, World Scientific, Singapore, 1998.
- R. Tao, ed., *Electrorheological Fluids: Mechanisms, Properties, Structure, Technology, and Applications*, World Scientific, Singapore, 1992.
- R. Tao, ed., *Electrorheological Fluids and Magneto-rheological Suspensions*, World Scientific, Singapore, 2000.
- R. Tao and G.D. Roy, eds., *Electrorheological Fluids: Mechanisms, Properties, Technology, and Applications*, World Scientific, Singapore, 1994.

ENVIRONMENTAL AND PEOPLE APPLICATIONS

HIROAKI YANAGIDA
 University of Tokyo
 Mutuno, Atsuta-ku, Nagoya, Japan

INTRODUCTION

There are two major centers of R&D on intelligent/smart materials in Japan. One is an academic assembly called the Forum for Intelligent Materials and the other is the industrial consortium known as Ken-materials Research Consortium. This article describes only the activities in the Forum. The activities of the consortium are described in an other article. The groundwork for R&D on intelligent materials was completed in 1999, with the Frontier Ceramics Project sponsored by the Science and Technology Agency of the Japanese government.

FORUM FOR INTELLIGENT MATERIALS

The Forum for Intelligent Materials was begun in 1990 to open the field of materials science to interdisciplinary academic research. The concept of intelligent materials was proposed to the Agency in November 1989, which led to the Forum being established. These requirement for intelligent materials was that they have functions such as sensor, processors, and actuators for feedback and/or feed forward control systems within the material itself. Several international workshops were held through the efforts of the Forum in Japan. The first was in Tsukuba in 1989, the second in 1992 in Oiso, and the third in Makuhan in 1998. Recently, a large seminar was held on January 14, 2000, in Tokyo. The topics during the proceeding covered the state-of-the-art of intelligent materials and perspectives on their future. Included were discussions of intelligent biomaterials, intelligent fibers, biomedical applications of intelligent surfaces, taste sensors by intelligent materials, and intelligent ceramics materials (1).

To open academic activities to the interdisciplinary interaction, with the Forum was organized to consist of members from various fields such as organic polymers, ceramics, biochemistry, metallurgy, electronics medical sciences, and pharmaceuticals. The R&D of intelligent materials was directed toward the improvement and advancement of human safety and welfare, the environment, and energy savings and resources.

Introduced were some typical intelligent materials as follows (2):

1. The FeRh alloy developed by Otani, Yoshimura, and Hatakeyama as a temperature-sensitive magnetic material for optothermo magnetic motors. The material has interesting properties such as the magnetization increases with temperature up to the transition temperature, as shown in Fig. 1. The temperature increase is very sharp, and the transition temperature can be modified from -100 to $+280^{\circ}\text{C}$ by the addition of small amounts of alien

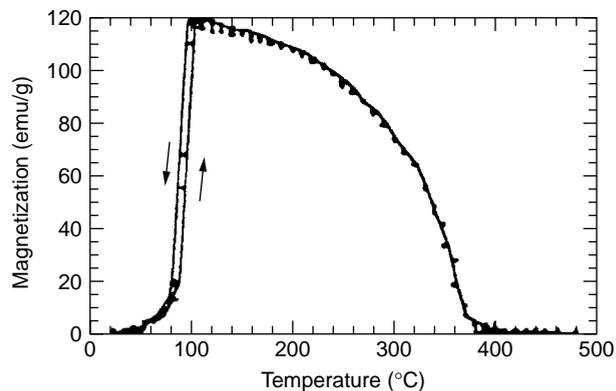


Figure 1. Temperature dependence of magnetization of FeRh.

elements. The mechanical stress or magnetic field can change the transition temperature. These characteristics are interesting examples of temperature sensors and actuators reflected by magnetic valves, magnetic fluid control switches and magnetic motors driven by temperature change in the form of pulsed light.

2. A chemomechanical system with polymers as polypyrrole(PPy) film developed by Okuzaki and Kunugi. The film shrinks or elongates with chemical environmental change. Its use may be to stimulate muscles in a living body.
3. An autonomous vibrating polymer gel with nonlinear reaction. In a living body there are some vibrational motions that are rhythmic. To these intelligent materials achieve mimesis of a living organ. The examples presented by Yoshida, Yamaguchi, and Ichijo are the Belousov-Zhabotinsky reaction of PIPAAm, poly-*N*-isopropylacrylamid, around its transition temperature between the reversible hydration-dehydration process.
4. The compatibility between strengthening and the capability of damage self-monitoring in CFGFRP, carbon-fiber and glass-fiber reinforced plastic bars developed by Yanagida et al.

FRONTIER CERAMICS PROJECT

Before 1999 there was a national Frontier Ceramics Project (FroC), instituted in 1994, that sought to analyze and design two-dimensional structures (3) and to observe interactions arising from the structures related to intelligent functions. The interfaces were found to promote novel nonlinear interactions between the two materials involving crystal axes orientations. The nonlinear interactions, were then categorized as shown in Fig. 2. The two-dimensional structures or their interfaces were compared according to whether the materials were the same or different, the structure was closed or open, and the transport phenomenon was taking place across or along the structure. This system gave rise to $2 \times 2 \times 2 = 8$ different interfaces.

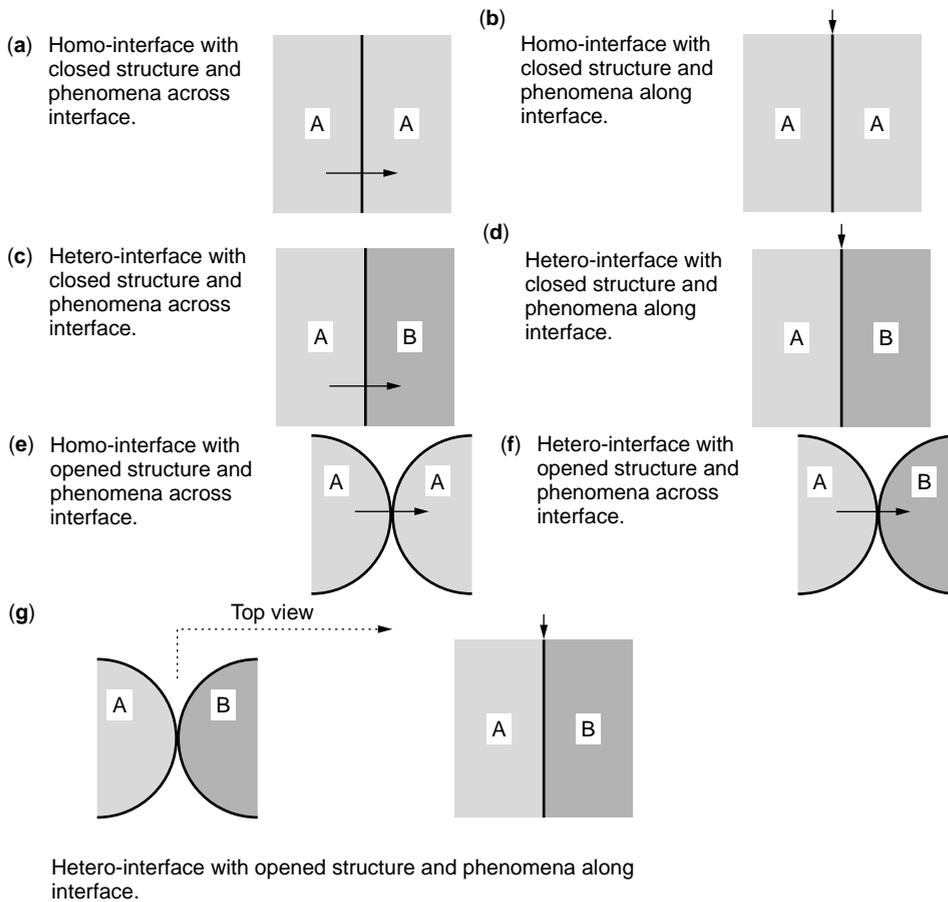


Figure 2. Classification of two-layer structures.

To discuss this study, we will denote the interfaces between the same materials as S , and the interfaces between different ones D . Closed interfaces are suffixed as “cl,” and open ones as “op.” Transport phenomena taking place across the interfaces are indicated as or \parallel (parallel) accordingly. The eight interfaces are as follows:

1. Closed interfaces between the same materials where transport phenomena occur: S_{cl+} .
2. Closed interfaces between the same materials where transport phenomena occur: $S_{cl\parallel}$.
3. Open interfaces between the same materials where transport phenomena occur: S_{op+} .
4. Open interfaces between the same materials where transport phenomena occur: $S_{op\parallel}$.
5. Closed interfaces between different materials where transport phenomena occur: D_{cl+} .
6. Closed interfaces between different materials where transport phenomena occur: $D_{cl\parallel}$.
7. Open interfaces between different materials where transport phenomena occur: D_{op+} .
8. Open interfaces between different materials where transport phenomena occur: $D_{op\parallel}$.

Of the eight types of interfaces type 5 was the most intensively investigated. An example of the interface is the

p–n junction. Among the typical cases, the transport phenomenon across the grain boundaries was categorized as type 1, diffusion along the grain boundaries as type 2, and the mechanism of chemical sensors by way of porous semiconductors as type 3. The p–n hetero-contact chemical sensors, which our group proposed and further investigated, was type 7, (4). The mechanism of humidity sensing and its effect upon acid-base mixtures was analyzed as type of 8. This system of categorization proved specially useful in our analyses of the working mechanisms of chemical sensors.

The interactions between different materials were considered as nonlinear when the structure was an open one, the ambient air give rise to some interesting phenomena.

The hetero-contact between an n-type semiconductor and a p-type semiconductor, namely between ZnO and CuO doped with Na^+ , proved to be an intelligent chemical sensor. This is because the structure consisted of different materials, was open, and the subjected to electric current flows across the interface with the current changing with the ambient air. Usually chemical sensors of porous semiconducting materials, such as SnO_2 or ZnO, cannot distinguish the among flammable gaseous species. For instance, it was very difficult to distinguish CO from H_2 until this experimental structure was developed. The sensitivity and selectivity of carbon monoxide (CO) can be modified by a bias between the hetero-contact as seen in

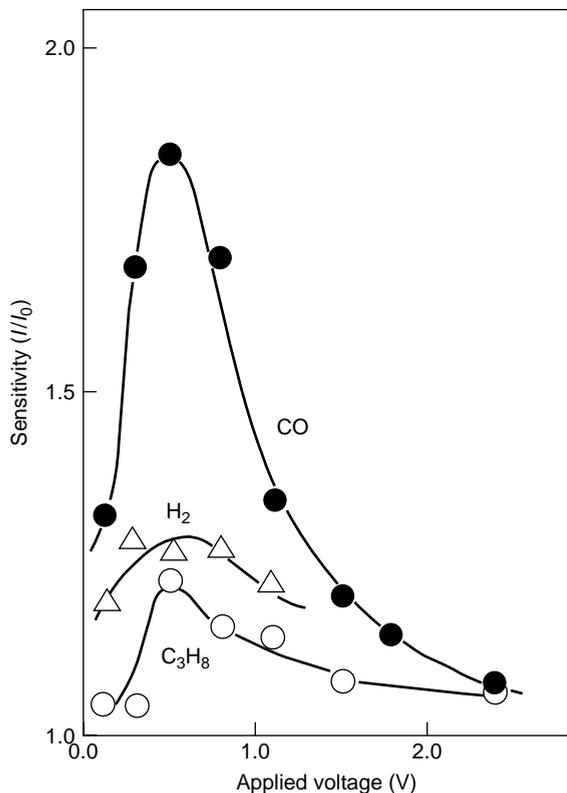


Figure 3. Relationship between the gas sensitivity and the forward applied voltage in CuO/ZnO hetero-contact.

Fig. 3. One of the intelligent functions has included was a tuning-capability as described in an introductory paper in 1988 by Yanagida on intelligent materials (5). The first success in finding a chemical sensor based on hetero-contact was back in 1979 when a humidity sensor was made of N_1O and ZnO (6). In the later FroC project, there was proposed (7) a similarly selective carbon monoxide gas sensor as shown in Fig. 4. The working mechanism was analyzed (8) as a bias-enhanced oxidation of absorbed CO or H_2 gas. This meant that catalytic activity can be controlled by changing the bias between the hetero-contact.

From the results of the FroC project it was clear, that out that the categorization had to be developed further. Since then well-known intelligent functions such as the ZnO varistor and PTCR, positive temperature coefficient of resistivity of $BaTiO_3$ have been analyzed considering the axial relations. The electric carrier transport of the zinc

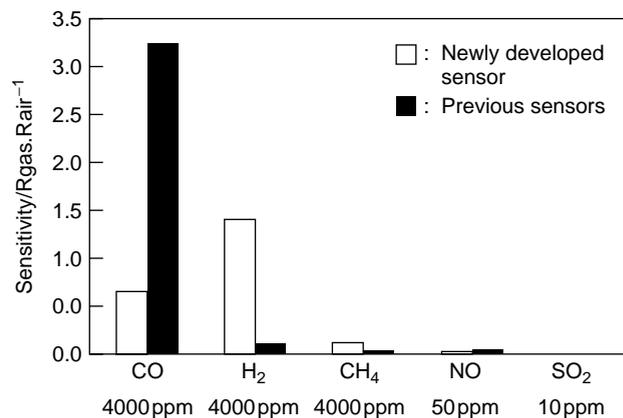


Figure 4. Selective molecular recognition of carbon monoxide from hydrogen.

oxide varistor was found to take place across the grain boundaries. However, it is now known that transport behaviors vary with crystal orientation and their relationships between the grains. When grains are the same, the interface will be alike between different materials. The relationship can be varied with poling in ferro-electric materials such as barium titanate.

Interfaces between the same materials but having a different crystal axis orientation must be considered as interfaces between different materials. In sum, intelligent functions were found to depend on the design of crystal axis orientation.

BIBLIOGRAPHY

1. By Science and Technology Agency, *Proc. Intelligent Materials Forum on Intelligent Materials*, January 14, 2000, Tokyo.
2. T. Takagi, *The Present State and the Future of Intelligent Material and Systems*. ICIM' 98. Makuhari, Japan.
3. H. Yanagida. *Kagaku to Kogyo* **39**: 831–833 (1986).
4. Y. Nakamura, T. Tsurutani, M. Miyayama, O. Okada, K. Koumoto, and H. Yanagida. *Nippon Kagaku Kaishi* (3): 477–483 (1987).
5. H. Yanagida. *Angewandte Chemie*. **100**(10): 1443–1446 (1988).
6. K. Kawakami and H. Yanagida. *Yogyo Kyokai Shi* **87**: 112–115 (1979).
7. *Nikkei Mechanical*. **2**(521): 88–89 (1998).
8. S.J. Jung, Y. Nakamura, A. Kishimoto, and H. Yanagida. *J. Ceram. Soc. Jpn.* **104**: 415–421 (1996).

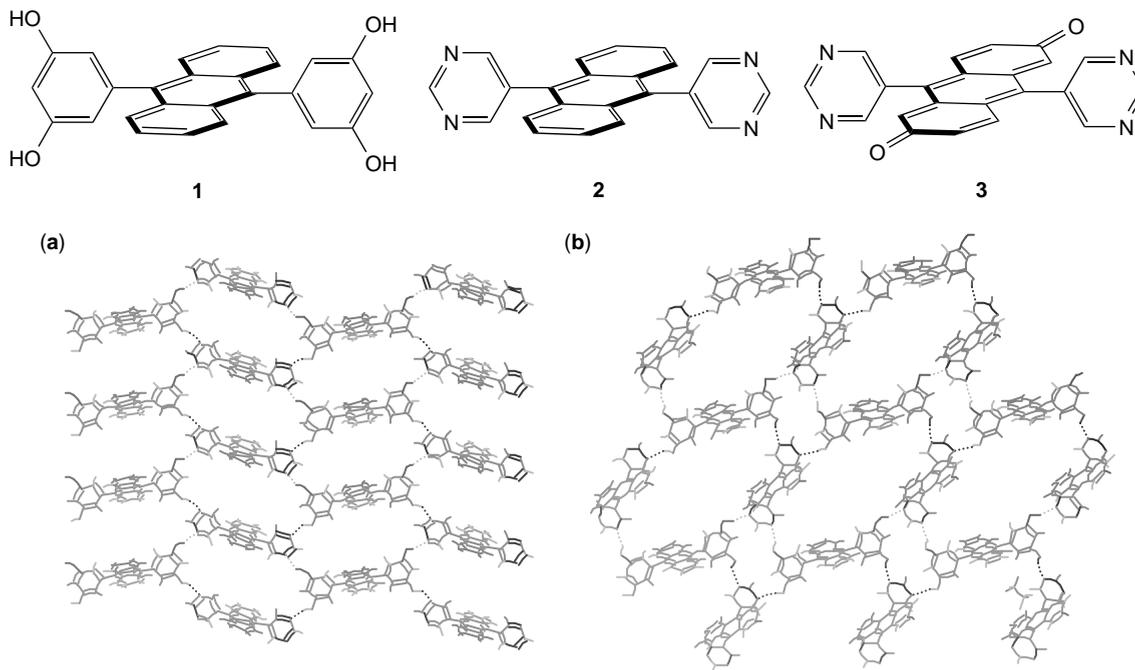


Figure 3. Crystal structures of (a) **1-2** complex and (b) **1-3** complex (solvent molecules have been omitted for clarity).

concentration-dependent. These assemblies range from a rigid-rod character at very dilute concentrations to a lyotropic liquid-crystalline gel at higher concentrations (5). Most interestingly, Meijer developed supramolecular polymers of type **5**, that are held together by quadruple hydrogen bonding between the ureidopyrimidone units. This material displays most, if not all, properties of macroscopic polymers based only on non-covalent connections (Fig. 4) (6).

Materials based principally on hydrogen bonding and other intermolecular interactions have been generated using low molecular weight organogelators (7). Recently, microcellular organic materials have been prepared by drying of organogels in supercritical CO_2 (8). The field of organogelation has evolved from molecules that have different structural and recognition properties to the rational design and fine-tuning of materials. In the following section, we describe advances in organogelation and the use of intermolecular interactions in developing these new supramolecular structures.

INTERMOLECULAR INTERACTIONS

Hydrogen Bonding

Hydrogen bonds are usually formed when a donor (D) that has an available acidic hydrogen is brought into close contact with an acceptor (A) that possesses a lone pair of electrons (Fig. 5).

Hydrogen bonding has been the subject of statistical investigations (9,10), X-ray diffraction analysis (11–13), and theoretical studies. The hydrogen bond can vary in strength from 1 kcal/mol for $\text{C-H}\cdots\text{O}$ hydrogen bonding (14,15) to 40 kcal/mol for the HF_2^- ion in the gas phase

(16,17). Hydrogen bonding has played a critical role in the development of areas such as self-assembly (18) and crystal engineering (19,20). Solid-state and solution studies of the hydrogen bond have provided evidence that this interaction is a highly ordered phenomenon, not a random event. In the solid state, Zaworotko demonstrated that the inorganic complex **6** that has four hydrogen bond donors oriented in a tetrahedral geometry can form hydrogen bonds to nitrogen and to π -systems (21). The crystal structures from this study show supramolecular diamond-like arrangements, where the size of the network cavities depend on the size of the hydrogen bond acceptor (see Fig. 6).

Hydrogen-bond strength is influenced by secondary electrostatic interactions. A particularly strong hydrogen-bonded complex is formed when a molecule which is comprised of all hydrogen-bond donors binds to an all hydrogen-bond acceptor (Fig. 7) (22). The calculated individual secondary electrostatic interactions in these complexes are ± 2.5 kcal/mol in chloroform (23). Schneider estimated the contribution of a related series of secondary interactions at ± 0.7 kcal/mol from a large number of examples in the literature (24).

The presence of a competitive hydrogen-bonding solvent can also influence the strength of hydrogen bonding. Lorenzi found that the dimerization constant of a small cyclic peptide is 80 M^{-1} in tetrachloromethane, whereas the same molecule does not dimerize in chloroform (25). Wilcox also studied the effects of the concentration of water on the binding free energy of hydrogen-bonding molecules in chloroform (26). In competitive solvents, such as water/methanol mixtures, guanidinium receptors and carboxylate substrates are highly solvated. Schmidtchen has observed that as the polarity of the solvent (higher percentage of water in methanol) increases, the enthalpic

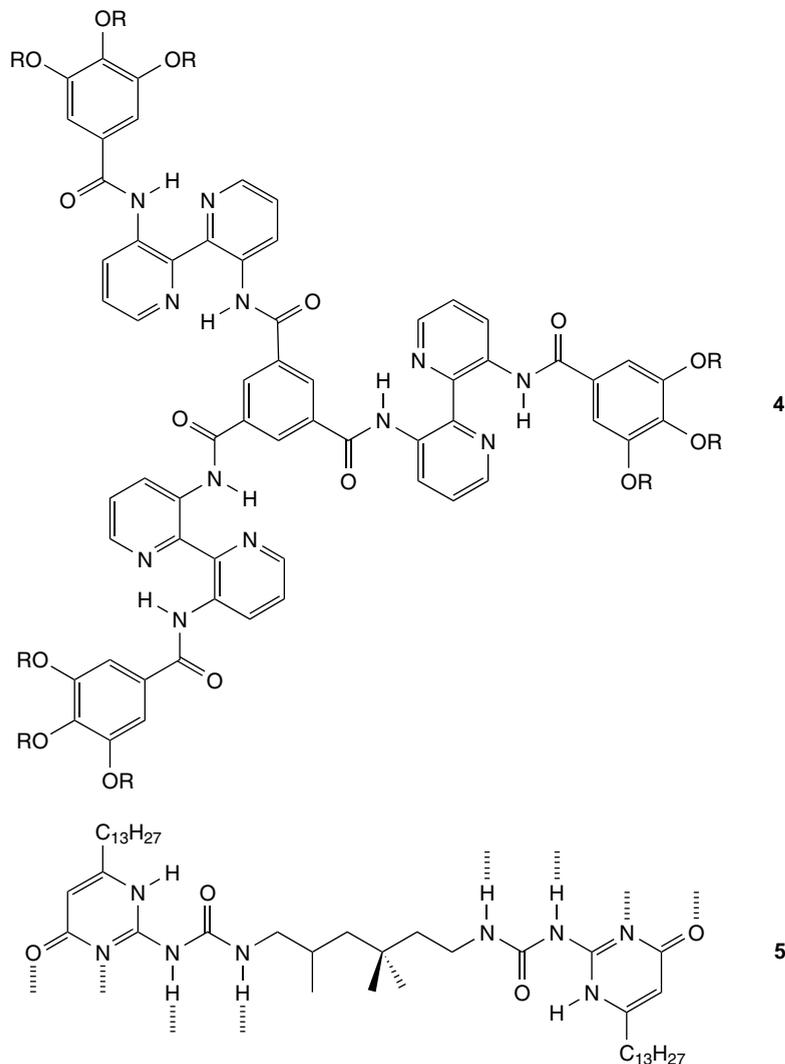


Figure 4. Examples of supramolecular materials: Molecule **4** forms liquid-crystal gels and **5** forms polymers.

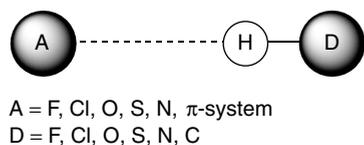


Figure 5. Hydrogen bond formed between an acidic hydrogen (D-H) and an acceptor (A).

component of the binding free energy and the binding affinity decreases (27).

π - π Interactions

Planar aromatic molecules, it is known, interact with one another in three possible geometrical arrangements: stacking (e.g., face-to-face overlap of duroquinone, Fig. 8); offset stacking (e.g., laterally shifted overlap in [18]annulene); and herringbone (e.g., T-shaped edge-to-face interactions in benzene, Fig. 8) (20). The greatest van der Waals interactive energy is found in the face-to-face overlap arrangement in which there is the highest number of C \cdots C intermolecular contacts. If van der Waals forces were solely to

determine the packing of flat aromatic molecules, the offset stack and herringbone arrangements would not be commonly observed. In effect, there is a barrier to face-to-face stacking due to $\pi \cdots \pi$ repulsions. As a result, the offset stack arrangement is the most commonly observed (28). Similarly, the herringbone interaction in many aromatic hydrocarbons offers evidence for the C(δ^-)-H(δ^+) nature of this interaction (29). The slightly electrostatic character of the herringbone interaction may predispose molecules during crystallization toward inclined geometries and reflects its character as a weak C-H \cdots π hydrogen bond (30).

van der Waals Interactions

Van der Waals interactions are dispersive forces caused by fluctuating multipoles in adjacent molecules that lead to attraction between them. In the solid state, the packing of aliphatic side chains is governed by van der Waals interactions. When the side chains are longer than five carbon atoms, H \cdots H interactions predominate. Similar effects have been observed in molecular recognition solution studies. Kataigorodskii's close-packing principle assumes that the potential energy of the system is minimized by

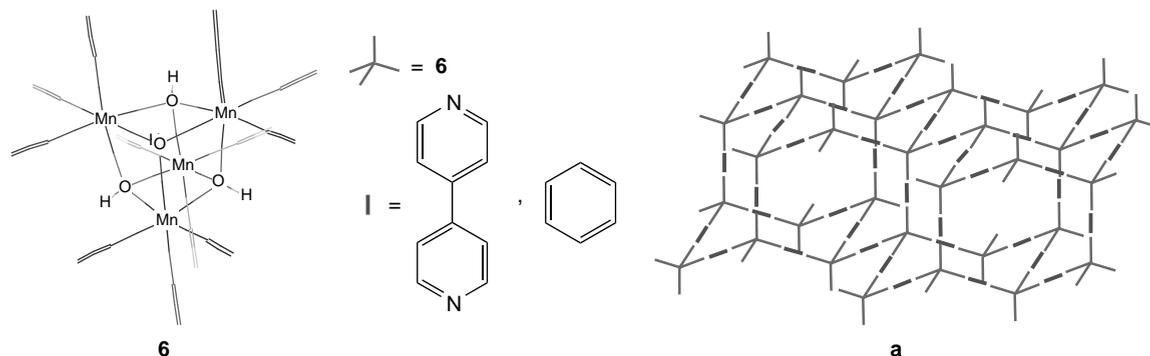


Figure 6. The hydroxyl groups of inorganic complex **6** can hydrogen bond to 4,4-dipyridine or benzene to form diamondoid networks (a).

molecules making a maximum number of intermolecular interactions (31). Therefore, because van der Waals interactions are nondirectional, the energy differences between alternative molecular arrangements are small. However, computational studies of rigid molecules that assemble into one-dimensional aggregates give insight into the importance of van der Waals and coulombic terms (32,33). The importance of van der Waals interactions in organogelation should not be underestimated because most organogelators have long alkyl chain groups.

ORGANOGELEATION

Many attempts have been made to define the phenomenon of gelation. In 1993, Kramer and colleagues proposed that use of the term “gel” be limited to systems that fulfill the following phenomenological characteristics: (1) They consist of two or more components, one of which is a liquid, and (2) they are soft, solid, or solid-like materials (34). The authors further described their definition of “solid-like” and also reviewed other existing definitions of gels. However, there is no precise definition for gelation, and in recent literature it is a phenomenon that is described rather than defined. Organogels are usually formed when an organogelator is heated and dissolved in an appropriate solvent. The mixture is allowed to cool to its gel transition temperature resulting in the formation of a matrix that traps the solvent due to surface tension. In general, the

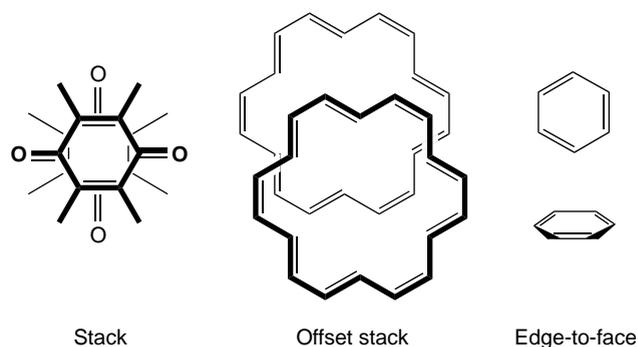


Figure 8. The most common π - π interactions are stacked, offset stacked, and edge-to-face.

amount of organogelator needed to gel a certain solvent is small with respect to solvent, and concentrations can be as low as 2% by weight.

Organogelators are usually divided into two distinct classes based on the nature of the chemical forces that stabilize them. Chemical organogelators are formed through covalent networks; examples include cross-linked polymer gels and silica gels. One common feature of these gels is that their formation is irreversible. In contrast, physical gels are stabilized by noncovalent forces that range from hydrogen bonding to π - π stacking interactions. These types of gels are thermoreversible from the gel phase to solution. The molecules that encompass this class of

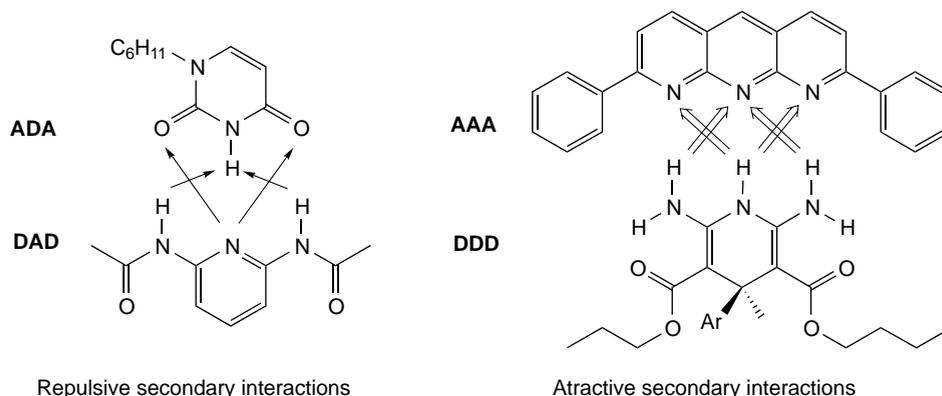


Figure 7. Repulsive and attractive secondary interactions in triple hydrogen-bonded donor-acceptor complexes.

Repulsive secondary interactions

Attractive secondary interactions

compounds range from peptides to carbohydrates to very simple organic molecules. We focus most of our discussion on these types of organogelators.

Several papers discuss the structures of organogelators and their properties. Terech and Weiss presented extensive research on anthryl and anthraquinone appended cholesterol derivatives (35). Extensive work on amide and urea organogelators is presented in van Esch and Feringa's 1999 book chapter (36). Hamilton and co-workers reported examples of organogelator design derived in many cases from molecular recognition and self-assembly (37).

Although the mechanism of gelation is not fully understood, there have been many attempts to understand the structure of gels. In a recent paper, Terech discussed and compared three methods for measuring phase transition temperatures in physical organogels (38). The three methods analyzed were the "falling ball" technique, nuclear magnetic resonance (NMR) spectroscopy, and rheology. The authors concluded that the rheology method is the most reliable.

Examples of Organogelators

Low molecular weight organogelators encompass a variety of compounds that can self-assemble into a fibrous matrix that can trap solvent molecules within its cavities. The noncovalent interactions that hold these structures together are various in nature. As a result, organogelators are usually classified by their chemical constitution. We will use this same type of classification previously employed by Terech, Weiss, van Esch, and Feringa.

Fatty Acid and Surfactant Gelators. Some of the first organogelators were based on substituted fatty acids. 12-Hydroxyoctadecanoic acid **7** and its monovalent salts form organogels in a variety of solvents (Fig. 9) (39,40). Observation of circular dichroism was used as evidence for the formation of supramolecular helical strands, although the

maximum of the signal depended on the solvent. For D-**7**, most of the helices were left-handed, whereas for L-**7** they were right-handed. Tetraalkylammonium derivatives such as compound **8** behave as surfactants and organogelators (41). Gemini (dimeric) surfactants formed by cetyltrimethylammonium ions (CTA) with various counterions gel organic solvents (see Fig. 9) (42). The most probable structure for the gels of **9** and **10** is an entangled network of long fibers that have polar groups at the core of the aggregate and long alkyl chains in contact with the solvent. These compounds gel chlorinated solvents most effectively at concentrations as low as 10 mM. However, they gel other solvents such as toluene, xylenes, chlorobenzene, and pyridine at concentrations from 20–30 mM.

Anthracene and Anthraquinone Derivatives. Anthracene and anthraquinone derivatives gel various alkanes, alcohols, aliphatic amines, and nitriles. These aromatic structures to form gels through π - π interactions. However, when the anthryl ring of **11** is partially hydrogenated, compounds **12** and **13** still form organogels (Fig. 10) (43). The interesting photochromatic properties of anthraquinones, such as **14**, has led to the study of substitution patterns on the ring. Dialkoxy-2,3-anthraquinone derivatives are the most effective agents (44). Both anthracene and anthraquinone substituents have been coupled to cholesterol groups for organogelation purposes. These are discussed later.

Amides and Ureas. The hydrogen bonding properties of amides and ureas have been the subject of solid-state and solution studies. Many low molecular weight organogelators have been designed by using this recognition element. The amide functional group can form eight-membered, hydrogen-bonded dimers and one-dimensional infinite arrays (Fig. 11). Primary and secondary amides (45) and pyrimidones (46) form cyclic dimers.

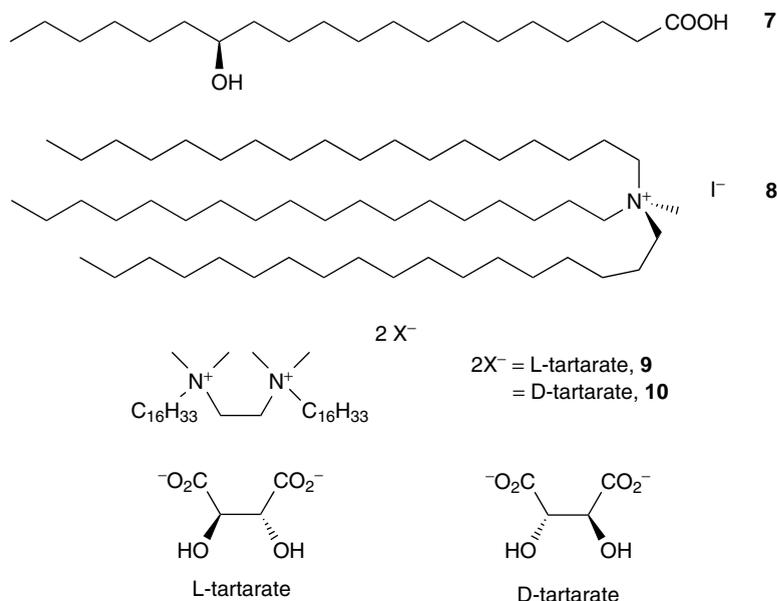


Figure 9. Examples of gemini surfactants and fatty acid organogelators.

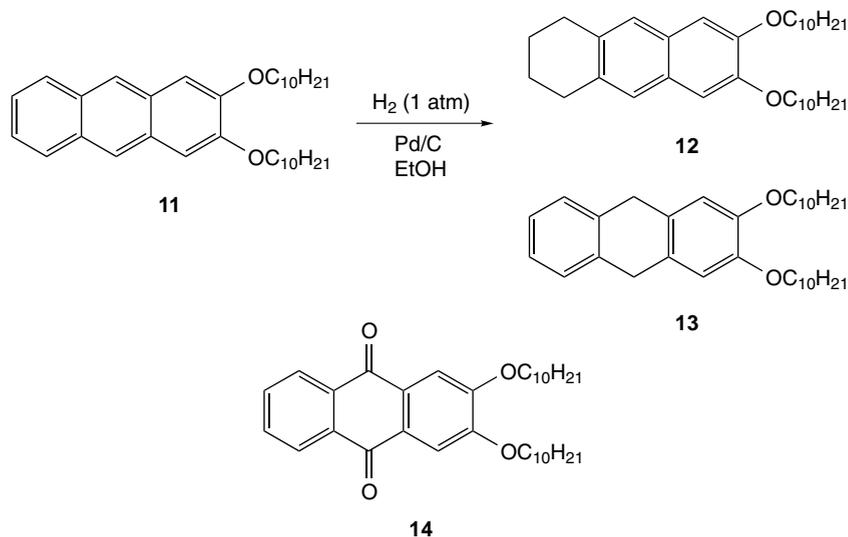


Figure 10. Examples of organogelators based on anthracene and anthraquinone derivatives.

Hanabusa reported that *trans*-cyclohexane-1,2-diamide **15** gels organic solvents, silicon oil, and liquid paraffin at concentrations as low as 2 g/L (47). Enantiomerically pure **15** produces stable gels, and circular dichroism indicates a chiral helical arrangement of the diamides. Electron micrographs of a gel produced from **15** in acetonitrile showed the presence of the helical superstructures. However, the racemic mixture of **15** and **16** produces only unstable gels. The authors believe that the helical superstructures could arise from stacked, hydrogen-bonded, infinite aggregates. Therefore, the orientation of the amide groups in a

antiparallel *trans* configuration (both equatorial) is critical for the complementary interaction. Interestingly, **17** which has *cis*-amide groups (one equatorial and one axial) cannot form this interaction favorably and is not observed to gel any solvents (Fig. 12). This type of stacked amide hydrogen bonding has been observed in the crystal structure of a cyclohexane-1,3,5-triamide reported by Hamilton (48).

Shirota reported similar amide-containing molecules in which the hydrogen-bonding groups are arranged around a rigid core. Compound **18** gels various solvents. To prove that hydrogen bonding is essential in the gelling ability of these molecules, the *N*-methyl analog **19** was synthesized, and no gelation was observed (49). Increasing the distance between the hydrogen-bonding groups does not have adverse effects on the gelling capacities of these compounds (Fig. 13), as was observed for **20** (50).

Hanabusa also reported long alkyl chain trisubstituted organogelators based on a flexible core (51). *cis*-1,3,5-Cyclohexanetricarboxamide derivatives **21–24** show a trend in improved gelation when the alkyl chains are longer (Fig. 14). These results suggest that intermolecular hydrophobic interactions among the alkyl chains are critical in stabilizing the gel network. These molecules increase the viscosity of solvents at very low concentrations. Compound **22** causes an increase in the viscosity of CCl_4 to 250.0 cP at 25°C from 0.908 cP in the absence of the gelling agent.

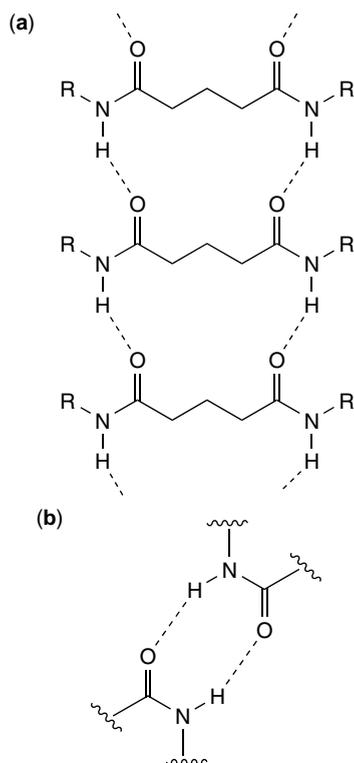


Figure 11. Amides can form linear array (a) or dimers (b) through hydrogen bonding.

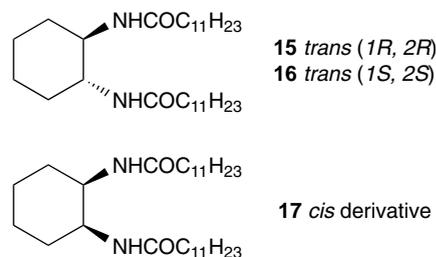


Figure 12. Organogelators based on 1,2-diaminocyclohexane derivatives.

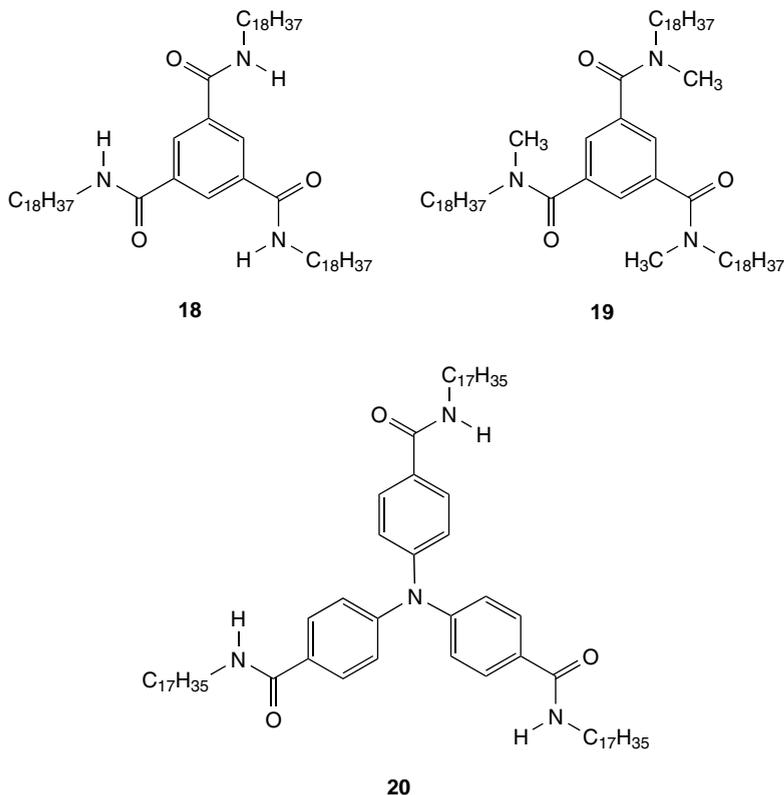


Figure 13. Low molecular weight organogelators based on amide hydrogen bonds.

Recent applications of organogelators have included trapping liquid crystals within the gel matrix. Kato used **15** to gel liquid crystals such as **25** and **26** (Fig. 15) in concentrations as low as 1 mol% (52). These gels were stable at room temperature for several months. Measurements of the response to an electric field were made on gels of **15** and **25**, and interestingly, the threshold voltage of the gel (5.0 V) is larger than that of the liquid crystal alone (1.1 V). The authors propose that the solvent (liquid crystal) is oriented within the gel and that the structure resembles the cartoon shown in Fig. 15. As a result of this property, these materials may have applications in electro-optical devices.

Organogels formed by hydrogen bonding of small molecules have also been stabilized by polymerization. For example, Masuda used amide hydrogen bonds as in 1-aldosamide **27** to template the position of diacetylene groups close to each other for polymerization (53a). IR stretching frequencies were consistent with additional hydrogen bonding between the aminosaccharides. Robust nanofibers are observed in **27** using electron microscopy. However, 1-galactosamide containing **28** forms amorphous solids presumably due to steric hindrance by the axial OAc group that leads to the formation of an infinite amide

hydrogen-bonded network. Polymerization of **27** was confirmed by UV absorption and gel permeation chromatography. In a similar example, Shinkai polymerized **29** *in situ* (53b). The absorption spectra of the gels before and after photoirradiation show a distinct change that is consistent with polymerization (Fig. 16).

Recently, Tamaoki published on the gelation and polymerization of **30** (see Fig. 17) (54). Compound **30** contains two cholesteryl ester units at the ends of a diyne spacer. **30** shows liquid crystal behavior when heated between 101 and 133°C and gels nonpolar solvents at low concentrations. Gels formed in cyclohexane were irradiated by UV light (500-W high-pressure Hg lamp), and their color changed from colorless to dark blue. The absorption spectra provided evidence for the presence of the exciton band of polydiacetylene. The T_{gel} for a gel at 5.3 mM was 45°C before polymerization. However, after UV irradiation, the gel maintained its shape even above the boiling point of cyclohexane (80.7°C).

Ureas have been studied in detail due to their hydrogen-bonding complementarity to carboxylates and other anions. Moran synthesized **31** (Fig. 18), which has additional hydrogen-bonding groups to promote stronger association (55). The bis-urea **32** developed by Rebek

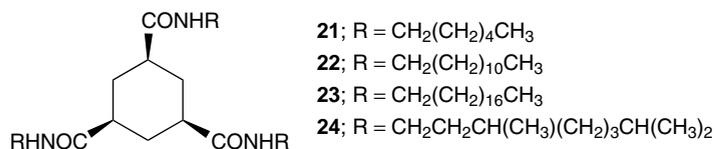


Figure 14. Long alkyl chain organogelators derivatives with hydrogen bonding groups around a cyclic core.

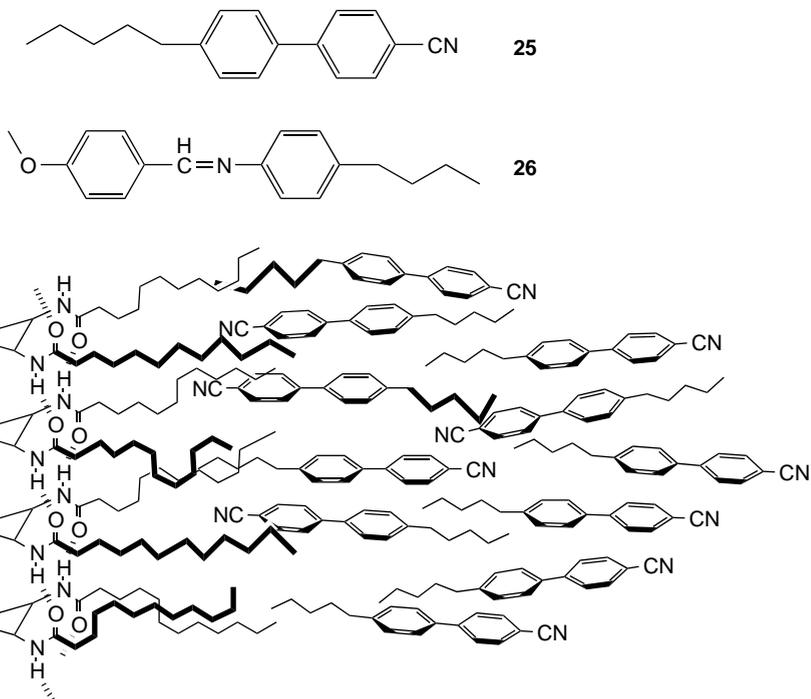


Figure 15. Compound 15 can gel liquid crystals 25 and 26. It is believed that the structure of the gels is as depicted in the illustration.

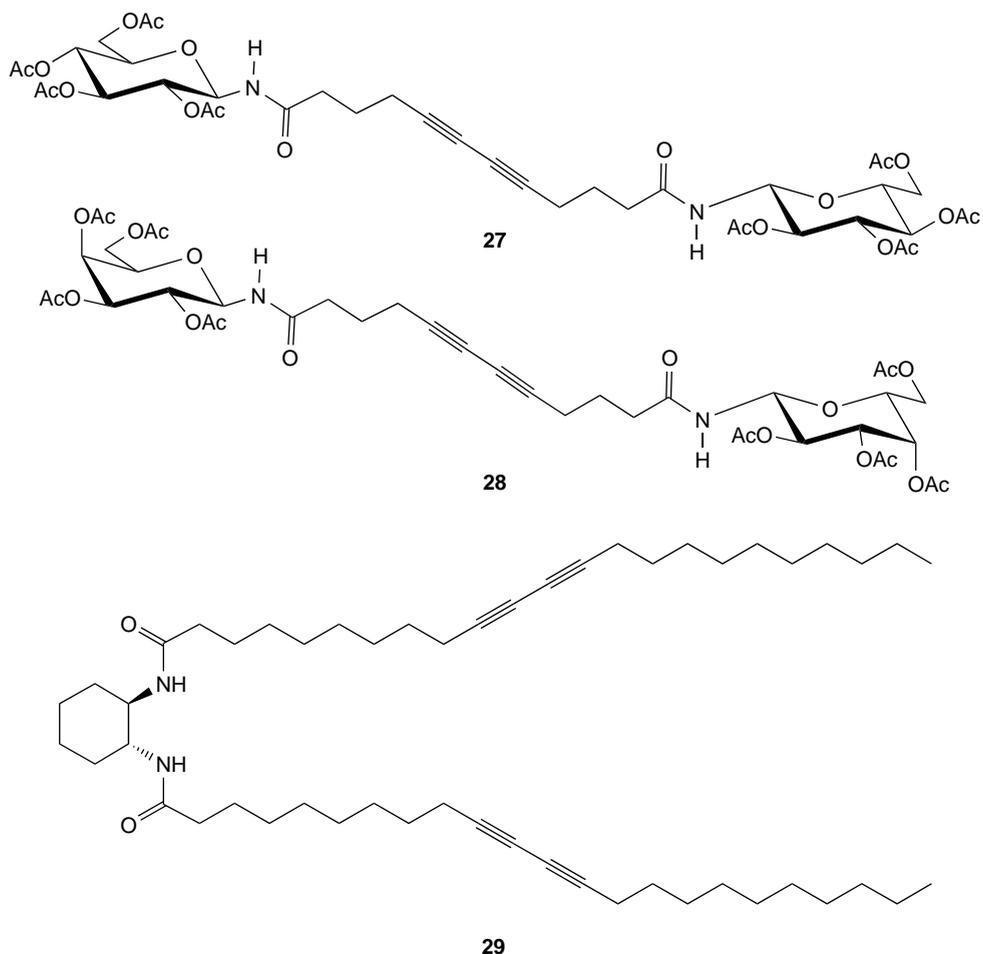


Figure 16. Aldosamide 27 is stabilized by one-dimensional hydrogen bonds, whereas 28 has axial acetyl groups that hinder the formation of such structures. Diamide 29 also forms one-dimensional strands and polymerizes *in situ*.

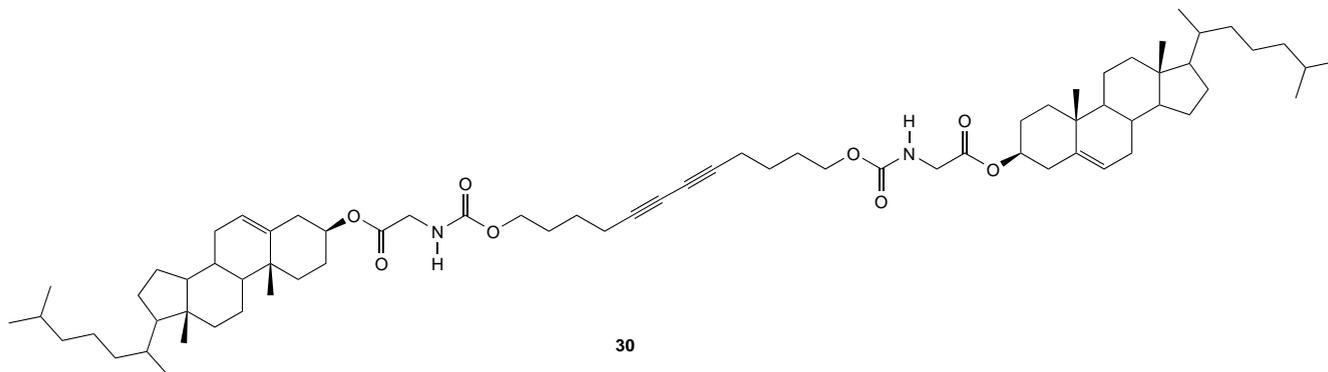


Figure 17. Polymerization of the gel formed by **30** in cyclohexane causes it to turn deep blue after 1 minute of photoirradiation.

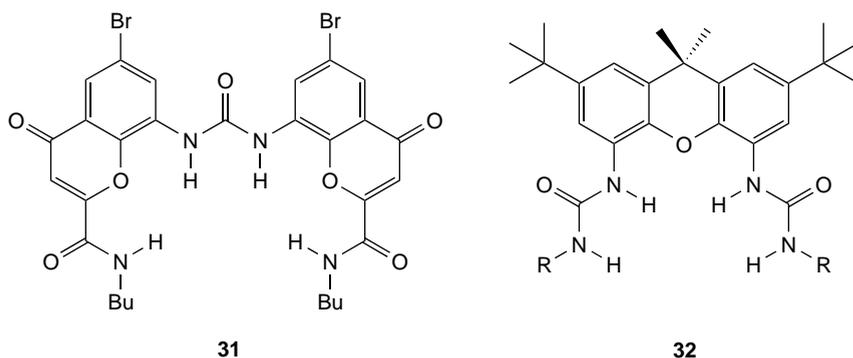


Figure 18. Urea molecules used for binding carboxylates. Molecule **31** has additional hydrogen-bonding sites, and **32** has chiral R groups for enantioselective recognition.

complexes carboxylates enantioselectively when R is chiral (56).

Etter and co-workers observed that bidentate hydrogen bonding predominates in bis-ureas in the solid state. Graph sets and hydrogen-bonding rules were derived from this data and used to predict hydrogen-bonding patterns in related structures (57). Lauher has published crystal structures of a family of ureylenedicarboxylic acids that form hydrogen-bonded sheets through the carboxylic acid dimer, as well as urea hydrogen-bonded one-dimensional strands (58,59). Examples are shown in Fig. 19.

Hanabusa synthesized and studied gelators based on the urea hydrogen-bonding group. Molecules that have rigid spacers between the urea functional groups **33** and **34** gel only toluene and tetrachloromethane, respectively (60). However, cyclic bis-ureas **35–37** show remarkable gelling properties in different organic solvents. Similar to the results with 1,2-bisamidocyclohexanes, these molecules show gelation that depends on the length of the alkyl chains. The antiparallel orientation of the bis-urea groups is also important because the cis analog does not gel any solvent. See Fig. 20.

Bis-urea molecules have been reported by Kellogg and Feringa (see Fig. 21) (61). The morphology of the dried gels observed is thin rectangular sheets. These compounds gel only a selective number of solvents at concentrations around 10 mg/mL. These gels are stable up to temperatures of 100°C and for months at room temperature.

Hamilton reported a family of bis-urea molecules, shown in Fig. 22, that gels mixtures of solvents at 5°C (62).

The crystal structure of **38** confirmed the formation of extensive hydrogen-bonded arrays by both urea groups. As seen in Fig. 22, all of the urea groups point in the same direction, which makes the aggregates chiral. In this particular case, chirality is translated to the entire crystal because all strands point in the same direction. The urea hydrogen bonding distances N–H...O are 2.18 and 2.23 Å, which are within the expected range.

Cyclic bis-ureas derived from trans-1,2-diaminocyclohexane and 1,2-diaminobenzene derivatives have been extensively studied by Kellogg and Feringa (63). They prepared polymerizable derivatives **39** and **40** (Fig. 23). It is interesting to note that **39** forms gels only in tetralin, but **40** gels a variety of solvents. After photoirradiation and polymerization of the methacrylate groups, there is a slight turbidity and stability increase in the gels. A highly porous material was obtained after removing the solvent by freeze-drying. Functional organogels can also be generated by introducing reactive groups in the spacer between the ureas. Thiophene **41** and bis-thiophene **42** show efficient charge transport within the organogels that they form (64). These gels have potential application as organic semiconductors and have been further investigated. Molecular modeling of 1,2-bisaminocyclohexane and 1,2-bisamidobenzene derivatives indicates that trans-antiparallel orientation of the bis-urea groups is the most favorable for forming an extensive hydrogen-bonding network. However, in crystal structure **43**, both urea groups are oriented in the same direction (65). The authors used a variety of techniques, such as infrared spectroscopy, differential scanning calorimetry

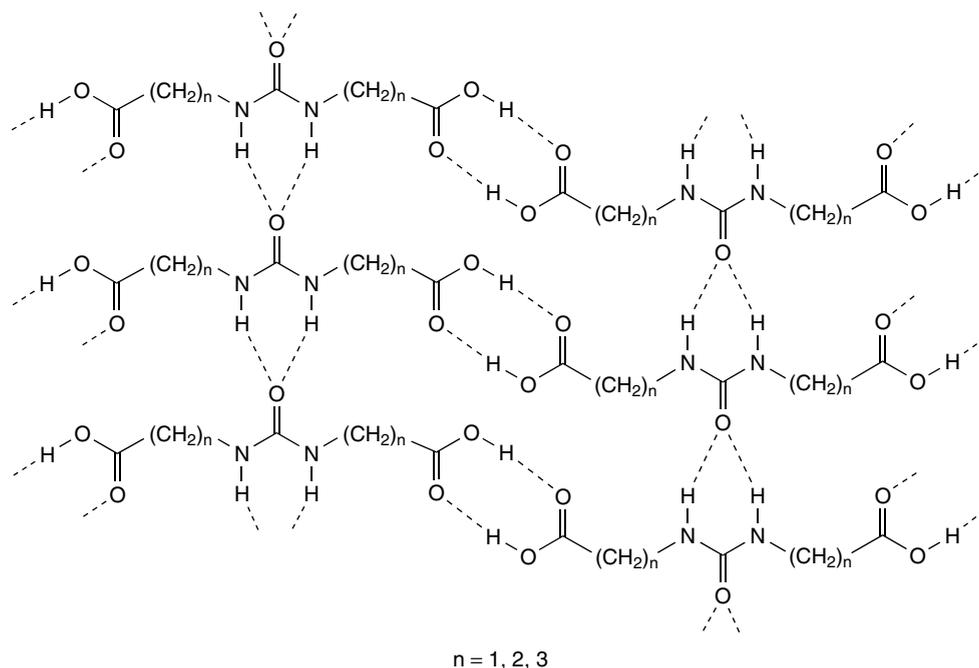


Figure 19. Formation of urea bidentate hydrogen-bonded chains prevail in the presence of other hydrogen-bonding groups such as carboxylic acids.

and electron microscopy, to elucidate the supramolecular structure of the gelators.

Barbiturate-Melamine and 2,6-Diaminopyridine-Barbiturate Organogelators. Hydrogen-bond complementarity between two different functional groups has been studied in many systems for recognition in host-guest chemistry and for the formation of infinite aggregates in solution and in the solid state. Two of these patterns that have been incorporated into organogelators are the barbiturate-melamine pair and 2,6-diaminopyridine-barbiturate.

Whitesides and Lehn exploited the hydrogen-bonding complementarity of melamine-barbituric/cyanuric acid (13,66). The focus of their research was to study the

preferential formation of cyclic or linear aggregates. Their approach was to use steric hindrance to enhance formation of cyclic aggregates over linear ones. When the melamine derivative has methyl substituents on the phenyl rings (**44**) or an *n*-butyl substituent at the 3-position on the melamine derivative (**45**), the aggregates cocrystallize with diethylbarbiturate **46** in a linear arrangement. However, when the substituents were bulky, as in *t*-butyl-substituted phenyl groups (**47**), the cocrystallized aggregates were cyclic (see Fig. 24).

These types of molecular recognition motifs have been coupled to long alkyl chains to generate two-component organogelators. Hanabusa used a 1:1 mixture of **48** and **49** to gel *N,N*-dimethylformamide, chloroform, tetrachloromethane and cyclohexane at concentrations as low as 0.04 mol/mL (Fig. 25) (67).

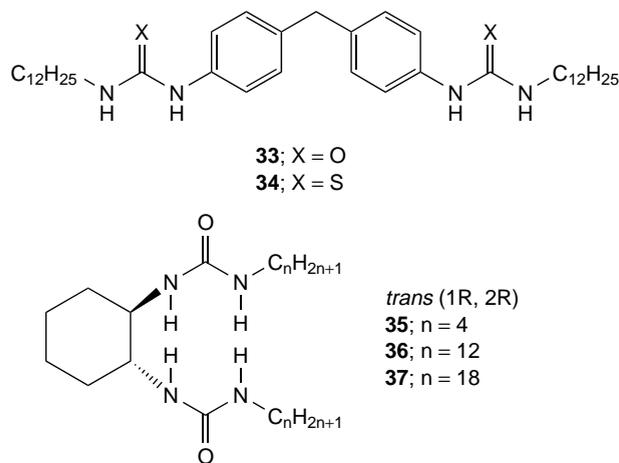


Figure 20. Organogelators based on bis-urea derivatives.

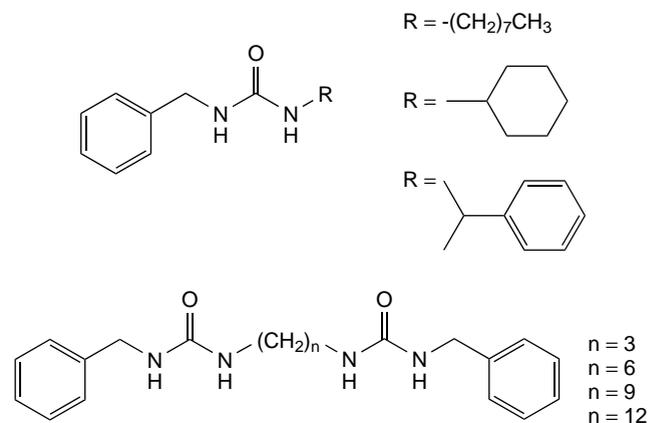


Figure 21. Simple alkyl bis-urea organogelators.

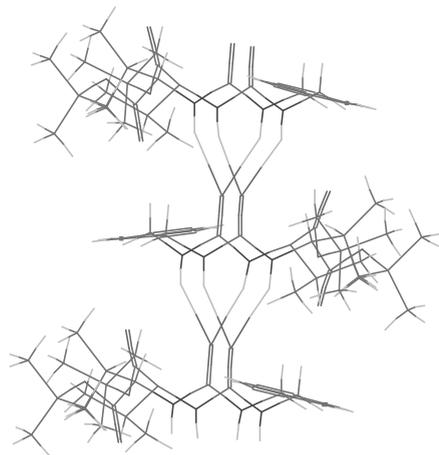
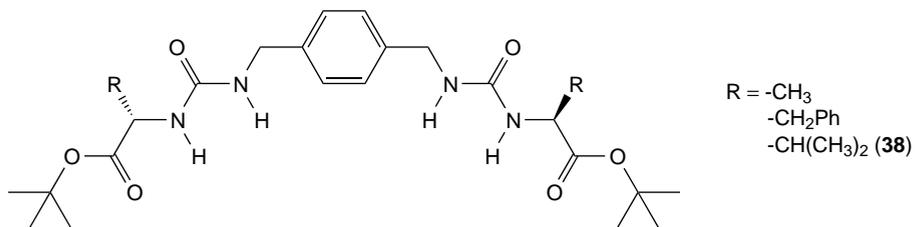
**38**

Figure 22. Crystal structure of **38**, a valine-bis-urea derivative. The urea groups are oriented in parallel and form the expected bidentate hydrogen bonds.

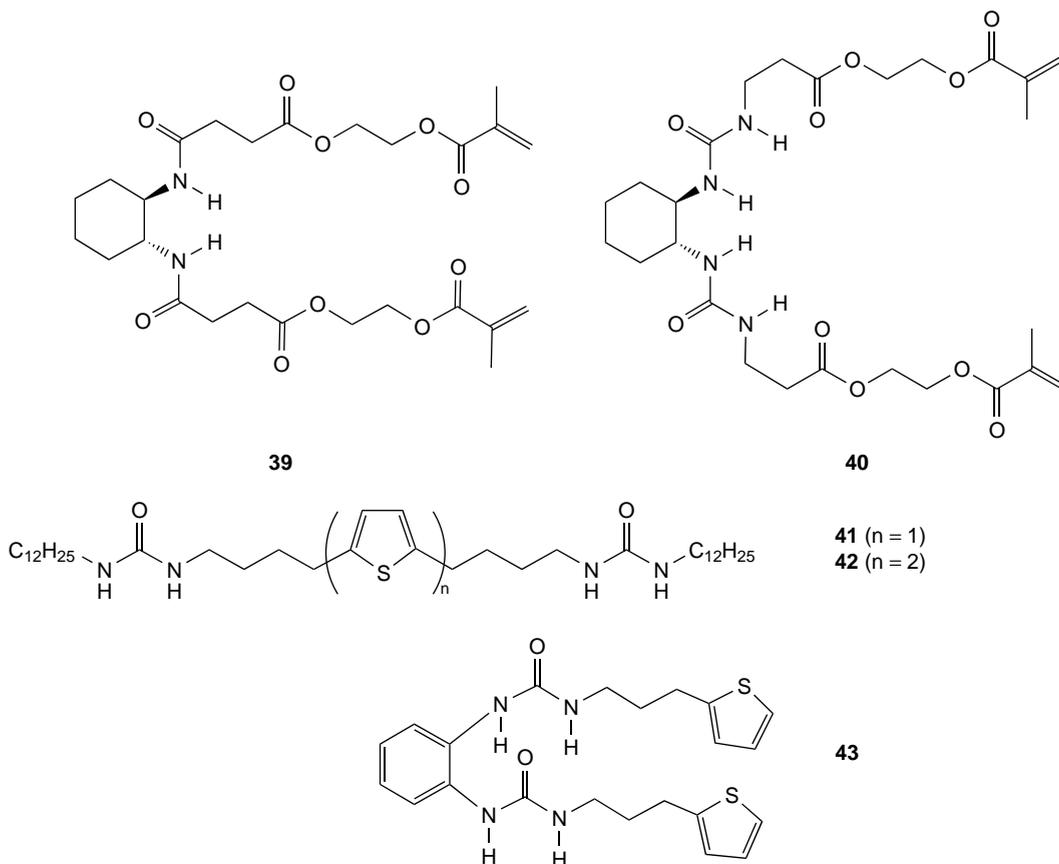


Figure 23. Examples of polymerizable (**39** and **40**) and charge transfer (**41**, **42**, and **43**) organogelators.

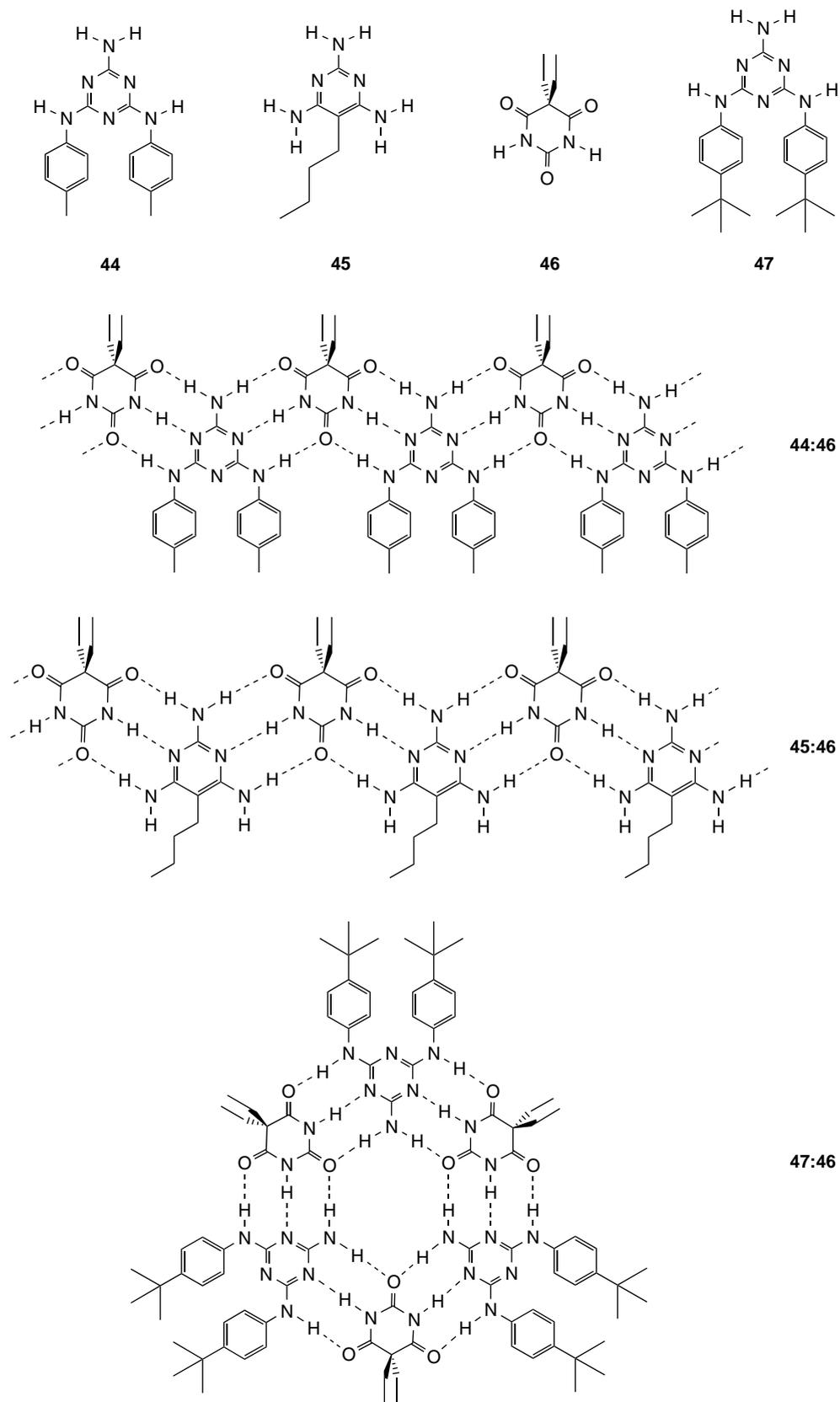


Figure 24. Hydrogen-bonded patterns formed in cocrystals of **46** with **44**, **45**, and **47**.

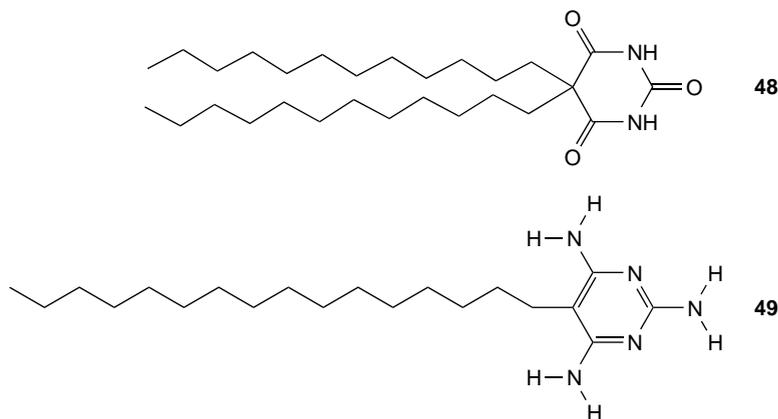


Figure 25. Two-component complementary organogelators based on melamine–barbiturate hydrogen bonding.

Hamilton has made extensive use of 2,6-diaminopyridines in designing receptors for barbiturates. This hydrogen-bonding group has been incorporated into macrocycles **50** and **51** (Fig. 26), which also have the appropriate size cavity to form a 1:1 host–guest complex with **46** (68). When coupled to a thiol nucleophile, these receptors show large increases in the rates of thiolysis reactions of barbiturate-active ester derivatives (69).

Shinkai synthesized host–guest organogelators based on cholesterol-substituted 2,6-diaminopyridines. Gels formed by **52** and **53**, when combined with **46**, are stabilized by different mechanisms (Fig. 27) (70). The complex of **46** with **52**, which has flexible spacers between hydrogen bonding groups forms gels by intermolecular stacking of the host–guest complex. Whereas, in **46:53** the alignment of the complex is not optimal, leaving free N–H and C=O groups for intermolecular hydrogen bonding, and leading to cross-linking of the aggregates to form gels. The linking of known organogel forming molecules to cholesterol substituents has been exploited in designing many organogelators. In the following section, we present organogelators based on structures that have already been discussed in previous sections.

Cholesterol Derivatives. In 1989, Weiss and co-workers published a report of a family of organogelators termed ALS (71) that contain an aromatic (A) and a steroidal (S) group linked by atoms (L). These compounds contain a 2-substituted anthracenyl group coupled to the C3 of the steroid group (see Fig. 28). Because there is no strong hydrogen-bonding group in these molecules, it is believed that the gels are stabilized by dipolar and van der Waals interactions. Nonetheless, these forces are strong enough for gels from 2% concentrations to retain their properties for several months. Results from fluorescence, circular dichroism, X-ray diffraction, and ^1H NMR studies indicate that **54** forms gels by a stacked helical arrangement of the molecules, where the anthracenyl group partially overlaps the aromatic region of neighboring molecules. The organogels formed by **54** were also studied in decane and butanol by neutron and X-ray scattering techniques (72). The results show that the aggregates are composed of long, rigid fibers. The diameter of the fibers is sensitive to the solvent: 160 Å in decane and 192 Å in butanol. The results from this and previous studies showed that interactions between the aromatic groups play an important role in determining gel structure, stability, and aggregate type.

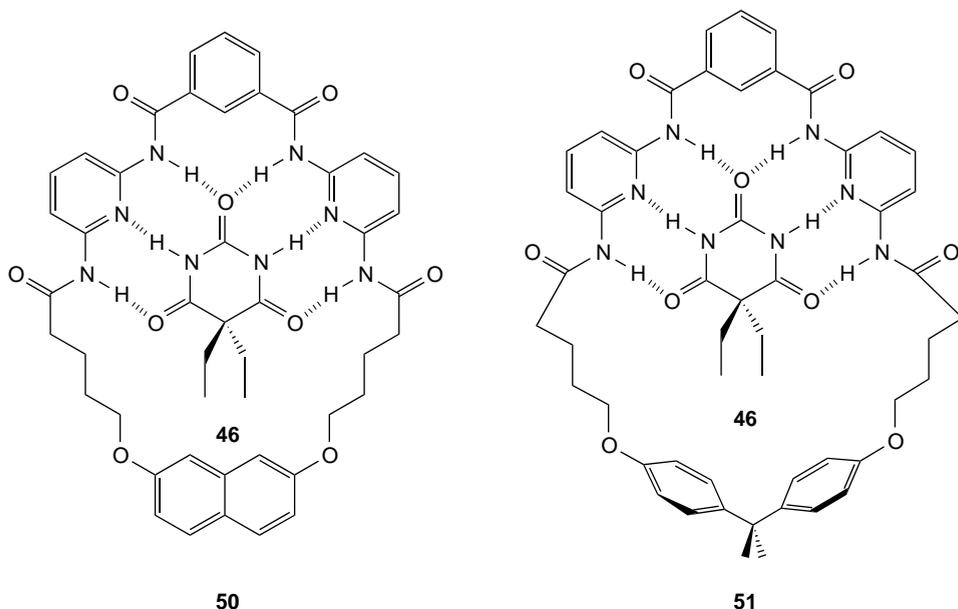


Figure 26. Examples of 1:1 host–guest recognition receptors for barbiturates.

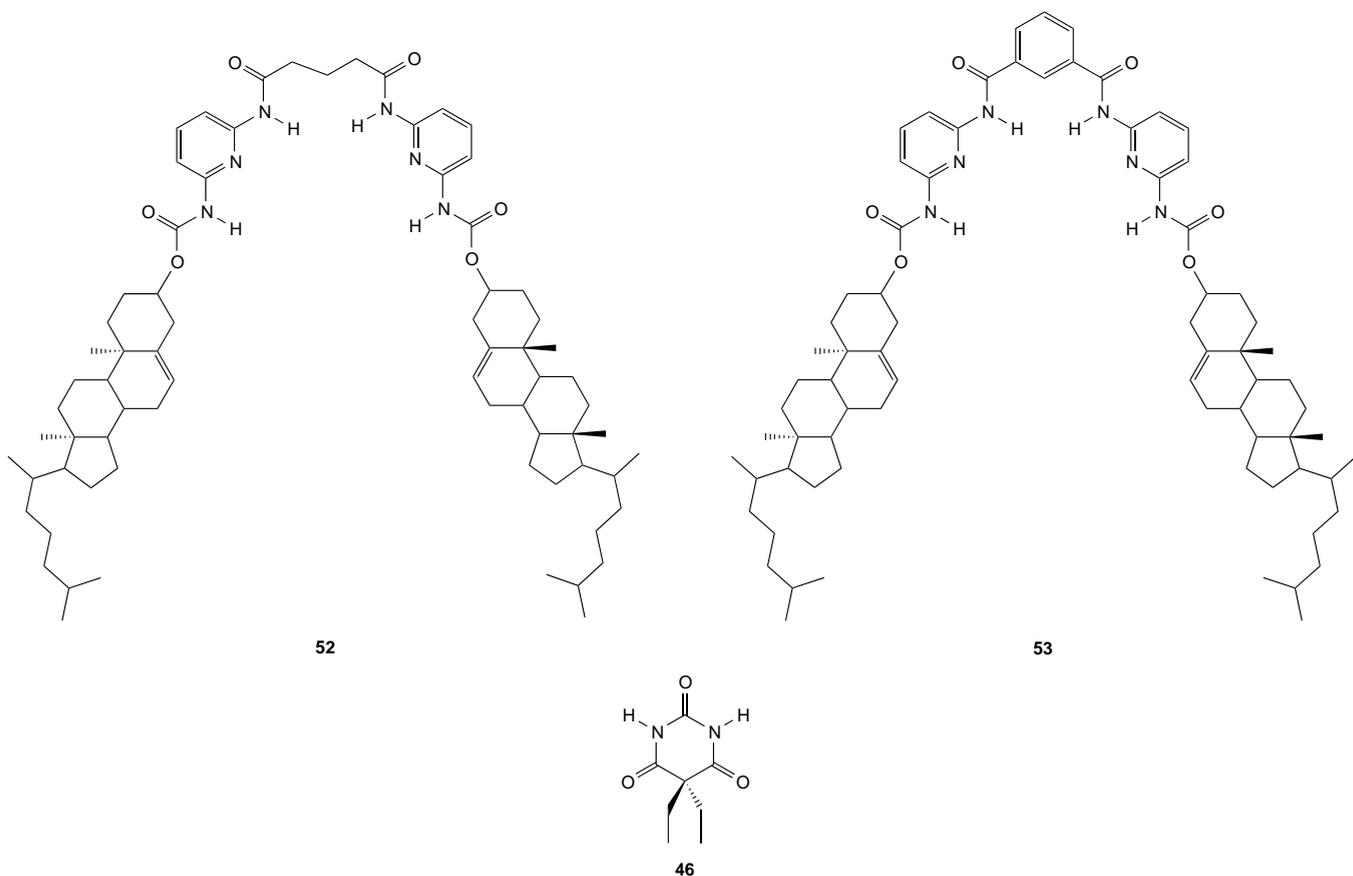


Figure 27. Cholesterol derivatives of 2,6-diaminopyridines **52** and **53** form hydrogen-bonded gels with barbital **46**.

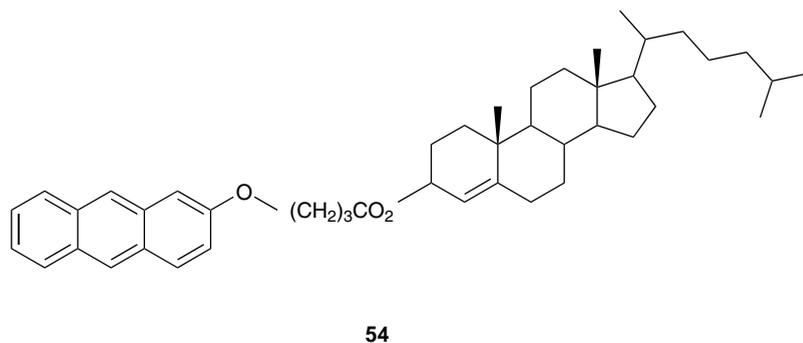
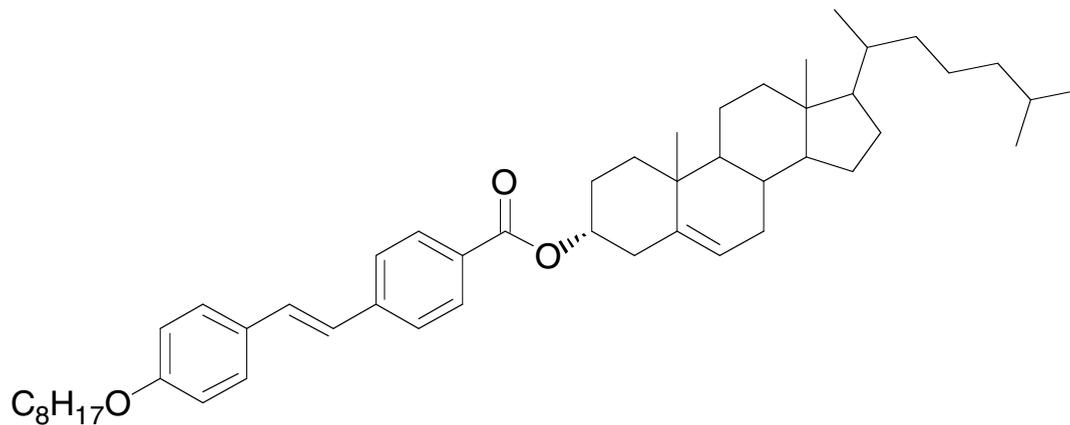


Figure 28. Example of an anthranil-cholesterol organogelator.

A fluorescence spectroscopic study of gels formed from cholesterol–stilbene and cholesterol–squaraine gelators showed similar fibrous structures (73). Studies using a variety of probes showed that the solvent is in a liquid-like phase in the gel matrix microenvironment. Furthermore, Swanson and Whitten captured a series of time transient images to monitor the sol-to-gel phase transition of **55** in 1-octanol (Fig. 29) on highly oriented pyrolytic graphite (74). Using AFM and the assumption that solvent molecules fill the separations between the fibers, they estimated that 30% of solvent molecules are inside the fibers and 70% are in the space between the fibers. The authors suggest that this finding provides a foundation for the selectivity that many organogelators

show among the solvents that they most effectively gel.

Shinkai studied thermal and photochemical control of cholesterol-based gelators that contain azobenzene groups coupled to the C3 of a steroid through an ester linkage (75). The results indicated that when the configuration at C3 is the naturally occurring (*R*), the gelators are effective in polar solvents, whereas when the stereocenter has (*S*) configuration, apolar solvents are gelled. Scanning electron microscopy established that the gelators form three-dimensional networks of helical fibrils. The gels prepared from **56** in cyclohexane showed a right-handed helical signal by circular dichroism (CD) when the configuration of the C3 center was *R*. The left-handed helix was observed



55

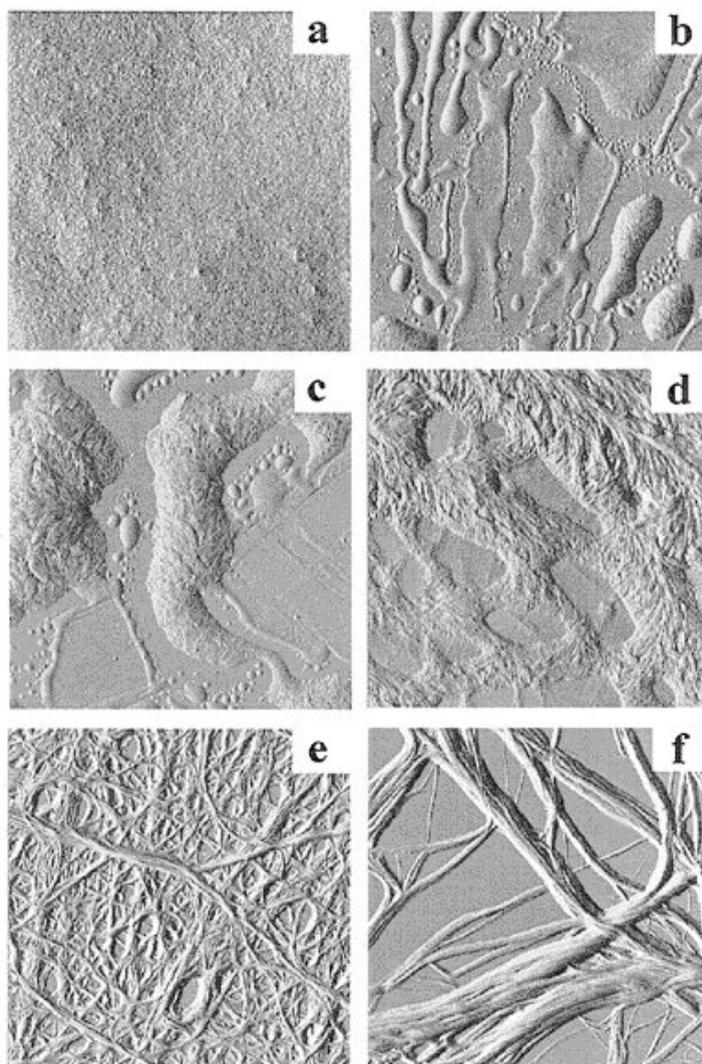


Figure 29. Time transient AFM images of sol-gel phase transition for **55**. These images were acquired after the heated solution was cooled to room temperature for (a) 0, (b) 10, (c) 15, (d) 18, (e) 21, and (f) 31 minutes. Images scale is $12 \times 12 \mu\text{m}$. (These images are reproduced with permission of the Journal of the American Chemical Society.)

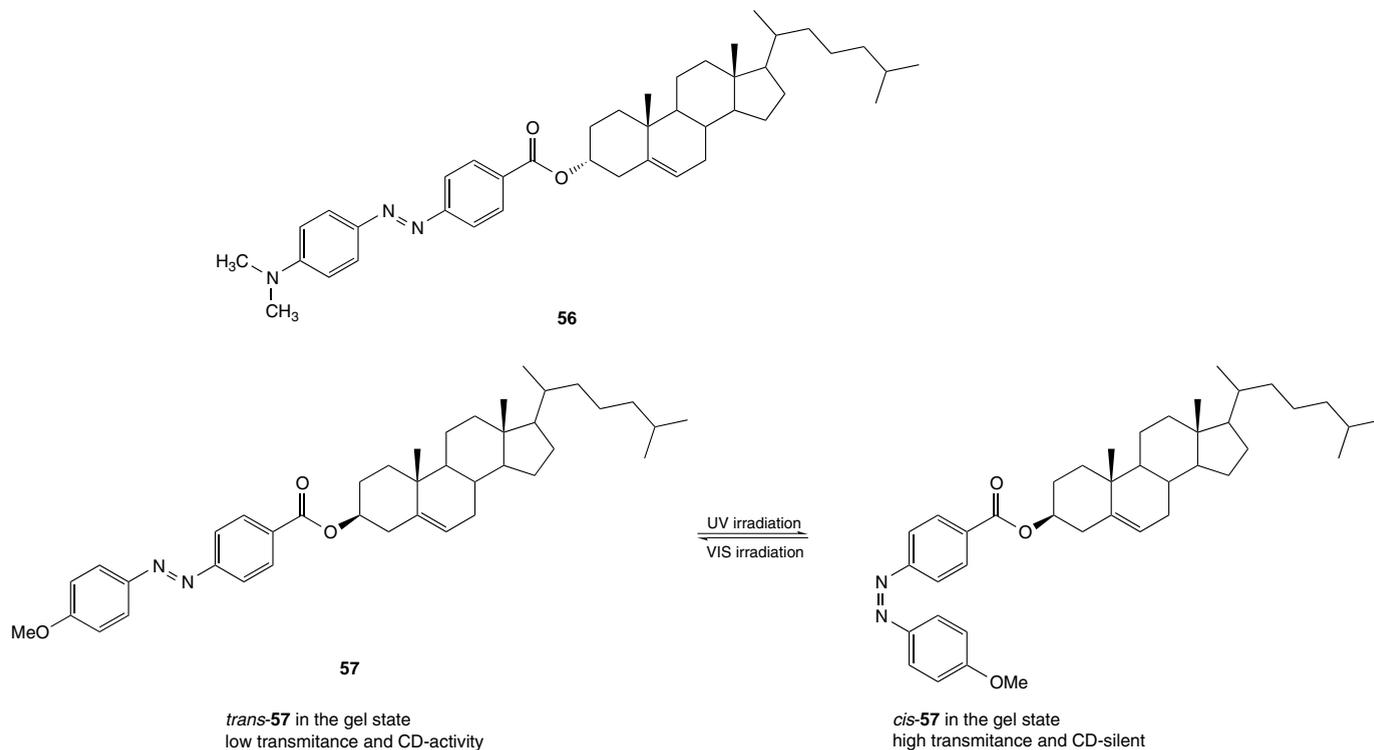


Figure 30. Gelation of azo-cholesterol derivatives can be photolitically controlled.

for the other stereoisomer. Another interesting result from this study is that the sol–gel phase transition of **57** was induced by photoresponsive *cis*–*trans* isomerism of the azobenzene group (see Fig. 30).

Shinkai exploited the gelation of cholesterol derivatives by generating a series of interesting organogelators. For example, the isocyanuric acid 2,3,6-triaminopyrimidine pair **58** was generated (Fig. 31) (76). Gelation in this system is remarkably dependent on the cooling rate of the mixture, as well as the nature of the solvents and

the concentrations used. Interestingly, the mismatched hydrogen-bonding combination leads to gelation, whereas the formation of molecular tapes results in precipitation.

Other interesting functional gels using the cholesteryl group have included a porphyrin substituent (**77**). The derivative with (*S*) C-3 configuration (**59**) effectively gelled a number of solvents, whereas the (*R*) isomer was ineffective with all solvents used. The absorption spectra showed a new λ_{\max} at 444 nm and its corresponding CD band only in the gel phase, indicating that gelation leads to ordered

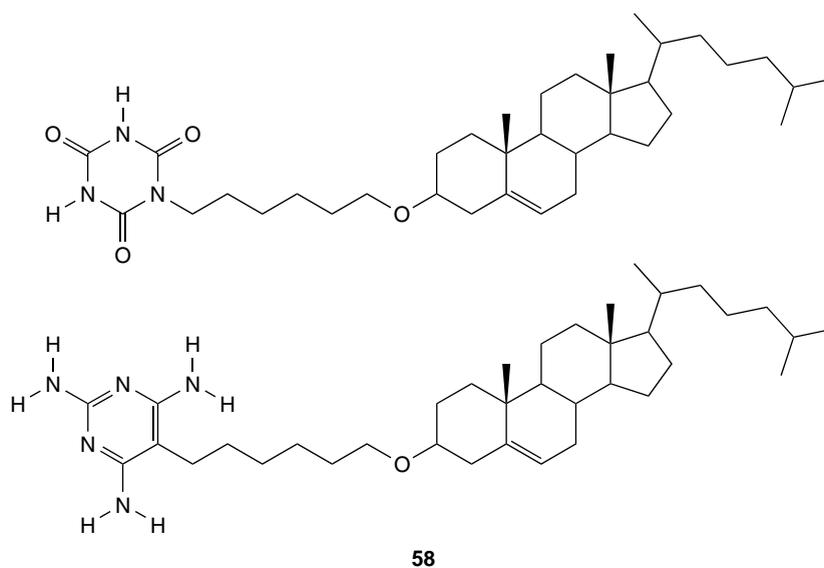


Figure 31. Cholesterol derivatives of isocyanuric acid and 2,4,6-triaminopyrimidine pair **58**.

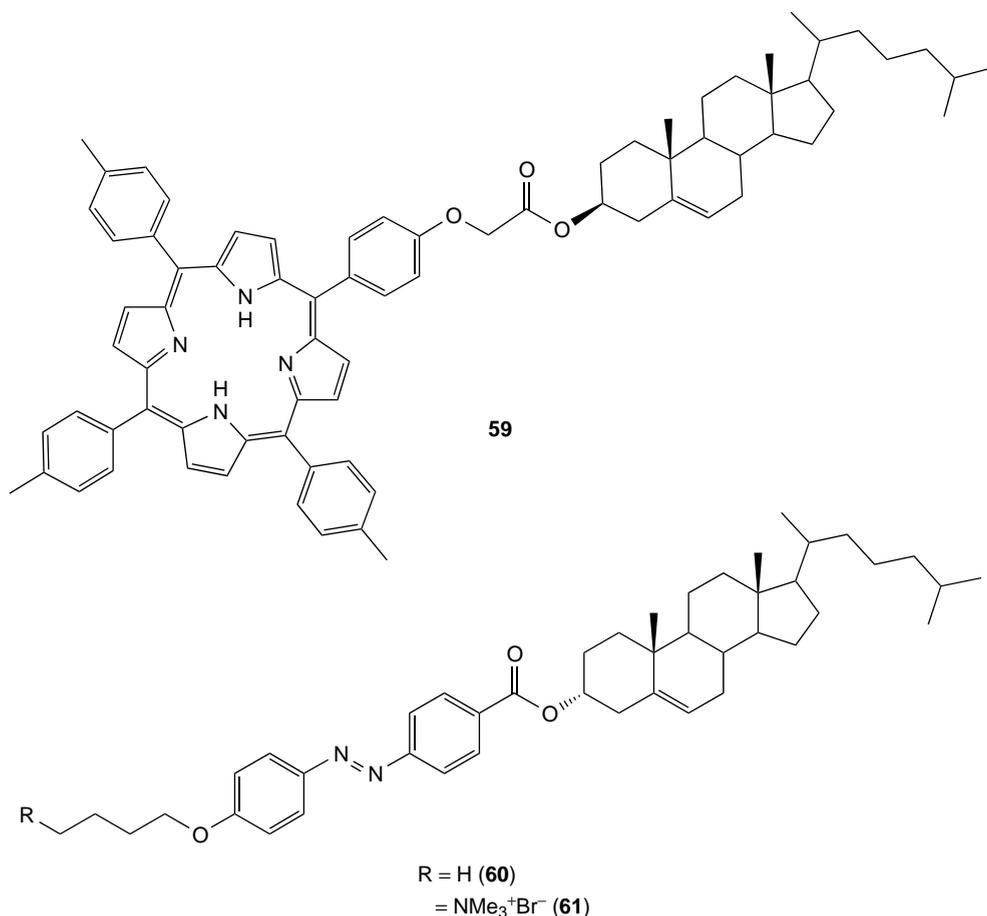


Figure 32. Porphyrin-cholesterol derivative **59** gels organic solvents and has potential photochemical properties. Mixtures of **60** and **61** are used as templates for preparing hollow fiber silica.

porphyrin assemblies. Shinkai also recently used organic cholesteryl gels as templates for preparing hollow fiber silica (78).

It was found that molecules such as **60** gel liquid silanol derivatives. The fibrous organic matrix can serve as a template in forming the silica gels. The pyrolytic removal of the organic matrix resulted in forming well-grown, fibrous silica whose tube edges contain cylindrical cavities 50–200 nm in diameter. When mixtures of **60** and **61** were used, a helical hollow fiber silica resulted (79). This finding is particularly interesting because it represents the transfer of chirality from a template into an inorganic matrix (see Fig. 32).

Maitra and co-workers generated a donor–acceptor organogelator by using bile acid derivatives, such as those shown in Fig. 33 (**80**). Compound **62** gels organic solvents (primarily alcohols) only in the presence of **63** (Fig. 33). The stoichiometry required for gelation is 1:1, and the gels are colored due to charge-transfer effects. The intensity of the charge-transfer band changes substantially during gelation suggesting that this interaction is important in the gelation process.

Amino Acid Gelators. Gelatin is a polypeptide that has a high content of glycine, proline, and 4-hydroxyproline.

It is usually obtained from denatured collagen. Therefore, it is not surprising that short peptide derivatives gel organic solvents. It is believed that the amide group in these systems plays a key hydrogen-bonding role.

N-Benzyloxycarbonyl-L-alanine-4-hexadecanoyl-2-nitrophenyl ester **64** can form thermoreversible gels at less than 1% concentration by mass in methanol or cyclohexane (81). Transmission and scanning electron microscopy revealed that the gels are formed by rod-like fibers that in turn are presumably assembled through N–H...O hydrogen-bonding between adjacent carbamate groups. Several analogs of **64** were studied by FT-IR and circular dichroism to determine the role of these interactions (82). The study suggests that π – π stacking, dipole–dipole interactions, and hydrophobic forces together with hydrogen bonding work cooperatively to form aggregates. The alkylamine of *N*-benzyloxycarbonyl-L-valyl-L-valine **65** can also gel a variety of solvents at very low concentrations (83). FT-IR and thermodynamic parameters suggest that the gels form by intermolecular hydrogen bonding between the N–H and C=O groups of the carbamate (see Fig. 34).

Hanabusa also reported diketopiperazines[cyclo(dipeptides)] as gelators for a variety of solvents, including edible oils, glyceryl esters, alcohols, and aromatic molecules

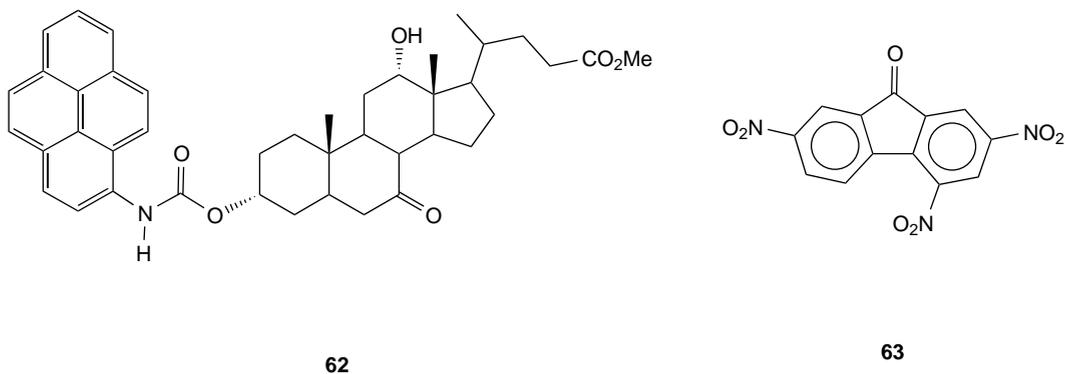


Figure 33. Donor–acceptor gels are formed by **62** in the presence of **63**.

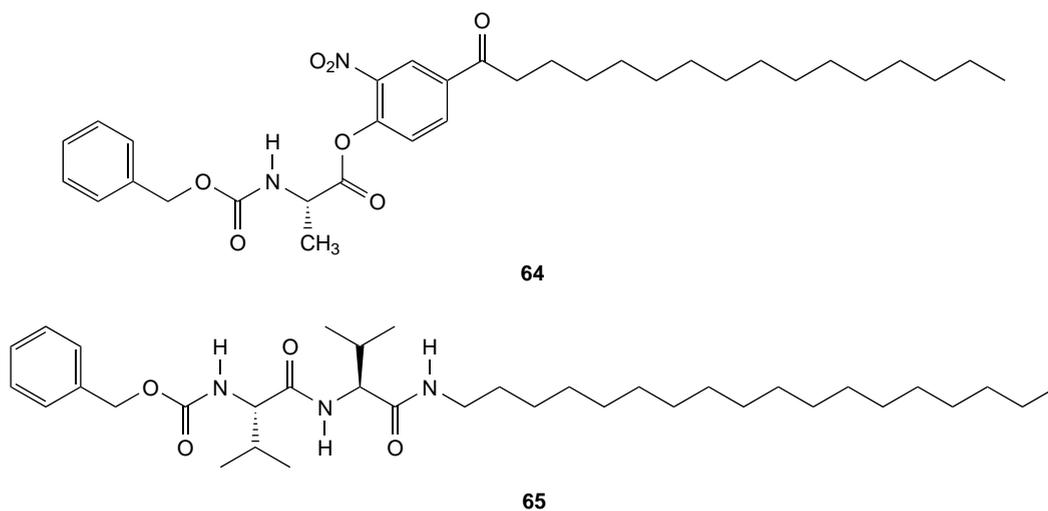


Figure 34. Examples of organogelators based on amino acid derivatives.

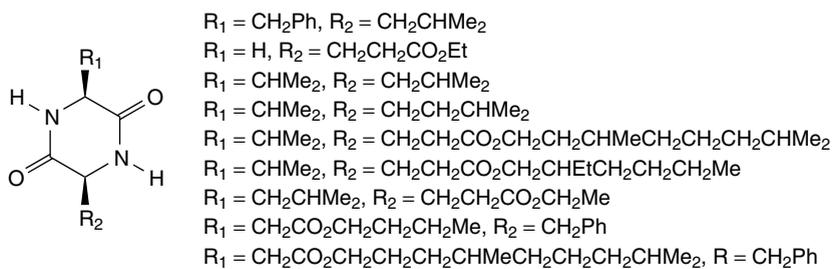


Figure 35. Cyclo(dipeptides) can gel a variety of organic solvents.

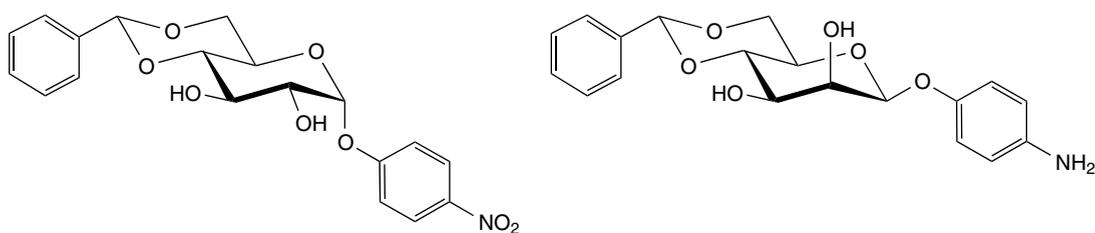


Figure 36. Examples of sugar-based organogelators.

(see Fig. 35) (84). An interesting feature is that even short alkyl chain derivatives are effective in gelling many solvents. However, introducing long chains does result in an increase in the gelation ability of the molecules.

Bhattacharya and co-workers reported a gelator family based on L-phenylalanine derivatives (85). Both mono and bis-carbamate derivatives of this amino acid were synthesized. This study shows a wide range of compounds. The investigation included many spacers between the carbamates. The gels were characterized by FT-IR, calorimetry, and X-ray diffraction studies.

Similar to the cholesteryl templates used to form hollow fiber silica, the amino acid derivative Z-L-Ile-NHC₁₈H₃₇ was used as a template in the polymerization of titanium tetraisopropoxide (86). Interestingly, after calcining the organic matrix, the result was porous titania that had fibrous structures.

Sugar-Based Gelators. The structure of the sugar component seems to play an important role in the gelation properties of sugar-based organogelators (87). In compounds that contain *p*-nitrophenyl groups, the chirality of the helical fibers formed by the gelator was monitored by CD spectroscopy (88). The *p*-aminophenyl derivatives were synthesized to have reinforce the formation of the gels through a metal coordination effect (Fig. 36) (89).

CONCLUSION

In this article, we have presented the development of organogelation from its initial stages of identifying intermolecular interactions through the development of functional organogelators. The applications of these materials range from the absorption of different solvents to a role as a template in forming fibrous silica. Microcellular organic materials formed by noncovalent interactions have also been prepared from gels by removing the solvent from their structure. The study of gels has been challenging, but the potential rewards in terms of understanding this form of matter between solid and liquid are considerable. The relative ease with which organogelation can be generated, the partial order that is present in their structure, and the potential for novel functionalization and reactivity all suggest that organogels will find many industrial and academic applications in the coming years.

ACKNOWLEDGEMENT

We thank the National Science Foundation for its support of our work in this area.

BIBLIOGRAPHY

1. R.B. Prince, T. Okada, and J.S. Moore, *Angew. Chem. Int. Ed.* **38**: 233 (1999).
2. M.S. Gin, T. Yokozawa, R.B. Prince, and J.S. Moore, *J. Am. Chem. Soc.* **121**: 2643 (1999).
3. Y. Aoyama, K. Endo, T. Anzai, Y. Yamaguchi, T. Sawaki, K. Kobayashi, N. Kanehisa, H. Hashimoto, Y. Kai, and H. Masuda, *J. Am. Chem. Soc.* **118**: 5562 (1996).
4. A.R.A. Palmans, J.A.J.M. Vekemans, E.E. Havinga, and E.W. Meijer, *Angew. Chem. Int. Ed. Engl.* **36**: 2648 (1997).
5. A.R.A. Palmans, J.A.J.M. Vekemans, R.A. Hikmet, H. Fischer, and E.W. Meijer, *Adv. Mater.* **10**: 873 (1998).
6. R.P. Sijbesma, F.H. Beijer, L. Brunsveld, B.J.B. Folmer, J.H.K.K. Hirschberg, R.F.M. Lange, J.K.L. Lowe, and E.W. Meijer, *Science* **278**: 1601 (1997).
7. J.H. Van Esch and B.L. Feringa, *Angew. Chem. Int. Ed.* **39**: 2263 (2000).
8. C. Shi, Z. Huang, S. Kilic, J. Xu, M. Enick, E.J. Beckman, A.J. Carr, R.E. Melendez, and A.D. Hamilton, *Science* **286**: 1540 (1999).
9. R. Taylor and O. Kennard, *Acta Crystallogr. Sect. B* **B39**: 133 (1983).
10. R. Taylor and O. Kennard, *Acc. Chem. Res.* **17**: 320 (1984).
11. M.C. Etter, *Acc. Chem. Res.* **23**: 120 (1990).
12. L. Leiserowitz, *Acta Crystallogr.* **B32**: 775 (1976).
13. J.C. MacDonald and G.M. Whitesides, *Chem. Rev.* **94**: 2383 (1994).
14. G.R. Desiraju, *Acc. Chem. Res.* **24**: 290 (1991).
15. G.R. Desiraju, *Acc. Chem. Res.* **29**: 441 (1996).
16. J. Emsley, *Chem. Soc. Rev.* **9**: 91 (1980).
17. T.B. McMahon and J.W. Larson, *J. Am. Chem. Soc.* **104**: 5848 (1982).
18. D. Philp and J.F. Stoddart, *Angew. Chem. Int. Ed. Engl.* **35**: 1154 (1996).
19. G.R. Desiraju, *Crystal Engineering, The Design of Organic Solids*. Elsevier, Amsterdam, 1989.
20. G.R. Desiraju, *The Crystal as a Supramolecular Entity*. Wiley, Chichester, 1995, Vol. 2.
21. M.J. Zaworotko, *Chem. Soc. Rev.* 283 (1994).
22. T.J. Murry and S.C. Zimmerman, *J. Am. Chem. Soc.* **114**: 4010 (1992).
23. W.L. Jorgensen and J. Pranata, *J. Am. Chem. Soc.* **112**: 2008 (1990).
24. J. Sartorius and H.-J. Schneider, *Chem. Eur. J.* **2**: 1446 (1996).
25. X. Sun and G.P. Lorenzi, *Helv. Chim. Acta.* **77**: 1520 (1994).
26. J.C. Adrian Jr. and C.S. Wilcox, *J. Am. Chem. Soc.* **113**: 678 (1991).
27. M. Berger and F.P. Schmidtchen, *Angew. Chem. Int. Ed.* **37**: 2694 (1998).
28. C.A. Hunter and J.K.M. Sanders, *J. Am. Chem. Soc.* **112**: 5525 (1990).
29. S.K. Burley and G.A. Petsko, *J. Am. Chem. Soc.* **108**: 7995 (1986).
30. T. Steiner, *J. Chem. Soc. Chem. Commun.* 95 (1995).
31. A.I. Kitaigorodskii, *Molecular Crystals and Molecules*. Academic Press, NY, 1973.
32. J. Perlstein, *J. Am. Chem. Soc.* **114**: 1955 (1992).
33. J. Perlstein, *J. Am. Chem. Soc.* **116**: 455 (1994).
34. K. Almdal, J. Dyre, S. Hvidt, and O. Kramer, *Polym. Gels Networks* **1**: 5 (1993).
35. P. Terech, R.G. Weiss, *Chem. Rev.* **97**: 3133 (1997).
36. J. van Esch, F. Schoonbeek, M. de Loos, E.M. Veen, R.M. Kellogg, and B.L. Feringa, in *Low Molecular Weight Gelators for Organic Solvents*, R. Ungaro and E. Dalcanale, eds., Kluwer Academic, Netherlands, 1999, p. 233.
37. R.E. Melendez, A.J. Carr, B.R. Linton, A.D. Hamilton, in *Controlling Hydrogen Bonding: From Molecular Recognition to Organogelation*, M. Fujita, ed., Springer-Verlag, 2000, Vol. 96, p. 31.

38. P. Terech, C. Rossat, and F. Volino, *J. Colloid Interface Sci.* **227**: 363 (2000).
39. A.T. Polishuk, *J. Am. Soc. Lubr. Eng.* **33**: 133 (1977).
40. Y. Uzu, *J. Jpn. Oil Chem. Soc.* **24**: 261 (1975).
41. L. Lu and R.G. Weiss, *Chem. Commun.* 2029 (1996).
42. R. Oda, I. Huc, and S.J. Candau, *Angew. Chem. Int. Ed.* **37**: 2689 (1998).
43. F. Placin, M. Colomès, and J.-P. Desvergne, *Tetrahedron Lett.* **38**: 2665 (1997).
44. G.M. Clavier, J.-F. Brugger, H. Bouas-Laurent, J.-L. Pozzo, *J. Chem. Soc. Perkin Trans. 2*: 2527 (1998).
45. L. Leiserowitz and G.M.J. Schmidt, *J. Chem. Soc.* 2372 (1969).
46. Y. Ducharme and J.D. Wuest, *J. Org. Chem.* **53**: 5787 (1988).
47. K. Hanabusa, M. Yamada, M. Kimura, and H. Shirai, *Angew. Chem. Int. Ed. Engl.* **35**: 1949 (1996).
48. E. Fan, J. Yang, S.J. Geib, T.C. Stoner, M.D. Hopkins, and A.D. Hamilton, *J. Chem. Soc. Chem. Commun.* 1251 (1995).
49. Y. Yasuda, E. Iishi, H. Inada, and Y. Shirota, *Chem. Lett.* 575 (1996).
50. Y. Yasuda, Y. Takebe, M. Fukumoto, H. Inada, and Y. Shirota, *Adv. Mater.* **8**: 740 (1996).
51. K. Hanabusa, A. Kawakami, M. Kimura, and H. Shirai, *Chem. Lett.* 191 (1997).
52. T. Kato, T. Kutsuna, K. Hanabusa, and M. Ukon, *Adv. Mater.* **10**: 606 (1998).
- 53a. M. Masuda, T. Hanada, K. Yase, T. Shimizu, *Macromolecules* **31**: 9403 (1998).
- 53b. K. Inoue, Y. Ono, Y. Kanekiyo, K. Hanabusa, and S. Shinkai, *Chem. Lett.* 429 (1999).
55. N. Tamaoki, S. Shimada, Y. Okada, A. Belaïssaoui, G. Kruk, K. Yase, and H. Masada, *Langmuir* **16**: 7545 (2000).
56. C. Raposo, M. Crego, M.L. Mussons, M.C. Caballero, and J.R. Moran, *Tetrahedron Lett.* **35**: 3409 (1994).
57. B.C. Hamann, N.R. Branda, and J. Rebek Jr., *Tetrahedron Lett.* **34**: 6837 (1993).
58. M.C. Etter, Z. Urbańczyk-Lipkowska, M. Zia-Ebrahimi, and T.W. Panunto, *J. Am. Chem. Soc.* **112**: 8415 (1990).
59. X. Zhao, Y.-L. Chang, F.W. Fowler, and J.W. Lauher, *J. Am. Chem. Soc.* **112**: 6627 (1990).
60. Y.-L. Chang, M.-A. West, F.W. Fowler, and J.W. Lauher, *J. Am. Chem. Soc.* **115**: 5991 (1993).
61. K. Hanabusa, K. Shimura, K. Hirose, M. Kimura, and H. Shirai, *Chem. Lett.* 885 (1996).
62. J. van Esch, R.M. Kellogg, and B.L. Feringa, *Tetrahedron Lett.* **38**: 281 (1997).
63. A.J. Carr, R. Melendez, S.J. Geib, and A.D. Hamilton, *Tetrahedron Lett.* **39**: 7447 (1998).
64. M. de Loos, J. van Esch, I. Stokroos, R.M. Kellogg, and B.L. Feringa, *J. Am. Chem. Soc.* **119**: 12675 (1997).
65. F.S. Schoonbeek, J.H. van Esch, B. Wegewijs, D.B.A. Rep, M. P. de Haas, T.M. Klapwijk, R.M. Kellogg, and B.L. Feringa, *Angew. Chem. Int. Ed.* **38**: 1393 (1999).
66. J. van Esch, F. Schoonbeek, M. de Loos, H. Kooijman, A. L. Spek, R.M. Kellogg, and B.L. Feringa, *Chem. Eur. J.* **5**: 937 (1999).
67. J.-M. Lehn, M. Mascal, A. DeCian, and J. Fischer, *J. Chem. Soc. Perkin Trans. 2*: 1992, 461–467 (1992).
68. K. Hanabusa, T. Miki, Y. Taguchi, T. Koyama, and H. Shirai, *J. Chem. Soc. Chem. Commun.* 1382 (1993).
69. S.-K. Chang and A.D. Hamilton, *J. Am. Chem. Soc.* **110**: 1318 (1988).
70. P. Tecilla, V. Jubian, and A.D. Hamilton, *Tetrahedron* **51**: 435 (1995).
71. K. Inoue, Y. Ono, Y. Kanekiyo, T. Ishi-i, K. Yoshihara, and S. Shinkai, *J. Org. Chem.* **64**: 2933 (1999).
72. Y.-C. Lin, B. Kachar, and R.G. Weiss, *J. Am. Chem. Soc.* **111**: 5542 (1989).
73. P. Terech, I. Furman, and R.G. Weiss, *J. Phys. Chem.* **99**: 9558 (1995).
74. C. Geiger, M. Stanescu, L. Chen, and D.G. Whitten, *Langmuir* **15**: 2241 (1999).
75. R. Wang, C. Geiger, L. Chen, B. Swanson, and D.G. Whitten, *J. Am. Chem. Soc.* **122**: 2399 (2000).
76. K. Murata, M. Aoki, T. Suzuki, T. Harada, H. Kawabata, T. Komori, F. Ohseto, K. Ueda, and S. Shinkai, *J. Am. Chem. Soc.* **116**: 6664 (1994).
77. S. Won and S. Shinkai, *Nanotechnology* **8**: 179 (1997).
78. H.J. Tian, K. Inoue, K. Yoza, T. Ishi-i, and S. Shinkai, *Chem. Lett.* 871 (1998).
79. Y. Ono, K. Nakashima, M. Sano, Y. Kanekiyo, K. Inoue, J. Hojo, and S. Shinkai, *Chem. Commun.* 1477 (1998).
80. Y. Ono, K. Nakashima, M. Sano, J. Hojo, and S. Shinkai, *Chem. Lett.* 1119 (1999).
81. U. Maitra, P.V. Kumar, N. Chandra, L.J. D'Souza, M.D. Prasanna, and A.R. Raju, *Chem. Commun.* 595 (1999).
82. K. Hanabusa, K. Okui, K. Karaki, T. Koyama, and H. Shirai, *J. Chem. Soc. Chem. Commun.* 137 (1992).
83. K. Hanabusa, K. Okui, K. Karaki, M. Kimura, and H. Shirai, *J. Colloid Interface Sci.* **195**: 86 (1997).
84. K. Hanabusa, J. Tange, Y. Taguchi, T. Koyama, and H. Shirai, *J. Chem. Soc. Chem. Commun.* 390 (1993).
85. K. Hanabusa, Y. Matsumoto, T. Miki, T. Koyama, and H. Shirai, *J. Chem. Soc. Chem. Commun.* 1401 (1994).
86. S. Bhattacharya and S.N.G. Acharya, *Chem. Mater.* **11**: 3121 (1999).
87. S. Kobayashi, K. Hanabusa, M. Suzuki, M. Kimura, and H. Shirai, *Chem. Lett.* 1077 (1999).
88. K. Yoza, Y. Ono, K. Yoshihara, T. Akao, H. Shinmori, M. Takeuchi, S. Shinkai, and D.N. Reinhoudt, *Chem. Commun.* 907 (1998).
89. N. Amanokura, K. Yoza, H. Shinmori, S. Shinkai, and D.N. Reinhoudt, *J. Chem. Soc. Perkin Trans. 2*: 2585 (1998).
90. N. Amanokura, Y. Kanekiyo, S. Shinkai, and D.N. Reinhoudt, *J. Chem. Soc. Perkin Trans. 2*: 1995 (1999).

GELS

ANTHONY M. LOWMAN
 THOMAS D. DZIUBLA
 Drexel University
 Philadelphia, PA

INTRODUCTION

Hydrogels are three-dimensional, water-swollen structures composed mainly of hydrophilic homopolymers or copolymers (1). They are rendered insoluble by chemical

or physical cross-links. The physical cross-links can be entanglements, crystallites, or weak associations such as van der Waals forces or hydrogen bonds. Cross-links provide the network structure and physical integrity. It is possible to design hydrogels whose swelling behavior depends on the external environment. During the last thirty years, there has been significant interest in the development and analysis of environmentally or physiologically responsive hydrogels (2).

Hydrogels are classified in a number of ways (1,3). They can be neutral or ionic based on the nature of the side groups. They can also be classified on the basis of network morphology as amorphous, semicrystalline, hydrogen-bonded structures; supermolecular structures; and hydrocolloidal aggregates. Additionally, in terms of their network structures, hydrogels can be classified as macroporous, microporous, or nonporous (1,3,4).

Since the development of poly(2-hydroxyethyl methacrylate) gels in the early 1960s, hydrogels have been considered for a wide range of applications, most notably for biomedical and pharmaceutical devices, due mainly to their high water content and rubbery nature which resembles natural living tissue more than any other class of synthetic biomaterials (1,5). Furthermore, the high water content gives these materials excellent biocompatibility. Some applications of hydrogels include contact lenses, biosensors, sutures, dental materials, and controlled drug delivery (1,5–9). Two of the most important characteristics in evaluating the ability of a polymeric gel to function in a particular controlled release application are the network permeability and the swelling behavior. The permeability and swelling behavior of hydrogels depend strongly on the chemical nature of the polymer(s) that composes the gel and on the structure and morphology of the network. As a result, there are different mechanisms that control the release of drugs from hydrogel-based delivery devices. Such systems are classified by their drug release mechanism as diffusional-controlled release systems, swelling-controlled release systems, chemically controlled release systems, and environmentally responsive systems (3).

Hydrogels may exhibit swelling behavior that depends on the external environment. Thus, in the last 30 years, there has been major interest in developing and analyzing these smart or responsive hydrogels (2). These hydrogels may drastically change their swelling ratios due to changes in their external pH, temperature, ionic strength, nature of the swelling agent, and electromagnetic radiation. Some advantages of environmentally or physiologically responsive hydrogels are relevant in numerous drug delivery applications, such as chemomechanical systems and drug targeting. In an exceptional new review in the field, am Ende and Mikos (10) offer a thorough physicochemical and mathematical interpretation of the conditions for diffusional release of various bioactive agents in hydrogels and other polymeric carriers; they emphasize the conditions of stability of peptides and proteins during delivery.

In this article, we present an overview of smart hydrogels and models that characterize the structure and

properties of these materials. Additionally, we review important contributions in this area and emphasize new synthetic methods that promise to lead to the development of exciting new structures for solutions to drug delivery problems.

STRUCTURE AND PROPERTIES OF HYDROGELS

To evaluate the feasibility of using a hydrogel in a particular application, it is important know the structure and properties of the polymer network. The structure of an idealized hydrogel is shown in Fig. 1. The most important parameters that define the structure and properties of swollen hydrogels are the polymer volume fraction in the swollen state $\nu_{2,s}$, the effective molecular weight of the polymer chain between cross-linking points \bar{M}_c , and the network mesh or pore size ξ (11).

The volume fraction of the polymer in the swollen gel is a measure of the amount of fluid that a hydrogel can incorporate in its structure:

$$\nu_{2,s} = \frac{\text{volume of polymer}}{\text{volume of swollen gel}} = \frac{V_p}{V_{\text{gel}}} = 1/Q. \quad (1)$$

This parameter can be determined by equilibrium swelling experiments (12). The molecular weight between cross-links is the average molecular weight of the polymer chains between chemical and physical junction points. This parameter provides a measure of the degree of cross-linking in the gel. This value is related to the degree of cross-linking in the gel X as

$$X = \frac{M_0}{2\bar{M}_c} \quad (2)$$

Here, M_0 is the molecular weight of the repeating units that make up the polymer chains.

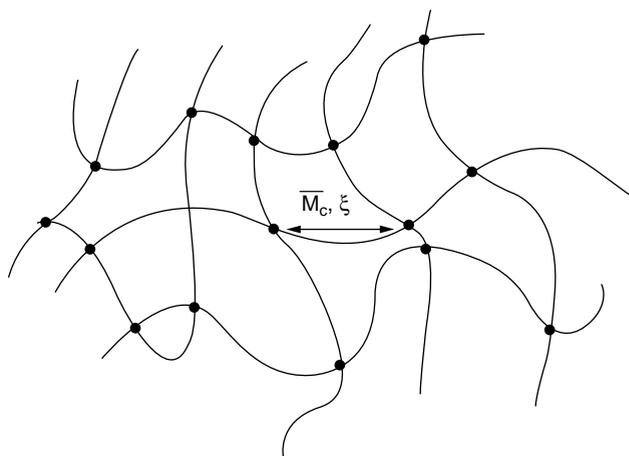


Figure 1. Schematic representation of the cross-linked structure of a hydrogel. \bar{M}_c is the molecular weight of the polymer chains between cross-links (\bullet), and ξ is the network mesh size.

The network mesh size represents the distance between consecutive cross-linking points and provides a measure of the porosity of the network. These parameters, which are not independent, can be determined theoretically or by a variety of experimental techniques.

Equilibrium Swelling Theories

Neutral Hydrogels. Flory and Rehner (13) developed the initial depiction of the swelling of cross-linked polymer gels using a Gaussian distribution of the polymer chains. They developed a model to describe the equilibrium degree of swelling of cross-linked polymers. The model postulated that the degree to which a polymer network swells is governed by the elastic retractive forces of the polymer chains and the thermodynamic compatibility of the polymer and the solvent molecules. In terms of the free energy of the system, the total free energy change upon swelling is written as

$$\Delta G = \Delta G_{\text{elastic}} + \Delta G_{\text{mix}}. \quad (3)$$

Here, $\Delta G_{\text{elastic}}$ is the contribution of the elastic retractive forces and ΔG_{mix} represents the thermodynamic compatibility of the polymer and the swelling agent. Upon evaluating each term at equilibrium, the swelling behavior for hydrogels cross-linked in the absence of a solvent can be described by the following equation:

$$\frac{1}{\bar{M}_c} = \frac{2}{\bar{M}_n} - \frac{(v/V_1)[\ln(1 - v_{2,s}) + v_{2,s} + \chi_1 v_{2,s}]}{\left(v_{2,s}^{1/3} - \frac{v_{2,s}}{2}\right)}, \quad (4)$$

where v is the specific volume of the polymer, V_1 is the molar volume of the swelling agent, χ_1 is the polymer solvent interaction parameter, and \bar{M}_n is the molecular weight of linear polymer chains prepared without a cross-linking agent under the same conditions.

In many cases, it is desirable to prepare hydrogels in the presence of a solvent. If the polymers were cross-linked in the presence of a solvent, the elastic contributions must account for the volume fraction density of the chains during cross-linking. Peppas and Merrill (14) modified the original Flory–Rehner to account for the changes in the elastic contributions to swelling. The equilibrium swelling equation for polymer gels cross-linked in the presence of a solvent is

$$\frac{1}{\bar{M}_c} = \frac{2}{\bar{M}_n} - \frac{(v/V_1)[\ln(1 - v_{2,s}) + v_{2,s} + \chi_1 v_{2,s}]}{v_{2,r}[(v_{2,s}/v_{2,r})^{1/3} - (v_{2,s}/2v_{2,r})]}. \quad (5)$$

Here, $v_{2,r}$ is the volume fraction of the polymer in the relaxed state. The relaxed state of the polymer is defined as its state immediately after cross-linking but before swelling or deswelling. By performing swelling experiments to determine $v_{2,s}$, the molecular weight between cross-links can be calculated for a particular gel by using this equation (12).

Ionic Hydrogels. Ionic hydrogels contain pendent groups that are either cationic or anionic. The side groups of

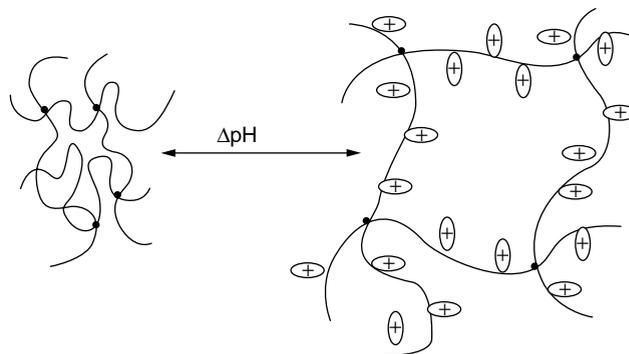


Figure 2. Expansion (swelling) of a cationic hydrogel due to ionization of pendent groups at specific pH values.

anionic gels are unionized below the pK_a , and the swelling of the gel is governed by the thermodynamic compatibility of the polymer and the swelling agent. However, above the pK_a of the network, the pendent groups are ionized, and the gels swell to a large degree due from the development of a large osmotic swelling force due to the presence of the ions. In cationic gels, the pendent groups are not ionized above the pK_b of the network. When the gel is placed in a fluid whose pH is less than this value, the basic groups are ionized, and the gels swell to a large degree (Fig. 2). The swelling behavior of ionic gels is depicted in Fig. 3. Note that this behavior can be completely reversible.

The theoretical description of the swelling of ionic hydrogels is much more complex than that of neutral hydrogels. Aside from the elastic and mixing contributions to swelling, the swelling of an ionic hydrogel is affected by the degree of ionization in the gel, the ionization equilibrium between the gel and the swelling agent, and the nature of the counterions in the fluid. As the ionic content of a hydrogel is increased in response to an environmental stimulus, increased repulsive forces develop, and the network becomes more hydrophilic. The result is a more highly swollen network. Because of Donnan equilibrium, the chemical potential of the ions inside the gel must be

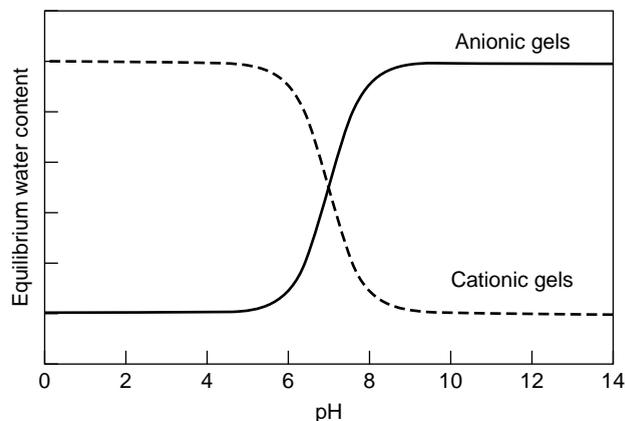


Figure 3. Equilibrium degree of swelling of anionic and cationic hydrogels as a function of the swelling solution pH.

equal to the chemical potential of the ions in the solvent outside the gel (15). An ionization equilibrium is established as a double layer of fixed charges on the pendent groups and the counterions in the gel. Finally, the nature of counterions in the solvent affects the swelling of the gel. As the valence of the counterions increases, they are more strongly attracted to the gel and reduce the concentration of ions needed in the gel to satisfy Donnan equilibrium conditions.

The swelling behavior of polyelectrolyte gels was initially described as a result of a balance between the elastic energy of the network and the osmotic pressure developed as a result of the ions (15–22). In electrolytic solutions, the osmotic pressure is associated with the development of a Donnan equilibrium. This pressure term is also affected by the fixed charges developed on the pendent chains. The elastic term is described by the Flory expression derived from assumptions of Gaussian chain distributions.

Models were developed for the swelling of ionic hydrogels by equating the three major contributions to the swelling of the networks. These contributions are due to mixing of the polymer and solvent, network elasticity, and ionic contributions. The general equation is given as

$$\Delta G = \Delta G_{\text{mix}} + \Delta G_{\text{el}} + \Delta G_{\text{ion}}. \quad (6)$$

In terms of the chemical potential, the difference between the chemical potential of the swelling agent in the gel and outside the gel is

$$\mu_1 - \mu_{1,0} = (\Delta\mu)_{\text{elastic}} + (\Delta\mu)_{\text{mix}} + (\Delta\mu)_{\text{ion}}. \quad (7)$$

The elastic contribution and mixing contributions of weakly charged polyelectrolytes will not differ from those of nonionic gels. However, for highly ionizable materials, ionization effects are significant, and $\Delta\mu_{\text{ion}}$ is important. At equilibrium, the elastic, mixing, and ionic contributions must sum to zero.

The ionic contribution to the chemical potential depends strongly on the ionic strength and the nature of the ions. Both Brannon-Peppas and Peppas (21,22) and Ricka and Tanaka (15) developed expressions to describe ionic contributions to the swelling of polyelectrolytes. Assuming that polymer networks under conditions of swelling behave similarly to dilute polymer solutions, activity coefficients can be approximated as one, and activities can be replaced by concentrations. Under these conditions, the ionic contribution to the chemical potential is described by the following:

$$(\Delta\mu)_{\text{ion}} = RTV_1 \Delta c_{\text{tot}}. \quad (8)$$

Here, Δc_{tot} is the difference in the total concentration of mobile ions within the gel. The difference in the concentration of mobile ions is due to the fact that the charged polymer requires that the same number of counterions remain in the gel to achieve electroneutrality. The difference in the total ion concentration could then be calculated from the equilibrium condition for the salt.

Brannon-Peppas and Peppas (21,22) developed expressions for the ionic contributions to the swelling of polyelectrolytes for anionic and cationic materials. The ionic contribution for an anionic network is

$$(\Delta\mu)_{\text{ion}} = \frac{RTV_1}{4I} \left(\frac{v_{2,s}^2}{\nu} \right) \left(\frac{K_a}{10^{-\text{pH}} + K_a} \right)^2. \quad (9)$$

The ionic contribution for a cationic network is

$$(\Delta\mu)_{\text{ion}} = \frac{RTV_1}{4I} \left(\frac{v_{2,s}^2}{\nu} \right) \left(\frac{K_b}{10^{\text{pH}-14} + K_b} \right)^2. \quad (10)$$

In these expressions, I is the ionic strength and K_a and K_b are the dissociation constants for the acid and base, respectively. It is significant to note that this expression has related the ionic contribution to the chemical potential to characteristics about the polymer/swelling agent that are readily determinable (e.g., pH, K_a , and K_b).

The equilibrium swelling of anionic polymer gels that were crosslinked in the presence of a solvent can be described by

$$\begin{aligned} \frac{V_1}{4I} \left(\frac{v_{2,s}^2}{\nu} \right) \left(\frac{K_a}{10^{-\text{pH}} + K_a} \right)^2 &= [\ln(1 - v_{2,s}) + v_{2,s} + \chi_1 v_{2,s}] \\ &+ \left(\frac{V_1}{\nu \bar{M}_c} \right) \left(1 - \frac{2\bar{M}_c}{\bar{M}_n} \right) v_{2,r} \left[\left(\frac{v_{2,s}}{v_{2,r}} \right)^{1/3} - \frac{v_{2,s}}{2v_{2,r}} \right]. \end{aligned} \quad (11)$$

The equilibrium swelling of cationic hydrogels prepared in the presence of a solvent is described by the following expression:

$$\begin{aligned} \frac{V_1}{4I} \left(\frac{v_{2,s}^2}{\nu} \right) \left(\frac{K_b}{10^{\text{pH}-14} + K_b} \right)^2 &= [\ln(1 - v_{2,s}) + v_{2,s} + \chi_1 v_{2,s}] \\ &+ \left(\frac{V_1}{\nu \bar{M}_c} \right) \left(1 - \frac{2\bar{M}_c}{\bar{M}_n} \right) v_{2,r} \left[\left(\frac{v_{2,s}}{v_{2,r}} \right)^{1/3} - \frac{v_{2,s}}{2v_{2,r}} \right]. \end{aligned} \quad (12)$$

The molecular weights of ionic hydrogels between crosslinks can be calculated by performing swelling experiments and applying Eqs. (11) (anionic gels) or (12) (cationic gels).

Rubber Elasticity Theory

Hydrogels are similar to natural rubbers; they can respond to applied stresses nearly instantaneously (23,24). These polymer networks can deform readily under low stresses. Following small deformations (less than 20%), most gels can also fully recover rapidly from the deformation. Under these conditions, the behavior of the gels can be approximated as elastic. This property can be exploited to calculate the cross-linking density or molecular weight between cross-links for a particular gel.

The elastic behavior of cross-linked polymers has been analyzed using classical thermodynamics, statistical thermodynamics, and phenomenological models. Based on classical thermodynamics and assuming an isotropic

system, the stress–strain behavior of rubbers can be expressed as (23,24)

$$\tau = \frac{\rho RT}{\bar{M}_c} \left(1 - \frac{2\bar{M}_c}{\bar{M}_n} \right) \left(\alpha - \frac{1}{\alpha^2} \right), \quad (13)$$

where τ is the stress, ρ is the density of the polymer, and α is the elongation ratio defined as the elongated length divided by the initial length of the sample.

Similar analysis can be applied to swollen hydrogels when the polymer volume fraction in the network is less than unity. The number of cross-links per unit volume of swollen gels is different from that of unswollen polymer networks. The equation of state for real, swollen polymer gels can be expressed as

$$\tau = \frac{\rho RT}{\bar{M}_c} \left(1 - \frac{2\bar{M}_c}{\bar{M}_n} \right) \left(\alpha - \frac{1}{\alpha^2} \right) \nu_{2,s}^{-1/3}. \quad (14)$$

This equation holds for networks cross-linked in the absence of any diluents. The equation of state of a swollen network cross-linked in the presence of a solvent can be written as (25)

$$\tau = \frac{\rho RT}{\bar{M}_c} \left(1 - \frac{2\bar{M}_c}{\bar{M}_n} \right) \left(\alpha - \frac{1}{\alpha^2} \right) \left(\frac{\nu_{2,s}}{\nu_{2,r}} \right)^{1/3}. \quad (15)$$

For elongation of a polymer network along a single axis, the stress is inversely proportional to the molecular weight between cross-links in the polymer network. Important structural information about polymer networks can be obtained from mechanical analysis under short deformations (26–28).

Network Pore Size Calculation

One of the most important parameters in controlling the rate release of a drug from a hydrogel is the network pore size ξ . This is particularly important for smart gels because pore size changes dramatically as environmental conditions shift. Pore size can be determined theoretically or by using a number of experimental techniques. Direct techniques for measuring this parameter are quasi-elastic laser-light scattering (29) and electron microscopy. Some indirect experimental techniques for determining the hydrogel pore size include mercury porosimetry (30,31), rubber elasticity measurements (26), and equilibrium swelling experiments (12,32). However, indirect experiments allow

calculating the porous volume (mercury porosimetry) and the molecular weight between cross-links (rubber elasticity analysis or swelling experiments).

Based on values for the crosslinking density or molecular weight between crosslinks, the network pore size can be determined by calculating the end-to-end distance of the swollen polymer chains between cross-linking points (12,23,32):

$$\xi = \alpha (r_0^{-2})^{1/2}. \quad (16)$$

In this expression, α is the elongation of the polymer chains in any direction and $(r_0^{-2})^{1/2}$ is the unperturbed end-to-end distance of the polymer chains between cross-linking points. Assuming isotropic swelling of the gels and using the Flory characteristic ratio C_n for calculating the end-to-end distance, the pore size of a swollen polymeric network can be calculated from the following equation (1):

$$\xi = \left(\frac{2C_n \bar{M}_c}{M_0} \right)^{1/2} l \nu_{2,s}^{-1/3} \quad (17)$$

where l is the length of the bond along the backbone chain (1.54 Å for vinyl polymers) and M_0 is the molecular weight of the monomeric repeating units.

CLASSIFICATIONS

Responsive hydrogels are unique. There are many different mechanisms for drug release, and many different types of release systems based on these materials. For instance, drug release in most cases occurs when the gel is highly swollen or swelling and is typically controlled by gel swelling, drug diffusion, or a coupling of swelling and diffusion. However, in a few instances, drug release occurs during gel syneresis by a squeezing mechanism. Drug release can also occur due to erosion of the polymer caused by environmentally responsive swelling.

Another interesting characteristic of many responsive gels is that the mechanism that causes the network structural changes can be entirely reversible. This behavior of a pH- or temperature-sensitive gel is depicted in Fig. 4. The ability of these materials to exhibit rapid changes in their swelling behavior and pore structure in response to changes in environmental conditions lends these materials favorable characteristics as carriers for bioactive agents, including peptides and proteins. This type of behavior may allow these materials to serve as self-regulated, pulsatile drug delivery systems. This type of behavior is shown

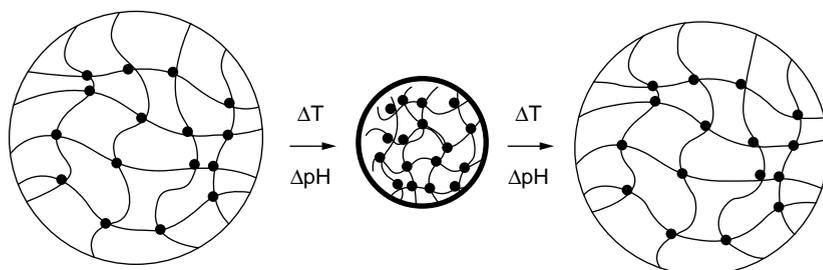


Figure 4. Swollen temperature- and pH-sensitive hydrogels may exhibit an abrupt change from the expanded (left) to the collapsed (syneresed) state (center) and then back to the expanded state (right), as temperature and pH change.

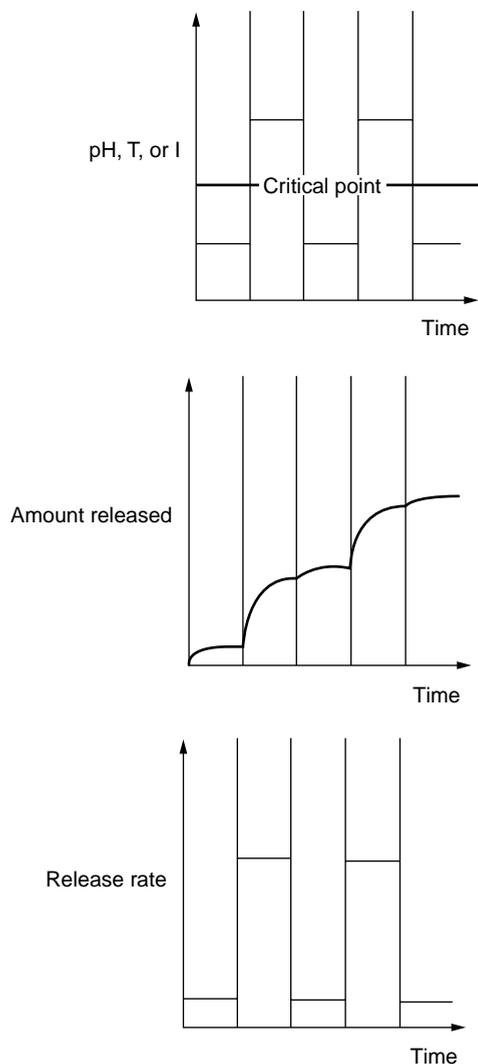


Figure 5. Cyclic change of pH, T or ionic strength I leads to abrupt changes in the drug release rates at certain time intervals in some environmentally responsive polymers.

in Fig. 5 for pH- or temperature-responsive gels. Initially, the gel is in an environment in which no swelling occurs. As a result, very little drug release occurs. However, when the environment changes and the gel swells, drug release is rapid. When the gel collapses as the environment changes, the release can be turned off again. This can be repeated through numerous cycles. Such systems could be of extreme importance in treating chronic diseases such as diabetes. Peppas (33) and Siegel (34) have presented detailed analyses of this type of behavior.

pH-Sensitive Hydrogels

One of the most widely studied types of physiologically responsive hydrogels is pH-responsive hydrogels. These hydrogels are swollen from ionic networks that contain either acidic or basic pendent groups. In aqueous media of appropriate pH and ionic strength, the pendent groups can ionize and develop fixed charges on the gel, as shown in Fig. 2 for

an ionic gel. The swelling behavior of these materials has been analyzed in a previous section.

There are many advantages of using ionic materials in neutral networks. All ionic materials exhibit sensitivity to pH and ionic strength. The swelling forces developed in these systems are increased compared to nonionic materials. This increase in swelling force is due to the localization of fixed charges on the pendent groups. As a result, the mesh size of the polymeric networks can change significantly from small pH changes. The drug diffusion coefficients and release rates in these materials will vary greatly as the environmental pH changes.

Temperature-Sensitive Hydrogels

Thermally sensitive polymers comprise another class of environmentally sensitive materials that are being targeted for drug delivery applications. This type of hydrogel exhibits temperature-sensitive swelling due to a change in the compatibility of the polymer/swelling agent across the temperature range of interest. Temperature-sensitive polymers typically exhibit lower critical solution temperatures (LCST), below which the polymer is soluble. Above this temperature, the polymers are typically hydrophobic and do not swell significantly in water (35). However, below the LCST, the cross-linked gel swells to significantly higher degrees because of increased compatibility with water. The rate of drug release of polymers that exhibit this sort of swelling behavior would depend on the temperature. The highest release rates would occur when the temperature of the environment is below the LCST of the gel.

Complexing Hydrogels

Some hydrogels may exhibit environmental sensitivity due to the formation of polymer complexes. Polymer complexes are insoluble, macromolecular structures formed by the noncovalent association of polymers that have affinity for one another. The complexes form due to the association of repeating units on different chains (interpolymer complexes) or on separate regions of the same chain (intrapolymer complexes). Polymer complexes are classified as stereocomplexes, polyelectrolyte complexes, and hydrogen bonded complexes by the nature of the association (36). The stability of the associations depends on such factors as the nature of the swelling agent, the temperature, the type of dissolution medium, the pH and ionic strength, the network composition and the structure, and the length of the interacting polymer chains.

Complex formation in this type of gel results in physical cross-links in the gel. As the degree of effective cross-linking is increased, the network mesh size and degree of swelling are significantly reduced. As a result, the rate of drug release in these gels decreases dramatically when interpolymer complexes are formed.

Materials Sensitive to Chemical or Enzymatic Reaction

More recently, researchers have proposed controlled release devices that can respond to specific chemicals or enzymes in the body (37,38). Under normal conditions, the structure of the hydrogel would be sufficient to prevent drug release. However, in the presence of specific

substances, chemical or enzymatic reactions could occur that may hydrolyze specific groups of the polymer chain and result in an increased pore structure and rate of drug release from the gel. If degradation of the gel occurs, the systems cannot be fully reversible.

In some instances, enzymes can be incorporated in the structure of the hydrogel. In the presence of specific chemicals, the enzyme could trigger a reaction which would change the microenvironment of the hydrogel. Changes in the local microenvironment (such as pH or temperature) could lead to gel swelling or collapse. In these situations, the release rates would be altered significantly. This type of system could be completely reversible.

Magnetically Responsive Systems

Magnetically responsive systems consist of polymers or copolymers that contain magnetic microbeads (39). These systems can be prepared from most polymers; however, the most commonly used copolymer for these systems is the hydrophobic polymer poly(ethylene-co-vinyl acetate). Such systems are not typically classified as hydrogels because they do not swell to any appreciable degree. The drug release mechanism of a typical magnetic system is shown in Fig. 6. The three-dimensional structure of these systems is such that no drug release can occur when no magnetic field is applied. However, when a magnetic field is applied, the microbeads pulsate and allow micropores to form.

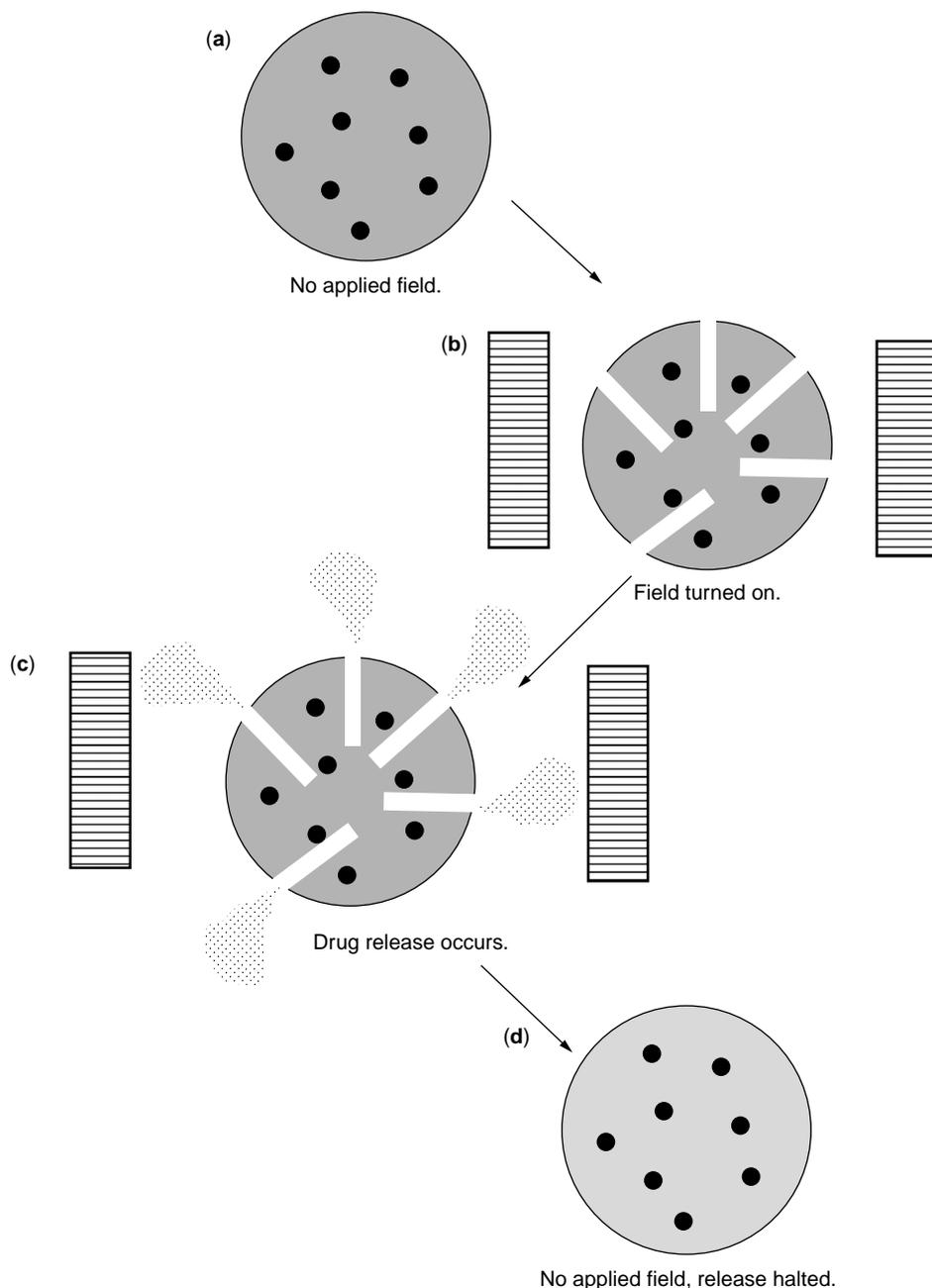


Figure 6. Schematic representation of drug release from a magnetically controlled system. Initially, when the field is off (a), no drug release occurs. When the field is turned on (b), the oscillations of the beads creates macropores through which drug release occurs (c). (d) Drug release halts when the field is switched off.

Additionally, the pulsation of the beads “squeezes” the drug out of the gel through these pores. When the field is removed, drug release halts rapidly. Such systems have been targeted for use as pulsatile drug delivery vehicles and, it has been shown, exhibit extremely reproducible behavior.

APPLICATIONS

pH-Sensitive Hydrogels

Hydrogels can respond to pH changes have been studied extensively over the years. These gels typically contain ionizable side groups such as carboxylic acids or amine groups (40,41). The most commonly studied ionic polymers include poly(acrylamide) (PAAm), poly(acrylic acid) (PAA), poly(methacrylic acid) (PMAA), poly(diethylaminoethyl methacrylate) (PDEAEMA), and poly(dimethylaminoethyl methacrylate) (PDMAEMA).

Cationic copolymers based on PDEAEMA and PDMAEMA have been studied by the Peppas and Siegel groups. The Siegel group focused on the swelling and transport behavior of hydrophobic cationic gels. Siegel and Firestone (42,43) studied the swelling behavior of hydrophobic hydrogels of PDMAEMA and poly(methyl methacrylate). Such systems were collapsed in solutions whose pH was higher than 6.6. However, in solutions less than pH 6.6, such systems swelled due to protonation of the tertiary amine groups. The release of caffeine from these gels was studied (44). No caffeine was released in basic solutions; however, in neutral or slightly acidic solutions, steady release of caffeine was observed for 10 days. More recently, Cornejo-Bravo and Siegel (45) investigated the swelling behavior of hydrophobic copolymers of PDEAEMA and PMMA. Additionally, Siegel (46) presented an excellent model of the dynamic behavior of ionic gels.

The Peppas groups studied the swelling behavior of more hydrophilic, cationic copolymers of P(DEAEMA-co-HEMA) and P(DMAEMA-co-HEMA) (47). These gels swelled in solutions less than pH 7 and were in the collapsed state in basic solutions. These materials swelled to a greater degree than those prepared by Siegel. These materials were used to modulate the release behavior of protein and peptide drugs (48). Schwarte and Peppas (49) studied the swelling behavior of copolymers of PDEAEMA grafted with PEG. The permeability of dextrans of molecular weight 4400 and 9400 was studied. The membrane permeabilities in the swollen membranes (pH 4.6) were two orders of magnitude greater than those of the collapsed membranes.

Anionic copolymers have received significant attention as well. The swelling and release characteristics of anionic copolymers of PMAA and PHEMA (PHEMA-co-MAA) have been investigated. The gels did not swell significantly in acidic media; however, in neutral or basic media, the gels swelled to a high degree due to ionization of the pendent acid group (50,51). Brannon-Peppas and Peppas also studied the oscillatory swelling behavior of these gels (52). Copolymer gels were transferred between acidic and basic solutions at specified time intervals. In acidic

solutions, the polymer swelled due to the ionization of the pendent groups. Rapid gel syneresis occurred in basic solutions. Brannon-Peppas and Peppas (53) modeled the time-dependent swelling response to pH changes by using a Boltzman superposition-based model. The pH-dependent release behavior of theophylline and proxiphylline from these anionic gels was also studied (54,55). Khare and Peppas (56) studied the pH-modulated release behavior of oxprenolol and theophylline from copolymers of PHEMA-co-MAA and PHEMA-co-AA. Drug release in neutral or basic media occurred rapidly by a non-Fickian mechanism. The release rate slowed significantly in acidic media. In another study, am Ende and Peppas (57) examined the transport of ionic drugs of varying molecular weight in PHEMA-co-AA. They compared experimental results to a free-volume based theory and found that deviations occurred due to interactions between the ionized backbone chains and pendent acid groups. The swelling and release behavior of interpenetrating polymer networks of PVA and PAA was also investigated (58,59). These materials also exhibit strong pH-responsive swelling. The permeability of these membranes was strongly dependent on the environmental pH and the size and ionic nature of the solute. Recent new studies have used ATR-FTIR spectroscopy to characterize the interactions between polyelectrolytes and solutes (59,60).

Heller et al. (61) studied the behavior of another type of pH-responsive hydrogel. In this work, they evaluated the pH-dependent release of insulin from degradable poly(ortho esters). Other researchers used chitosan (CS) membranes for drug delivery applications. These materials exhibited pH-dependent swelling due to gelation of CS upon contact with anions (62). Interpenetrating networks of CS and PEO have been proposed as drug delivery devices due to their pH-dependent swelling behavior (63,64). More recently, Calvo et al. (65) prepared novel CS-PEO microspheres. These systems were used to provide continuous release of entrapped bovine serum albumin for 1 week. In another study, methotrexate, an anticancer drug, was encapsulated in microspheres of pH-sensitive CS and alginate (66). Zero-order release of the drug was observed from the microspheres in pH 1.2 buffer for more than 1 week. Okano's group (67) studied pH-responsive calcium alginate gel beads. In these systems, modulated release of dextran was achieved by varying the pH and ionic strength of the environmental solution. Such systems may be promising for protein and peptide delivery applications.

An exciting application of pH-sensitive hydrogels is being pursued by Narayani et al. (68). They developed both anionic and cationic hydrogels using either alginate or chitosan as a comonomer with HEMA, respectively. They were able to demonstrate high loading efficiencies of the cancer agent methotrexate.

Temperature-Sensitive Hydrogels

Some of the earliest work on temperature-sensitive hydrogels was done by the Tanaka's group (69). They synthesized using cross-linked poly(*N*-isopropylacrylimide) (PNIPAAm) and determined that the LCST of the PNIPAAm gels was 34.3°C. Below this temperature,

significant gel swelling occurred. The transition about this point was reversible. They discovered that the transition temperature was raised by copolymerizing PNIPAAm with small amounts of ionic monomers. Beltran et al. (70) also worked with PNIPAAm gels containing ionic comonomers. They observed results similar to those achieved by Tanaka.

The earliest investigators who studied PNIPAAm gels discovered that the response time of the materials to temperature changes was rather slow. Future studies focused on developing newer materials could collapse/expand more rapidly. Dong and Hoffman (71) prepared heterogeneous gels containing PNIPAAm that collapsed at significantly faster rates than homopolymers of PNIPAAm. Kabra and Gehrke (72) developed a new method for preparing PNIPAAm gels that resulted in significant increases in the swelling kinetics of the gels. They prepared gels below the LCST to produce a permanent phase separated microstructure in the gels. These gels expanded at rates 120 times faster and collapsed at rates 3000 times faster than homogeneous PNIPAAm gels. Okano's group (73,74) developed an ingenious method for preparing comb-type graft hydrogels of PNIPAAm. The main chain of the cross-linked PNIPAAm contained small molecular weight grafts of PNIPAAm. Under conditions of gel collapse (above the LCST), hydrophobic regions were developed in the pores of the gel that resulted in rapid collapse. These materials could collapse from a fully swollen conformation in less than 20 minutes, whereas comparable gels that did not contain graft chains required up to a month to collapse fully. Such systems show major promise for rapid and abrupt or oscillatory release of drugs, peptides, and proteins.

Thermoresponsive polymers may be particularly useful in a wide variety of drug delivery applications (35,75). Okano et al. (76) studied the temperature-dependent permeability of PNIPAAm gels. They used these gels as "on-off" delivery devices in response to temperature fluctuations. This type of behavior was useful for controlling the release of insulin (77) and heparin (48). These gels were also used as "squeezing" systems (78). More recently, conjugates of PNIPAAm with trypsin have been developed (79). Such systems could be extremely useful in the targeted delivery of enzymes. Another promising application of these systems was explored by Vernon et al. (80). In this work, islets of Langerhans were entrapped by thermal gelation of PNIPAAm for use as a rechargeable artificial pancreas. Fukumori et al. (81,82) also worked with microcapsules coated in a thermosensitive layer comprised of either hydroxypropyl cellulose (HPC) or ethyl cellulose that contained nanoparticles of PNIPAAm hydrogels. These two systems result in two entirely different temperature-sensitive drug release schemes. Drug release occurs readily below the LCST for the the HPC coating due to the more open mesh size. However, above this temperature, the network collapses, hindering solute transport. In the other system, the drug release occurs only at temperatures above the LCST. At these temperatures, the nanohydrogels are collapsed, leaving voids that allow rapid release of solute. These systems, it was, are fully reversible, resulting in an "on/off" release behavior.

Other materials that possess a LCST near physiological conditions have also been pursued. Yuk et. al. (83,84)

proposed another temperature-sensitive comonomer system comprised of DMAEMA and acrylamide (Aam). By changing the comonomer feed ratio, the LCST of these systems varied from 28°C to 50°C. They also developed a mathematical model to describe solute transport as a function of temperature and network swelling kinetics (85). Ogata has also worked on hydrogels composed of the nucleic acid, uracil (86). These hydrogels, it has been shown, rapidly change volume at 35°C.

Another application of thermally sensitive hydrogels is in injectable localized drug delivery systems. Cui and Messersmith (87) used an aqueous solution of sodium alginate and temperature-sensitive liposomes containing Ca^{2+} and a drug. Once the solution is injected, the temperature of the body causes the liposomes to release the calcium ions and the drug. The calcium ions then crosslink the alginate to form a hydrogel, and controlled release of the drug to the surrounding tissue is possible. Hoffman et al. (88) created networks of chitosan grafted with Pluronic side chains. This copolymer system remains a solution until the temperature is raised to 37°C, and the side chains for hydrophobic domains cross-link the network to form a hydrogel.

pH- and Temperature-Sensitive Hydrogels

During the last 10 years, researchers have developed a novel class of hydrogels that exhibits both pH- and temperature-sensitive swelling behavior. These materials may prove extremely useful in enzymatic or protein drug delivery applications. Hydrogels were prepared from PNIPAAm and PAA that exhibited dual sensitivities (89,90). These gels responded rapidly to both temperature and pH changes. Kim's group investigated such systems as carriers of insulin (91) and calcitonin (92). In general, these hydrogels exhibited strong temperature-sensitive swelling behavior only when large amounts of PNIPAAm were in the gel. Cationic pH- and temperature-sensitive gels were prepared using poly(amines) and PNIPAAm (93). These systems were evaluated for local delivery of heparin.

Chen and Hoffman prepared new graft copolymers of PAA and PNIPAAm that responded more rapidly to external stimulus than previously studied materials (94). These new materials exhibited increased temperature sensitivity due to the presence of the PNIPAAm grafts. Such systems were evaluated for prolonged mucosal delivery of bioactive agents, specifically peptide drugs (95).

Brazel and Peppas (96) studied the pH- and temperature-responsive swelling of gels containing PNIPAAm and PMAA. These materials were used to modulate the release of streptokinase and heparin in response to pH and temperature changes (97). Baker and Siegel (98) used similar hydrogels to modulate glucose permeability. However, only large amounts of PNIPAAm were needed to observe large temperature sensitivities. The Peppas group developed novel pH- and temperature-sensitive terpolymers of PHEMA, PMAA, and PNIPAAm (99). These systems were prepared to contain PNIPAAm-rich blocks, and as a result, these materials exhibited strong temperature sensitivity at only 10% PNIPAAm in the gel. Using these materials, they were effectively able to modulate release

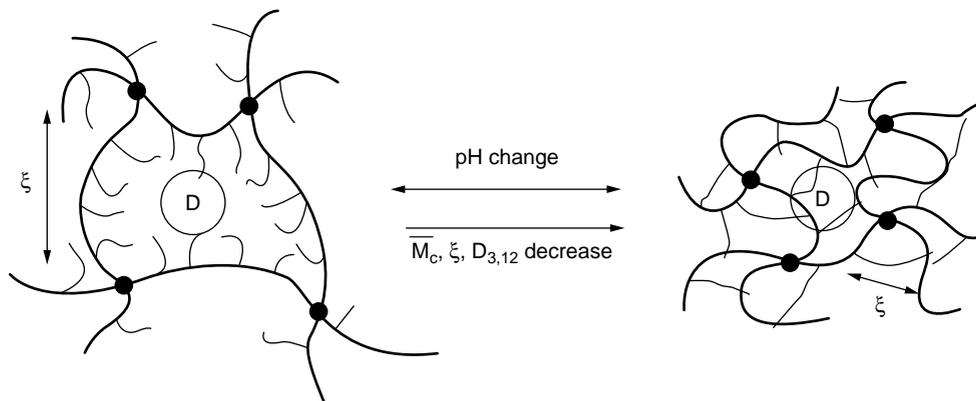


Figure 7. The effect of interpolymer complexation on the correlation length ξ and on the effective molecular weight between crosslinks \bar{M}_c , in P(MAA-*g*-EG) graft copolymer networks that have permanent, chemical cross-links (•).

kinetics of streptokinase (100). Cationic pH/temperature-sensitive polymer networks have also been analyzed. Park and Hoffman (101,102) studied gels based on NIPAAm and *NN'*-dimethylaminopropyl methacrylamide (DMAPMAAm), and Gutowska et al. (103,104) looked at networks that consisted of the more commonly used DEAEEMA and NIPAAm. They demonstrated significant pH swelling dependence at 37°C.

A novel approach to pH- and temperature-sensitive drug release was developed by the Orkisz et al. (105). They developed a system that consists of two compartments, one a drug reservoir and the other a pH- temperature-sensitive hydrogel comprised of carboxylic acid functionalized hydroxypropyl cellulose. During periods of neutral pH and low temperature, the gel is swollen and prevents release of the drug from the reservoir. However, when the temperature increases or the pH decreases, the hydrogel barrier collapses, thus allowing drug release. Kaetsu et al. (106–109) also developed a similar technology on a silicon wafer chip. They photoetched pits into the wafer, and these pits were subsequently covered by a layer of polyethylene terephthalate mesh filled with a polyelectrolyte gel. These gels covered the holes in the silicone wafer. Enzymes can be immobilized on the gel and act as the sensing mechanism. When the internal pH of the gel changes, it collapses, thus releasing the drug contained in the hole on the silicone wafer. The advantage of such a system is that a multitude of drug reservoir can be placed on a single chip, thus allowing for localized complex drug delivery.

Complexing Hydrogels

Another promising class of hydrogels that exhibits responsive behavior consists of complexing hydrogels. Osada studied complex formation in PMAA hydrogels (110). In acidic media, the PMAA membranes collapsed in the presence of linear PEG chains due to the formation of interpolymer complexes between the PMAA and PEG. The gels swelled when placed in neutral or basic media. The permeability of these membranes was strongly dependent on the environmental pH and PEG concentration (111). Similar results were observed for hydrogels of PAA and linear PEG (112). Interpenetrating polymer networks of

PVA and PAA that exhibit pH- and weak temperature-sensitive behavior due to complexation between the polymers were prepared, and the release behavior of indomethacin was studied (113,114).

The Peppas group developed a class of copolymer gels of PMAA grafted with PEG (P(MAA-*g*-EG) (115–123). These gels exhibited pH- dependent swelling due to the presence of acidic pendent groups and the formation of interpolymer complexes between the ether groups on the graft chains and protonated pendent groups. Complexation in these covalently cross-linked, complexing P(MAA-*g*-EG) hydrogels resulted in the formation of temporary physical cross-links due to hydrogen bonding between the PEG grafts and the PMAA pendent groups. The physical cross-links were reversible and depended on the pH and ionic strength of the environment. As a result, complexing hydrogels exhibit drastic changes in their mesh size across small pH changes, as shown in Fig. 7. One particularly promising application for these systems is oral delivery of protein and peptide drugs (123). As shown in Fig. 8, these copolymers

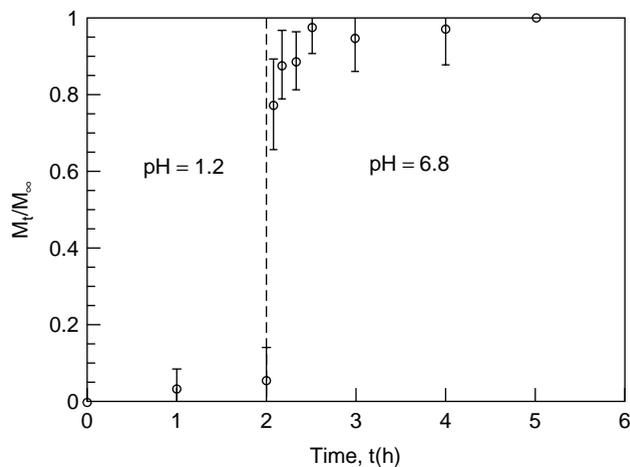


Figure 8. Controlled release of insulin *in vitro* from P(MAA-*g*-EG) microparticle simulated gastric fluid (pH 1.2) for the first 2 hours and phosphate buffered saline solutions (pH 6.8) for the remaining 3 hours at 37°C (172).

severely limit the release of insulin in acidic environments like those found in the stomach. However, insulin release occurred rapidly in conditions similar to those in the intestines.

Another form of complexation-dependent drug delivery is in drug-polymer complexation schemes. For example, Hoffman et al. (124,125) made hydrogels of NIPAAm and methacryloyl-polyoxyethyl phosphate. These phosphate-containing networks formed complexes with positively charged proteins such as lysozyme. At pH 7.4, this complex becomes destabilized, thus allowing for a pH- and ionic-strength-dependent release. The advantage of such a system is in the high loading efficiencies possible due to the capacity of the protein to form a polymer complex.

Materials Sensitive to Chemical or Enzymatic Reaction

Major developments have been reported in using environmentally responsive hydrogels as glucose-sensitive systems that could serve as self-regulated delivery devices for treating diabetes. Typically, these systems have been prepared by incorporating glucose oxidase into the hydrogel structure during polymerization. In the presence of glucose, glucose oxidase catalyzes the reaction between water and glucose to form gluconic acid. Gluconic acid lowers the pH of the gel's microenvironment.

The first such systems developed by Kost et al. (126) consisted of glucose oxidase immobilized in hydrogels based on PHEMA and PDMAEMA. These systems exhibited glucose-sensitive swelling. In the presence of glucose, gluconic acid is formed, resulting in a decrease in the local pH. As a result, the cationic-based gels swelled to larger degrees in the presence of glucose due to the production of gluconic acid. Glucose-responsive swelling allowed control of insulin permeation in these membranes by adjusting the environmental glucose concentrations (127,128). The kinetics of gel swelling and insulin release from cationic, glucose-sensitive hydrogels was also studied (129).

Glucose-responsive systems that were based on anionic hydrogels were proposed (130,131). Ito et al. (130) prepared systems of porous cellulose membranes containing an insulin reservoir. The pores of these devices were grafted with PAA chains functionalized by glucose oxidase. In the presence of glucose, the decrease in environmental pH caused the PAA chains to collapse, opening the pores to allow insulin release. More recently, glucose-sensitive complexation gels of P(MAA-*g*-EG) were developed by the Peppas group (131). As the pH decreased in these gels in response to elevated glucose concentrations, interpolymer complexes formed, resulting in rapid gel syneresis. The rapid collapse resulted in insulin release due to a "squeezing" phenomenon.

Other glucose-responsive systems have been developed that take advantage of the formation of complexes between glucose molecules and polymeric pendent groups. Lee and Park (132) prepared erodible hydrogels containing allyl glucose and poly(vinyl pyrrolidone). These systems were crosslinked by the noncovalent associations between concanavalin-A (Con-A) and the glucose pendent groups. In the presence of free glucose, Con-A was

bound to the free glucose, and the gels dissolved due to disruption of the physical cross-links. Newer materials developed by Okano's group exhibited glucose-responsive swelling and insulin release (133–135). These gels were based on phenylboronic acid (PBA) and acrylamides. Another class of glucose-sensitive gels was prepared that contained PBA, PNIPAAm, and PVA (136). These gels allow the release of insulin at physiological pH and temperature.

Similar to the glucose-sensitive hydrogels, Eremeev and Kukhtin made hydrogels that are immobilized by urease instead of glucose oxidase (137). Hence, these networks swell in the presence of urea due to urea hydrolysis.

A number of investigators concentrated on developing environmentally responsive gels that are biodegradable. This can be achieved by a number of synthetic methods. Kopecek and associates (138) developed biodegradable hydrogels by incorporating azo compounds. Bae and associates (139) synthesized very promising biodegradable carriers by preparing eight-arm, star-shaped, block copolymers that contain poly-lactic acid (PLA) and PEO. Another potentially useful biodegradable system is a photo-cross-linked polymer based on poly(L-lactic acid-*co*-L-aspartic acid) (140) which could be prepared *in situ* for delivering anti-inflammatory drugs following surgery.

Kurisawa and Yui investigated several systems that are sensitive to two enzymes. Interpenetrating networks (IPN) of dextran and methacryloyl-glycylglycylglycyl-terminated PEG are hydrogels that degrade only in the presence of both papain and dextranase (141–143). Without both of these enzymes present, the network is not degraded and hence cannot release its solute. Karisawa worked on a similar dual-stimuli system that is an IPN of gelatin and dextran (144,145). In this system, the two enzymes needed for degradation are alpha-chymotrypsin and dextranase.

BIBLIOGRAPHY

1. N. Peppas, *Hydrogels in Medicine and Pharmacy*, Vol. I: *Fundamentals*. CRC, Boca Raton, FL, 1986.
2. N.A. Peppas, *J. Bioactivity Compatible Polym.* **6**: 241–246 (1991).
3. R. Langer and N. Peppas, *J. Macromol. Sci. Rev. Macromol. Chem. Phys.* **C23**: 61–126, (1983).
4. N.A. Peppas and D.L. Meadows, *J. Membrane Sci.* **16**: 361–377 (1983).
5. B.D. Ratner and A.S. Hoffman, in *Hydrogels for Medical and Related Applications*, Andrade and Andrade, eds., American Chemical Society, Washington, DC, 1976, pp. 1–36.
6. K. Park, *Controlled Release: Challenges and Strategies*. American Chemical Society, Washington, DC, 1997.
7. N.A. Peppas, *Curr. Opin. Colloid Interface Sci.* **2**: 531–537 (1997).
8. N.A. Peppas and R. Langer, *Science* **263**: 1715–1720 (1994).
9. R. Baker, *Controlled Release of Biologically Active Agents*. Wiley, NY, 1987.
10. M.T. am Ende and A.G. Mikos, in *Protein Delivery: Physical Systems*, L.M. Sanders and R.W. Hendren, eds., Plenum Press, Tokyo, 1997, pp. 139–165.

11. N.A. Peppas and A.G. Mikos, in *Hydrogels in Medicine and Pharmacy*, N.A. Peppas, ed., Boca Raton, FL, 1986, Vol. 1, pp. 1–25.
12. N.A. Peppas and B.D. Barr-Howell, *Hydrogels in Medicine and Pharmacy*, in N.A. Peppas, ed., CRC, Boca Raton, FL, 1986, Vol. 1, pp. 27–56.
13. P.J. Flory and J. Rehner, *J. Chem. Phys.* **11**: (1943).
14. N.A. Peppas and E.W. Merrill, *J. Polym. Sci. Polym. Chem. Ed.*, **14**: 441–457 (1976).
15. J. Ricka and T. Tanaka, *Macromolecules* **17**: (1984).
16. A. Katchalsky, *Experientia* **5**: 319–320 (1949).
17. T. Tanaka, *Polymer* **20**: 1404–1412 (1979).
18. M. Rubinstein, R.H. Colby, A.V. Dobrynin, and J.F. Joanny, *Macromolecules* **29**: 398–426 (1996).
19. U.P. Schroder and W. Opperman, in *The Physical Properties of Polymeric Gels*, J.P.C. Addad, ed., Wiley, NY, 1996, pp. 19–38.
20. R. Skouri, F. Schoessler, J.P. Munch, and S.J. Candau, *Macromolecules* **28**: 197–210 (1995).
21. L. Brannon-Peppas and N.A. Peppas, *Chem. Eng. Sci.* **46**: 715–722 (1991).
22. L. Brannon-Peppas and N.A. Peppas, *Stud. Polym. Sci.* **8**: 67–102 (1990).
23. P.J. Flory, *Principles of Polymer Chemistry*. Cornell University Press, Ithaca, NY, 1953.
24. L.R.G. Treloar, *The Physics of Rubber Elasticity*. Clarendon Press, Oxford, 1967.
25. N.A. Peppas and E.W. Merrill, *J. Appl. Polym. Sci.* **21**: 1763–1770 (1977).
26. A.M. Lowman and N.A. Peppas, *Macromolecules* **30**: 4959–4965 (1997).
27. K.S. Anseth, C.N. Bowman, and L. Brannon-Peppas, *Biomaterials* **17**: (1996).
28. J.E. Mark, *Adv. Polym. Sci.* **44**: 1–26 (1982).
29. R.S. Stock and W.H. Ray, *J. Polym. Sci. Phys. Ed.* **23**: 1393–1447 (1985).
30. D.N. Winslow, ed., *Surface and Colloid Science*. Plenum Press, NY, 1984, pp. 259–282.
31. A.G. Mikos, Y. Bao, L.G. Cima, D.E. Ingber, J.P. Vacanti, and R. Langer, *J. Biomed. Mater. Res.* **27**: 183–189 (1993).
32. T. Canal and N.A. Peppas, *J. Biomed. Mater. Res.* **23**: 1183–1193 (1989).
33. N.A. Peppas, in *Pulsatile Drug Delivery*, R. Gurny, H.E. Junginger, and N.A. Peppas, eds., Wissenschaftliche Verlagsgesellschaft, Stuttgart, 1993, pp. 41–56.
34. R.A. Siegel, in K. Park, ed., American Chemical Society, Washington, DC, 1997, pp. 1–27.
35. S.W. Kim, in *Advanced Biomaterials in Biomedical Engineering and Drug Delivery Systems*, N. Ogata, et al., eds., Springer, Tokyo, 1996, pp. 125–133.
36. E.A. Bekturov and L.A. Bimendina, *Adv. Polym. Sci.* **43**: 100–147 (1981).
37. J. Heller, in J. Kost, ed., CRC, Boca Raton, FL, 1990, pp. 93–122.
38. R. Langer, *Science* **249**: 1527–1533 (1990).
39. E.R. Edelman, J. Kost, H. Bobech, and R. Langer, *J. Biomed. Mater. Res.* **19**: 67–74 (1985).
40. A.B. Scranton, B. Rangarajan, and J. Klier, *Adv. Polym. Sci.* **120**: 1–54 (1995).
41. W. Oppermann, in *Polyelectrolyte Gels: Properties, Preparation, and Applications*, R.S. Harland and R.K. Prud'homme, eds., American Chemical Society, Washington, DC 1992, pp. 159–170.
42. R.A. Siegel and B.A. Firestone, *Macromolecules* **21**: 3254–3259 (1988).
43. B.A. Firestone and R.A. Siegel, *Polym. Commun.* **29**: 204–208 (1988).
44. R.A. Siegel, M. Falamarzian, B.A. Firestone, and B.C. Moxley, *J. Controlled Release* **8**: 179–182 (1988).
45. J.M. Cornejo-Bravo and R.A. Siegel, *Biomaterials* **17**: 1187–1193 (1996).
46. R.A. Siegel, in *Pulsed and Self-Regulated Drug Delivery*, J. Kost, ed., CRC, Boca Raton, 1990, pp. 129–155.
47. D. Hariharan and N.A. Peppas, *Polymer* **37**: 149–161 (1996).
48. A. Gutowska, Y.H. Bae, J. Feijen, and S.W. Kim, *J. Controlled Release* **22**: 95–104 (1992).
49. L.M. Schwarte and N.A. Peppas, *Polym. Prepr.* **38**: 596–597 (1997).
50. L. Brannon-Peppas and N.A. Peppas, *Biomaterials* **11**: 635–644 (1990).
51. J.H. Kou, G.L. Almindon, and P.I. Lee, *Pharm. Res.* **5**: 592–597 (1988).
52. L. Brannon-Peppas and N.A. Peppas, *Int. J. Pharm.* **70**: 53–57 (1991).
53. L. Brannon-Peppas and N.A. Peppas, *J. Controlled Release* **16**: 319–329 (1991).
54. R. Bettini, P. Colombo, and N.A. Peppas, *J. Controlled Release* **37**: 105–111 (1995).
55. L. Brannon-Peppas and N.A. Peppas, *J. Controlled Release* **8**: 267–274 (1989).
56. A.R. Khare and N.A. Peppas, *J. Biomater. Sci. Polym. Ed.* **4**: 275–289 (1993).
57. M.T. am Ende and N.A. Peppas, *J. Controlled Release* **48**: 47–56 (1997).
58. L.F. Gudeman and N.A. Peppas, *J. Membrane Sci.* **107**: 239–248 (1995).
59. N.A. Peppas and S.L. Wright, *Macromolecules* **29**: 8798–8804 (1996).
60. M.T. Am Ende and N.A. Peppas, *Pharm. Res.* **12**: 2030–2035 (1995).
61. J. Heller, A.C. Chang, G. Rodd, and G.M. Grodsky, *J. Controlled Release* **11**: 193–201 (1990).
62. R. Bodmeier and O. Paeratakul, *J. Pharm. Sci.* **78**: 964–969 (1989).
63. S. Shiraiishi, T. Imai, and M. Otagiri, *J. Controlled Release* **25**: 217–223 (1993).
64. K. Yao, T. Peng, M.F.A. Goosen, J.M. Min, and Y.Y. He, *J. Appl. Polym. Sci.* **48**: 343–348 (1993).
65. P. Calvo, C. Remunán-López, J.L. Vila-Jato, and M.J. Alonso, *J. Appl. Polym. Sci.* **63**: 125–132 (1997).
66. R. Narayani and K.P. Rao, *J. Appl. Polym. Sci.* **58**: 1761–1769 (1995).
67. A. Kikuchi, M. Kawabuchi, M. Sugihara, Y. Sakurai, and T. Okano, *J. Controlled Release* **47**: 21–29 (1997).
68. R. Narayani, R.S. Hermes, and K.P. Rao, *Proc. IUPAC Int. Symp. Adv. Polym. Sci. Technol.*, 1998, p. 579–582.
69. S. Hirotsu, Y. Hirokawa, and T. Tanaka, *J. Chem. Phys.* **87**: 1392–1395 (1987).
70. S. Beltran, J.P. Baker, H.H. Hooper, H.W. Blanch, and J.M. Prausnitz, *Macromolecules* **24**: 549–551 (1991).
71. L.C. Dong and A.S. Hoffman, *J. Controlled Release* **13**: 21–31 (1990).

72. B.G. Kabra and S.H. Gehrke, *Polym. Commun.* **32**: 322–323 (1991).
73. R. Yoshida, K. Uchida, Y. Kaneko, K. Sakai, A. Kikuchi, Y. Sakurai, and T. Okano, *Nature* **374**: 240–242 (1995).
74. Y. Kaneko, K. Saki, A. Kikuchi, Y. Sakurai, and T. Okano, *Macromol. Symp.* **109**: 41–53 (1996).
75. A.S. Hoffman, *J. Controlled Release* **6**: 297–305 (1987).
76. T. Okano, Y.H. Bae, H. Jacobs, and S.W. Kim, *J. Controlled Release* **11**: 255–265 (1990).
77. Y.H. Bae, T. Okano, and S.W. Kim, *J. Controlled Release* **9**: 271–276 (1989).
78. R. Yoshida, Y. Kaneko, K. Sakai, T. Okano, Y. Sakurai, Y.H. Bae, and S.W. Kim, *J. Controlled Release* **32**: 97–102 (1994).
79. M. Matsukata, T. Aoki, K. Sanui, N. Ogata, A. Kikuchi, Y. Sakurai, and T. Okano, *Bioconjugate Chem.* **7**: 96–101 (1996).
80. B. Vernon, A. Gutowska, S.W. Kim, and Y.H. Bae, *Macromol. Symp.* **109**: 155–167 (1996).
81. H. Ichikawa, A. Ohdoi, K. Fujioka, and Y. Fukumori, *World Congr. Part. Technol.* **3**, 1998, pp. 1225–1234.
82. H. Ichikawa and Y. Fukumori, *J. Controlled Release* **107**–119 (2000).
83. S.H. Yuk, S.H. Cho, and H.B. Lee, *Proc. 1997 Spring ACS Meet*, San Francisco, 1997, v. 76.
84. S.H. Yuk, S.H. Cho, and S.H. Lee, *Macromolecules* **30**: 6856–6859 (1997).
85. M. Grassi, S. Hong Yuk, and S. Hang Cho, *J. Membrane Sci.* **241**–249 (1999).
86. N. Ogata, *6th Int. Conf. Adaptive. Struct.*, 1996, pp. 54–60.
87. H. Cui and P.B. Messersmith, in *ACS Symp. Ser.* **203**–211 (1998).
88. A.S. Hoffman, J.E. Matura, X. Wu, and W.R. Gombotz, *Proc. Int. Symp. Controlled Release Bioactive Mater.*, 1997, pp. 126–127.
89. L.C. Dong, Q. Yan, and A.S. Hoffman, *J. Controlled Release* **19**: 171–178 (1992).
90. H. Feil, Y.H. Bae, and S.W. Kim, *Macromolecules* **25**: 5528–5530 (1992).
91. Y.H. Kim, Y.H. Bae, and S.W. Kim, *J. Controlled Release* **28**: 143–152 (1994).
92. A. Serres, M. Baudyš, and S.W. Kim, *Pharm. Res.* **13**: 196–201 (1996).
93. Y. Nabeshima, Z.L. Ding, G.H. Chen, A.S. Hoffman, H. Taira, K. Kataoka, and T. Tsuruta, in *Advanced Biomaterials in Biomedical Engineering and Drug Delivery Systems*, in N. Ogata et al., eds., Springer, Tokyo, 1996, pp. 315–316.
94. G.H. Chen and A.S. Hoffman, *Nature* **373**: 49–52 (1995).
95. A.S. Hoffman, G.H. Chen, S.Y. Kaang, Z.L. Ding, K. Randeri, and B. Kabra, *Advanced Biomaterials in Biomedical Engineering and Drug Delivery Systems*, in N. Ogata et al., eds., Springer, Tokyo, 1996, pp. 62–66.
96. C.S. Brazel and N.A. Peppas, *Macromolecules* **28**: 8016–8020 (1995).
97. C.S. Brazel and N.A. Peppas, *J. Controlled Release* **39**: 57–64 (1996).
98. J.P. Baker and R.A. Siegel, *Macromol. Rapid Commun.* **17**: 409–415 (1996).
99. S.K. Vakkalanka and N.A. Peppas, *Polym. Bull. (Berlin)* **36**: 221–226 (1996).
100. S.K. Vakkalanka, C.S. Brazel, and N.A. Peppas, *J. Biomater. Sci. Polym. Ed.* **8**: 119–129 (1996).
101. T.G. Park and A.S. Hoffman, *J. Appl. Polym. Sci.* **46**: 659–671 (1992).
102. T.G. Park, *Biomaterials* **517**–521 (1999).
103. A. Gutowska and S.W. Kim, *212th ACS Nat. Meet., Orlando, FL, August 25–29*, 1996, p. OLY-011.
104. A. Gutowska, Y. Seok Bark, I. Chan Kwon, Y. Han Bae, Y. Cha, and S. Wan Kim, *J. Controlled Release* **141**–148 (1997).
105. E.S. Ron, M.E. Schiller, E. Roos, M. Orkisz, and A. Staples, *PCT Int. Appl. Gel Sciences, Inc.*, 1998.
106. I. Kaetsu, K. Uchida, H. Shindo, S. Gomi, and K. Sutani, *Radiat. Phys. Chem.* **193**–201 (1999).
107. I. Kaetsu, K. Uchida, and K. Sutani, *Radiat. Phys. Chem.* **673**–676 (1999).
108. I. Kaetsu, K. Uchida, K. Sutani, and S. Sakata, *Radiat. Phys. Chem.* **465**–469 (2000).
109. I. Kaetsu, Y. Morita, O. Takimoto, M. Yoshihara, A. Ohtori, and M. Andoh, *Proc. 18th Program Int. Symp. Controlled Release Bioactive Mater.*, 1991, pp. 449–450.
110. Y. Osada, *J. Polym. Sci. Polym. Lett. Ed.* **18**: 281–286 (1980).
111. Y. Osada, K. Honda, and M. Ohta, *J. Membrane Sci.* **27**: 339–347 (1986).
112. S. Nishi and T. Kotaka, *Macromolecules* **19**: 978–984 (1986).
113. J. Byun, Y.M. Lee, and C.S. Cho, *J. Appl. Polym. Sci.* **61**: 697–702 (1996).
114. H.S. Shin, S.Y. Kim, and Y.M. Lee, *J. Appl. Polym. Sci.* **65**: 685–693 (1997).
115. J. Klier, A.B. Scranton, and N.A. Peppas, *Macromolecules* **23**: 4944–4949 (1990).
116. N.A. Peppas and J. Klier, *J. Controlled Release* **16**: 203–214 (1991).
117. C.L. Bell and N.A. Peppas, *J. Biomater. Sci. Polym. Ed.* **7**: 671–683 (1996).
118. C.L. Bell and N.A. Peppas, *Biomaterials* **17**: 1203–1218 (1996).
119. C.L. Bell and N.A. Peppas, *J. Controlled Release* **39**: 201–207 (1996).
120. C.L. Bell and N.A. Peppas, *Adv. Polym. Sci.* **122**: 125–175 (1995).
121. A.M. Lowman and N.A. Peppas, *Polymer* **41**: 73–80 (1999).
122. A.M. Lowman and N.A. Peppas, *ACS Symp. Ser.* **728**: 30–42 (1999).
123. A.M. Lowman, N.A. Peppas, M. Morishita, and T. Nagai, *ACS Symp. Ser.* **709**: 156–164 (1998).
124. K. Kato, M. Furukawa, Y. Kanzaki, A.S. Hoffman, and K. Nakamae, *Kobunshi Ronbunshu* **353**–358 (1998).
125. K. Nakamae, T. Nizuka, T. Miyata, T. Urugami, A.S. Hoffman, and Y. Kanzaki, *Stimuli-sensitive release of lysozyme from hydrogel containing phosphate groups*, *Adv. Biomater. Biomed. Eng. Drug Delivery Syst., [5th Iketani Conf. Biomed. Polym.]*, 1996, pp. 313–314.
126. J. Kost, T.A. Horbett, B.D. Ratner, and M. Singh, *J. Biomed. Mater. Res.* **19**: 1177–1133 (1984).
127. K. Ishihara, M. Kobayashi, and I. Shinohara, *Polym. J.* **16**: 625–631 (1984).
128. G. Albin, T.A. Horbett, and B.D. Ratner, *J. Controlled Release* **2**: 153–164 (1985).
129. M. Goldraich and J. Kost, *Clin. Mater.* **13**: 135–142 (1993).
130. Y. Ito, M. Casolaro, K. Kono, and Y. Imanishi, *J. Controlled Release* **10**: 195–203 (1989).

131. C.M. Hassan, F.J. Doyle, III, and N.A. Peppas, *Macromolecules* **30**: 6166–6173 (1997).
132. S.J. Lee and K. Park, *Polym. Prepr.* **35**: 391–392 (1994).
133. D. Shiino, Y. Murata, A. Kubo, Y.J. Kim, K. Kataoka, Y. Koyama, A. Kikuchi, M. Yokoyama, Y. Sakurai, and T. Okano, *J. Controlled Release* **37**: 269–276 (1995).
134. D. Shiino, A. Kubo, Y. Murata, Y.J. Kim, Y. Koyama, K. Kataoka, A. Kikuchi, Y. Sakurai, and T. Okano, *J. Biomater. Sci. Polymer Ed.* **7**: 697–705 (1996).
135. T. Aoki, Y. Nagao, K. Sanui, N. Ogata, A. Kikuchi, Y. Sakurai, K. Kataoka, and T. Okano, *Polym. J.* **28**: 371–374 (1996).
136. Hisamitsu, K. Kataoka, T. Okano, and Y. Sakurai, *Pharm. Res.* **14**: 289–293 (1997).
137. N.L. Eremeev and A.V. Kukhtin, *Anal. Chim. Acta* **27–34** (1997).
138. Ghandehari, P. Kopecková, P.Y. Yeh, and J. Kopecek, *Macromol. Chem. Phys.* **197**: 965–980 (1996).
139. Y.K. Choi, S.W. Kim, and Y.H. Bae, *Polymer Prepr.* **37**: 109–110 (1996).
140. J. Elisseeff, K. Anseth, R. Langer, and J.S. Hrkach, *Macromolecules* **30**: 2182–2184 (1996).
141. M. Kurisawa, M. Terano, and N. Yui, *Macromol. Rapid Commun.* **16**: 663–666 (1995).
142. M. Kurisawa, M. Terano, and N. Yui, *J. Biomater. Sci. Polymer Ed.*, 691–708 (1997).
143. M. Kurisawa and N. Yui, *Proc. Int. Symp. Controlled Release Bioactive Mater.* 1996, pp. 765–766.
144. M. Kurisawa, N. Yamamoto, and N. Yui, *Proc. Int. Symp. Controlled Release Bioactive Mater.*, 1996, pp. 761–762.
145. M. Kurisawa and N. Yui, *J. Controlled Release* **191–200** (1998).

GIANT MAGNETOSTRICTIVE MATERIALS

HIROSHI EDA
LIBO ZHOU
IBARAKI University
Nakanarusawa, Japan

INTRODUCTION

The importance of an interface that associates the macro world and the microworld is indicated in Fig. 1. The interface is composed mainly of actuators, sensors, and micro-computers. As shown in Fig. 2, it enables users to feel the microscopic phenomenon as if they were in the macro world.

In life sciences and medical and biological engineering, there are increasing demands for implementing fabrication, assembly, inspection, modification, and/or evaluation at the micron or submicron scale (1,2). To achieve this, the development of new materials and devices that can realize the required functions and movements at a microscopic level becomes mandatory. For instance, great power output, large displacement, and quick response are often required for functional materials, and safety, stability, and flexibility are the primary concern for structural applications. The gap between the macro world and the

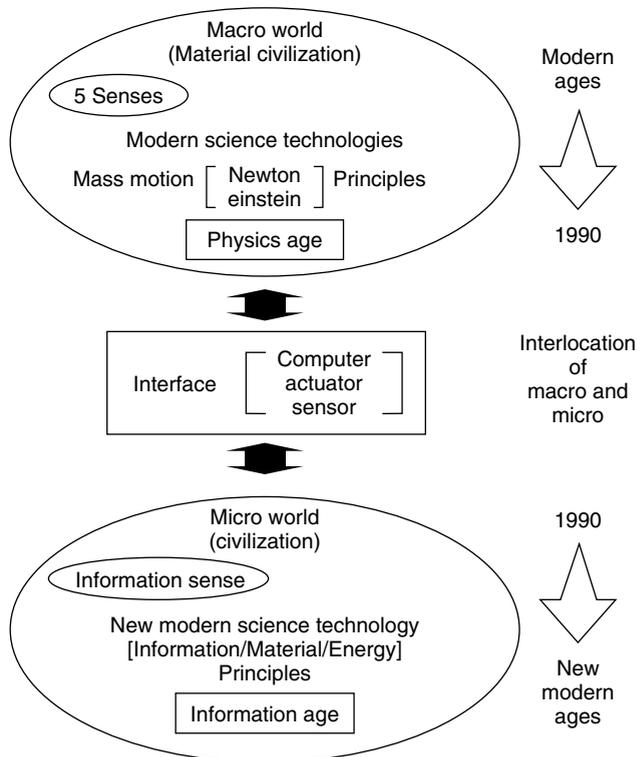


Figure 1. The transition between the macro world and the microworld.

microworld needs to be bridged naturally by using advanced devices and smart materials. In these respects, it is believed that giant magnetostrictive material (GMM) is one of the most promising materials for satisfying such requirements (Fig. 3).

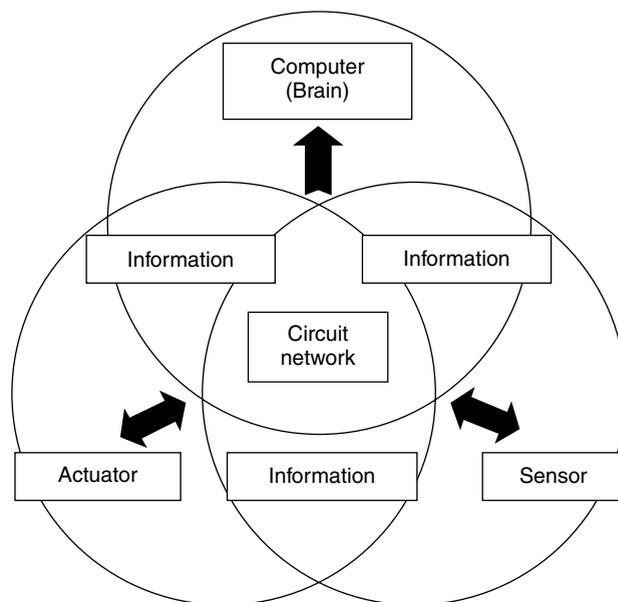


Figure 2. An intelligent system proposed for bridging the gap between the macro world and the microworld.

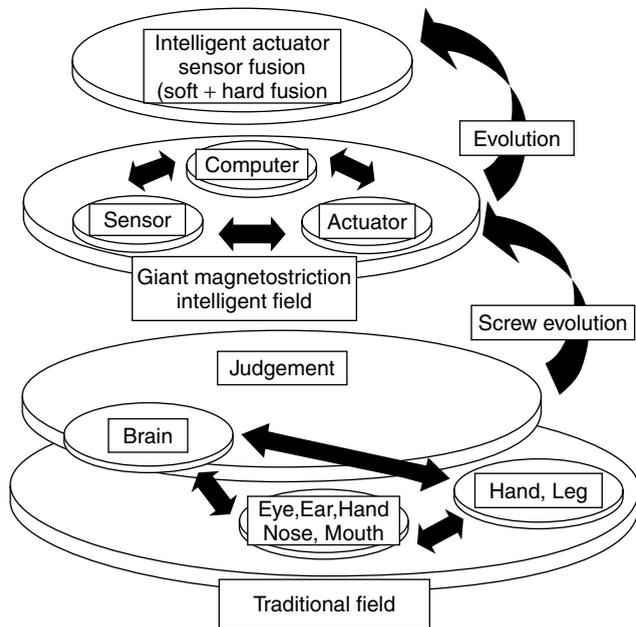


Figure 3. Evolutionary process in a dynamic intelligent system illustrated by unitizing microcomputers, actuators, and sensors, where the GMM plays a very important role.

GIANT MAGNETOSTRICTIVE MATERIALS

What Is a GMM

The magnetostrictive effect is any change in the dimensions of a magnetic material caused by a change in its magnetic state (3). GMMs are a group of materials that produce magnetostriction of more than several thousands ppm (parts per million). GMM was invented by A. E. Clark and his fellow researchers in 1963, when they first discovered the extraordinary magnetostrictive properties of the rare-earth element terbium at an extremely low temperature close to absolute zero. During the next two decades, they found that this giant magnetostrictive behavior exists at room temperature in terbium-iron (TbFe_2) and the magnetic anisotropic compensated alloy $\text{Tb}_x\text{Dy}_{1-x}\text{Fe}_{1.9}$ (or Terfenol-D where $x = 0.3$) for further temperature stability.

GMMs are unique in the sense that their magnetostrictions are as large as 0.2%, and they can sustain external loads as large as 500–600MPa. Unlike conventional materials, the response speed of GMMs is also very fast, of the order of 10^{-6} sec. Other attributes of GMMs include their low impedance and possible noncontact delivery of power. The former, in particular, enables a low voltage drive of GMM which may be of vital significance for applications that are used in inflammable atmospheres. Although there are some drawbacks in GMMs such as expensive material cost, rather poor corrosion resistance, heat generation by the driving coil, and eddy currents at high frequency excitation, however, fairly stable and robust performance can be obtained by choosing driving conditions according to the applications to be used.

Physics of GMMs

GMMs produce large strains up to approximately 2000 ppm in an imposed magnetic field. Magnetostriction is a result of the rotation of small magnetic domains that cause internal strains in the materials. These strains result in a positive expansion of a GMM in the direction of a magnetic field. As the magnetic field is increased, more domains rotate and become aligned until full saturation is achieved. If the magnetic field is reversed, the domains reverse their direction but align again in the field direction and also result in a length increase. The concept of the magnetostrictive effect is sketched in Fig. 4, where the magnetostrain $S(= \Delta l/l)$ is plotted versus the magnetic field H ; here l is the original length of the sample GMM, and the magnetic field intensity $H = nI$, where I is the current through a surrounding coil of N turns across a length l where $n = N/l$.

As the current I is increased, the magnetic field increases, and the strain $\Delta l/l$ increases to 1400 ppm at a field strength of 1000 Oe. According to the curve in this region, the slope $\Delta S/\Delta H$ is nearly constant across a very wide range of strains. For a larger field, the increase in the curve reaches a plateau where almost all domains are essentially lined up in the direction of the field, and thus there is no further increase in length as the result of saturation. If a negative current is used instead of a positive current starting at $H = 0$, the mirror image curve on the left is followed

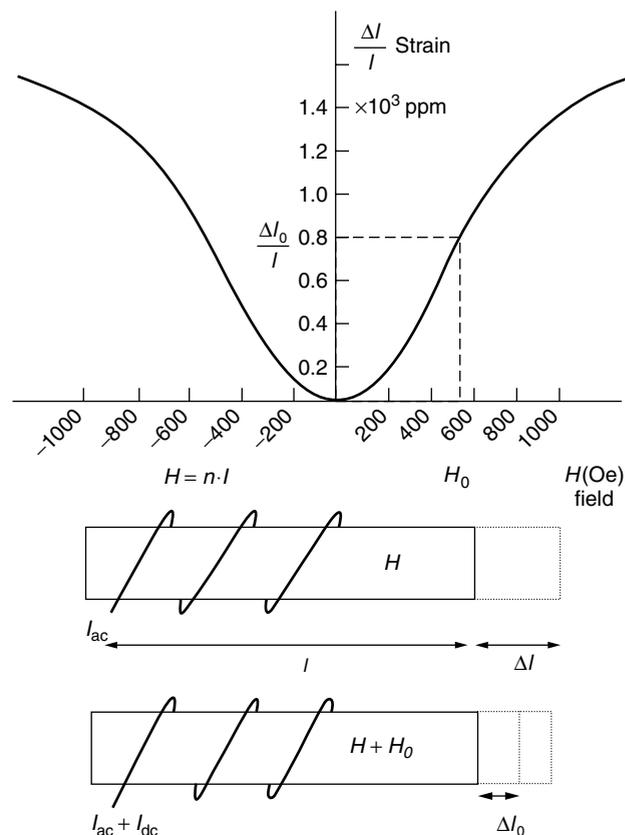


Figure 4. Strain versus magnetic field for a GMM (3).

and again there is an expansion of the GMM rod. Thus the negative cycle of a current oscillation also produces a positive expansion, just the same as the positive cycle does. This results in a mechanical motion twice the frequency of the input current. This behavior, if not appropriate, is avoided by introducing a DC biasing field H_0 to expand the GMM rod initially to position Δl_0 which is approximately one-half the total linear expansion limit. As a positive AC current is applied, the magnetic field causes additional expansion. In the negative half of the cycle, the field cancels the DC bias, and the displacement from the neutral position Δl_0 is reduced. The bias may be introduced by a permanent magnet or a DC current that uses a second coil.

Hysteresis should be taken into consideration in being applied into a precision positioning device because the S - H curves are slightly different, depending on whether the magnetic field is increased from zero to saturation or decreased from saturation to zero. The butterfly curves in Fig. 5 illustrate the hysteresis effect. The flux density B versus magnetic field H is shown in Fig. 5a. As H is reduced to zero, there is a remnant flux density value B_0 . In Fig. 5b, the strain S is shown against the flux density B , where no hysteresis is observed. However, when these two curves are combined, the hysteresis turns up in the resulting curve of the strain versus the magnetic field in Fig. 5c. Although the hysteresis effect is relatively small, it could be important in an accurate positioning device. A closed-loop feedback scheme is preferable for such applications.

Dynamics of GMMs

Magnetostriction is a transduction process in which electrical energy is converted into mechanical energy. Reciprocally, GMMs also convert mechanical energy to electrical energy. The material behavior follows a piezomagnetic law (4), and the linear region can simply be expressed by its first-order approximation as

$$B = dT + \mu^T H, \quad H = nI, \quad (1)$$

$$S = s^H T + dH, \quad (2)$$

where, as before, B is the magnetic flux density, H is the strength of the magnetic field, S is the strain, T is the mechanical stress, μ^T is the constant- T permeability corresponding to the slope of the curve in the first quadrant of the B - H curve in Fig. 5(a), s^H is the constant- H elastic modulus or the reciprocal of Young's modulus, d is the piezomagnetic constant corresponding to the slope of the linear region of the S - H curve in Fig. 4, and I is the electric current. The interrelations that are well expressed in Fig. 6, which shows the relationship between the magnetic variables B and H , the mechanical variables S and T , and also the interconnection between the magnetic and mechanical fields.

Note that Eqs. (1) and (2) are valid only if other nonlinear factors such as eddy current loss or nonuniform flux density can be omitted. To accommodate certain degrees of nonlinearity, Carmen and Mitrovic (6) extended Eqs. (1) and (2) by using the Gibbs free energy and representing T

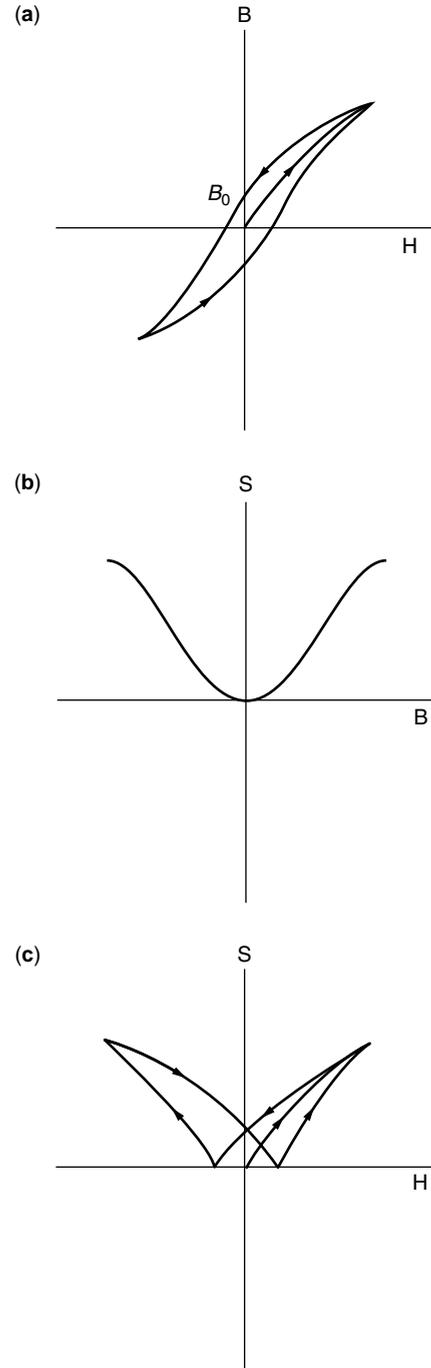


Figure 5. Butterfly curve of S - H .

and H by Taylor series:

$$S = s^H T + \frac{1}{2} d_{m-s} H^2 + \frac{1}{2} s_{m-e} T H^2, \quad (3)$$

$$B = d_{m-s} T H + \mu^T H + \frac{1}{2} s_{m-e} T^2 H, \quad (4)$$

where

$$d_{m-s} = \frac{d_{T=0}}{H_0}, \quad (5)$$

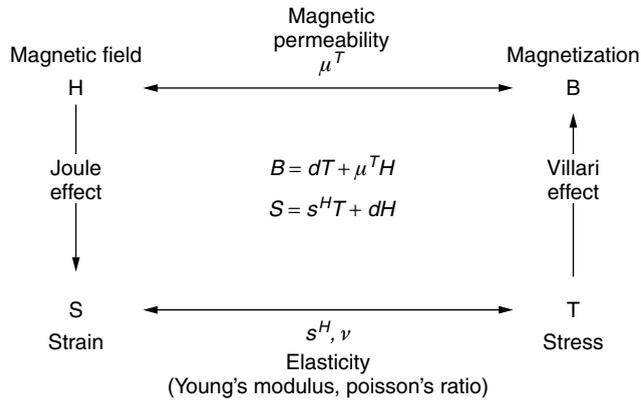


Figure 6. Interrelationships among variables (by ETREMA).

$$s_{m-e} = \frac{dT_{=T_0} - dH_0}{T_0 H_0}, \quad (6)$$

and H_0 is the magnetic field strength under no load. For instance, Carmen and Mitrovic (6) obtained d_{m-s} and s_{m-s} for Terfenol-D ($Tb_{0.3}Dy_{0.7}Fe_{1.93}$) by taking nonlinear terms (7) into account as follows:

$$\begin{aligned} d_{m-s} &= 0.18 \times 10^{-12} \text{ m}^2/\text{A}^2, \\ s_{m-e} &= 0.20 \times 10^{-20} \text{ m}^2/\text{A}^2 \cdot \text{Pa}. \end{aligned}$$

These values show better agreement with reality when the compressive prestress becomes greater than 40 MPa. Compliance and permeability used in the nonlinear constitutive relations [Eqs. (3) and (4)] are

$$\begin{aligned} s^H &= 55 \times 10^{-12} \text{ Pa}^{-1}, \\ \mu^T &= 63 \times 10^{-7} \text{ N/A}^2. \end{aligned}$$

Among these, the most frequently used is the **Joule Effect**, which is responsible for the dimensional change of a

GMM rod when the longitudinal magnetic field varies. If the mechanical stress T is zero, the magnetostrain of a free GMM rod is proportional to the magnetic field to which the GMM is subject, given as

$$\begin{aligned} S &= dH = dnI, \\ B &= \mu^T H = \mu^T nI, \quad \text{for } T = 0. \end{aligned} \quad (7)$$

Therefore, the elongation Δl of a GMM is

$$\Delta l = l \cdot S = dnI = dNI, \quad (8)$$

where N is the number of coil turns over a length l . Equation (8) addresses the way electrical energy is converted to mechanical energy, which is fundamental for designing most GMM actuators. When compressed (a prestress T is applied), a GMM produces significantly more strains. Figure 7 shows strain versus magnetic field up to 20 MPa. It can also be seen that a larger field is necessary to attain a given strain as the prestress is increased. For sinusoidal operation, as described previously, the bias field must be increased somewhat to obtain the same strain.

The inverse Joule effect, also known as the “**Villari effect**,” is the basic principle for sensor design. Equation (1) states that the mechanical stress or the internal strain of a GMM in a biased magnetic field causes a proportional change in the magnetic flux density, which in turn induces voltage changes in a pickup coil that surrounds the GMM rod. Now, suppose that the GMM rod is biased and used as a sensor for sinusoidal input force $F = F_0 \sin \omega t$ at its end. The output voltage V generated at the open circuit, according to Faraday–Noiman’s law, is given as follows:

$$V = -N \frac{d\phi}{dt} = -NA \frac{dB}{dt}, \quad (9)$$

where ϕ is the magnetic flux, A is the cross-sectional area of the GMM rod, and N is the number of turns of the pickup coil. When a circuit is open, $I = 0$, and consequently

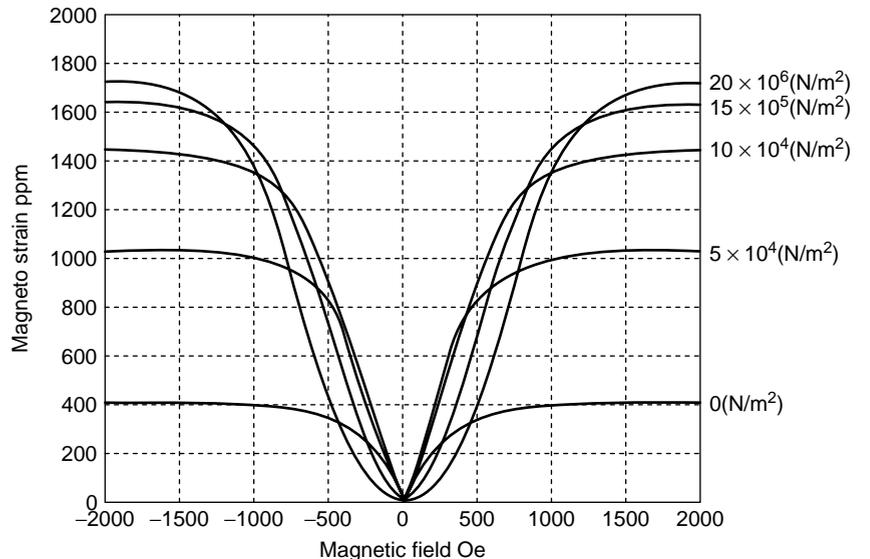


Figure 7. S – H curve at different prestress levels.

$H = 0$; the fundamental equation set, Eqs. (1) and (2), are simplified to

$$\begin{aligned} B &= d \cdot T, \\ S &= s^H T, \quad \text{for } H = 0. \end{aligned} \quad (10)$$

Thus, the flux ϕ is determined as

$$\begin{aligned} \phi &= B \cdot A = d \cdot T \cdot A \\ &= d \cdot F, \end{aligned} \quad (11)$$

and the output voltage

$$V = -N \frac{d\phi}{dt} = -d \cdot N \frac{dF}{dt} = -\omega d N F_0 \cos \omega t, \quad (12)$$

where $\omega = 2\pi f$ and f is the frequency of the force imposed. Note that the sensitivity increases as frequency increases. This increase in sensitivity continues up to the first resonance of the GMM rod. If we use the amplitude of the stress T_0 to replace the force amplitude, Eq. (12) is rewritten as

$$V = -\omega \cdot d \cdot n \cdot l \cdot A \cdot T_0 \cos \omega t, \quad (13)$$

which shows that the voltage sensitivity is also proportional to the product of the number of turns per unit length n and the volume $A \times l$ of the GMM rod.

A GMM may also be used to convert a mechanical force or stress to a current rather than a voltage. In this case, a short circuit or very low impedance ($B = 0$) is applied to Eq. (1) to achieve $H = -Td/\mu^T$. As a result, the current I generated for a mechanical stress T is given as

$$I = \frac{H}{n} = -\frac{d}{\mu^T} \cdot \frac{T}{n}, \quad (14)$$

which shows a theoretical basis for designing a GMM power generator.

Other useful GMM effects include

- I. **ΔE Effect:** Young's modulus proportionally changes with the magnetic field H and the acoustic velocity also shows a proportional change in it as well.
- II. **Wiedemann Effect:** At two dimensional change takes place in a GMM ring when a closed magnetic field along its circumference and a longitudinal magnetic field along its axis are superimposed simultaneously. In this case, as shown in Fig. 8, the magnetization changes its direction to a spiral. This combined magnetic field results in a torsional displacement of a GMM (also known as the "**Wertheim Effect**"). Supplying the GMM ring with a twist force in a forcible way induces a current in the surrounding coil, which is called the "**Inverse Wiedemann Effect.**"
- III. **Birkhauzen Effect:** Small discontinuities are observed in the magnetization curve of ferromagnetic materials, which may occur when the magnetizing force is being increased smoothly.

Most GMM applications developed so far have used energy transduction between magnetic and mechanical

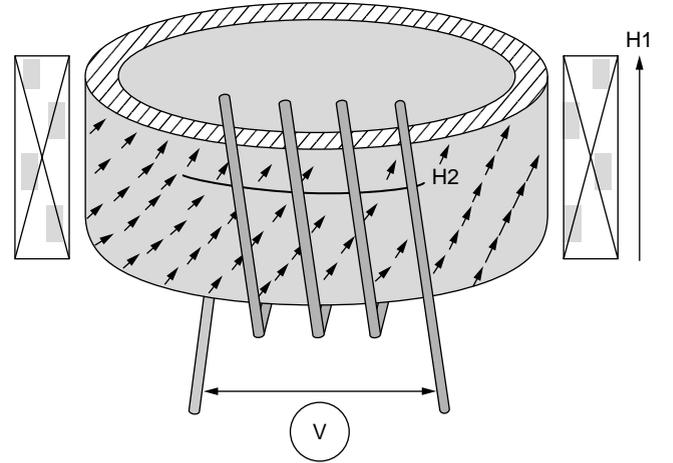


Figure 8. Wiedemann effect.

energy. The coupling coefficient is an important factor for a transducer because the square of the coupling coefficient k^2 is equal to the ratio of the energy converted to the total energy stored. For instance, when the coupling coefficient $k = 0$, there is no energy conversion at all, whereas $k = 1$ implies full conversion of energy from one form to another. That is, $0 < k < 1$. The greater the coupling coefficient, the greater the electrical efficiency for a given electrical loss.

The coupling coefficient may be obtained most simply from the fundamental equation set, Eqs. (1) and (2). The procedure is similar to that obtained for the coupling coefficient of a transformer, that is, k^2 is given by the ratio of the cross products:

$$k^2 = \frac{d^2}{\mu^T s^H}. \quad (15)$$

To convert more than half of the electrical energy to mechanical energy, the coupling coefficient has to be at least 0.707. The important physical properties of a GMM are shown in Table 1 in comparison with piezoelectric materials (PZT) and pure nickel. Table 1 shows that a GMM can convert more than 50% of the energy stored, whereas PZT material brings over only about 35%. The 15% difference grants a GMM a significant improvement in the functionality of sensors and actuators.

Table 1. Physical Property Comparison

Property	GMM	PZT	Pure Nickel
Density (10^3 kg/m ³)	9.25	7.49	9.5
Young's modulus (10^{10} N/m ²)	2.5–3.5	7.5	22.5
Electric resistance ($\mu\Omega \cdot \text{cm}$)	60	–	6.6
Saturating strain (ppm)	1500–2000	100	–40
Stress output (kgf/mm ²)	3.0	1.5	0.1
Energy density (J/m ³)	14000–25000	930	30
Coupling coefficient (%)	70–75	65	30
Acoustic velocity (m/s)	1720	3130	4940
Permeability (emu)	4.5–9.3	–	60

Other features of GMM are summarized here. Advantages:

1. Large magnetostriction (more than 2000 ppm)
2. Choice of positive or negative magnetostriction controlled by material compositions
3. High Curie temperature (380°C)
4. Noncontact drive via magnetic field
5. Low voltage drive (due to low impedance)
6. Outstanding energy density, hence large force output
7. Small hysteresis
8. Fast response
9. Controllable temperature characteristics
10. Flexible net shape (if the powder metallurgy method is used)
11. High elasticity, high stiffness
12. Large coupling coefficient

Disadvantages;

1. Necessity for a driving mechanism to supply the magnetic field
2. Joule heat loss from a driving coil
3. Eddy current heat loss under a high-frequency drive
4. Poor corrosion resistance

Other important values of the GMM properties not listed in Table 1 are compressive strength of 800 Mpa, tensile strength of 100 Mpa, thermal expansion rate of $12 \times 10^{-6} \text{ }^{\circ}\text{C}^{-1}$ and saturation magnetization of 10 kG.

GMM MANUFACTURING PROCESS

There are several methods of producing a GMM alloy; four have been used on at least a near-production basis: free stand zone melting (FSZM), modified Bridgman (MB) (8), powder metallurgical compaction (PM), and polymer matrix composites. The unidirectional solidification techniques of FSZM and MB have the advantage of producing rods that have high magnetostrains and power density. In both cases, secondary machining is required to generate complex geometries beyond the right-angle cylinder. The powder techniques can use the powder metallurgical process to produce complex geometries. The composite materials possess substantially high electrical resistivity which would be advantageous in high-frequency applications where eddy current loss is a problem. Of these, the MB and PM are currently the most typical processes for making bulk materials. The process flows of MB and PM are compared in Fig. 9, and a brief description of each process is given here.

Modified Bridgman Method

As noted, The dimensional or structural changes of a GMM are subject to the externally applied magnetic field as well as the material crystal structure, including the

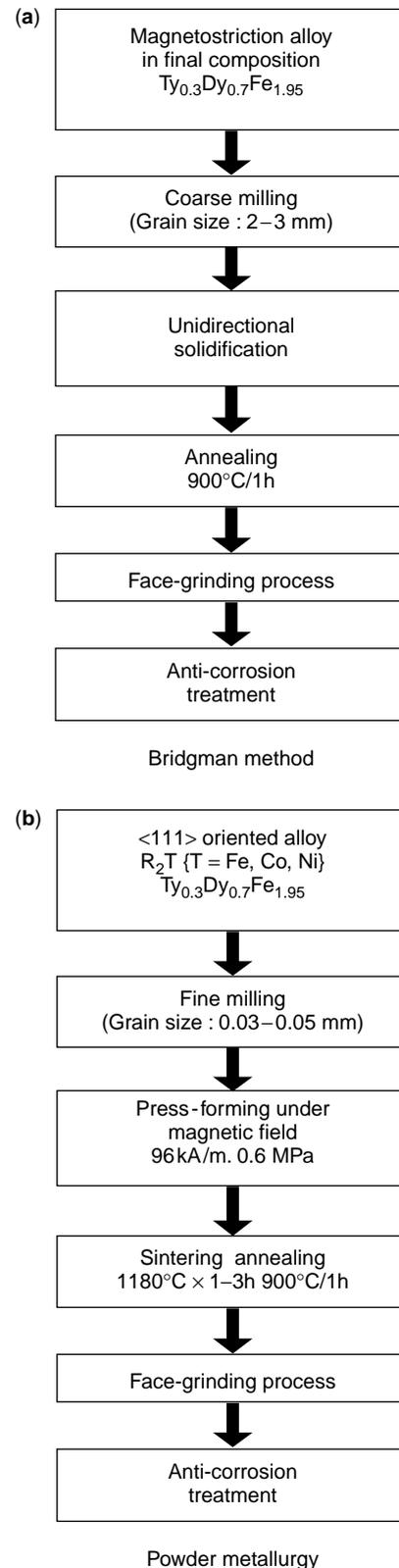


Figure 9. GMM manufacturing processes (10). (a) Bridgman method; (b) powder metallurgy.

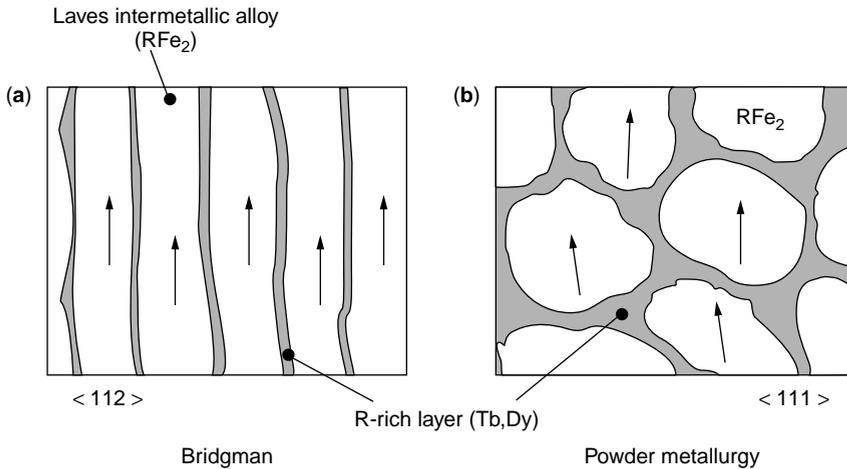


Figure 10. Crystal structure developed.

atomic vacancies, the spin magnetic moment, and the orbital magnetic moment. As a well-established process, the MB method (9) controls the crystal orientation by unidirectionally solidifying the melt of a specified composition. The raw materials used for MB are the Dy, T_y, and Fe grains in the ingredient ratio of 3:7:20 and about 2–3 mm in size. The grains are first melted in a ceramic crucible and then poured via a hole in the bottom of the crucible into preheated molds. The molds are withdrawn from the secondary furnace sections in a standard Bridgman manner. As the material solidifies, the rate of withdrawal and direction of the heat flow are controlled to produce crystallographically aligned drivers. The drivers then go through annealing (900°C/1h), secondary machining, and an anti-corrosion treatment to obtain the final products.

The crystal structures made by different processes are schematically shown in Fig. 10. In a MB ideal driver, the crystal are all aligned to $\langle 112 \rangle$, which is deviated 19° from the rod axis $\langle 111 \rangle$. Because the domains are rotated more fully, the GMM rod potentially risks an initial “jump.” If the rod is not fully aligned, the deviation angle is greater. As a result, the initial “jump” of the domain will not be as great, but the rod normally shows a reduction in magnetostrictive performance.

Powder Metallurgical Method (11)

As shown in the process flow of Fig. 9, the PM process starts with preparation of the alloy composite by milling the alloy ingots to a particle size of 0.03–0.05 mm. Heat treatment is employed before and after the milling process, 14 hours at 950°C for crystal grain homogenization and 2.5 hours for the residual stress relief. Three different types of milled grains that have the following compositions are then blended together and preformed in the presence of a magnetic field of 96 kA/m and under a pressure of 0.6 MPa:

Alloy 1: $Tb_{0.4} Dy_{0.6} Fe_{1.95} + R_2O_3 + R_2C$

Alloy 2: $Dy_2 (Fe_{0.2}Co_{0.5}) + R_2O_3 + R_2C$

Raw material: $Fe + Fe_2O_3$

where, R = Dy, Tb.

The structure developed is obtained in the form of $Tb_{0.3} Dy_{0.7} (Fe_{1.85}Co_{0.06})$. The preformed material is further

sintered under pressure of 6–12 ton/cm² and temperature of 1180°C for 1–3 hours in argon gas at a flow rate of 2 l/min, followed by annealing at 900°C for 1 hour. As shown in Fig. 10, the powder metallurgical GMM takes a polycrystalline form where the main phase RFe₂ grains are wrapped up with the R-rich layer and the oxide impurity is the grain boundary. Because it is formed in the presence of a magnetic field, most of the crystals of the powder metallurgical GMM are oriented in the $\langle 111 \rangle$ direction. Figure 11 is an output chart of X-ray diffraction in which the $\langle 111 \rangle$ oriented crystals amount to 38%. This means that an anisotropic composition has been created in an appropriate crystal orientation along which effective magnetostriction takes place.

Figure 12 compares the *S–H* curves obtained by PM and MB methods. Regardless of the prestress, the magnetostrictions obtained are quite comparable, and so are the hystereses as well. It is particularly noteworthy that there is no initial “jump” observed in the powder metallurgical GMM. The other material properties are summarized in Table 2.

In addition to the manufacturing cost which is as low as one-third that of conventional means, the great advantage of powder metallurgy is the capability of near-net-shape

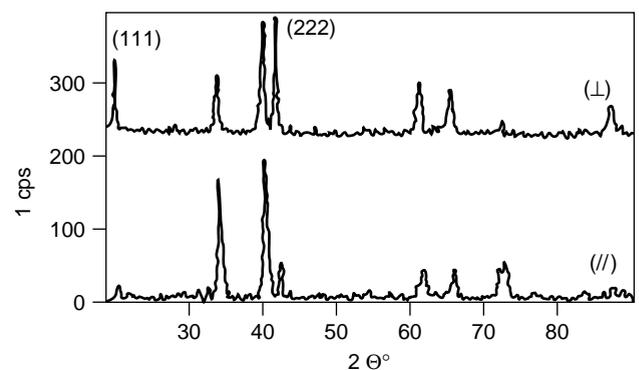


Figure 11. X-ray diffraction of a GMM made by PM.

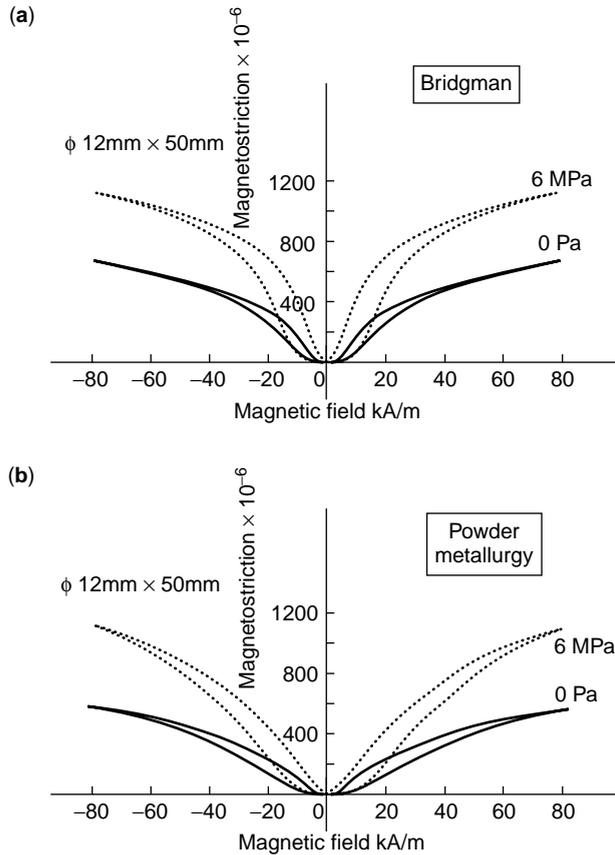


Figure 12. Magnetostriction versus applied magnetic field.

forming. Unlike the MB made GMM, which is available only as a right-angle cylinder, the powder metallurgical GMM can be configured into any desired form, as long as the die and mold are applicable.

Apart from the processes mentioned, it is worth noting that thin film GMM using sputtering techniques (12) has recently widened its applications in micromachining, MEMS, and microsystem technology (MST).

APPLICATIONS

Sensor Design

GMM sensors basically transduce mechanical energy to electrical energy. Their design deals with changes in flux density B and the strength of the magnetic field H under

Table 2. Physical Properties of GMM Produced

Property	Bridgman	Power Metallurgy
Density ρ (kg/m ³)	9.20×10^3	7.56×10^3
Elastic E_r (N/m ²) ^a	$3.8\text{--}4.0 \times 10^{10}$	3.7×10^{10}
Modulus E_a (N/m ²) ^a	$5.1\text{--}5.3 \times 10^{10}$	4.2×10^{10}
Bending strength (kg/mm ²)	5.2	4.6
Coupling coefficient k	0.56	0.46–0.61
Sonic C_r m/s ^a	2088	2065
Velocity C_a m/s ^a	2426	2227
Cost	1	1/3

^ar: parallel to anisotropic direction; a: vertical to anisotropic direction.

prestress and basically follows Eqs. (13) and (14). Jiles (13) and Kvarnsjö (14) extended linear Eqs. (1) and (2) to get better agreement with hysteresis:

$$B = 0.5 \sin(\tan^{-1}\{\tan[62^\circ e^{-0.085\sigma(\text{MPa})}] \times e^{C|H_B|}\}) + 3\mu_0 H, \quad (16)$$

where μ_0 is the permeability at zero bias. In an open circuit of $H = 0$, the output voltage is given as

$$\begin{aligned} V &= -N \frac{d\phi}{dt} = -NA \frac{dB}{dt} \\ &= -\frac{NA}{2} \frac{d}{dt} [\sin(\tan^{-1}\{[62^\circ e^{-0.085\sigma(i)(\text{MPa})}] \times e^{C|H_B|}\})] \end{aligned} \quad (17)$$

for a short circuit or very low impedance of $B \approx 0$; on the other hand, the output current is given as

$$I = H \frac{l}{N} = \frac{\sin\{\tan^{-1}[62^\circ e^{-0.085\sigma(i)(\text{MPa})}] \times e^{C|H_B|}\}}{6\mu_0} \frac{l}{N} \quad (18)$$

Vibration Sensor (Accelerometer) (15). Figure 13 shows a schematic drawing of the structure of a vibration sensor

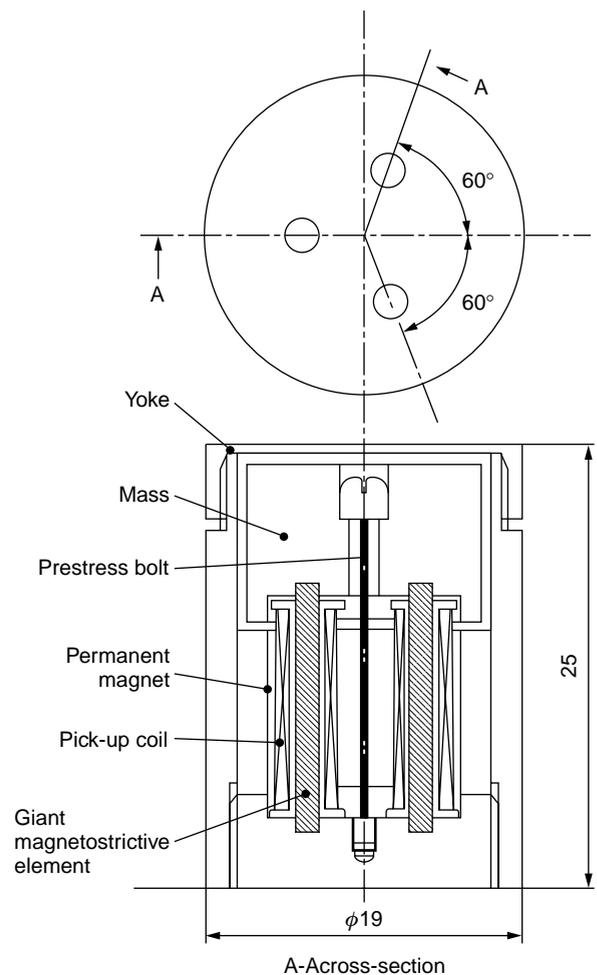


Figure 13. Sectional view of a vibration sensor.

that is 19 mm in diameter and 25 mm long. Three magnetostrictive rods are deployed equally spaced and concentrically. Each rod is surrounded by a pickup coil. As described before, both a mechanical stress bias and a magnetic bias are needed to obtain magnetostriction effectively. This is especially true for applications of reverse transduction because, without magnetic bias, no flux variation would be caused by the varying stress.

A linear model is appropriate for a harmonic small-signal operation around the biased neutral point (H_0, T_0) where H_0 and T_0 are, respectively, the magnetic bias of the permanent magnet and the prestress. When ignoring non-linear factors such as eddy current, hysteresis, and a non-ideal magnetic circuit and leakage, the induced output voltage in an open circuit is given by Eq. (12), which indicates that the induced voltage is proportional to the number of coil turns and the frequency and magnitude of the force amplitude.

Pressure/Force Sensor. Unlike dynamic vibration, static pressure does not cause any change in the flux density of an open circuit. Therefore, there is no output voltage from Eqs. (12) or (13). The static pressure/force sensor requires an alternative design. Because of the coupled nature of the GMM transducer, however, electrical properties are affected by the mechanical boundary conditions on the rod, and likewise, mechanical properties are affected by the boundary conditions on the electrical wires. That is the so-called ΔE effect, which means that a different elastic modulus $1/s^H$ is measured that depends on whether the coil wires are left in an open circuit or a short circuit. The measured permeability and inductance would also be different if the GMM rod is left free or clamped rigidly at its end. The relationship between various parameters is through the coupling coefficient in the form,

$$\mu^S = \mu^T (1 - k^2)$$

and

$$s^B = s^H (1 - k^2). \quad (19)$$

The superscript S means zero strain, the so-called clamped condition, whereas the T indicates a free bar (zero stress). Likewise, the superscript B means zero flux or a short circuit, and H is an open circuit.

When static pressure $T = T_0$ is applied to a GMM rod, the basic Eq. (1) is rewritten as

$$B = dT_0 + \mu^{T_0} \cdot n \cdot I, \quad (20)$$

where $\mu^S < \mu^{T=T_0} < \mu^{T=0}$. Thus, the induced voltage is,

$$V = -NA \frac{dB}{dt} = -n^2 \cdot \mu^{T_0} \cdot l \cdot A \cdot \frac{dI}{dt}. \quad (21)$$

To obtain the output voltage, an AC current $I = I_0 \sin \omega t$ is needed to excite the pickup coil. The detecting voltage is given by

$$V = -n^2 \cdot \mu^{T_0} \cdot l \cdot A \cdot I_0 \cdot \omega \cdot \cos \omega t \quad (22)$$

which is can be represented by its virtual value $n^2 \cdot l \cdot A \cdot I_0 \cdot \omega \cdot \mu^{T_0} / \sqrt{2}$. The sensitivity of the measured virtual value

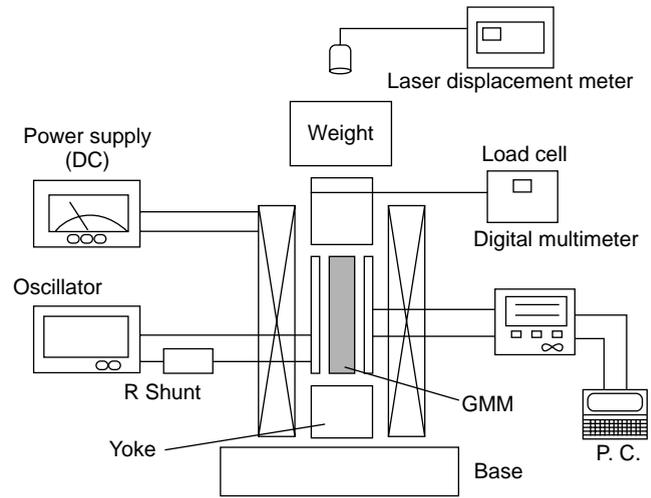


Figure 14. Structure of pressure/force sensor.

is proportional to the product of the square of the number of turns per unit length n , the volume of the GMM rod $l \cdot A$, the excitation current I_0 , and its frequency $\omega = 2\pi f$. The variation in the virtual voltage responding to the applied stress T_0 is subject to the permeability μ^{T_0} . For instance, the difference in the virtual voltage between the free bar and the bar undergoing T_0 is

$$\Delta V = \frac{n^2 \cdot l \cdot A \cdot I_0 \cdot \omega}{\sqrt{2}} (\mu^{T=0} - \mu^{T=T_0}). \quad (23)$$

Figure 14 shows the structure of a pressure/force sensor 100 mm in diameter and 50 mm long. The GMM rod deployed is a right cylinder shape $\phi 6 \times 25$ mm and coated with 30 μm of Teflon at its ends. The rod is surrounded by the bias coil which is sized about OD = 100 mm, ID = 30 mm and is produced by 1180 turns of $\phi 1.0$ mm UEW wire. Seated in between the GMM rod and bias coil is the pickup coil, which uses $\phi 0.2$ wire wound in 570 turns and formed to OD = 7.7 mm and ID = 6.5 mm.

Figure 15 shows the measured virtual voltage as the force varies from 20 to 200 kgf/cm² at 400 Oe bias and 5 V, 1 kHz AC current for the excitation.

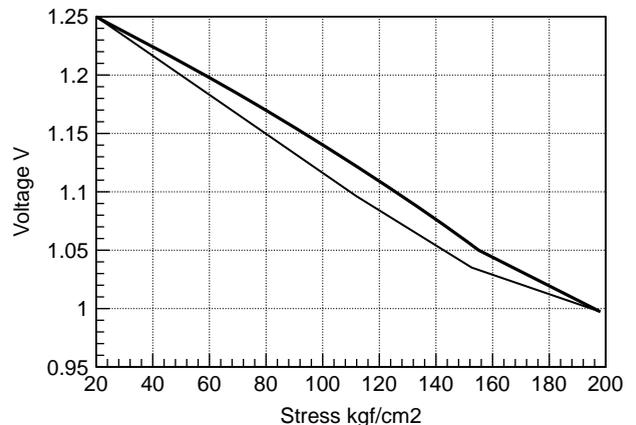


Figure 15. Performance of pressure/force sensor.

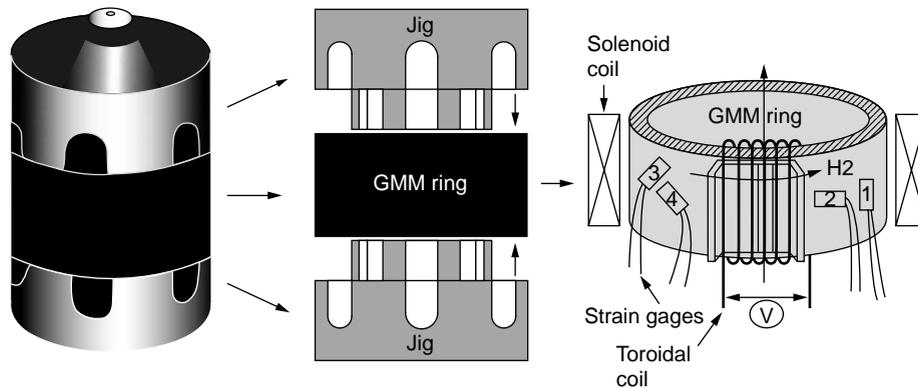


Figure 16. Torque sensor.

Torque Sensor (16). The design of a torque sensor is based on the principle of the inverse Wiedemann effect. As schematically shown in Fig. 16, a torque sensor is simply structured by a set of jigs where torque is applied and a GMM ring is sandwiched. The dimensions of the GMM ring used are OD = 25 mm, ID = 20 mm, and L = 15 mm. Using the Bridgman method, it is difficult to give the GMM a shape other than a cylinder or rectangular solid. On the other hand, the geometry of a powder metallurgical GMM is subject only to which molds are applied. This creates great flexibility in shaping GMM products. Using the recently developed MEMS technology like micromolding, a GMM element is also available on a micrometer scale.

A toroidal coil is made by winding 300 turns of $\phi 0.3$ mm wire around the GMM ring. It is used to generate a magnetic field circumferentially and also as the pickup coil to sense the voltage variation induced by the torque applied. A solenoid coil, which is made of 400 turns of $\phi 0.3$ mm wire, is placed outside the GMM ring to apply a vertical magnetic bias. In addition, four strain gauges are placed on the periphery of the GMM ring in 0° , 45° , 90° and 135° directions to monitor the twist strain applied to the GMM ring by torque. When a torque T_r is applied to a tube like the GMM ring, the sharing stress τ in the sensor is given as

$$\tau = r \frac{T_r}{I_p}, \quad (24)$$

where I_p is the polar moment of the inertia of area for the GMM ring expressed by $\pi(OD^4 - ID^4)/32$. Equation (24) indicates that the sharing stress τ increases proportionally with radius r and reaches its maximum when the radius becomes $r = OD/2$. The varied stress, however, can be simply represented by the average sharing stress at $r = (OD + ID)/2$ if $(OD - ID)$ is relatively small.

Like the scheme of pressure/force sensor, the sharing stress causes a change in the permeability of the GMM rod, which in turn induces an output voltage in the pickup coil. For such a torque sensor, similarly, the pickup coil requires an AC excitation current, and the output voltage is evaluated by its virtual value. In an experiment carried out under 200 N preload, 5 V and 1 kHz excitation current, it was found that the mechanical torque also changed

the permeability μ^T of the GMM material, which in turn changed the flux density.

Applied torque varying from 0 Nm to 1 Nm in 0.2 Nm increments and the corresponding outputs are plotted in Figure 17. The linearity is quite acceptable except that a little hysteresis exists.

Power Generator (17). When converting mechanical energy to electrical energy, a GMM can also output a current rather than voltage. Equation (14) theoretically describes the relationship between the mechanical stress and generated current and shows the possibility of building a GMM power generator.

Figure 18 is a drawing of a prototype that stores the electrical energy generated by mechanical force in a condenser. The GMM rod used for this device is 6 mm in diameter and 50 mm long and is preloaded at 0.2 MPa. The impact F is delivered directly to the rod via the movable yoke. The bias coil has 2560 turns and generates a 24-kA/m magnetic field and an 800-mA DC current. When a 36-g steel ball drops 10 cm to hit the movable yoke, the GMM rod generates 27 V from a 300-turn pickup coil, 86 V from 900 turns, and 700 V from 6750 turns.

The voltage induced by the impact is, as shown in Fig. 19a, a kind of AC current, thus needs to be rectified

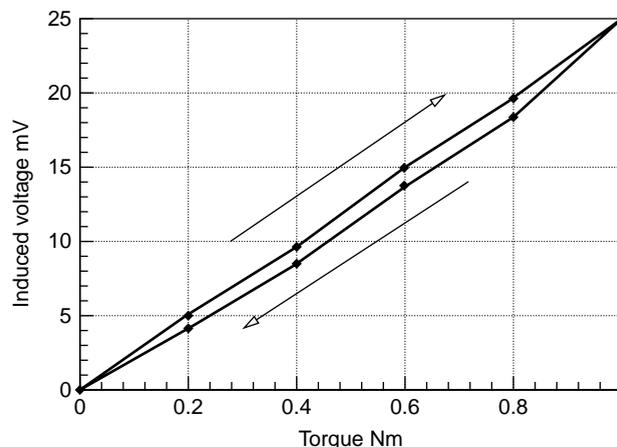


Figure 17. Torque applied versus voltage induced.

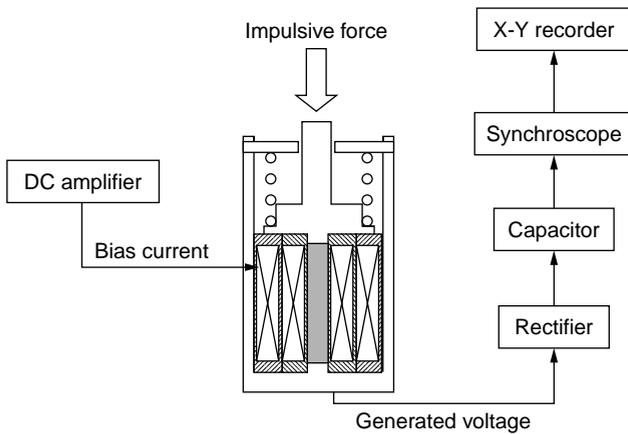


Figure 18. GMM power generator.

to waveform (b) before being stored in the condenser. In this case, each impact induces 45 V, equivalent to 2.2 mJ. The ratio of converted energy to the total kinetic energy of the steel ball is about 5%. Figure 20 shows the step of energy storage done by every impact of the steel ball falling from 15 cm above. The electrical energy after 20 impacts is sufficient to flash the strobe light of a compact camera (Fig. 21).

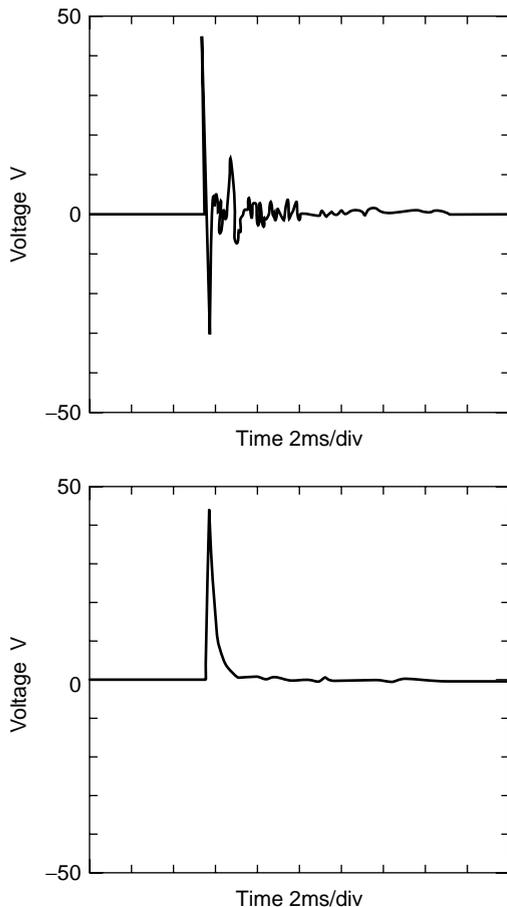


Figure 19. Voltage induced by impact.

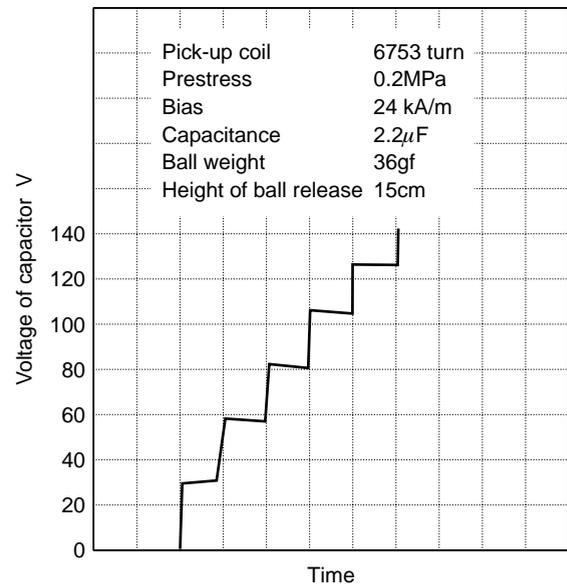


Figure 20. Process of energy storage.

Actuator Design

If only positive displacements are required, no magnetic bias is required. However, a bias is needed for linear actuation. The magnetic bias is normally introduced by permanent magnets or a DC bias current. In the sample curve of Fig. 4, a field of 500 Oe (40 kA/m) is chosen as a bias because it is midway between the minimum and maximum fields. This field $H_0 = nI$ where n is the number of turns per unit length and I is the DC current. Thus, for a current of 4 A, 10,000 turns per meter are necessary.

The performance of a GMM actuator is significantly improved under compressive mechanical bias (preload or prestress). The compressive bias also allows the material to be subjected to greater tension and to sustain its strength under shock. A compressive bias is normally given by using a high-strength stainless bolt. A simple arrangement using one bolt is illustrated in Fig. 22 where the

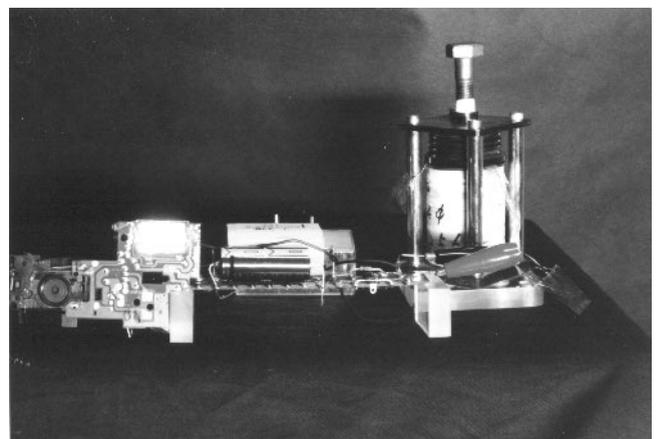


Figure 21. Camera strobe light flash.

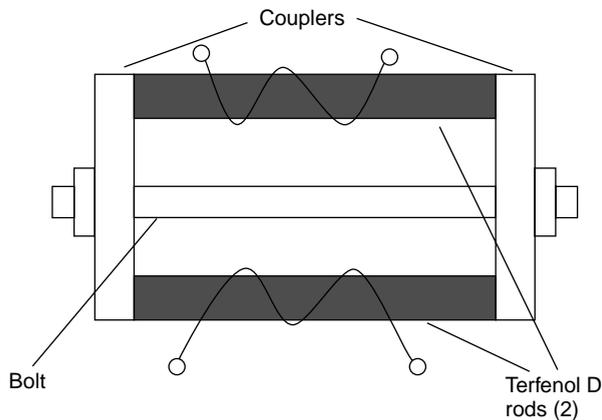


Figure 22. Mechanical bias.

two end pieces on the rods also provide a closed magnetic path.

An actuator is a device that puts things into motion. The cross-sectional view of a general giant magnetostriction actuator is shown in Fig. 23. The actuator is composed mainly of a magnetostrictive rod, yokes and a driving coil to make a magnetic circuit, and two permanent magnets to impose a bias magnetic field. SS41 steel is used for the yoke, and SmCo for the permanent magnet. A water-cooling device, made of a copper tube 2 mm in diameter, is also incorporated to isolate the heat between the driving coil and the magnetostrictive rod.

The distribution of magnetic flux in the magnetostrictive rod in connection with the bias condition is an important factor to be considered in designing an actuator. Figure 24 shows the distribution of magnetic flux at the inside of the solenoid coil in the various cases; (A) magnetic flux generated by the solenoid coil only, (B) magnetic flux generated by the SmCo magnet, and (C) magnetic flux in a

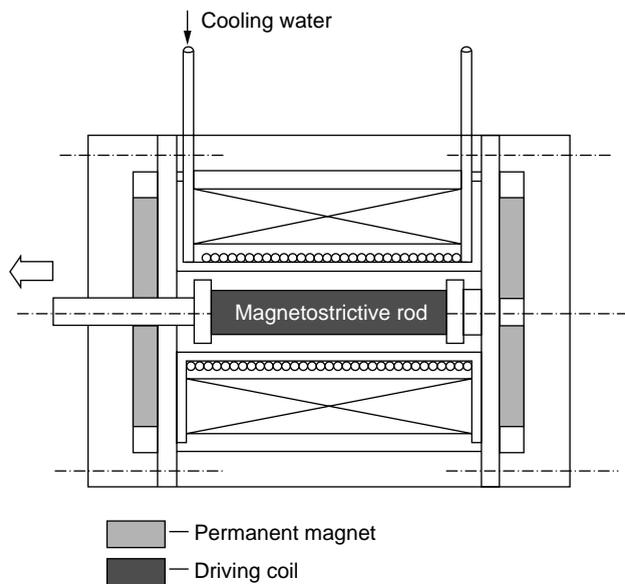


Figure 23. GMM actuator.

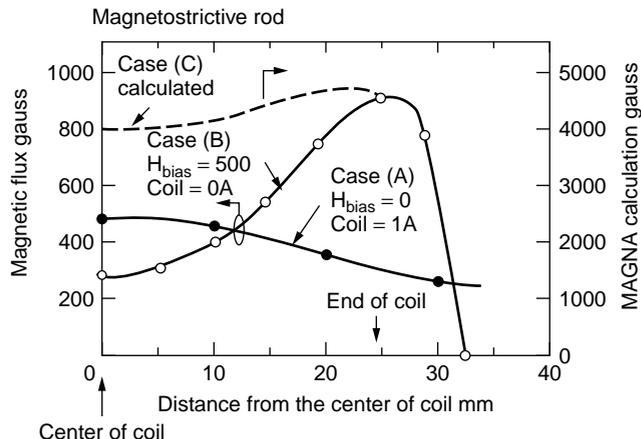


Figure 24. Distribution of magnetic flux.

rod under the real driving condition, which is computed by MAGNA code under the following conditions:

1. The permeability of a magnetostrictive rod is constant irrespective of driving conditions.
2. The permeability of both the permanent magnets for bias and the solenoid coil can be adjusted under any driving condition.

Each magnetic flux is shown by both curves of (A) (B) and (C) in Fig. 24 and depends on the distance from the center of the coil in the direction of the coil axis. On the other hand, the distribution of magnetic flux in a magnetostrictive rod is quite uniform under real conditions, as shown by curve. This result means that the displacement coefficient is almost constant in every position of the magnetostrictive rod.

It is also important in designing an actuator to note that the displacement hysteresis of the giant magnetostrictive rod depends remarkably on the bias magnetic field generated by the permanent magnet. The bias magnetic field dependence of the displacement hysteresis is shown in Figure 25, which is obtained for a full stroke $\pm 1 \mu\text{m}$.

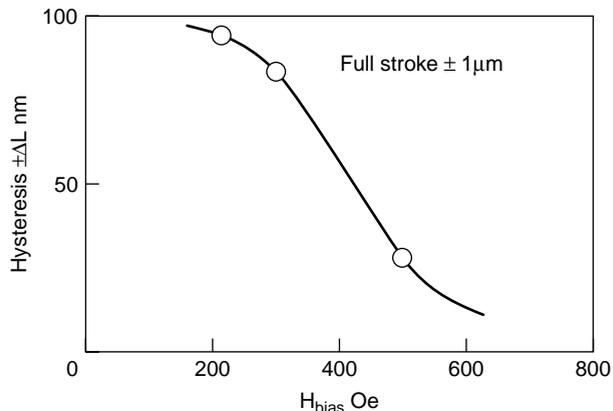


Figure 25. Bias magnetic dependence of displacement hysteresis.

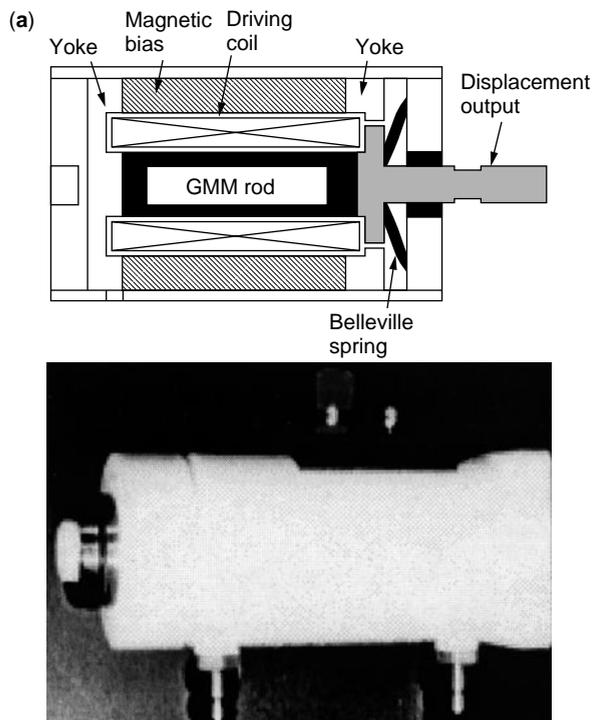


Figure 26. Linear positioner: (a) Cross-sectional view; (b) external view.

The hysteresis decreases markedly as the bias magnetic field increases. Therefore, the bias magnetic field should be properly selected to restrain the hysteresis.

Positioner (18). A DC field causes a GMM rod to extend by a specific displacement. Based on the dynamics of the GMM actuator mentioned before, an actuator has been designed to position the cutting tool for a diamond lathe. As shown schematically in Fig. 26, the positioner consists of a GMM rod, a solenoid coil for actuation, a Belleville spring for the mechanical preload, a set of permanent magnets and yoke for magnetic bias, and an output piston. The GMM rod is shaped into a cylinder of $\phi 6 \text{ mm} \times 40 \text{ mm}$ and is mechanically compressed by the Belleville spring under a 6-MPa preload. The magnetic bias applied is 400 Oe. A solenoid coil that has 1500 turns, can generate a 500-Oe magnetic field for every 1A of current input. The photo in Fig. 26 is an external view of the GMM positioner.

The response of the GMM positioner, at 100 nm/step, 15 nm/step, and 3 nm/step is shown in Figure 27. It confirms that the GMM positioner can deliver a large displacement ($40 \mu\text{m}$) without any assistance from the magnifier, while maintaining a resolution as fine as 3 nm. The GMM positioner is used to control the depth of cut (DOC) to achieve ductile mode cutting/grinding of hard, brittle materials such as silicon substrates, glasses, and ceramics.

Linear actuators are basically transduction devices but are usually confined to operation at low frequencies and produce greater motions than acoustic transducers. Accordingly, the design considerations for linear actuators may be somewhat different from acoustic transducers.

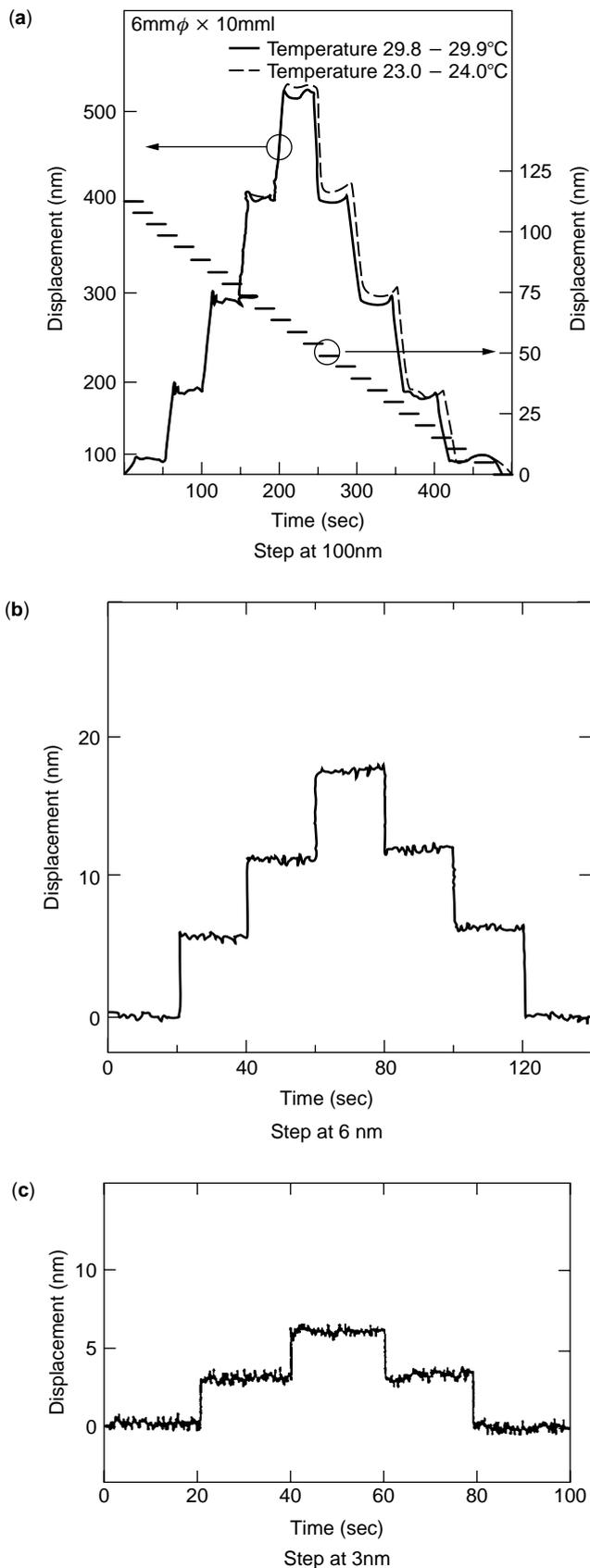


Figure 27. Step response of the positioner: (a) Step at 100 nm; (b) step at 6 nm; (c) step at 3 nm.

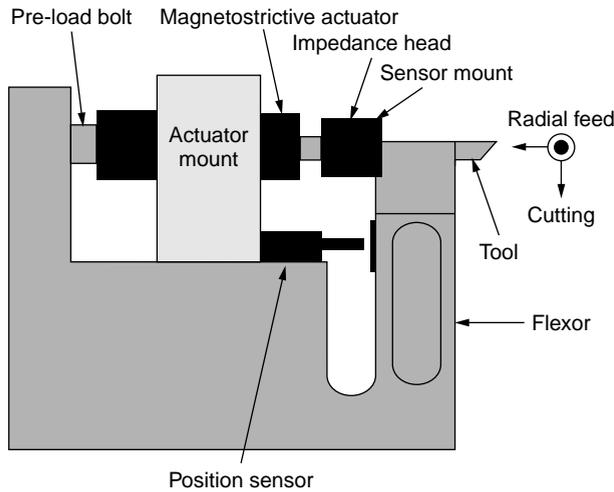


Figure 28. Machining application (19).

Increasing the length of the GMM rod increases the total displacement for a given current, and increasing the cross-sectional area increases the force of the rod for a given current. Therefore, the total work done by a GMM rod is proportional to the product of the force and the displacement, which is in turn subject to the volume of the GMM rod.

Other applications of a linear actuator include the active positioner (Fig. 28), which is used for nonround shape generation and counterchattering, the smart wing (Fig. 29) (19), which dynamically changes wing shape according to the flying speed to reduce petrol consumption, and the hydraulic actuator (Fig. 30), which is modularized and delivers high power output.

Pump (20). A GMM pump has been developed for a medical/biological application in delivering a microliter flow of liquid. As shown in Fig. 31, it consists of a GMM rod, a solenoid coil for magnetic field generation, a Belleville spring for preload, an external yoke for magnetic bias, a piston, check valves, and chambers. The GMM rod is shaped into a cylinder of $\phi 12 \text{ mm} \times 50 \text{ mm}$ and seats in the center of the solenoid coil. It is compressed by a Belleville

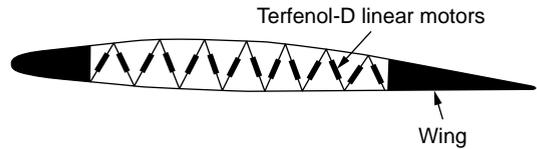


Figure 29. Smart wing (19).

spring with a 6-MPa preload against the external yoke. The coil, which has 5200 turns, is 50 mm long and 28.5 mm in diameter. For every ampere current input, the coil generates a 1.2-kA/m magnetic field.

Unlike a rotary pump driven by a motor, the GMM pump operates by converting its magnetostrain directly into reciprocating motion for the plunger. Figure 32 shows the operating cycle of the pump. When AC power input is in a sine wave form, the magnetic field generated is also synchronized to the sine wave. As the magnetic field (current) increases in (1), the GMM rod is expanded to push the piston, allowing the pump to discharge. When the magnetic field (current) decreases in (2), the GMM rod contracts. Meanwhile the Belleville spring pushes the piston back and makes the pump charge the liquid. Because the GMM rod responds positively to a negative magnetic field, the pump repeats the discharge-and-charge in the second half cycle (3) and (4). For every cycle of the input sine wave current, therefore, the pump reciprocates twice.

The flow rate Q_{th} (m^3/min) and the outlet pressure P_{th} (Pa) of the pump are, respectively, given as

$$Q_{th} = 120 \cdot f \cdot S_p \cdot \Delta L \quad (25)$$

$$P_{th} = P_m \cdot S_m / S_p \quad (26)$$

where f is the current frequency, S_p is the cross sectional area of the piston, ΔL is the piston stroke, and, P_m , S_m are the created stress and the cross-sectional area of the GMM rod. The actual measured flow rate–outlet pressure is plotted in Fig. 33 for the operating conditions of $H = 32 \text{ kA/m}$ and $f = 50 \text{ Hz}$. The test was carried out continuously for 10 days (4.3×10^7 cycles), and no failure or damage was found in the powder metallurgical GMM pump.

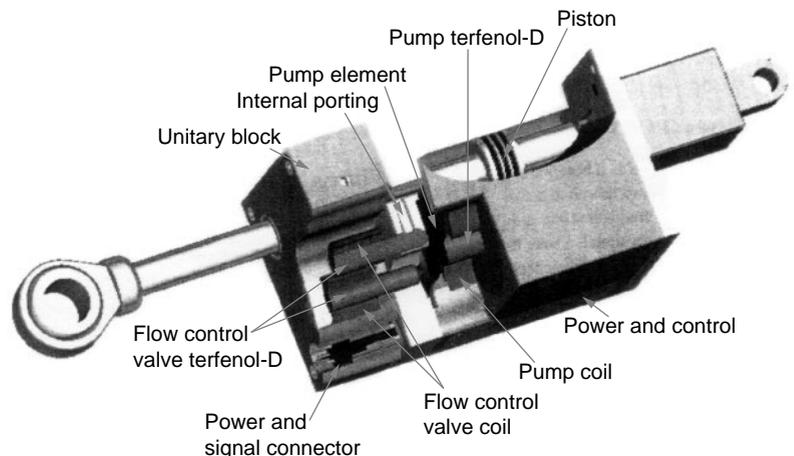


Figure 30. Hydraulic application (19).

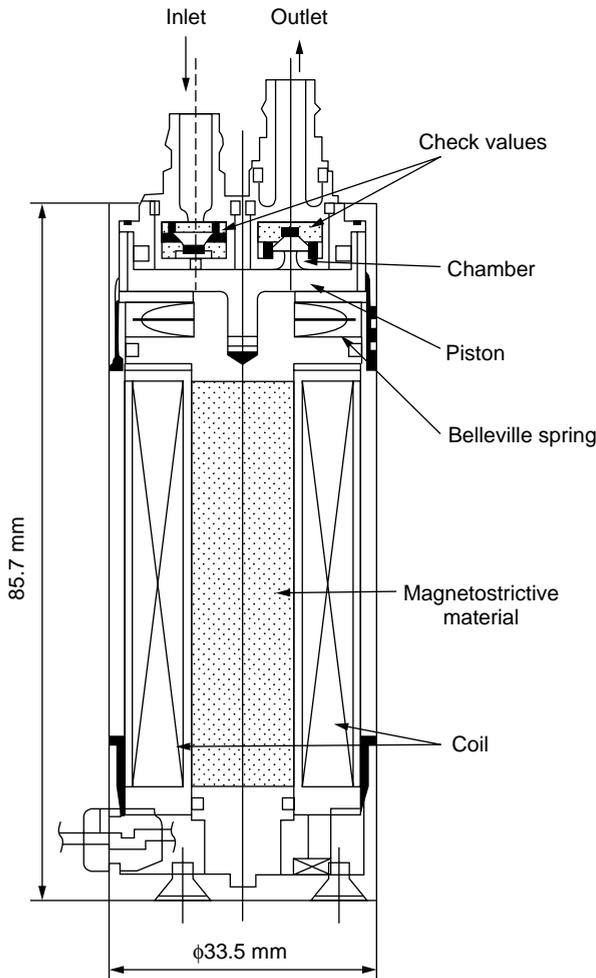


Figure 31. GMM pump.

Suppression (11). The GMM material is characterized by high power, large displacement, and fast response, as well as safety, long life and flexibility, which are often required by an aerospace structure. Vibration suppression for an aerospace structure has been explored as one application of a giant magnetostrictive actuator. A prototype of the vibration suppression device is shown schematically in Fig. 34. The beam is made of aluminum, the first resonant frequency is about 3 Hz or lower, and the acceleration at the tip is 1G or less. The beam is held vertically by a single-end supporter. The mass is added to the beam tip accordingly. Before starting suppression, necessary excitation is applied to the device. The suppression device consists of a pair of giant magnetostrictive actuators, electronic equipment for control, and a measuring instrument. A pin hinge is built in the beam supporter. The controllable force around the pin hinge is generated by the paired giant magnetostrictive actuator. Parameters measured include acceleration of the beam tip, strain of the beam root, and displacement of the magnetostrictive actuator; and any of them can be used as an input signal to be fed back to the control system.

The vibration suppression experiment was carried out with the prototype device without the feedback control, but

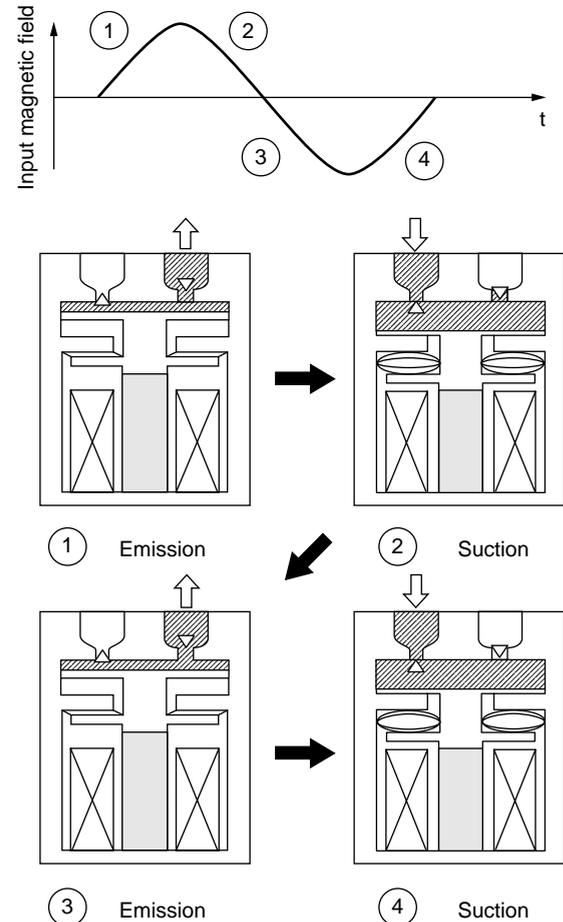


Figure 32. Operating cycle of GMM pump.

inputting a proper sine waveform corresponding to the resonance frequency and in the opposite phase of the vibration detected from the signal of the accelerometer. The control was continued until the vibration (the maximum acceleration) was constantly below a specified level. The beam was first excited from standstill by applying a sine waveform current (maximum 3A) that had the first-order frequency of the beam to the magnetostrictive actuators. The suppression was then started from the steady state of vibration, by

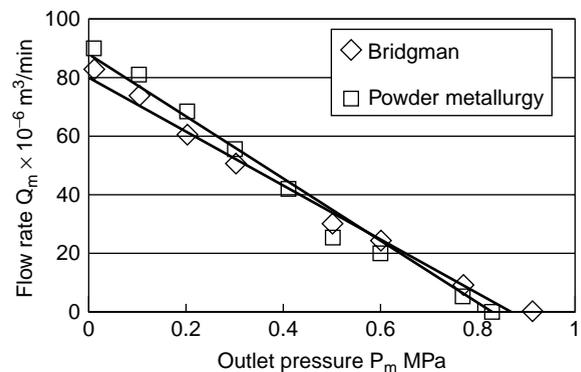


Figure 33. Flow rate versus outlet pressure (measured).

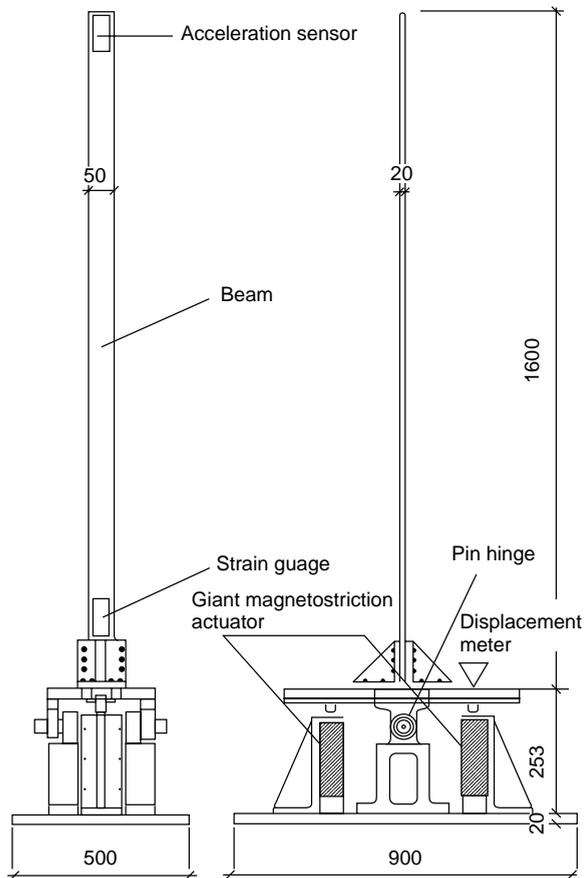


Figure 34. Prototype for space structure suppression.

giving the actuators a maximum 4-A current whose phase was inverted from that of the input current. To compensate for the phase deviation taking place during suppression, the phase of the suppression current was corrected every few seconds according to information from the accelerometer. Figure 35 shows the history of the acceleration at the

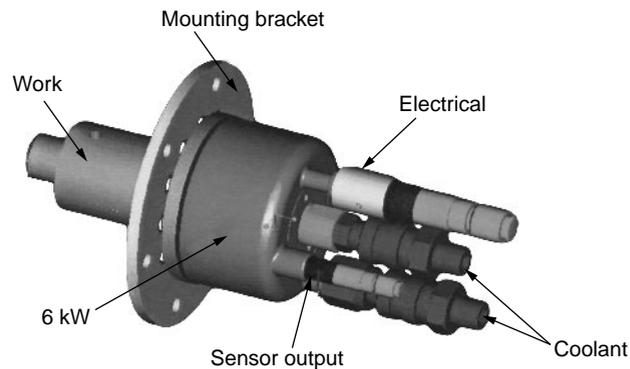


Figure 36. Ultrasonic transducer.

tip of the beam and the suppression current input to the actuators. The beam tip acceleration dramatically decreased about 10 seconds after suppression started.

Vibration Source (19). Another major application of a GMM actuator is as a vibration source, where the GMM rod is operated at its resonant frequency by properly selecting the AC driving current. There are two resonances in magnetostrictive transducers: the maximum and the minimum impedance resonance f^B , which occur under an open circuit and a short circuit, respectively. For a free bar of length l_1 the fundamental half wavelength resonance frequency conditions are $f^{Hl_1} = 845 \text{ kHz} \cdot \text{mm}$ or $f^{Bl_1} = 1225 \text{ kHz} \cdot \text{mm}$. For a free 2.5-inch GMM bar, the resonant frequency would be 13.32 kHz for an open circuit and 19.28 kHz for a short circuit. Operation at these frequencies would lead to considerable eddy current loss and reduced permeability and coupling coefficient. The critical frequency for GMM is written as $f_c = 0.0687/D^2$ where D is in meters. For a $D = 6.35\text{-mm}$ diameter rod, the critical frequency is 1.7 kHz. Reducing the eddy current losses is achievable either by increasing the length or reducing the diameter of the rods.

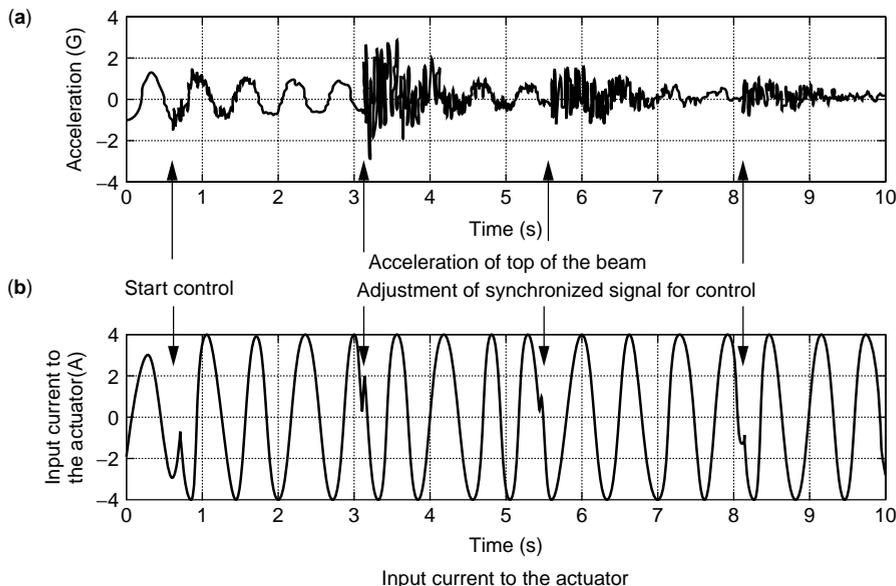


Figure 35. An example of a vibration suppression experimental result: (a) Acceleration of top of the beam; (b) input current to the actuator.

An ultrasonic transducer developed by ETREMA Products Inc. is shown in Figure 36. This transducer uses a $\phi 75 \text{ mm} \times 100 \text{ mm}$ Terfenol-D rod and can be continuously operated at $20 \text{ kHz} \pm 1 \text{ kHz}$ and output power of 6 kW. In the unloaded condition, the output displacement is $20 \mu\text{m}$ peak to peak at resonance.

This kind of vibration source has wide application in aerospace, medicine, sonochemical processing, and the naval and petroleum industries. Some of the commercialized products are listed here:

- Active noise and vibration control system
- Ultrasonic device
- Ultrasonic knives
- Surgical/dental tool
- Speakers—audio/warning system
- Cleaning device
- Hearing assistance/diving communication

BIBLIOGRAPHY

1. M. Madou, *Fundamentals of Micro Fabrication*, CRC Press, Boca Raton, FL, 1997.
2. H. Eda, New technoworld of evolution on machine tool and production processes (giant magnetostriction material towards a revolution on intelligent actuator and sensor). *Trans. Jpn. Soc. Mech. Eng.* **63** (614, C), 3321 (1997) (in Japanese).
3. A.E. Clark and H. Eda, *Giant Magnetostrictive Materials: Application on Micro-system and Actuators*, p. 159. Nikkan-Kogyo Shinbun Press, Japan, 1995, (in Japanese).
4. J.L. Butler, *Application Manual for the Design of ETREMA TERFENOL-D Magnetostrictive Transducers*, EDGE Technologies, 1988.
5. H. Eda, Y. Yamamoto et al., Giant magnetostrictive composite materials and its applications to actuators. *Proc. ECCM-8*, Italy, 1998, Vol. 3, pp. 549–556.
6. G.P. Carmen and M. Mitrovic, Nonlinear constitutive relationship for magnetostrictive materials with application to 1-D problems. *J. Intell. Mater., Syst. Struct.* **6**, 673 (1995).
7. M.B. Mofectteral, Characterization of TERFENOL-D for magnetostrictive transducers. *J. Acousti. Soc. Am.* **89** (3), 1448 (1991).
8. J.D. Snodgrass and O.D. McMaster, Optimized TERFENOL-D manufacturing processes. *J. Alloys Compd.* **258**, 24 (1997).
9. E.D. Gibson, J.D. Verhoeven et al., Continuous method for manufacturing grain-oriented magnetostrictive bodies. U.S. Pat. 4,770,704 ().
10. T. Mori, *TDK Product Literature*, Materials Research Center, TDK Corporation.
11. H. Eda, Y. Tomita, T. Mori, Y. Okada, J. Shimizu, and L. Zhou, Research and development of smart giant magnetostriction composite materials using powder metallurgical technology and their application to advanced devices. *Ext. Abstr., ICCM-12*, France, 1999, p. 751.
12. T. Honda, K.I. Arai, and M. Yamaguchi, Fabrication of magnetostrictive actuators using rare-earth(Tb,Sm)-Fe thin films. *Jpn. J. Appl. Phys.* **76**, 6994–6999 (1994) (in Japanese).
13. D.C. Jiles, Theory of ferromagnetic hysteresis. *J. Magn. Mater.* **61**, 43 (1986).
14. L. Kvarnsjö, A. Bergvist, and L. Engdahl, Application of stress-dependent magnetic Preisach hysteresis model on a simulation model for terfenol-D. *IEEE Trans.* **28** (5), 2623 (1992).
15. Y. Yamamoto, H. Eda, T. Mori, and A. Rathore, Three-dimensional magnetostrictive vibration sensor; Development, analysis and applications. *J. Alloys Compd.* **258**, 107 (1997).
16. H. Eda, T. Mori, L. Zhou, K. Kubota, and J. Shimizu, Powder metallurgical giant magnetostrictive material and its applications in micro-actuator and micro-sensors. *Proc. SPIE* **3875**, 114 (1999).
17. H. Eda, H. Nakamura, and Y. Yamamoto, Development of the giant magnetostriction electric generator prototype—Application of the reverse magnetostriction effect. *J. Jpn. Soc. Precis. Eng.* **63**, (5), 706 (1997) (in Japanese).
18. H. Eda, E. Ohmura, M. Sasaki, and T. Kobayashi, Ultra precise machine tool equipped with a giant magnetostriction actuator—Development of new materials, $\text{Tb}_x \text{Dy}_{1-x} (\text{Fe}_y \text{Mn}_{1-y})_n$ and their application. *Ann. CIRP* **41** (1), 421 (1992).
19. L. Langnau, *Actuator Solutions to Linear Problems, Linear Motion Technology Update*, Peton Media, 1996.
20. T. Mori, H. Eda, Y. Yamamoto, K. Suzuki, H. Nakamura, and C. Arai, Powder-metallurgical giant magnetostrictive material and its application to actuators—Development of magnetostrictive pump. *J. Jpn. Soc. Precis. Eng.* **62** (6), 891, (1996) (in Japanese).

HEALTH MONITORING (STRUCTURAL) USING WAVE DYNAMICS

DARRYLL J. PINES
University of Maryland
College Park, MD

INTRODUCTION

The extraction of the modal parameters of a structure for model-based damage detection methods has received a considerable amount of attention in the literature (1). Many of these model-based methods are widely used for problems of structural health monitoring of aerospace, civil, and mechanical systems. However, studies have shown that these methods have many limitations, including sensitivity to model uncertainty, changes in boundary conditions, insensitivity to incipient structural defects, and sensor and actuator placement (2,3). To counter these inherent problems, attention has focused on the development of alternative modeling approaches to describe the vibratory motion of structures under external excitation. One such approach is to represent the response of a structure in terms of the superposition of traveling waves that can traverse individual elements of a structure and reflect off boundaries to establish standing waves from constructive interference. These standing waves constitute the resonance or reverberant response of the structure and lead to reverberated transfer functions (RTF) from force excitation of sensor outputs. Common attributes of such frequency domain responses are characteristic pole-zero patterns for collocated and noncollocated transfer functions. Using these frequency response functions, many methods (4) have been developed to obtain modal properties of a variety of structures. However, often neglected in the process is that these RTF responses are comprised of the steady-state superposition of incident and reflected traveling waves propagating along the structure. These traveling wave components can be used to identify local damage in structural systems.

The idea of identifying damage using the scattering of transient structural waves (5,6) is directly related to the dereverberated response of a structure. Whereas the reverberated transfer function (RTF) of a structure measures the response when travelling waves interfere constructively and destructively with each other, the dereverberated transfer function (DTF) is a measure of the steady-state response of a structure when incident waves do not reflect off internal or external boundaries. Thus, the characteristic pole-zero behavior commonly found in RTF responses is suppressed in favor of the mean or “backbone” response of the RTF transfer function. The DTF response essentially represents the steady-state incident power that flows through the structure from each actuator to each sensor.

Initial attempts to suppress broadband vibration of structural waves involved adhoc feed-forward and local feedback control laws (7–12). MacMartin and Hall

(9) found that the optimal collocated feedback controller had noncausal properties that were not physically realizable by using conventional analog electronic components. However, many researchers (11,12) have developed a series of approximate collocated controllers that were physically realizable but not optimal in terms of energy dissipation. Such collocated controllers, it was demonstrated experimentally reduce the resonant response of uniform structures across a fairly broad frequency range that involves several modes of a structure. These approximate controllers were later used in an initial study (13) to infer the presence of delamination damage qualitatively in several tapered composite rotorcraft flexbeam specimens using the DTF response. This article develops the precise mathematical formalism for determining the matched terminating controllers required between successive nonuniform structural elements to obtain the DTF from the RTF response at every degree of freedom (DOF). The DTF is used to infer the presence and location of damage by examining the relative phase lag error between successive degrees of freedom. A damage index method is introduced to quantify the amount and type of damage (mass or stiffness) that occurs in a particular structural element. One-dimensional examples illustrate that the presence, location, and amount of local damage can be determined in discrete and continuous structures.

In the next section of this article, the concept of dereverberation is introduced, followed by a discussion of the way to obtain the DTF response of discrete and continuous structural elements. In the following section, the DTF response is obtained from simulated nonuniform discrete and continuous structures. Finally, in the last section, a damage detection method using the DTF response and a phase damage index method illustrate the effectiveness of the approach.

DEREVERBERATED TRANSFER FUNCTIONS OF STRUCTURAL ELEMENTS

Discrete Spring-Mass Elements

Discrete spring-mass structures can be constructed from two types of elements: symmetrical and asymmetrical (14). A symmetrical element is shown in Fig. 1a. It has identical masses m at both ends of the element connected by a spring whose stiffness is k in the middle. An asymmetrical element, shown in Fig. 1b, has only one mass attached to either end of the spring. A generic spring-mass element is shown in Fig. 1c. Note that when $m_1 = m_2$, the generic elements are symmetrical. When $m_1 = 0$, they are asymmetrical elements. This section introduces an approach for obtaining dereverberated transfer functions for asymmetrical spring-mass elements and for generic spring-mass elements.

Asymmetrical Spring-Mass Element. The equations of motion of the asymmetrical spring-mass element shown in

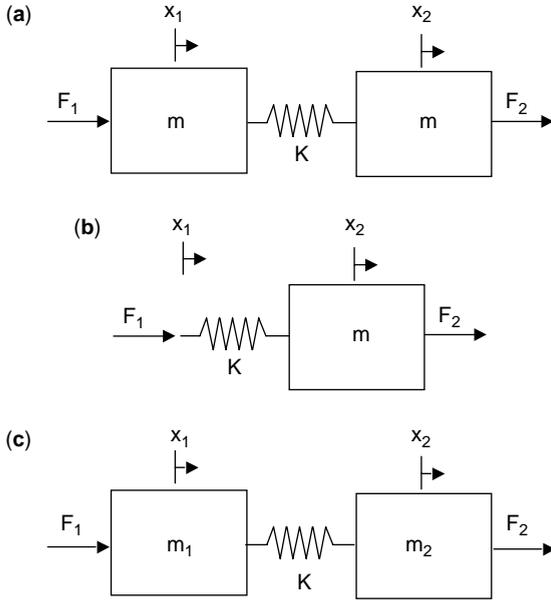


Figure 1. Discrete spring-mass elements: (a) symmetrical (b) asymmetrical (c) generic.

Fig. 1b are

$$\begin{cases} k(x_1 - x_2) = F_1 \\ m\ddot{x}_2 + k(x_2 - x_1) = F_2. \end{cases} \quad (1)$$

Assuming a solution of the following form:

$$x_n(t) = Ae^{i\omega t - n\mu} \quad n = 1, 2, \quad (2)$$

where n is the DOF of interest, leads to the following expressions for the wave numbers μ_l and μ_r :

$$\begin{aligned} \mu_l = \mu &= \cosh^{-1} \left(\frac{2k - m\omega^2}{2k} \right), \\ \mu_r = -\mu &= -\cosh^{-1} \left(\frac{2k - m\omega^2}{2k} \right). \end{aligned} \quad (3)$$

The subscripts l and r refer to the leftward and rightward propagation coefficients. Therefore, the spectral representations for the longitudinal displacements of each degree of freedom are

$$\begin{aligned} \hat{x}_1(\omega) &= W_{rp}(\omega)e^{-\mu} + W_{lp}(\omega)e^{\mu}, \\ \hat{x}_2(\omega) &= W_{rp}(\omega)e^{-2\mu} + W_{lp}(\omega)e^{2\mu}. \end{aligned} \quad (4)$$

Assuming that energy is input from the left where $F_1 \neq 0$, a controller can be added at the right to extract energy. The control force is designed to have the form $F_{2c} = G_d x_2$ and must satisfy two objectives: (1) equilibrium of forces at degree of freedom x_2 , and (2) prevent energy reflection between the elements. Therefore, the control force F_{2c} , must satisfy the following equations:

$$F_{2m} + F_{2c} = 0,$$

and

$$W_{lp}(\omega) = 0, \quad (5)$$

where F_{2m} is the motion-dependent force. From the first part of Eq. (5),

$$m\ddot{x}_2 + k(x_2 - x_1) + G_d x_2 = 0. \quad (6)$$

Transformed into the frequency domain, Eq. (6) becomes

$$-m\omega^2 \hat{x}_2 + k(\hat{x}_2 - \hat{x}_1) + \hat{G}_d \hat{x}_2 = 0. \quad (7)$$

Now let us substitute Eq. (5) in Eq. (7) and set $W_{lp} = 0$. After some algebra, the control gain is determined to be

$$\hat{G}_d = m\omega^2 - k + ke^{\mu}. \quad (8)$$

If the RTF matrix of an asymmetrical element is available, transfer functions \hat{x}_1/\hat{F}_1 and \hat{x}_2/\hat{F}_1 can be dereverberated by using the following expression:

$$\begin{bmatrix} \hat{x}_1 \\ \hat{F}_1 \\ \hat{x}_2 \\ \hat{F}_1 \end{bmatrix}_{DTF} = \left(\text{RTF}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \hat{G}_d \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \quad (9)$$

RTF and DTF responses of \hat{x}_1/\hat{F}_1 and \hat{x}_2/\hat{F}_1 are shown in Fig. 2a, b, respectively. As expected, the effect of the control force F_{2c} , is to remove the resonance and antiresonance from the response. The control gain for the asymmetrical structural elements are plotted versus ω in Fig. 2c. Note that this compensator has an unusual magnitude and phase behavior that is not consistent with conventional linear pole-zero representations. The magnitude of behavior for the compensator has a derivative type characteristic up until the cutoff frequency. However, the phase is not necessarily equal to 90° . Thus, such a compensator would be difficult to implement by using conventional electronic components.

Generic Structural Elements. Considering the generic element shown in Fig. 3a, energy is input from the left in terms of $F_1 \neq 0$. A control force F_{2c} is applied at the right end to prevent energy reflection. This generic element can be separated into two parts, shown in Fig. 3b.

The first part of Fig. 3b displays the force boundary condition. The equation of motion at DOF x_1 is given by

$$F_1 = m_1 \ddot{x}_1 + F_w, \quad (10)$$

where F_w , is the force that generates a wave propagating through k_2 and m_2 . The control force $F_{2c} = G_d x_2$ prevents energy that arrives at m_2 from being reflected. From the discussion in the previous section, the controller gain G_d is determined only by parameters m_2 and k_2 . The control gain is independent of F_w , which means that it is independent of m_1 . m_1 , joined by F_1 , determines the speed and amplitude of energy passing through the element but has nothing to do with the controller applied at the right end to obtain the nonresonant transfer functions. If the conventional

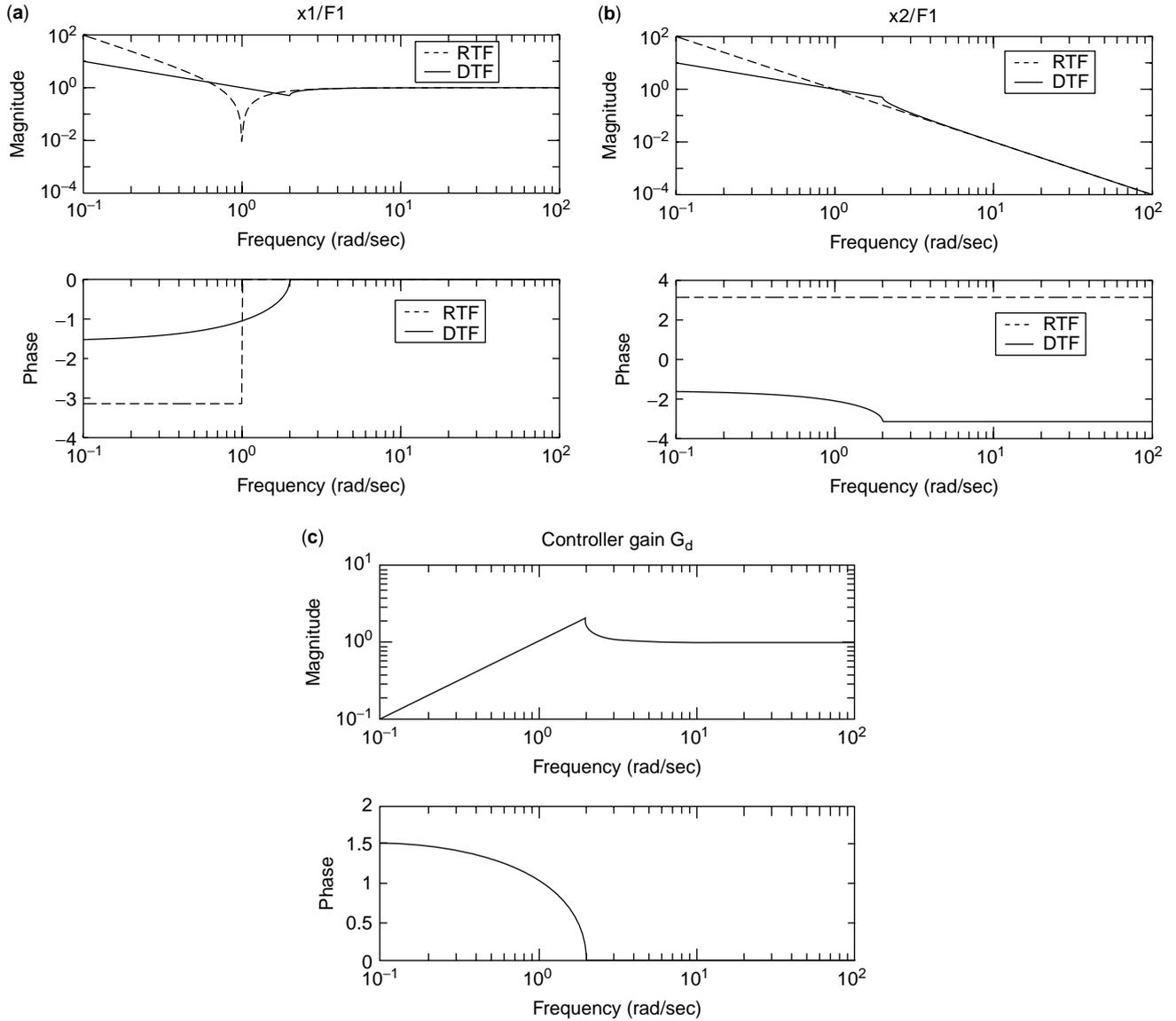


Figure 2. Asymmetrical element: $k_2 = 1 \times (1 + 0.00001i)$; $m_2 = 1$. (a) RTF and DTF responses of \hat{x}_1/\hat{F}_1 ; (b) RTF and DTF responses of \hat{x}_2/\hat{F}_1 ; (c) control gain \hat{G}_d .

reverberated transfer function matrix (RTF) of the generic element is known, transfer functions \hat{x}_1/\hat{F}_1 and \hat{x}_2/\hat{F}_1 can be dereverberated using the following expression:

$$\begin{bmatrix} \hat{x}_1 \\ \hat{F}_1 \\ \hat{x}_2 \\ \hat{F}_1 \end{bmatrix}_{\text{DTF}} = \left(\text{RTF}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \hat{G}_d \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \quad (11)$$

The DTF responses \hat{x}_1/\hat{F}_1 and \hat{x}_2/\hat{F}_1 that correspond to different m_1 and m_2 values are shown in Fig. 4.

Continuous Structural Elements

Although discrete spring-mass structural elements can be used to model a variety of complex structures, continuous structural elements are also being used to model structures

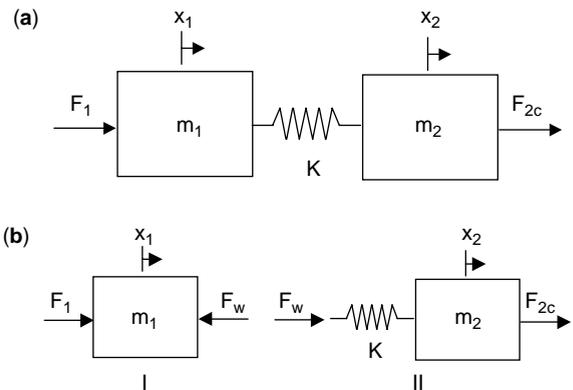


Figure 3. Generic structural element.

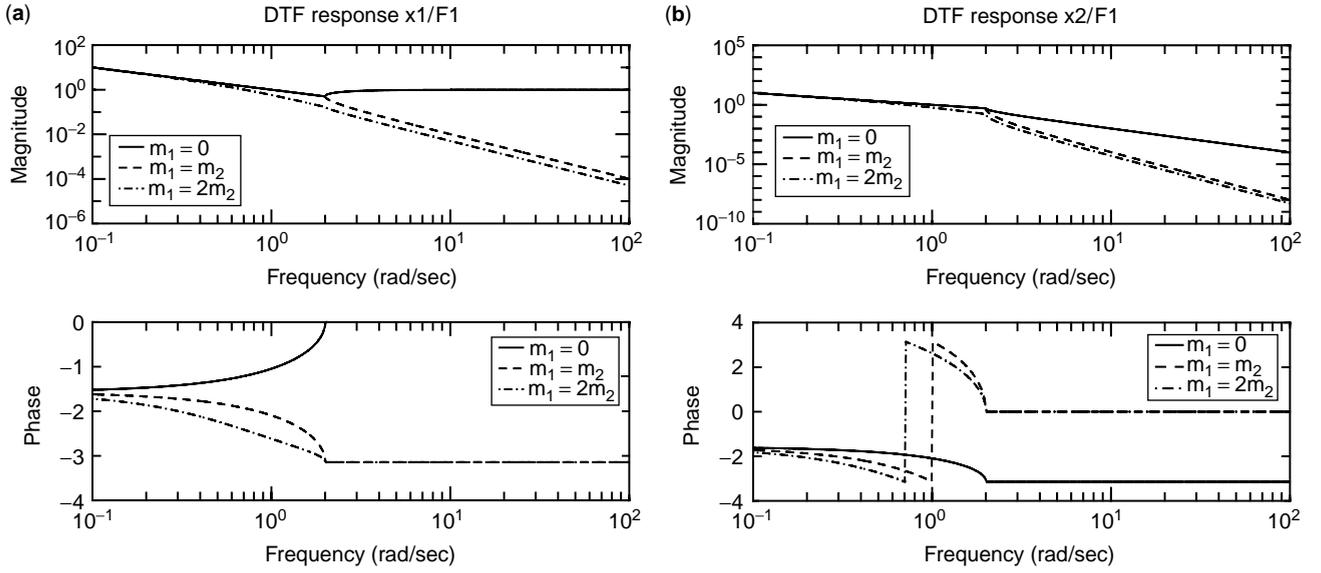


Figure 4. DTF responses of generic elements.

such as space trusses, rotor blades, and buildings. In this section, models of rod and beam elements are developed. Using the concept of the virtual controller, the dereverberated response is obtained for continuous rod and beam spectral elements.

Uniform Rod Element. A uniform rod element is shown in Fig. 5. The equation of motion for a rod is given by

$$EA \frac{\partial^2 u(x, t)}{\partial x^2} - \rho A \frac{\partial^2 u(x, t)}{\partial t^2} = f(x, t). \quad (12)$$

Transformed into the frequency domain, Eq. (12) becomes

$$EA \frac{\partial^2 \hat{u}(x, \omega)}{\partial x^2} + \rho A \omega^2 \hat{u}(x, \omega) = \hat{f}(x, \omega). \quad (13)$$

The spectral representation (15) for the longitudinal deflection of the rod is given by

$$\hat{u}(x, \omega) = W_{rp}(\omega) e^{-i\mu x} + W_{lp}(\omega) e^{-i\mu(L-x)}, \quad (14)$$

where $\mu = \omega \sqrt{\rho A / EA}$ is the longitudinal wave number. Assume that energy is input from the left end where $F_1 \neq 0$. As the incident energy arrives at the right end, a control force $F_{2c} = G_r u_2$ is designed to satisfy two objectives: (1) equilibrium of forces at the right end and (2) prevent

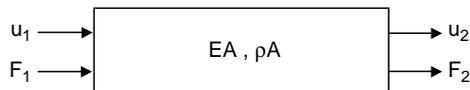


Figure 5. Uniform rod element.

energy reflection at the right end. Therefore, the control force must satisfy

$$F_{2m} + F_{2c} = 0,$$

and

$$W_{lp}(\omega) = 0, \quad (15)$$

where $F_{2m} = EA \frac{\partial \hat{u}}{\partial x} |_{x=L}$ is the motion-dependent force. After some algebra, the control gain is determined as

$$\hat{G}_r = i\mu EA. \quad (16)$$

If the RTF matrix of a uniform rod element is available, the dereverberated transfer functions \hat{u}_1 / \hat{F}_1 , \hat{u}_2 / \hat{F}_1 can be obtained by using the following expression:

$$\begin{bmatrix} \hat{u}_1 \\ \hat{F}_1 \\ \hat{u}_2 \\ \hat{F}_1 \end{bmatrix}_{DTF} = \left(\text{RTF}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \hat{G}_r \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \quad (17)$$

The DTF and RTF responses of \hat{u}_1 / \hat{F}_1 , \hat{u}_2 / \hat{F}_1 are shown in Fig. 6.

It should be pointed out that transfer functions \hat{u}_1 / \hat{F}_2 , \hat{u}_2 / \hat{F}_2 are not dereverberated by Eq. (17). These transfer functions can be dereverberated by

$$\begin{bmatrix} \hat{u}_1 \\ \hat{F}_2 \\ \hat{u}_2 \\ \hat{F}_2 \end{bmatrix}_{DTF} = \left(\text{RTF}^{-1} + \begin{bmatrix} \hat{G}_r & 0 \\ 0 & 0 \end{bmatrix} \right)^{-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \quad (18)$$

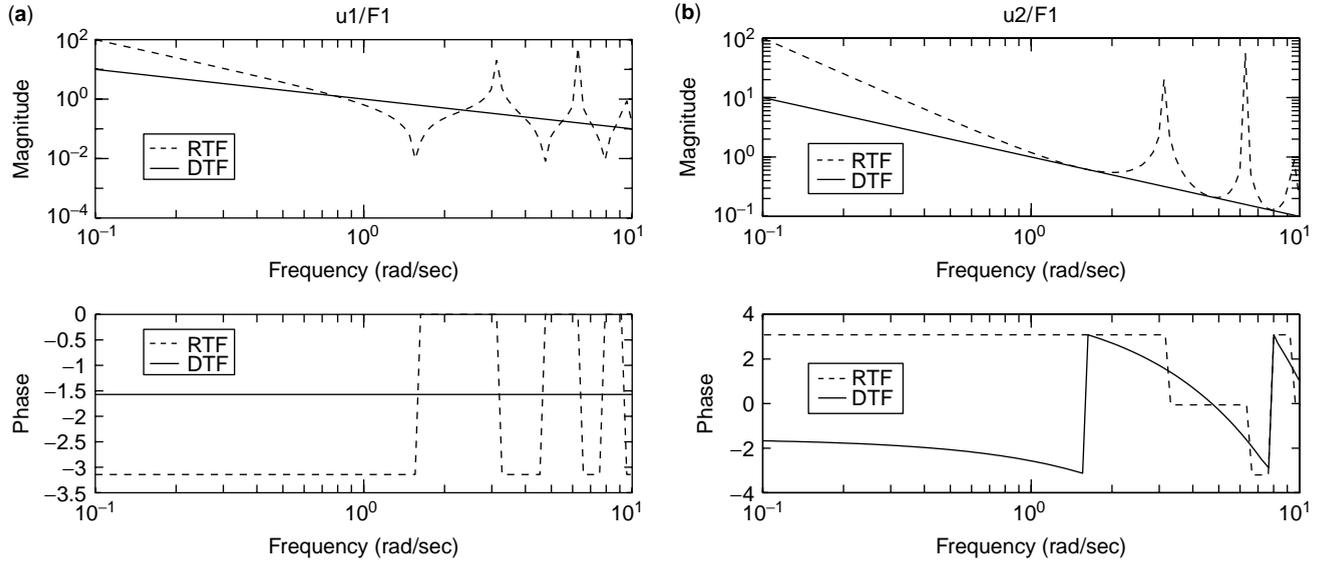


Figure 6. RTF and DTF responses of uniform rod element where $EA = 1 \times (1 + 0.000001i)$, $\rho A = 1$, $L = 1$. (a) \hat{u}_1/\hat{F}_1 , (b) \hat{u}_2/\hat{F}_1 .

Uniform Beam Element. A uniform beam element is shown in Fig. 7. The equation of motion for a Bernoulli–Euler beam is given by

$$EI \frac{\partial^4 v(x, t)}{\partial x^4} + \rho A \frac{\partial^2 v(x, t)}{\partial t^2} = f(x, t). \quad (19)$$

Transformed into the frequency domain, Eq. (19) becomes

$$EI \frac{\partial^4 \hat{v}(x, \omega)}{\partial x^4} - \rho A \omega^2 \hat{v}(x, \omega) = \hat{f}(x, \omega). \quad (20)$$

The spectral representation for the beam's flexural vibration is

$$\begin{cases} \hat{v} = W_{rp} e^{-i\mu x} + W_{re} e^{-\mu x} + W_{lp} e^{-i\mu(L-x)} + W_{le} e^{-\mu(L-x)} \\ \hat{\theta} = W_{rp}(-i\mu)e^{-i\mu x} + W_{re}(-\mu)e^{-\mu x} + W_{lp}(i\mu)e^{-i\mu(L-x)} \\ \quad + W_{le}(\mu)e^{-\mu(L-x)}. \end{cases} \quad (21)$$

where v is vertical displacement and θ is angular displacement. $\mu = (\frac{\rho A}{EA} \omega^2)^{1/4}$ is the bending wave number.

Different sign conventions lead to different expressions of nodal displacements and nodal forces. The sign convention applied here is shown in Fig. 7. Nodal displacements are given by

$$\begin{bmatrix} v_0 \\ \theta_0 \\ v_1 \\ \theta_1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & e^{-i\mu L} & e^{-\mu L} \\ -i\mu & -\mu & i\mu e^{-i\mu L} & \mu e^{-\mu L} \\ e^{-i\mu L} & e^{-\mu L} & 1 & 1 \\ -i\mu e^{-i\mu L} & -\mu e^{-\mu L} & i\mu & \mu \end{bmatrix} \begin{bmatrix} W_{rp} \\ W_{re} \\ W_{lp} \\ W_{le} \end{bmatrix}. \quad (22)$$

Nodal forces $F_0 = EI \frac{\partial^3 v}{\partial x^3} |_{x=0}$, $M_0 = -EI \frac{\partial^2 v}{\partial x^2} |_{x=0}$, $F_1 = -EI \frac{\partial^3 v}{\partial x^3} |_{x=L}$, and $M_1 = EI \frac{\partial^2 v}{\partial x^2} |_{x=L}$, are given by

$$\begin{bmatrix} F_0 \\ M_0 \\ F_1 \\ M_1 \end{bmatrix} = EI \mu^2 \begin{bmatrix} i\mu & -\mu & -i\mu e^{-i\mu L} & \mu e^{-\mu L} \\ 1 & -1 & e^{-i\mu L} & -e^{-\mu L} \\ -i\mu e^{-i\mu L} & \mu e^{-\mu L} & i\mu & -\mu \\ -e^{-i\mu L} & e^{-\mu L} & -1 & 1 \end{bmatrix} \begin{bmatrix} W_{rp} \\ W_{re} \\ W_{lp} \\ W_{le} \end{bmatrix} \quad (23)$$

Assuming that energy is input from the left, control forces

$$\begin{bmatrix} F_{1c} \\ M_{1c} \end{bmatrix} = G_{br} \begin{bmatrix} v_1 \\ \theta_1 \end{bmatrix} \quad (24)$$

are applied at the right end to satisfy the following equations:

$$\begin{bmatrix} F_{1c} \\ M_{1c} \end{bmatrix} + \begin{bmatrix} F_1 \\ M_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

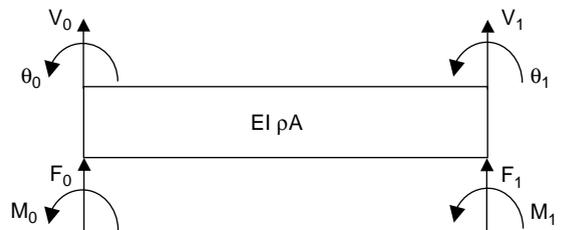


Figure 7. Uniform beam element.

and

$$\begin{bmatrix} W_{lp}(\omega) \\ W_{le}(\omega) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \quad (25)$$

Solving for the control gain matrix leads to

$$\hat{G}_{br} = \begin{bmatrix} (i-1)EI\mu^3 & iEI\mu^2 \\ iEI\mu^2 & (1+i)EI\mu \end{bmatrix}. \quad (26)$$

Using this matrix control law, the dereverberated transfer functions can be obtained from

$$\begin{bmatrix} \frac{\hat{v}_0}{\hat{F}_0} & \frac{\hat{v}_0}{\hat{M}_0} \\ \frac{\hat{\theta}_0}{\hat{F}_0} & \frac{\hat{\theta}_0}{\hat{M}_0} \\ \frac{\hat{v}_1}{\hat{F}_0} & \frac{\hat{v}_1}{\hat{M}_0} \\ \frac{\hat{\theta}_1}{\hat{F}_0} & \frac{\hat{\theta}_1}{\hat{M}_0} \end{bmatrix}_{DTF} = \left(\text{RTF}^{-1} + \begin{bmatrix} 0_{4 \times 4} & 0_{4 \times 4} \\ 0_{4 \times 4} & \hat{G}_{br} \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}. \quad (27)$$

The DTF and RTF response for \hat{v}_0/\hat{F}_0 , $\hat{v}_0/\hat{M}_0 \dots \hat{\theta}_1/\hat{F}_0$, $\hat{\theta}_1/\hat{M}_0$ are shown in Fig. 8.

If the energy is from the right, different control forces

$$\begin{bmatrix} F_{0c} \\ M_{0c} \end{bmatrix} = \hat{G}_{bl} \begin{bmatrix} v_0 \\ \theta_0 \end{bmatrix} \quad (28)$$

The control gains are given by

$$\hat{G}_{bl} = \begin{bmatrix} (i-1)EI\mu^3 & -iEI\mu^2 \\ -iEI\mu^2 & (1+i)EI\mu \end{bmatrix}. \quad (30)$$

The dereverberated transfer functions are obtained from

$$\begin{bmatrix} \frac{\hat{v}_0}{\hat{F}_1} & \frac{\hat{v}_0}{\hat{M}_1} \\ \frac{\hat{\theta}_0}{\hat{F}_1} & \frac{\hat{\theta}_0}{\hat{M}_1} \\ \frac{\hat{v}_1}{\hat{F}_1} & \frac{\hat{v}_1}{\hat{M}_1} \\ \frac{\hat{\theta}_1}{\hat{F}_1} & \frac{\hat{\theta}_1}{\hat{M}_1} \end{bmatrix}_{DTF} = \left(\text{RTF}^{-1} + \begin{bmatrix} \hat{G}_{bl} & 0_{4 \times 4} \\ 0_{4 \times 4} & 0_{4 \times 4} \end{bmatrix} \right)^{-1} \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}. \quad (31)$$

Remember that the controllers attached at both ends of rod elements are the same. But for a beam, the controller gain matrices \hat{G}_{br} and \hat{G}_{bl} have a sign difference in the off-diagonal terms because the sign convention that was used for the beam did not create a symmetrical element stiffness matrix.

DTF RESPONSES OF NONUNIFORM STRUCTURES

Discrete Spring-Mass Structures

Energy Is Input from the End. A five DOF spring-mass structure is shown in Fig. 9. The properties of this structure are listed in Table. H. Complex stiffness is used to add a small amount of damping to the structure and to overcome potential numerical problems. The right end of this structure is fixed, and the left end is free. Energy is input from the left in terms of $F_1 \neq 0$. The reverberated transfer function matrix RTF of this structure is given by

$$\begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \\ \hat{x}_3 \\ \hat{x}_4 \\ \hat{x}_5 \end{bmatrix} = \text{RTF} \begin{bmatrix} \hat{F}_1 \\ \hat{F}_2 \\ \hat{F}_3 \\ \hat{F}_4 \\ \hat{F}_5 \end{bmatrix} = \begin{bmatrix} k_1 - m_1\omega^2 & -k_1 & 0 & 0 & 0 \\ -k_1 & k_1 + k_2 - m_2\omega^2 & -k_2 & 0 & 0 \\ 0 & -k_2 & k_2 + k_3 - m_3\omega^2 & -k_3 & 0 \\ 0 & 0 & -k_3 & k_3 + k_4 - m_4\omega^2 & -k_4 \\ 0 & 0 & 0 & -k_4 & k_4 + k_5 - m_5\omega^2 \end{bmatrix}^{-1} \begin{bmatrix} \hat{F}_1 \\ \hat{F}_2 \\ \hat{F}_3 \\ \hat{F}_4 \\ \hat{F}_5 \end{bmatrix}. \quad (32)$$

are applied at the left end to satisfy the following equations:

$$\begin{bmatrix} F_{0c} \\ M_{0c} \end{bmatrix} + \begin{bmatrix} F_0 \\ M_0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

and

$$\begin{bmatrix} W_{rp}(\omega) \\ W_{re}(\omega) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \quad (29)$$

To obtain the dereverberated transfer functions with respect to external force F_1 , the structural part to the right of mass m_5 will be ignored for two reasons. First, the original wave has already passed all DOFs after it arrives at m_5 . Structural components behind m_5 along the wave length will be ignored because DTF deals with the incident wave. Second, the fixed boundary condition should be ignored; otherwise, it will generate a reflected wave. The control scheme to obtain DTF responses is shown in Fig. 10. F_w is derived from the force boundary condition at the left. Local controllers are used to match the impedance of adjacent elements. There is no energy reflection at the interface

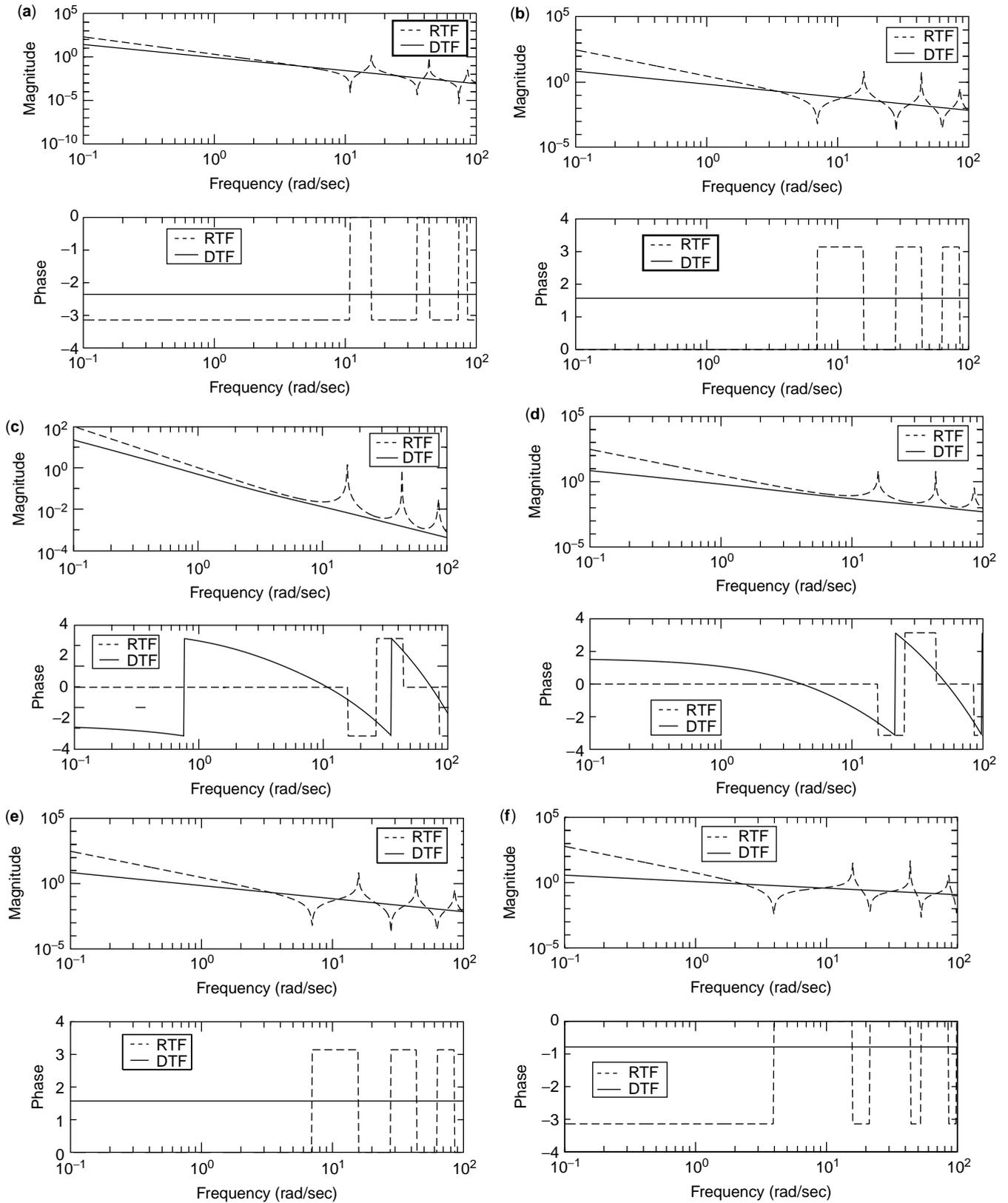


Figure 8. RTF and DTF responses: (a) \hat{v}_0 / \hat{F}_0 ; (b) $\hat{\theta}_0 / \hat{F}_0$; (c) \hat{v}_1 / \hat{F}_0 ; (d) $\hat{\theta}_1 / \hat{F}_0$; (e) \hat{v}_0 / \hat{M}_0 ; (f) $\hat{\theta}_0 / \hat{M}_0$; (g) \hat{v}_1 / \hat{M}_0 ; (h) $\hat{\theta}_1 / \hat{M}_0$.

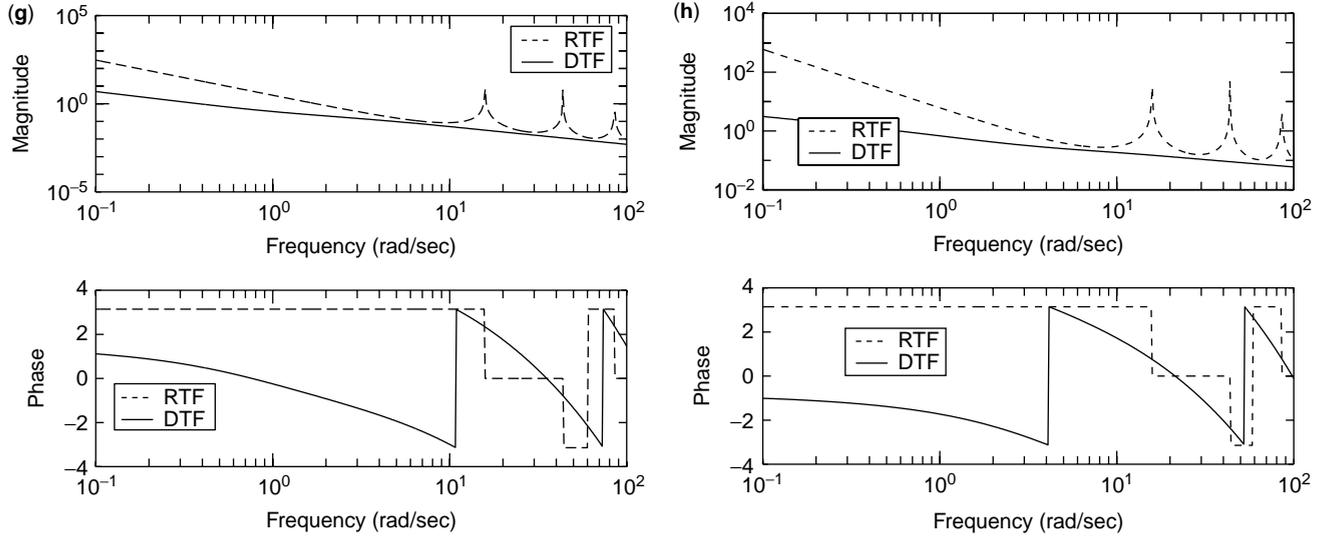


Figure 8. Continued.

of different elements. Mathematically, the dereverberated transfer functions for $\hat{x}_1/\hat{F}_1, \hat{x}_2/\hat{F}_1 \dots \hat{x}_5/\hat{F}_1$ can be found by using the following expression:

$$\begin{bmatrix} \hat{x}_1 \\ \hat{F}_1 \\ \hat{x}_2 \\ \hat{F}_1 \\ \hat{x}_3 \\ \hat{F}_1 \\ \hat{x}_4 \\ \hat{F}_1 \\ \hat{x}_5 \\ \hat{F}_1 \end{bmatrix}_{\text{DTF}} = \left(\text{RTF}^{-1} + \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & G_{2d} - G_{3d} & 0 & 0 & 0 \\ 0 & 0 & G_{3d} - G_{4d} & 0 & 0 \\ 0 & 0 & 0 & G_{4d} - G_{5d} & 0 \\ 0 & 0 & 0 & 0 & G_{5d} - k_5 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad (33)$$

where $G_{id} = m_i \omega^2 - k_{i-1} + k_{i-1} e^{\mu_i}$ and $\mu_i = \cosh^{-1} \left(\frac{2k_{i-1} - m_i \omega^2}{2k_{i-1}} \right)$. Now, we know the DTF responses of $\hat{x}_1/\hat{F}_1, \hat{x}_2/\hat{F}_1 \dots \hat{x}_5/\hat{F}_1$. For damage detection, we are more interested in how an individual element affects a wave passing through it. This is demonstrated by DTF responses $\frac{\hat{x}_{i+1}}{\hat{x}_i}, (i = 1, 2, \dots, 4)$ of successive degrees of freedom. These transfer function responses are shown in Fig. 11.

$$\text{RTF} = \begin{bmatrix} k_1 - m_1 \omega^2 & -k_1 & 0 & 0 & 0 \\ -k_1 & k_1 + k_2 - m_2 \omega^2 & -k_2 & 0 & 0 \\ 0 & -k_2 & k_2 + k_3 - m_3 \omega^2 & -k_3 & 0 \\ 0 & 0 & -k_3 & k_3 + k_4 - m_4 \omega^2 & -k_4 \\ 0 & 0 & 0 & -k_4 & k_4 + k_5 - m_5 \omega^2 \end{bmatrix}^{-1}. \quad (34)$$

Energy Is Input from the Middle. A five-DOF spring-mass structure that has energy input at the middle ($F_3 \neq 0$) is

shown in Fig. 12a. Spring stiffness k_5 and the fixed end will be ignored to obtain DTF responses. Hence, this structure can be thought of as constructed from two parts, shown in

Fig. 12b. Each part reveals the path of an incident wave propagating in both directions, respectively. Adding these two parts together, an alternate structure is used to obtain the DTF of the original structure with respect to input F_3 (see Fig. 12c).

The RTF response of the original structure shown in Fig. 12a is given by

Note that the alternate structure is different from the original structure in two ways: (1) the spring attached to

Table 1. Properties of the Five-DOF Spring-Mass Structure

i	1	2	3	4	5
k_i	$20 \times (1 + 0.0001i)$	$10 \times (1 + 0.0001i)$	$10 \times (1 + 0.0001i)$	$20 \times (1 + 0.0001i)$	$30 \times (1 + 0.0001i)$
m_i	2	2	1	3	1

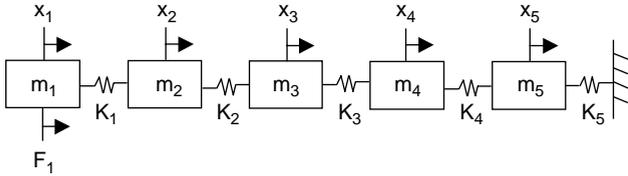


Figure 9. Five-DOF spring-mass structure.

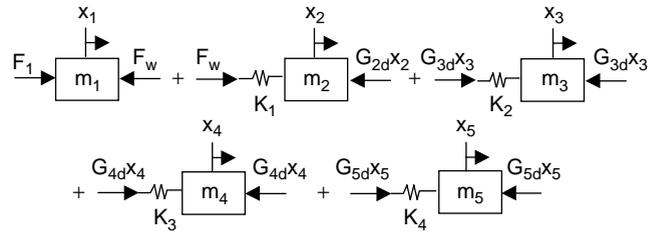


Figure 10. Control scheme to obtain DTF responses for the structure shown in Fig. 9.

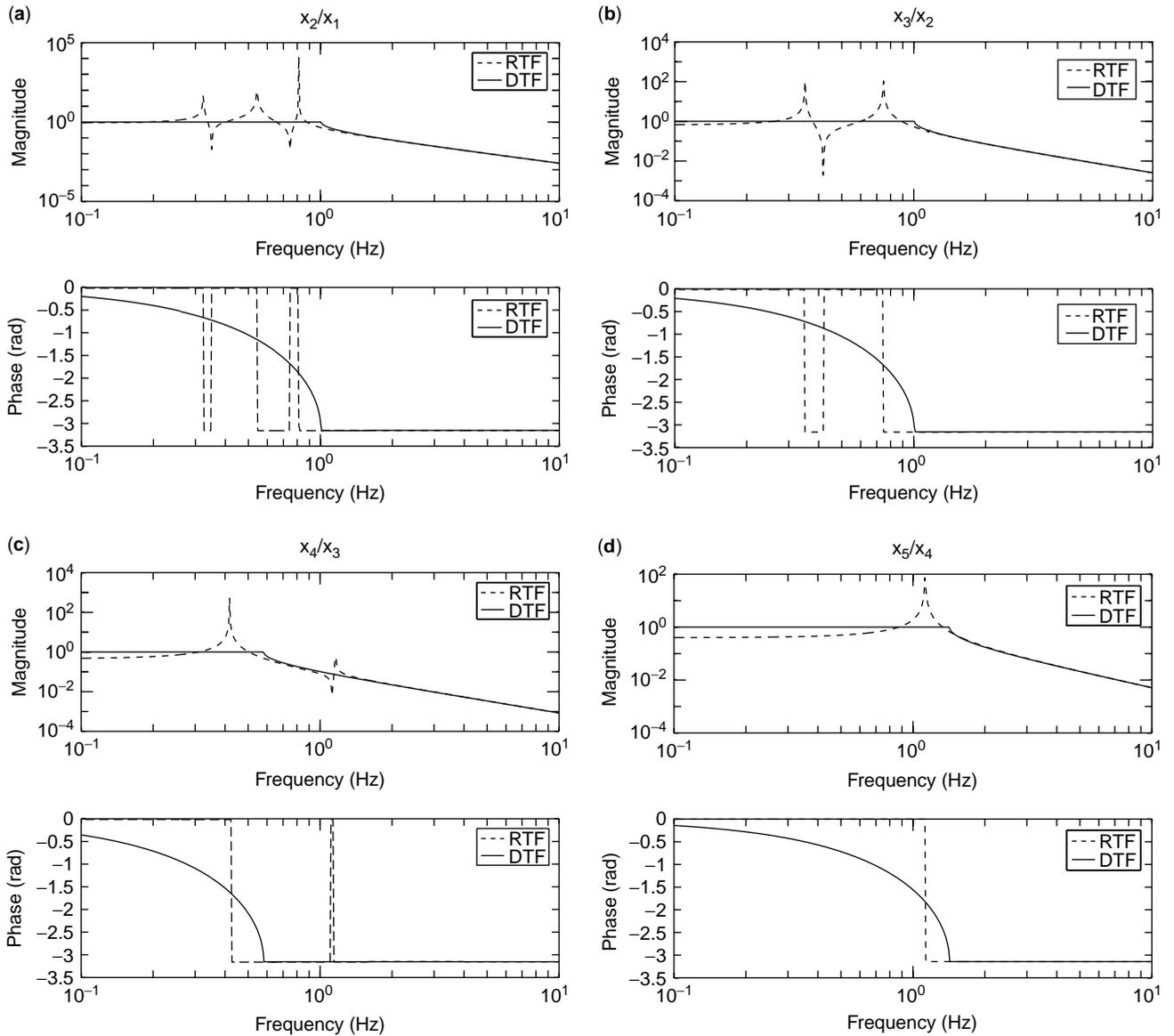


Figure 11. DTF and RTF responses: (a) \hat{x}_2/\hat{x}_1 ; (b) \hat{x}_3/\hat{x}_2 ; (c) \hat{x}_4/\hat{x}_3 ; (d) \hat{x}_5/\hat{x}_4 .

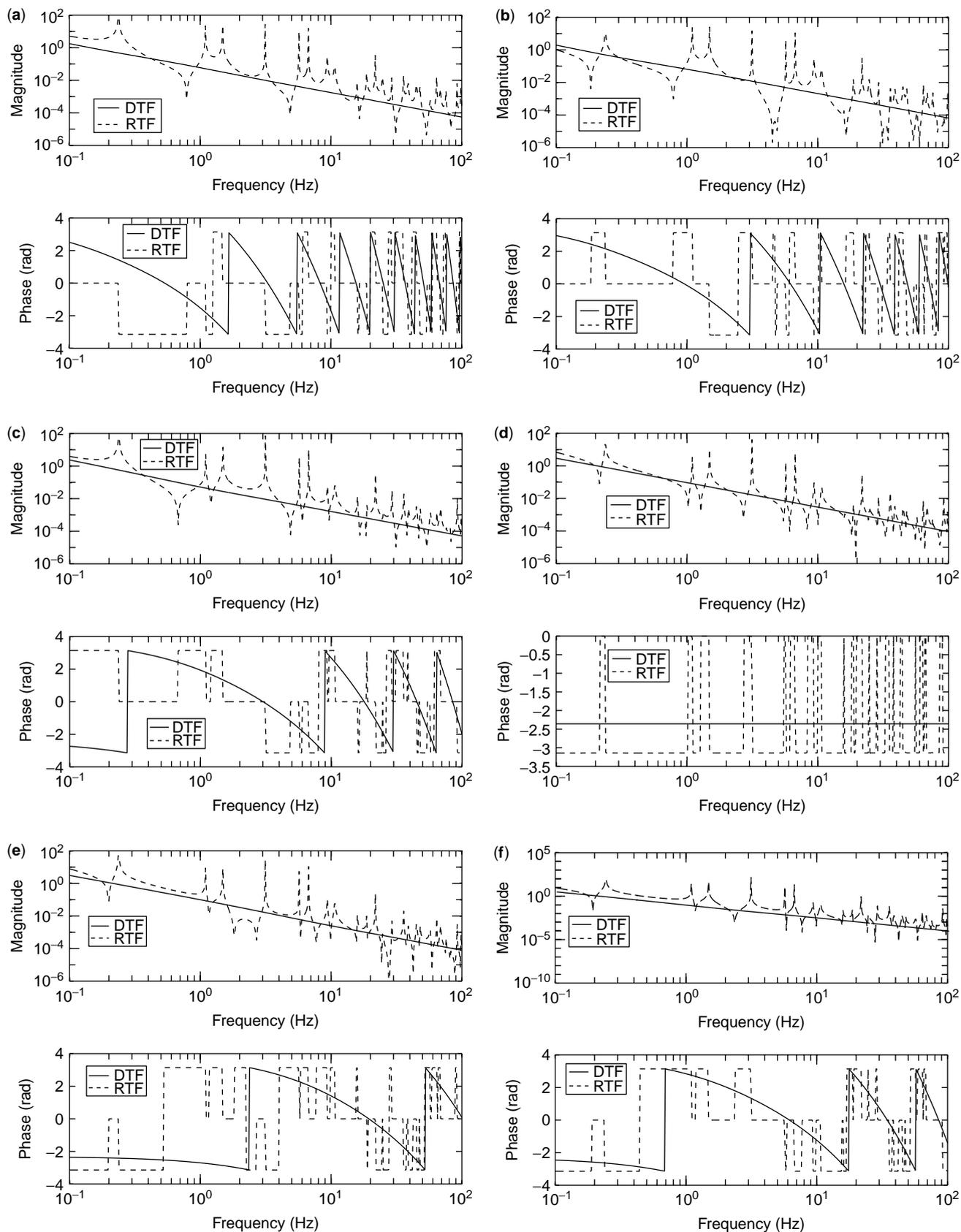


Figure 16. DTF and RTF response of free-free beam structure: (a) v_0/F_3 ; (b) v_1/F_3 ; (c) v_2/F_3 ; (d) v_3/F_3 ; (e) v_4/F_3 ; (f) v_5/F_3 .

Table 2. Properties of the Beam Structure

Element Number	1	2	3	4	5
EI	$4 \times (1 + 0.0001i)$	$1 \times (1 + 0.0001i)$	$10 \times (1 + 0.0001i)$	$2 \times (1 + 0.0001i)$	$3 \times (1 + 0.0001i)$
ρA	2	1	5	3	1
Length	2	3	3	1	1

Consider the fixed-free beam shown in Fig. 17. The left end of the beam is fixed, and the right end is free. Energy is input into this beam by $F_3 \neq 0$. The conventional reverberated transfer function matrix of this five-element beam can be obtained experimentally. The inverse of the RTF is the dynamic stiffness matrix of the beam.

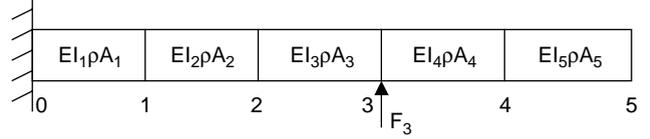
$$\text{RTF}^{-1} = \begin{bmatrix} K_4^1 + K_1^2 & K_2^2 & & & 0 \\ K_3^2 & K_4^2 + K_1^3 & K_2^3 & & \\ & K_3^3 & K_4^3 + K_1^4 & K_2^4 & \\ & 0 & K_3^4 & K_4^4 + K_1^5 & K_2^5 \\ & & & K_3^5 & K_4^5 \end{bmatrix}. \quad (40)$$

As previously shown for the free-free beam, the vertical displacement and angular displacement of each node are fed back into local controllers to generate control forces that can eliminate wave reflection at each node. Controllers can be attached at all nodes except for the node at the fixed end of the fixed-free beam. The DTF responses with respect to input F_3 can be obtained from the following expression:

$$\begin{bmatrix} \frac{v_1}{F_3} \\ \frac{\theta_1}{F_3} \\ \frac{v_2}{F_3} \\ \frac{\theta_2}{F_3} \\ \frac{v_3}{F_3} \\ \frac{\theta_3}{F_3} \\ \frac{v_4}{F_3} \\ \frac{\theta_4}{F_3} \\ \frac{v_5}{F_3} \\ \frac{\theta_5}{F_3} \end{bmatrix}_{\text{DTF}} = \left(\text{RTF}^{-1} + \begin{bmatrix} -K_4^1 + G_{2l} & & & & \\ & G_{3l} - G_{2l} & & & \\ & & 0_{2 \times 2} & & \\ & & & G_{4r} - G_{5r} & \\ & & & & G_{5r} \end{bmatrix} \right)^{-1} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad (41)$$

where K_4^1 can be obtained because the physical properties of the first element are known. The RTF and DTF responses are shown in Fig. 18

Comparing the expressions in Eqs. (39) and (41), the difference between the free-free beam and the fixed-free beam


Figure 17. Fixed free beam that has five elements.

dereverberated responses depends on the way the boundary conditions are treated. A controller is needed for the free-free beam to extend the free end to infinity. However, for the fixed-free beam, element 1 is ignored, and element 2 needs to be extended to the left to infinity. To achieve this, K_4^1 is the affecting part from element 1, and G_{2l} is from element 2.

DAMAGE DETECTION APPROACH BASED ON DTF RESPONSE

Detecting the Presence and Location of Damage

In the previous section, collocated noncausal controllers were developed to obtain the DTF responses for discrete

spring-mass structural elements and spectral rod and beam finite elements. In this section, a methodology for detecting damage in one-dimensional structures is developed based on the characteristic changes in the undamaged DTF responses. For an undamaged structure, collocated

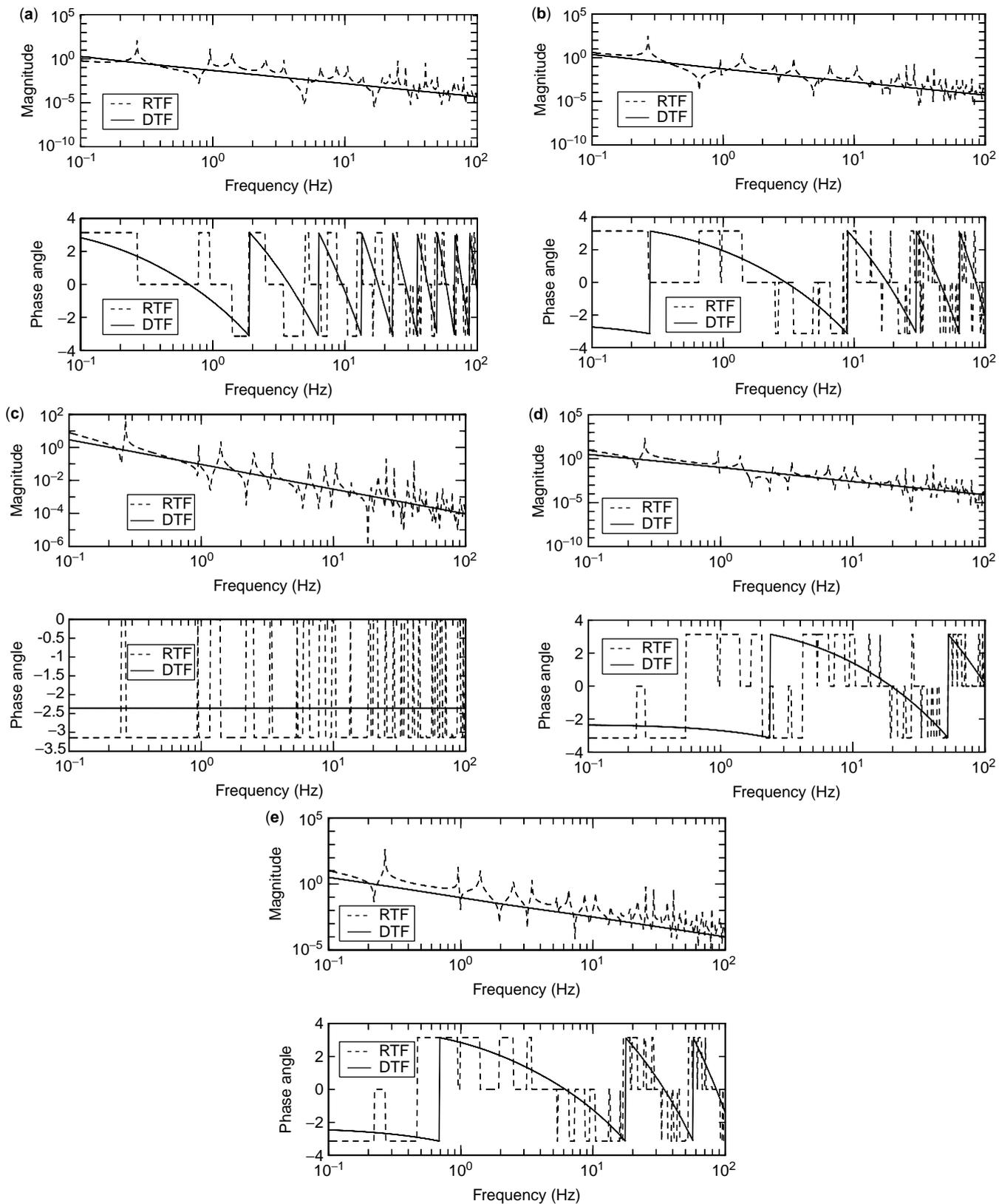


Figure 18. DTF and RTF response of fixed-free beam structure: (a) v_1/F_3 ; (b) v_2/F_3 ; (c) v_3/F_3 ; (d) v_4/F_3 ; (e) v_5/F_3 .

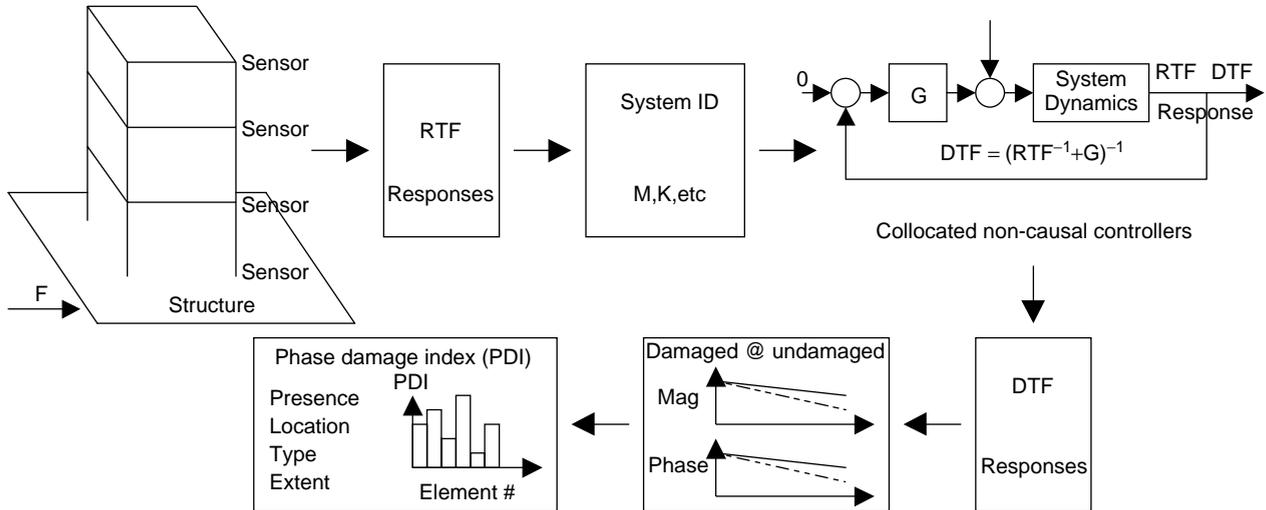


Figure 19. Damage detection methodology based on the DTF response.

noncausal controllers can be developed based on an identified model of the structural system. Implementing these noncausal controllers off-line in a computer will yield the DTF response from each actuator to each sensor. To infer the presence of damage, these undamaged controllers can be repeatedly applied to the identified RTF responses of a structural system. Thus, as damage appears in the structure, its presence will be revealed in the magnitude and phase plots of the DTF responses. Therefore, by tracking relative magnitude and phase errors between the undamaged and damaged structure's DTF response, it is possible to infer both the presence and location of damage. Furthermore, to determine damage in a particular structural element, one needs only to track the transmission of incident energy through the structure. This can be done by computing the ratio of the DTF responses for sequential degrees of freedom. This damage detection methodology is illustrated in Fig. 19.

Using examples of a free-free and fixed-free beam, the concept of a phase damage index is introduced following to illustrate how the presence, location, type, and amount of damage can be determined. This approach is later demonstrated on a model of a civil building structure.

Free-Free Beam. Consider the free-free beam structure previously shown in Fig. 14. Energy is input at node 3 where external force $F_3 = 0$. Figure 20 shows the DTF responses for the following transmission ratios of sequential degrees of freedom ($v_1/v_0, v_2/v_1, v_3/v_2 \dots$). The solid line refers to the undamaged case, and the dashed line refers to the damaged case. Each part of the figure displays the respective magnitude and phase response of the transmission ratio. It is assumed that damage in the form of a loss of stiffness is simulated in element number 3. Careful inspection of Fig. 20 allows two observations: First, in the case of stiffness damage, the phase of the DTF response will wrap around $-\pi$ earlier than in the undamaged case. Secondly, there is not a significant change in the DTF magnitude response due to damage in the structure.

To explain these observations in more detail, one has to consider the effect of damage on the structure's wave propagation response. When a leftward propagating wave passes through a damaged element such as element number 3, the effect of a stiffness loss is to increase the wave number and slow down the propagation of the transmitted wave. This introduces an additional phase lag in the DTF response. Undamaged structural elements have no effect on propagating the transmitted wave. In terms of the magnitude response, the noncausal controllers are designed to prevent wave reflection at the interfaces between structural elements. However, when damage occurs, wave reflection is not prevented because the controllers can no longer provide a match terminating boundary condition at each end. Nevertheless, because the simulated damage is small, the magnitude of the DTF displays modest sensitivity to damage. Therefore, a phase damage index (PDI) has been developed to locate the element that contains the damage. The PDI is based on the relative phase error from successive DOFs and is given by

$$\text{PDI}_i = \sum_{\omega_l \leq \omega \leq \omega_u} \left| \text{phase} \left[\left(\frac{v^i}{v^{i-1}}(\omega) \right)_{\text{DTF}}^{\text{undamaged}} \right] - \text{phase} \left[\left(\frac{v^i}{v^{i-1}}(\omega) \right)_{\text{DTF}}^{\text{damaged}} \right] \right|, \quad (42)$$

where $i = 1, 2, \dots, 5$ and refers to the i th element. ω_l, ω_u are the lower and upper bounds of the frequency range of interest. Computing the value of the PDI for each structural element indicates which element is damaged. This is evident from the PDI computed for the free-free beam example described earlier in this section. Note that the PDI is significantly larger for element number 3 in which simulated stiffness damage was assumed (see Fig. 20). Figure 21 displays the PDI for assumed stiffness damage in each of the five structural elements. Note that in all cases, the PDI that has the largest absolute value indicates, the structural element in the free-free beam example that is damaged.

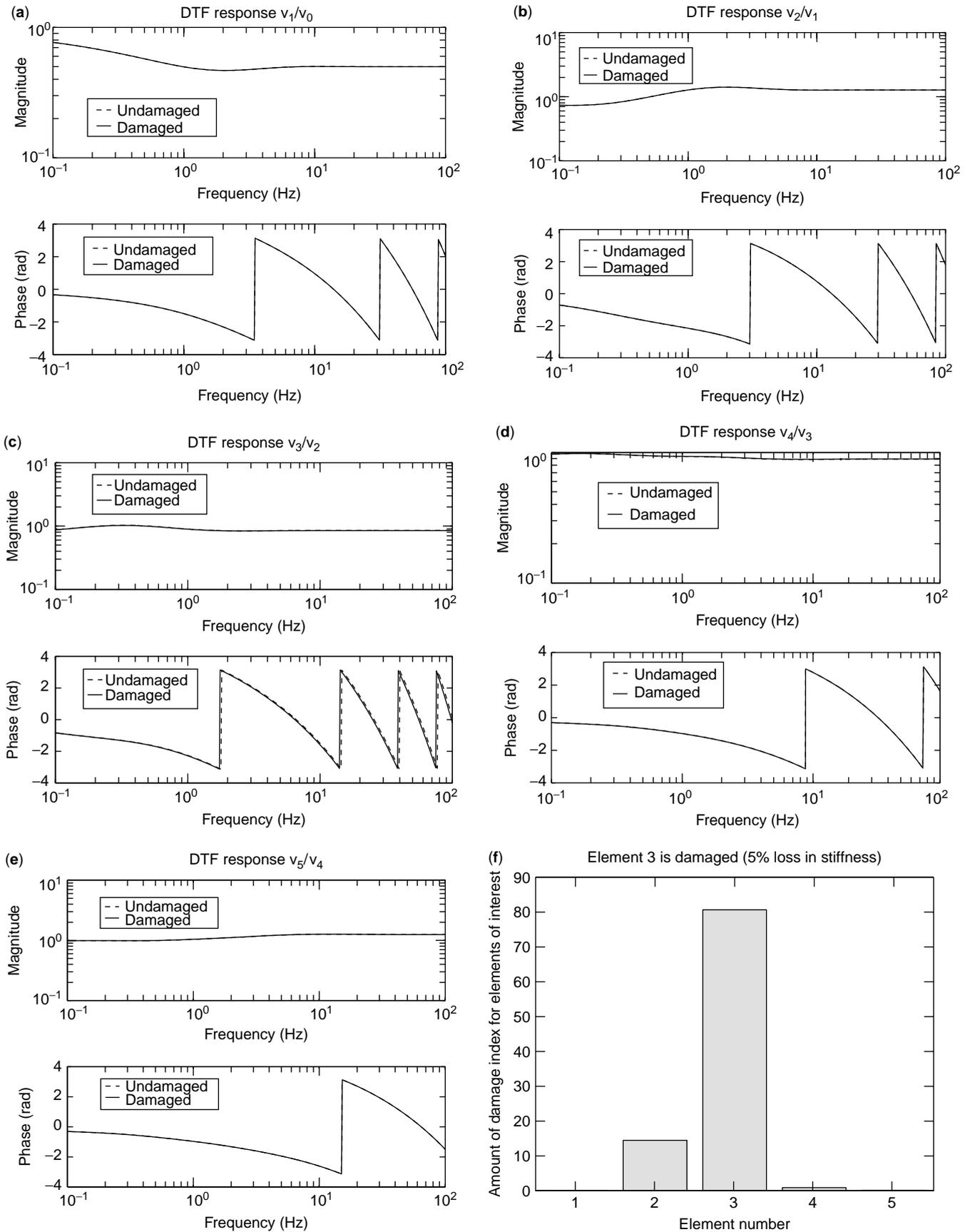


Figure 20. Element 3 is damaged in terms of 5% loss of stiffness. (a) DTF response v_1/v_0 ; (b) DTF response v_2/v_1 ; (c) DTF response v_3/v_2 ; (d) DTF response v_4/v_3 ; (e) DTF response v_5/v_4 ; (f) damage index vs. element number.

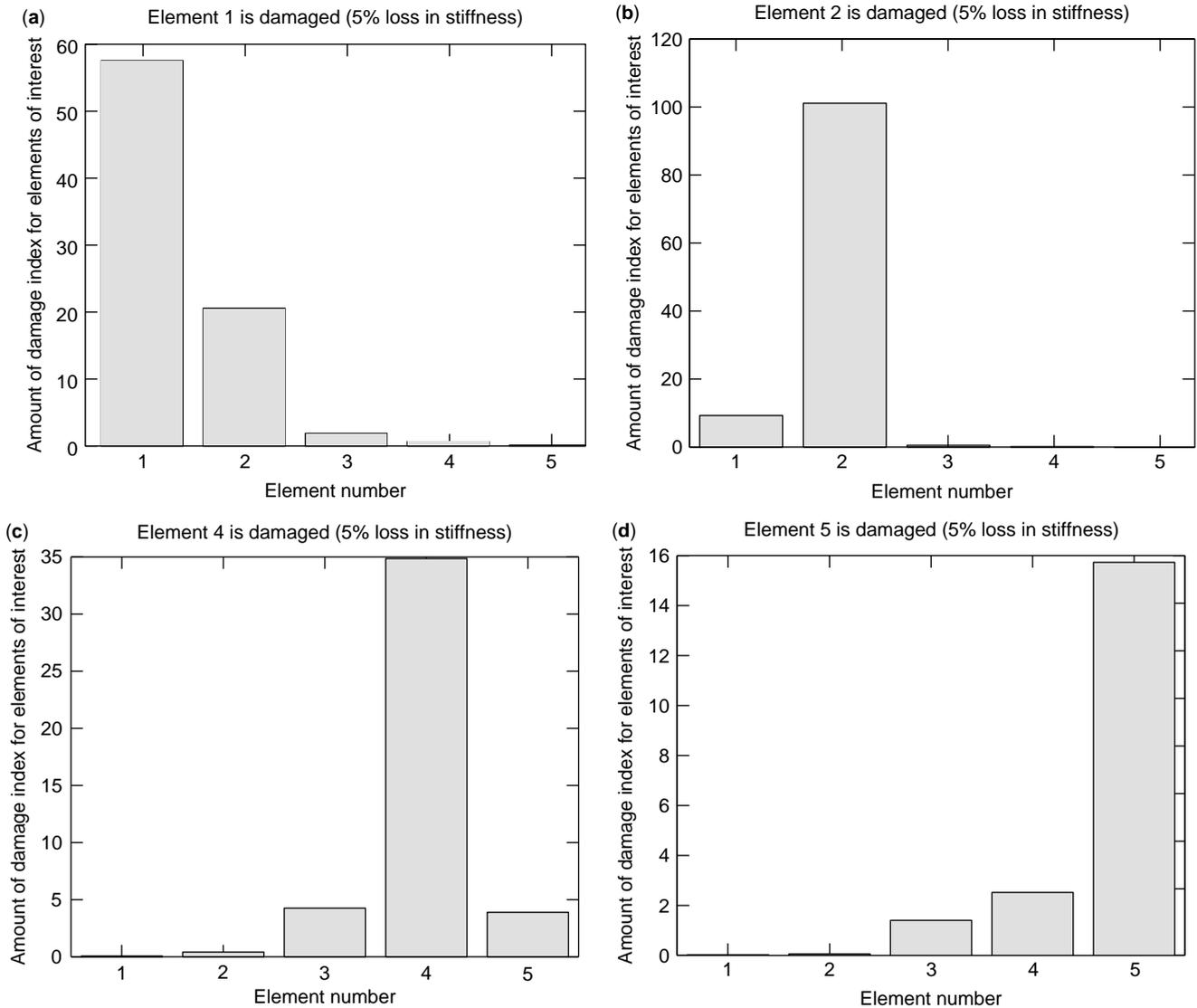


Figure 21. Damage index vs. element number. (a) Element 1 is damaged (5% stiffness loss). (b) Element 2 is damaged (5% stiffness loss). (c) Element 4 is damaged (5% stiffness loss). (d) Element 5 is damaged (5% stiffness loss).

Fixed-Free Beam. For a fixed-free beam, $\begin{pmatrix} v_0 \\ \theta_0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$. Only $i = 2, 3, 4, 5$ can be substituted in Eq. (42), and four damage indices, $PDI_2, PDI_3, PDI_4, PDI_5$, can be formulated. They are associated with elements 2, 3, 4, and 5, respectively. Five damage cases are simulated. Figure 22a shows damage indexes versus element number when element 1 is damaged. Remember that no damage index is formulated for element 1. However, from Fig. 22a, the index for element 2, PDI_2 , is large. This implies that damage occurred in element 1, demonstrated by the damage index of element 2. Of course, from Fig. 22b, damage in element 2 leads to a blowup of PDI_2 , too. Therefore, if the damage index of element 2, PDI_2 , exhibits some large value, both element 1 and element 2 need to be checked to see whether they are damaged. The damage indexes of elements 3, 4, and 5 can locate the damage of the associated elements as shown in Fig. 22c–e.

Damage Type and Extent

Most structural damage can be categorized as mass or stiffness related. In stiffness loss, the phase curve of the DTF response of a damaged structure tends to wrap at π or $-\pi$ earlier than in the undamaged case. In mass loss, the phase of the DTF response of a damaged structure tends to wrap later than in the undamaged case. This pattern in the phase behavior of stiffness or mass damage is primarily a result of the way wave number variations affect the transmission properties of an incident wave through a structural element. Thus, by simply examining the phase behavior of undamaged and damaged DTF responses, it is possible to ascertain the type of damage, albeit stiffness or mass. In addition, damping effects can also be identified.

Although the type of damage in a structural system can be determined from the phase response of the DTF,

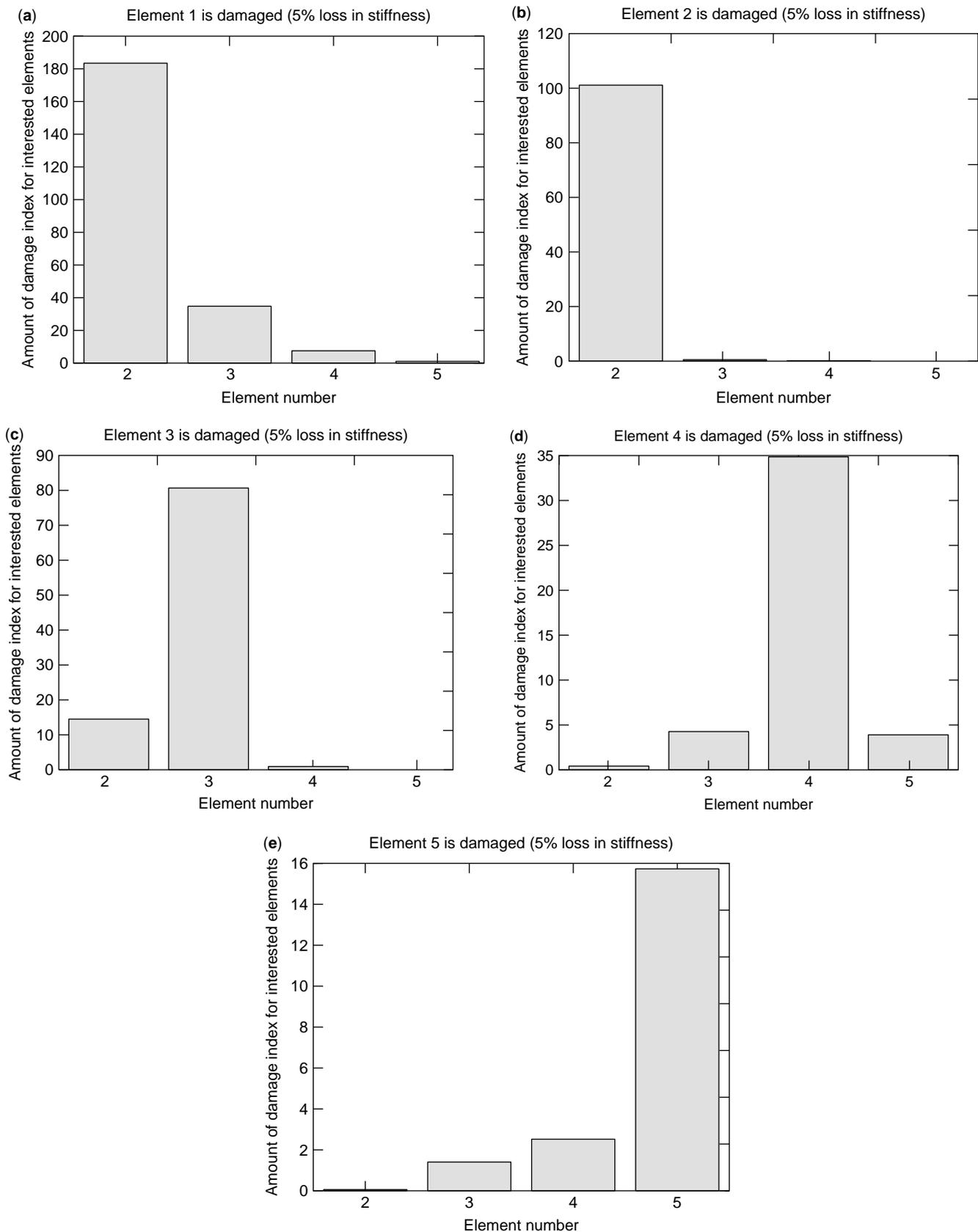


Figure 22. Damage indices vs. element number. (a) Element 1 is damaged (5% stiffness loss). (b) Element 2 is damaged (5% stiffness loss). (c) Element 3 is damaged (5% stiffness loss). (d) Element 4 is damaged (5% stiffness loss). (e) Element 5 is damaged (5% stiffness loss).

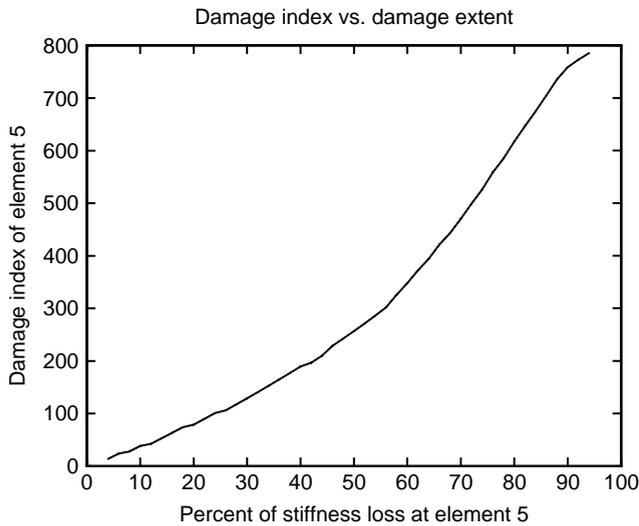


Figure 23. Fixed-free beam; the fifth element is damaged.

determining the extent of damage in a structural element is a bit more difficult. The phase damage index (PDI) described before represents a qualitative measure of damage in a structure. However, by conducting a numerical study of the structural model identified, one can simulate structural damage in each structural element and correlate the PDI with percentage damage. Thus, a lookup table or curve-fitted database can be developed for each structural element to provide an estimate of the amount of damage in that element. Figure 23 displays such a correlation for the

fixed-free beam example developed earlier. Note that the PDI increases as structural damage increases in the structural element of interest. Similar trends exist for other elements.

DAMAGE DETECTION IN A BUILDING STRUCTURE USING DTF

Consider the discrete three-DOF model of the three-story framed building structure displayed in Fig. 24. The properties of the structure are given in Table 3. The equation of motion for the building structure is

$$M\ddot{X} + KX = q \quad (43)$$

or

$$\begin{bmatrix} m_1 & 0 & 0 \\ 0 & m_2 & 0 \\ 0 & 0 & m_3 \end{bmatrix} \begin{Bmatrix} \ddot{x}_1 \\ \ddot{x}_2 \\ \ddot{x}_3 \end{Bmatrix} + \begin{bmatrix} k_1 + k_2 & -k_2 & 0 \\ -k_2 & k_2 + k_3 & -k_3 \\ 0 & -k_3 & k_3 \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \end{Bmatrix} = \begin{Bmatrix} k_1 x_g \\ 0 \\ 0 \end{Bmatrix}. \quad (44)$$

Transformed into the frequency domain and assuming that $x = Xe^{i\omega t}$, Eq. (44) becomes

$$[-M\omega^2 + K]X = Q \quad (45)$$

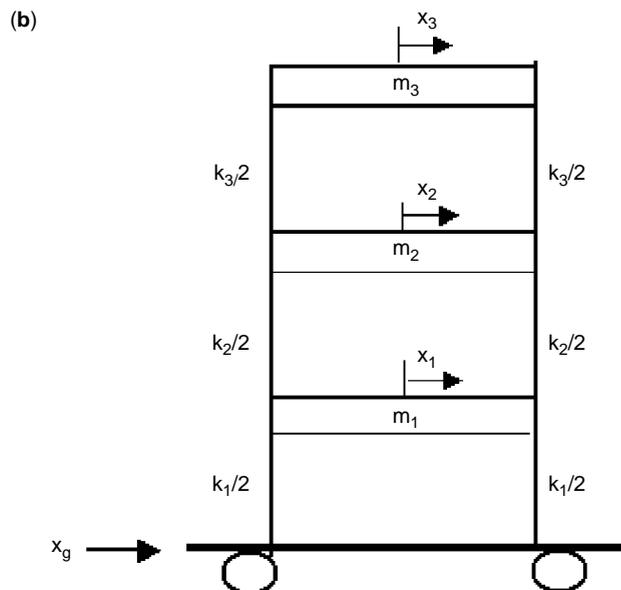
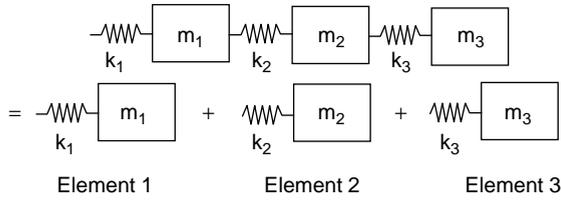


Figure 24. (a) Framed building structure. (b) Discrete three-DOF model.

Table 3. Properties of the structure

Mass	$m_1 = 16.7676$	$m_2 = 16.7676$	$m_3 = 16.7676$	unit: kg
Stiffness	$k_1 = 9158.7$	$k_2 = 13084$	$k_3 = 36635$	unit: N/m

**Figure 25.** Summing asymmetrical spring-mass elements to construct a building structure.

or

$$\begin{bmatrix} k_1 + k_2 - m_1\omega^2 & -k_2 & 0 \\ -k_2 & k_2 + k_3 - m_2\omega^2 & -k_3 \\ 0 & -k_3 & k_3 - m_3\omega^2 \end{bmatrix} \begin{Bmatrix} X_1 \\ X_2 \\ X_3 \end{Bmatrix} = \begin{Bmatrix} -k_1\omega^2 X_g \\ 0 \\ 0 \end{Bmatrix}. \quad (46)$$

From Eq. (46), note that energy is input into the building structure from ground motion by $F_1 = -k_1\omega^2 X_g(\omega)$ or $k_1 \ddot{X}_g(\omega)$.

A model of the three-DOF three-story building structure in Fig. 24 can be constructed by simply adding three asymmetrical spring-mass elements (see Fig. 25). The DTF of the whole building structure can be obtained by first adding a virtual controller to the right end of element 1, then adding elements 2 and 3 that have similar virtual controllers attached to both ends. This process is shown in Fig. 26. The control architecture is shown in Fig. 27. The controller gains G_{1r} , G_{2l} , G_{2r} , G_{3l} , G_{3r} are obtained by using Eq. (8) and substituting different k , m , and μ values for the different structural elements. The RTF and DTF are shown in Fig. 28.

Considering the structure shown in Fig. 26, energy is input from the left end of the structure. After passing element 1, controller G_{1r} will prevent the energy from being reflected back into element 1. A portion of the energy will continue transmitting into element 2. At the end of element 2, controller G_{2r} acts like controller G_{1r} to ensure that no energy is reflected back. The same effect will be achieved by controller G_{3r} . The result is that wave reflection does

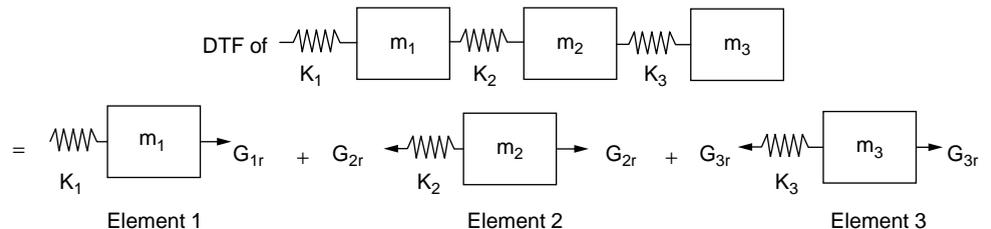
not occur across the structure. If damage is present in a structural element, the idealized local controllers can not prevent reflection.

Damage in civil structures is usually categorized as stiffness loss. The following damage case studies simulate one-quarter damage loss for each stiffness of the three-DOF three-story building. This study attempts to show the characteristic change in the DTFs before and after damage. Further, a methodology is developed to locate damage. Six sets of figures are displayed for each case study, and each figure has two parts: the magnitude and phase of a particular DTF response. Emphasis is placed on examining the incident path of energy traveling through the structure. As the energy is transmitted across a damaged structural element, either the magnitude or the speed of wave propagation will change compared to the undamaged element. This change in local structural properties is revealed in the magnitude and phase plot of the DTF response.

Case Study: One-Quarter Stiffness Loss for k_1

Considering k_1 damaged in terms of stiffness loss, the energy is still input from the left end of the structure. Remember that G_{1r} is determined by m_1 and k_1 . The controller gains G_{1r} can prevent wave reflection perfectly at element 1. But for the damaged element, we do not know how much change there is in k_1 . If we still use G_{1r} as a virtual controller for the damaged element, wave reflections will occur at the end of element 1. The DTF response of \ddot{x}_1/\ddot{x}_g will change before and after damage, as is illustrated in Fig. 29 a. Strictly speaking, \ddot{x}_1/\ddot{x}_g is no longer the DTF after damage because a wave is reflected back into element 1. But that is not the case for elements 2 and 3. Controller gains G_{2r} and G_{3r} will do their jobs perfectly. This leads to identical DTF responses of \ddot{x}_2/\ddot{x}_1 , \ddot{x}_3/\ddot{x}_1 , and \ddot{x}_3/\ddot{x}_2 , illustrated in Fig. 29d–f before and after damage.

Now, consider Fig. 29a. In the low-frequency range, the DTF phase of an undamaged element is smaller than that of a damaged element. This means that the wave number of an undamaged element is smaller than that of a damaged element. The reason for the increased wave number is either stiffness loss and/or mass increase. Mass increase is unlikely in the event of damage in a civil structure. Thus, the cause of damage is assumed to be stiffness.

**Figure 26.** Applying controllers to obtain the DTF of a building structure.

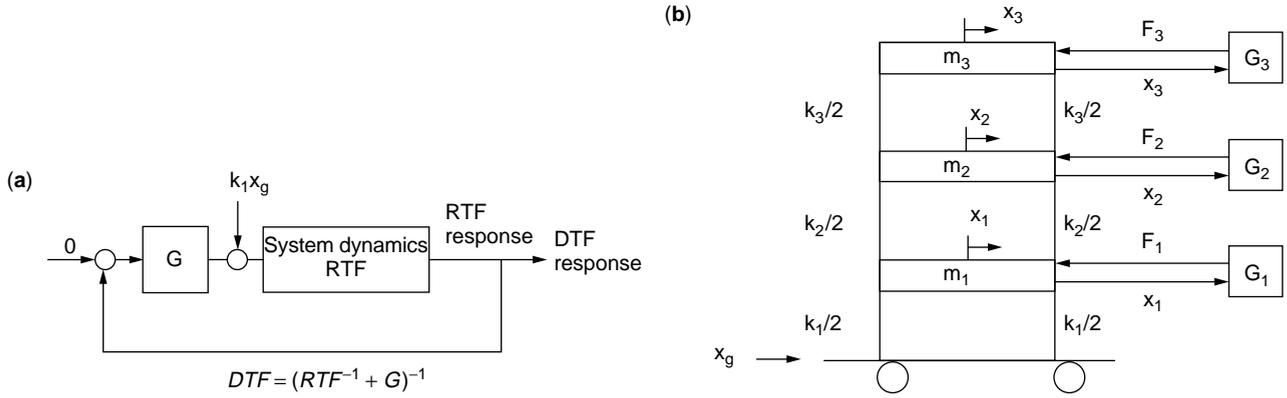


Figure 27. Control architecture.

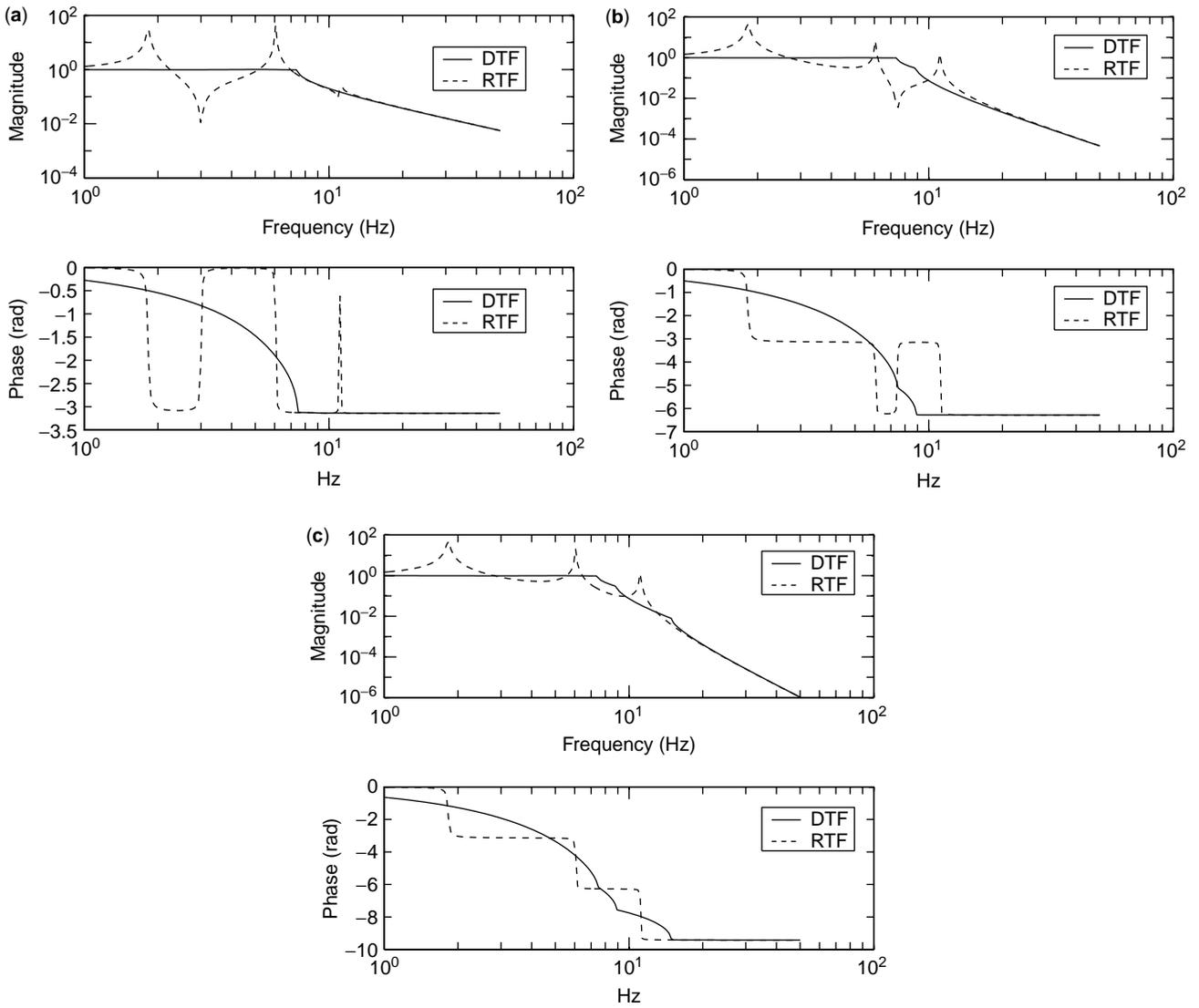


Figure 28. RTF and DTF response of a building structure: (a) \dot{x}_1/\dot{x}_g ; (b) \dot{x}_2/\dot{x}_g ; (c) \dot{x}_3/\dot{x}_g .

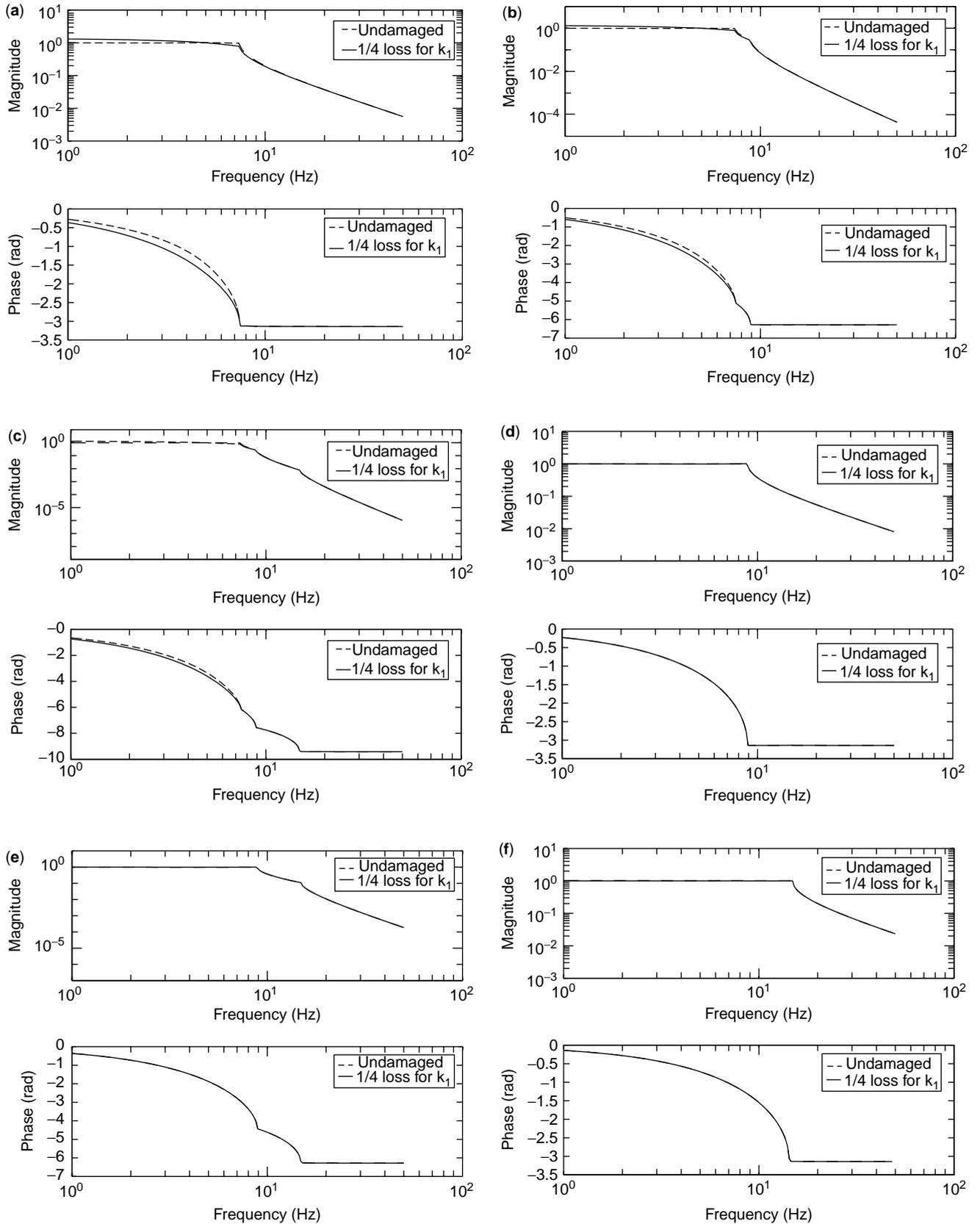


Figure 29. DTF response: (a) \ddot{x}_1/\ddot{x}_g ; (b) \ddot{x}_2/\ddot{x}_g ; (c) \ddot{x}_3/\ddot{x}_g ; (d) \dot{x}_2/\dot{x}_1 ; (e) \ddot{x}_3/\dot{x}_1 ; (f) \ddot{x}_3/\ddot{x}_2 .

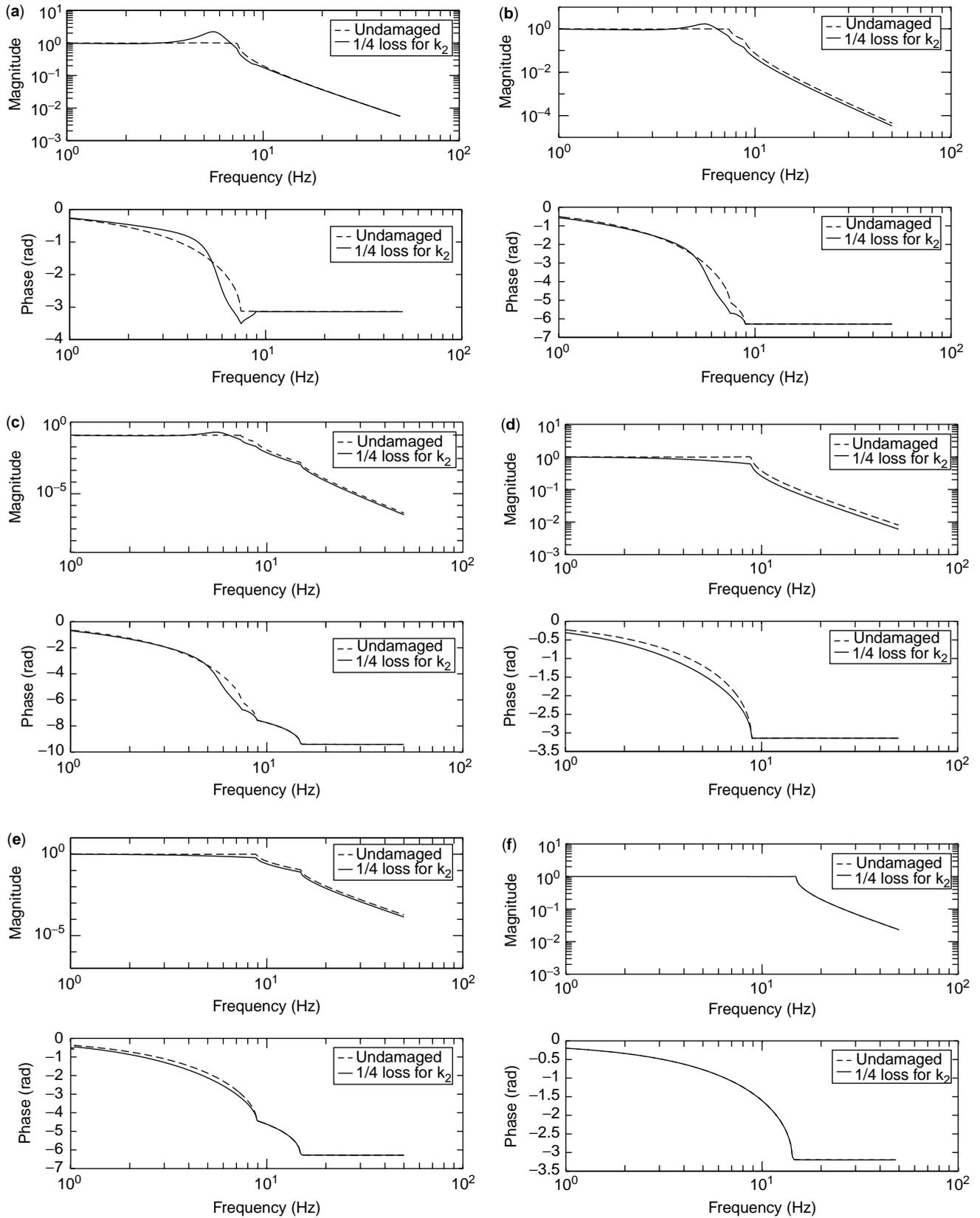


Figure 30. DTF response: (a) \dot{x}_1/\dot{x}_g ; (b) \ddot{x}_2/\ddot{x}_g ; (c) \ddot{x}_3/\ddot{x}_g ; (d) \ddot{x}_2/\dot{x}_1 ; (e) \ddot{x}_3/\dot{x}_1 ; (f) \ddot{x}_3/\ddot{x}_2 .

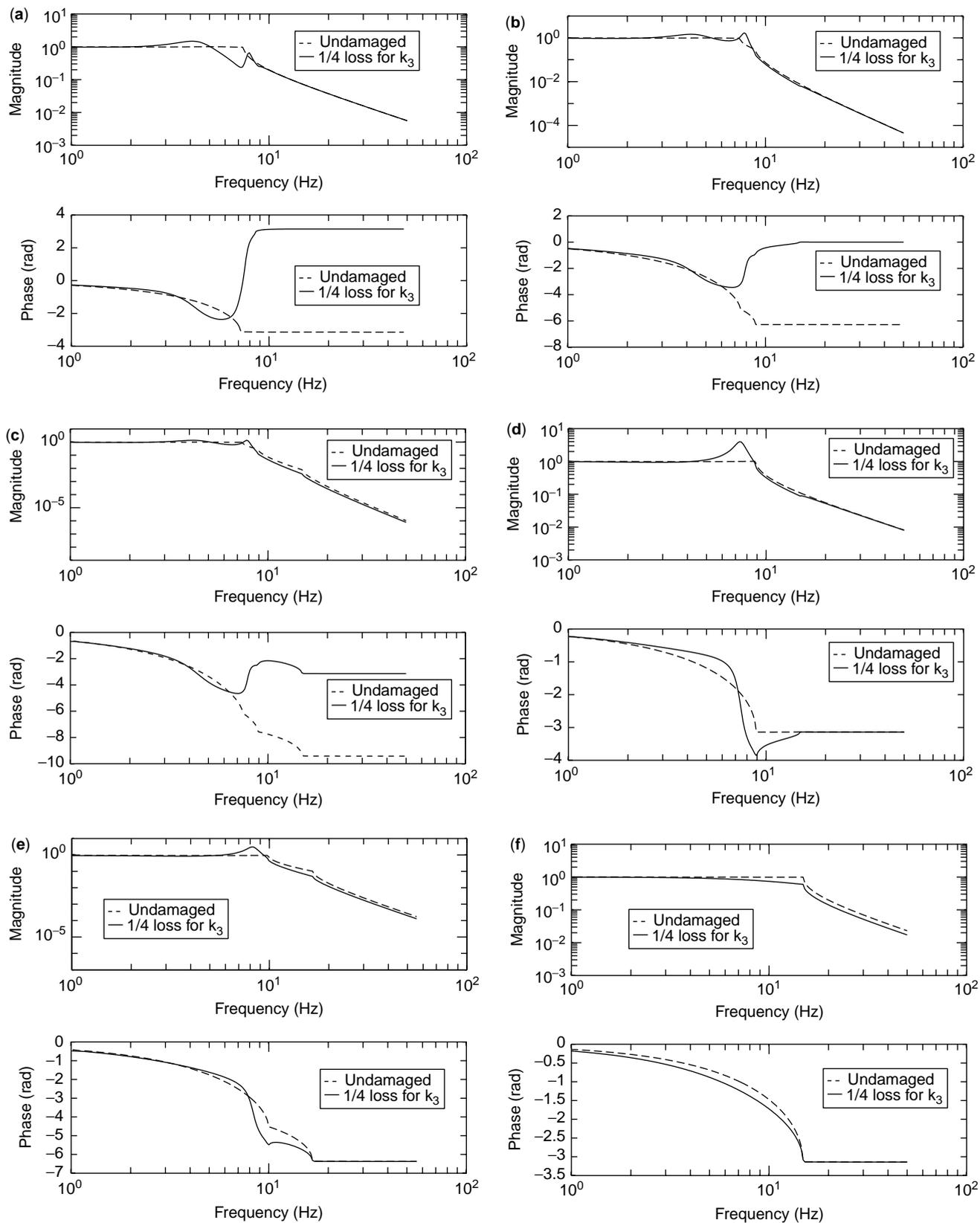


Figure 31. DTF response: (a) \dot{x}_1/\dot{x}_g ; (b) \dot{x}_2/\dot{x}_g ; (c) \dot{x}_3/\dot{x}_g ; (d) \dot{x}_2/\dot{x}_1 ; (e) \dot{x}_3/\dot{x}_1 ; (f) \dot{x}_3/\dot{x}_2 .

Now, it is evident that the damage occurs at element 1 in terms of stiffness loss. This led to the malfunction of controller G_{1r} . Using Eqs. (4) and (8), reducing the value of k_1 to k_{1r} , a new controller G_{1r} will be found to obtain a perfect DTF response for each element. Thus, the damage extent, $\Delta k_1 = k_1 - k_{1r}$, can be obtained.

Case Study: One-Quarter Stiffness Loss for k_2

From Fig. 30f, \ddot{x}_3/\ddot{x}_2 is identical before and after damage. This shows immediately that the damage happened at element 2. Now, only the DTF response of \ddot{x}_2/\ddot{x}_1 needs to be examined. This is shown in Fig. 30d. Applying similar reasoning, as earlier, a stiffness loss for k_2 is concluded.

Case Study: One-Quarter Stiffness Loss for k_3

From Fig. 31, no identical DTF before and after damage shows up. This shows that G_{3r} did not perform correctly. The wave reflection at the right end of the structure degrades all DTFs. We can conclude that damage must have happened at element 3, and the same reasoning leads to a stiffness loss for k_3 .

SUMMARY AND CONCLUSIONS

This article has introduced a wave propagation approach for computing the DTF responses of nonuniform structures. By using DTF responses, boundary effects are ignored in favor of the incident path that the energy takes to travel through a structure. It has been shown that the DTF responses, associated with an individual element, are sensitive to physical parameter changes which are directly in the load path from the input force to the measured sensor response. The DTF provides direct information regarding the source, location, and amount of damage.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under grant CMS9625004, and Dr.'s S.C. Liu and William Anderson served as contract monitors.

BIBLIOGRAPHY

1. S.W. Doebbling, C.R. Farrar, and M.B. Prime, and D.W. Shevitz, Los Alamos National Laboratory Report No. LA-13070-MS, Los Alamos, NM, 1995.
2. C.R. Farrar and D.A. Jauregui, *Smart Mater. Struct.* **7**, (5): 704–719 (1998).
3. C.R. Farrar and D.A. Jauregui, *Smart Mater. Struct.* **7**, (5): 720–731 (1998).
4. F.K. Chang, *Proc. 2nd Int. Workshop Struct. Health Monitoring*, Stanford University, Stanford, CA, Sept. 8–10, 1999.
5. J.F. Doyle, *Exp. Mech.* **35** : 272–280 (1995).
6. K.A. Lakshmanan and D.J. Pines, *J. Intelligent Mater. Syst. Struct.* **9**: 146–155 (1998).
7. A.H. von Flowtow and B. Schafer, *AIAA J. Guidance*, **19**: 673–680 (1986).

8. D.J. Pines, and A.H. von Flowtow, *J. Sound Vibration* (1990).
9. D.G. MacMartin and S.R. Hall, *J. Guidance*, **14**: 521–530 (1991).
10. D.W. Miller and S.R. Hall, *J. Guidance*, **14**: 91–98 (1990).
11. K. Matsuda and H.A. Fujii, *J. Guidance Control Dynamics* **19**: 91–98 (1996).
12. R.S. Betros, O.S. Alvarez-Salazar, and A.J. Bronowicki, *Proc. 1993 Smart Struct. Intelligent Syst. Conf.* 1993, Vol. 1917 pp. 856–869.
13. A. Purekar and D.J. Pines, *Smart Mater. Struct.*, in press.
14. L. Brillouin, *Wave Propagation in Periodic Structures*. 2e, Dover, NY, 1946.
15. J.F. Doyle, *Wave Propagation in Structures*. 2e, Springer, NY, 1997.

HIGHWAYS

KAMBIZ DIANATKHAH
Lennox Industries
Carrollton, TX

INTRODUCTION

The number of automobiles has increased dramatically as a result of population and job growth during the past several decades. During the same period, the commuting distance has also increased (1). This in turn has resulted in congestion in many suburban areas. This increase in traffic flow translates into higher cost for accident expenses, a rise in fuel consumption, and air pollution. The Department of Transportation has estimated that the volume of traffic will increase by 50% in next 25 years (2). The loss of time and productivity and health issues caused by increased carbon monoxide and dioxide are the predominant factors that call for building smart highway systems.

SMART MATERIALS

Smart material technology is progressively becoming one of the most important new research areas for engineers, scientist, and designers. Increased use of smart materials will undoubtedly influence our daily lives fundamentally in the near future. Presently, the emergence of smart materials and smart structures has resulted in new applications that change the way we think about materials, sensors, actuators, and data processing. Smart materials are defined as materials whose properties alter predictably in response to external stimuli. Smart materials can be divided into several categories:

1. Shape-memory alloys: Polymers or alloys that remember their original shape under an applied load and temperature through phase transformation. Typical alloys are Ti–Ni (Nitinol) and TiNi–Cu (K-alloy).
2. Piezoelectric materials where strain results from an applied load or voltage (electric field), for example, polyvinylidene fluoride (PVDF) polymer.

3. Electrorheological fluids that change their viscosities according to the intensity of an electric field.
4. Photochromic glasses whose transparency changes with the intensity of light (photographic lens).

In this article, the advantages of smart materials are described for application to smart highways, structures and intelligent transportation systems (ITS). Through the application of new technology, there is potential to integrate multiple modes of travel and to focus on demand as well as transportation supply throughout the world. The field of smart highways and structures consist of many areas of innovation in developing superhighways, bridges, modern cars that have built-in computer-aided navigation equipment, and a central control unit within each highway system to assist in traffic management.

OBJECTIVES OF SMART HIGHWAYS

The objectives of building smart highways are safety, low maintenance, and conveyance. By building smart highways, more vehicles can be on the road thereby reducing congestion and eliminating the need for building additional lanes. For safety, computers installed in an automobile will perform all driving tasks, and this will enhance safety because most traffic fatalities are due to human error. For general safety purposes, the following are required for a modern highway: (1) piezo MTLs that connect to electric heaters and therefore, do not ice up, (2) "glow in the dark" surface material, and (3) standards for traffic flow. Maintenance is a major issue in establishing engineering design parameters for estimating the life of the road versus replacement cost. The convenience features of smart highways should include optical sensors on-site, cars that have Global Positioning System (GPS)/road maps and bar code road signs, and autopilot features. Environmental issues also play a critical role in designing a smart highway system. Finally, development of road surfaces that can break down pollutants such as nitrogen oxide gas, will be a major focus of future research for metropolitan areas such as Los Angeles and Denver that have high pollution.

Smart Structures

Smart structures are nonbiological physical structures that have the following attributes: (1) a definite purpose and (2) means and an imperative to achieve that purpose.

The functional aspects of a smart highway are to integrate the normal design features and provide means of controlling traffic to optimize the traffic flow and human safety. Smart highway structures are designed for normal and abnormal events. Normal design conditions are dead-weight, thermal expansion and cyclic traffic loads (3). A smart structure is a subset of many intelligent structures that is complex and made of innovative materials, control laws, and communications. Smart structures have sensors and or actuators to help them function. Smart structures generally should be light, take advantage of new computer technology, integrated sensors, actuators, and contain some sense of intelligence to attain structural performance capabilities.

America's bridges have deteriorated during the past two decades due to lack of funding and neglect. Many collapses have occurred on shorter spans during this period, and many are also riddled with cracks and weak spots. Concrete bridges have been more susceptible to failures than steel because their flaws are often less apparent. Because bridges are a critical part of any highway system, particularly a modern one, continuous monitoring and innovative technology are required to construct and modernize bridges to modern standards. In the construction of modern bridges and structures, smart materials must be capable of warning of potential failures for operations or of enhancement by structural health monitoring of civil infrastructures and marine structures.

The most significant areas of concern for smart structures are performance, cost, sensing technology, engineering integration, structural stress and the need for wireless technology. Smart structures provide useful tangible benefits, and adaptability is achieved by knowing the limits, constraints, and compatibility with existing designs, methodologies, and the ability to learn. The drawbacks of building smart structures lie in the capital intensive nature of the projects, lack of understanding what data to collect and how to interpret the data, lack of an integrated team approach, reluctance to change, lack of consistency, and dealing with regulations. Major areas of concern in design, construction, and maintenance are monitoring and evaluating structures of bridges, soil, and concrete as a result of stress or natural disasters such as earthquakes that will also be a major factor in constructing an intelligent highway system. Monitoring a bridge by fiber-optic deformation sensors or Doppler vibrometers to detect disbonded composites is currently being studied to prevent catastrophic failures(1,4).

Monitoring the structural integrity of highways and bridges for safety and ease of repair presents an emerging field of study to find new ways to support the infrastructure of these systems. Smart materials are beginning to play an important role in civil engineering designs for dams, bridges, highways, and buildings. Sensors embedded throughout a concrete and composite structure can sense when any structural area is about to degrade and notify maintenance personnel to prepare for repair or replacement. Smart materials will be used to improve reliability, longevity, performance, and reduce the cost of operating smart highways and structures (5). The application of sensors and actuators, diagnostic monitoring, structural integrity/repair, damage detection, and active hybrid vibrational control will be the major areas of discussion in building intelligent highways and structures using smart materials. A smart structure using smart technology may include the use of fiber-reinforced (FR) concrete and optical-fiber sensors. Glass or carbon fibers in a cement matrix (6) are used instead of steel to increase the strength of concrete. Steel tends to corrode in salt, water, and corrosive deicing compounds, but reinforced concrete is not susceptible.

Shape-memory alloys that are important elements of intelligent (smart) materials can be used to build smart sensing structures, for example, a damping device made of shape-memory alloys can absorb seismic energy and

reduce its force. Due to its unique properties, this type of alloy can return a bridge to its original position after seismic activity (7). Optical-fiber sensors can also be embedded in composite beams along their length to monitor any stresses from traffic loads, cold, and wind. The change in wavelength of light reflections from optical sensors is compared to a set of baseline wavelength data (6). Old bridges can also be reconstructed and reequipped with small amounts of carbon fibers in the concrete mix and by using an electrode at each end of the bridge and measuring its resistance. Should developing cracks disintegrate the fibers that can conduct electricity thereby increasing the electrical resistance. The fractures can also be correlated with reduction in the strength of steel and concrete and also to determine whether serious damage has occurred (8).

Materials technology, specifically, the use of composites and shape-memory alloys in new structures and highways will have a great impact on human society, including the creation of new industries, extension of the women frontier to space, high-speed transportation, and earthquake-resistant and disaster-preventing construction.

A major area of focus in building a smart highway structure is the road pavement smoothness that creates better driving conditions and increases the life of the road (9). Equipment for road smoothness in all states will be required to set a minimum standard. The quality of cement and concrete also plays a significant role in developing a high quality road. Paving material quality plays an important role in durability and safety of roads. Use of recycled materials and polymer-modified binders have been considered for the durability of paving systems in some California highways (10), compared with traditional asphalt pavement. The major obstacles to using recycled or polymer material in pavement materials are extreme loads introduced by heavy trucks that may impact the integrity of the road and the driving performance of the truck.

However, these new materials cost far less than ordinary asphalt and may also assist in design, construction, and easier maintenance of the roads. Another concern that recycled materials and polymers must address is the material's performance in variable weather conditions and major fluctuations in temperature.

Highway fatalities have declined about 20% within the past decade from 47,000 to 41,000 annually as a result of safety improvement (9). Road condition plays a critical role in highway safety.

Liability issues and cost-effectiveness will be significant factors in the development of modern highways in upcoming years. A successful, low-cost system for modern automobiles that can reduce fatalities will be a key initial step to globalization of this system.

Sensors

Sensors must have properties that enable them to detect small changes in a structure (8), that is, changes in strain and capability for a measurable output signal. The response time of a sensor is a critical issue in monitoring crack growth within a structure; although it will not be as critical in observing stiffness changes from fatigue.

Piezoelectric ceramics are a common type of embedded sensor and are used for noise and vibrational sensing (II). Mechanical and viscoelastic properties and compatibility with the surrounding structure are the primary factors in selecting a sensor because it is usually embedded in a polymeric composite. Polymers at high frequency and low temperatures are stiff due to lack of molecular motion and because they are in a glassy state (8). At high temperatures, polymers are glassy due to high viscosity because atoms can move more easily. The behavior of a polymer is a function of T_g , the glass transition temperature at which a polymer is in a glassy state. The strain properties of a polymer are directly related to the state of the polymer.

As mentioned previously, carbon fibers are used as reinforcement in smart structures for strain sensing (9) and can replace the need for strain gauges and optical sensors. However, the important factor in considering carbon fibers is their electromechanical properties, namely, the electrical resistivity of the fibers under load in composite, polymer, and concrete structures. The electrical resistivity and the modulus of elasticity are affected by tensile and compressive forces (9).

Fiber-optic sensors can provide information on any strain fluctuations as a result of stress and early warning of a flaw in a joint or within the concrete. Fiber-optic sensors, as related to smart structures, can reduce the risk of failure in an aging infrastructure. Although fiber-optic sensors are not smart and lack actuating capability, they are the predominant technology that is discussed in relation to smart structures. Structures instrumented with fiber-optic sensors can respond to or warn of impending failure and indicate the health of a structure after damage. Applications include the instrumentation of bridges, highways, dams, storage tanks, oil tankers, and buildings. Systems that can measure strain or vibration have been tested in the United States and Germany. The cost and complexity of such optical systems and the limited benefit will make this a slow market to develop. Larger scale and longer term demonstrations will be required to gain acceptance from the engineering community.

Vibrational measurements could provide information on any earthquake activity within a region or the deterioration of a structure. Electromagnetic sensors that take advantage of steel's magnetic permeability as a function of its internal stress also present tools for monitoring bridge cables and prestressed concrete structures (12). In this method, the internal stresses of highly elastic steel are measured by determining its permeability, which can be measured indirectly by its inductance (12). Strain sensors will be a key tool for monitoring crack initiation sites and a good indicator of structural failure. Typical sensors, in addition to the those mentioned before, include strain gauge sensors, displacement transducers, accelerometers, anemometers, electrical time domain reflectometry (for stress/strain sensing), and temperature sensors.

Many highway systems enforce weight restrictions on large truck to reduce road damage. The use of weight in-motion sensors such as piezoelectric polyvinylidene fluoride (PVDF) polymer can reduce the damage caused by heavy trucks. This type of polymer embedded in elastomeric

material placed in a groove on the highway can detect the weight of a passing vehicle and translate the weight into a voltage output (13). The disadvantages of this type of sensor are temperature variation and physical damage under extreme loads.

SMART HIGHWAYS

A smart highway system consists of sensors, computers, and communication tools to enable all driving functions to work. The smart highway program is designed to make travel smoother on highways by quickly alerting motorists to traffic accidents, icy stretches of road, and other hazards, and by posting the best alternate routes. Today's highway systems are long, steep grades and sharp curves that present problems especially for high traffic volume and bad weather conditions. More than half of all traffic accidents occur during foggy, rainy, icy, and snowy conditions. More than two-third of truck accidents occur on curves or slopes. One way to reduce traffic accidents is to use advanced communication systems via satellites and place warning signs. The following must be considered in building a smart highway:

- Traffic control centers assisted by a computer networking system where controllers adjust traffic flow. In this center, video-image signals, which are sent by cameras and video cameras, mounted on poles and building, are converted to digital maps.

- Construction of many ramp meters where traffic lights are installed in critical entrance ramps to control the flow of merging traffic.

- Placement and design of narrow poles that support signs and light on fixed objects such as a bridge (14) for safety enhancement.

- Installation of hundreds of sensors in the pavement to count cars as they pass and to estimate and transfer this information to the control center.

- Broadcast of critical traffic information to alert drivers to slow down ahead and advise an alternate route.

- Automatic toll collection where sensors read optical cards on dashboards.

- Sufficient safety enhancement.

In addition to building and monitoring highways systems, control centers that are assisted by computer-networking system are also required to manage traffic and construct intelligent transportation systems. Central units are a way to communicate to drivers and law enforcement officers to reduce routine accidents by improving visibility at night or in bad weather by early warning to drivers (15).

An intelligent highway system that has an electronic communication system should be capable of the following tasks: (1) automatically regulate the flow of traffic, (2) provide drivers with up-to-the-minute information, (3) perform most driving tasks, (4) ease carpooling, and (5) manage and guide commercial fleets (14).

To build a modern highway, loop detectors and video cameras will be installed in critical areas to monitor traffic flow, speed of vehicles, and identify bottlenecks to a central communication headquarters. An intelligent highway system will consist of information processing, communications, control, and electronics to transmit critical information to drivers.

Satellites flying in low orbit will be required to transmit data to a central unit. Satellites must have a broad range and can collect, monitor, store, and selectively transmit data to process information.

Smart highways could also be used for smart public transportation, allowing bus drivers control over passenger traffic, manage traffic flow, and reduce delays. Smart transit systems can be organized for expensive share-ride taxis and to assemble carpools and vanpools for daily operation (16), for smart goods movement systems to assist companies to transport goods more cheaply and with less energy and resource use, using streamlined truck inspections and better routing through traffic and linking road-rail transport systems. Computer process technology can also be used to improve manufacturer information and communications to reduce the need for long-distance shipping and provide faster delivery systems to encourage purchases from home or local stores.

To minimize fuel consumption and improve fuel efficiency, formulas must be developed based on factors such as the lane miles of roadway, vehicle miles traveled, the level of mail routes, the population and the size of the state in square miles and transmitted to a central computer system.

ADVANCED AUTOMOBILES

In conjunction with smart highways, smart vehicles will provide complete control of nearly all driving functions. The major tasks of driving consist of navigation, braking, steering, throttle control, and avoiding accidents; the vehicle will automatically control traffic light management. Most automated driving tasks have already been implemented in pilot vehicles by major auto manufacturers.

Today's automobiles carry more advanced semiconductor technologies than they did in the early 1980s. Until now, chip technology has been used to enhance the performance of engines and to control airbags and antilock brake systems. However, navigation systems such as the global satellite (GPS) systems will dominate the new generation of telecommunication advances in congested traffic areas. Modern car manufacturers have developed navigation systems to pinpoint a driver's location and are also developing systems that activate warnings for avoiding objects in the blind spots. Collision avoidance units are also under development to steer, brake, or accelerate a vehicle automatically (15). A unique feature of the modern automobile is the card key for opening car doors and driver information identification that will enable identifying a car's speed and taking care of tolls. The microchip-embedded card that is slightly thinner than a normal credit card can operate within three feet of the vehicle. A Siemens Smart Card is

already available as an option on new the Mercedes-Benz S-Class.

Personal communicators and computers are presently under development to receive data from Global Positioning Satellite (GPS) signals. Receivers work with embedded systems and translate and correlate data with multiple satellites to determine the positioning of vehicles. A video screen using GPS signals can determine the location on a road or within a city map. In combination with sensors, computers will be able to monitor traffic and road conditions and even control the distance between cars for increased safety.

To build smarter highways and structures, we may also need to build smarter and more sophisticated automobiles. Smart sensors and devices can be used to control traction, steering, and suspension and monitor tire pressure and sense and orient a car automatically to road conditions. Sensors that can control the speed, vibration, and temperature of vehicles could be used in conjunction with road sensors to optimize most critical functions of an automobile. Sensors could also be used in the rear and front of a vehicle to warn drivers that they are getting too close to another vehicle or are being approached too closely by another automobile in addition to lane changing. Optical sensors, based on the misbonding effect and speckle phenomenon technology, can be used to identify a vehicle type and its speed and also to monitor traffic flow and count vehicles on the road. These sensors can be placed inside the asphalt layer of the road surface (17). The major area of focus for automobiles besides safety is the use of sensors for automation and using exotic materials such as composites to substitute for steel to improve fuel efficiency. The use of new materials such as composites provides a multitude of potentials and degrees of freedom for materials design that involve increased strength, creating new functions and expanding to multiple functionalities. Smart materials consist of composites that indicate exactly the direction of the future development of materials engineering and represent a change from "supporting" to "working" to build up a new materials application system that integrates structures, functions, and information.

Smart sonic traffic sensors placed on acoustic sensors is another alternative (18) to magnetic-loop sensors to detect vehicles from the sounds that they make. More sophisticated cruise control could be tied in with sensors to reduce or increase speed instantaneously to avoid accidents. General Motors and Ford have tested computerized navigation systems to pinpoint a driver's location and to warn drivers of potential obstacles by using detection systems to steer clear of objects in blind spots and avoid collisions.

Modern automobiles for smart highways also being considered where by drivers can take their hands off the wheel and eyes off the road enabling advanced cruise controls take charge of the driving (19). Smart highways and uniform speed are the major requirements for this futuristic idea before such cars can be used. In addition to cruise control features, this type of vehicle is equipped with radar fields. Sensors emit a beep if the car is about to hit something. This type of technology is currently available

in the Ford Windstar and some other commercial vans (19).

The combination of sensors and the advanced cruise control will enable vehicles to take charge of most routine driving tasks. Examples of modern cars under consideration include General Motors Buick Division where magnetic pegs are inserted on the road and vehicles are then capable of riding on their own. Front sensing radar plays a key role in maintaining the distance from other cars; one of the key features of this type of automobile is communication between cars equipped with similar technology. Mitsubishi has also built a futuristic car where the vehicle had dual mode driving capability; the first mode is the traditional driving mode where the driver is in command. In the second mode, the driver uses the passenger seat, and the car uses sensors and HR6 technology to take complete care of all driving tasks. This type of vehicle is equipped with the latest communication tools where the driver can monitor the traffic and weather conditions, access the Internet, and check e-mail. In the automatic mode, the vehicle body turns into an aerodynamic type. Mitsubishi has also developed cars that have multiple sensors to detect steep roads, curves, and hazardous signs, and the vehicle adjusts after detecting any upcoming warning sign. Ford also uses a new technology of light beam output where the size and the shape of the beam are calibrated with the speed and the type of road, and radar is installed in the vehicle. This technology provides ideal speed control for safety and fuel efficiency. Jaguar progress has been in night vision technology where infrared technology gives the driver the ability to monitor any object that cannot be observed during darkness. Mercedes-Benz in cooperation with Boeing is also developing a limousine equipped with the latest electronic features that can drive on its own and is also equipped with GPS technology.

The future car for the twenty-first-century will look like a rolling recreation room and a source of entertainment as manufacturers progress in developing new technology. This in turn will reduce traffic commuting between home and office and also will require far less attention from drivers. Future automobiles will be designed to represent true mobility rather than a transportation tool. The new generation of cars will possess more revolutionary and innovative electronic features to ease driving tasks and access communication networks for weather, news, the worldwide web, and satellite or cellular networks. Thus far, the United States has lagged behind other industrialized nations; in 1998, of all vehicles equipped with navigation systems, less than 5% were purchased in the United States and more than 90% were sold in Europe and Japan, where there is higher demand for communication technology.

Recent developments in automobile manufacturing consist of using a laptop computer to interface with all sensors to warn and control the devices in a car. The new laptop computers offer ample processing power and disk space and can operate on 12 Vdc power. On new futuristic cars, the laptop computer is likely to become standard, and the cost of remaining associated hardware is expected to fall significantly in the near future.

Global Positioning System (GPS)

GPS is a technology for improving the accuracy of positioning information. Greater accuracy is potentially useful in such ways as improving the accuracy of maps, enhancing search and rescue efforts, improving navigation on crowded highways and waterways, and helping planes land in bad weather. Present technology has made it possible to improve greatly the accuracy of global positioning information available from satellites. This technology, called Differential Global Positioning Systems, allows pilots, surveyors, and others using satellite positioning information to determine their positions on earth to within a few meters—or even a few centimeters. Normal GPS can provide latitude and longitude, speed, and direction of travel. GPS is beneficial to improve safety for trucks by installing receivers and sensors in the trailer section. The load of the truck then can be monitored in addition to determining tax and fuel rates (9). The driver can also receive real-time traffic and navigational information.

The development of a central GPS unit for nationwide use is critical for managing multiple functions for entire smart highways within all states. This system could be used on land as well as in the air and on the sea. Development of common equipment standards, technical feasibility and accessibility, and organizational structures will be the key issues for coordinating this system. Global positioning data can presently be provided from a network of Department of Defense (DOD) satellites. Planes, boats, vehicles, and mapping and survey teams can determine their positions on earth by using equipment that receives and interprets signals from these satellites. For smart highway applications, the satellites provide a signal that is accurate to about 100 meters without the use of GPS. Coordination between the federal government and local states may be needed to enhance joint development or sharing of Differential Global Positioning Systems equipment, facilities, and information for future use. The limitation of existing GPS technology lies in highly populated areas that have large buildings and trees. GPS is not functional inside a tunnel or any enclosed area (16). Present GPS technology relies on a satellite signal whose signal is received and translated by a receiver (9). The system works perfectly in an uninhabited area where it may not be as useful.

The price of a GPS system has fallen dramatically in recent months provided that the automobile is equipped with a portable computer. Earthmate sells for less than \$180 and is a high-performance, easy-to-use receiver that links to the satellite navigation technology of the Global Positioning System (GPS).

UPDATE ON SMART HIGHWAY PROJECTS UNDER CONSTRUCTION

Major smart highway development has been underway in the states of California and New Jersey. Thus far, major problems consist of major delays in completing construction and some minor accidents due to the extreme weight of signs that require support. New Jersey's Route 80 from the George Washington Bridge to its connection with 287 in Morris County and Routes 95, 23, 46, 4, 17, 202, 287,

and 280 use radar, pavement sensors, and closed circuit TV and cameras. This highway system was designed to provide real-time information about traffic, ice, upcoming accidents, and weather (20). Besides major construction delays, problems appear to be the variable sizes of signs throughout the highway that make reading them difficult. Another obstacle is a potential design flaw where the strengths of structure may be underestimated for strong storms and abnormal weather conditions. Most of the problems thus far have been related to scheduling, lack of coordination for use, and timing of installation. It appears that a pilot smart highway may be required where extreme weather conditions are present before additional major superhighway construction begins. Chrysler Corporation has developed vehicles, particularly large trucks for smart highways, that are presently being tested without any drivers. However, the company is not betting that any major smart highway projects will be started soon. Chrysler believes that there are many old cars on the road that may interfere with the general concept of fully automated highways. This presents a case for a two-tier highway system, one for modem vehicles and one for cars that are not equipped with smart computers. The associated costs and capital for building new highways must be considered relative to potential revenues. Another dilemma concerns turning over the control of human lives on such highways to a major corporation or the government (20).

SMART HIGHWAYS IN JAPAN

Japan has been far ahead of most industrialized nations in developing and using smart materials; therefore, it is beneficial to review the recent progress of smart cars and highways that is a model for the rest of the world.

Traffic fatalities in Japan are approximately 14,000 per year (21) at an annual cost of \$120 billion. Population density is also 12 times higher than that in the United States. Therefore, the benefits of constructing an ITS system will have a tremendous impact on productivity. The total annual budget for an intelligent transportation system is estimated at 700 million, proportionally higher than that in the United States (21). Japan has more than 3800 miles of toll roads and development is underway to automate a toll collection system fully.

Although traffic control systems have been used in Japan for a number of years to ease traffic, the major objective in automating a traffic system in Japan by using a smart highway system is for safety enhancement and reduction of traffic fatalities. Another objective is to enhance communication between vehicles, particularly commercial vehicles and public transit, by using a central traffic management system. Today, Japan has more than 112 miles of smart highways, which consist of 2,077 vehicle detectors to monitor the number of vehicles and speed. These smart roads also have graphic displays and television cameras. Presently, Japanese auto manufacturers offer 40 different models of navigation systems; approximate sales are one million units per year.

In Japan, ITS development began in the 1960s and 1970s by construction of a road system, the Electronic

Route Guidance Systems similar to those in the United States and Germany (21). In the 1980s, due to microprocessing technology development and the lower cost of computer chips, work has been under way to set up a system for communicating between the road and the automobile. The final phase will be implementation of smart highway and smart cars to build an intelligent transportation system (ITS). Although United States lags behind Japan in smart highway development, the implementation of fully automated highways has not yet materialized in Japan or Europe.

SUMMARY

Fully automated highways may still be a long-range vision, but one must recognize the tremendous social and cost implications of converting our basic transportation structure to a more interactive, customer-oriented "smart" system. The critical obstacles besides cost are whether the public really wants some level of external control over its driving behavior, even if that control means increased safety and efficiency. As for cost, a fully implemented "smart" highway or transit system is probably akin to building another interstate system, not to mention the increased cost of smart vehicles to consumers. Although ITS will probably be implemented incrementally, the areas that need to be considered are social and cost implications for the public and agreement on some long-range vision of an ITS futuristic design. Congress and local states will make the final decision; however, because the United States reliance on cars is expected to continue for at least the next 50 years, development of smart cars and smart highways will be required to reduce the traffic and improve safety.

Future operational prototypes of smart highways have been decided; the actual test and evaluation phase nationwide is planned between 2002 and 2006 (22). The challenge for future intelligent transportation systems (ITS) will be to maximize safety and efficiency and reduce traffic congestion and associated costs. Auto manufacturers have already begun to build collision avoidance devices, electronic brakes, and steering and sensors to automate driving. Future evaluation of ITS will be based on reduction of traffic and accidents, energy efficiency, and reduction of cost and travel time compared to the present highway system.

ACKNOWLEDGMENTS

I thank Professor James Harvey, who introduced me to the field of smart materials and structures and encouraged me to broaden my knowledge in this field. I also express my gratitude to Mary B. Taylor who read the manuscript and contributed and suggested great futuristic ideas on how to build smarter highways and structures.

BIBLIOGRAPHY

1. D. Wills, *Transp. Res. Rec.* **1234**: 47 (1989).
2. *Cybermautic Digest*, Vol. 3, Number 3: *Transportation*, KFH, 1996.

3. M.M. Ettouney, R. Daddazio, and A. Hapij, 78–89.
4. M.W. Lin, A.O. Abatan, and W.M. Zhang, 297–304.
5. S.C. Liu and D.J. Pines, *SPIE Conf.*, 1999.
6. Intelligent Sensing for Innovative Structures (ISIS): (204) 474-8506, Smart Materials Smart Bet for the Future of Engineering.
7. Y. Adachi and S. Unjoh, 31–42.
8. S.P. Marra, K.T. Ramesh, and A.S. Douglas, *SPIE Conf. San Diego*, Vol. 3324, pp. 94–95.
9. *Road & Bridges*, **37**(11): 1999.
10. Summaries from 1998 Westech's Virtual Job Fair & High Technology Careers.
11. A.E. Glazounov, Q.M. Zhang, and C. Kim, *SPIE Conf. San Diego*, Vol. 3324, pp 82–91.
12. N. Lhermet, F. Claeysen, and P. Bouchilloux, 46–52.
13. R.K. Panda, P.J. Szary, A. Mahr, and A. Safari, *SPIE Conf. Smart Mater. Technol.*, March 1998, Vol. 3324, pp. 127–134.
14. J.C. Wu and J.N. Yang, 23–34.
15. *V. Tech Mag.* **14**(I): 1991.
16. M.A. Replogle, member of the U.S. Department of Transportation Intelligent Vehicle and Highway Systems Architecture Task Force. © 1994 Environmental Defense NY.
17. P. Suopajarvi, M. Heikkinen, P. Karioja, V. Lyori, R.A. Myllyla, S.M. Nissila, H.K. Kopola, and H. Suni, 222–229.
18. R. Klashinsky and J. Lee, *AT&T SmartSonic Traffic Surveillance System*, Lucent Technologies, Inc.
19. R. Konrad, Smart Highways, Detroit Free Press, April 28, 1999.
20. P.R. Gilbert, January 19, 1998.
21. H. Tokuyama, Intelligent Transportation Systems in Japan, 1999.
22. N. Congress, Smart Road, Smart Car: The Automated Highway System, 1999.

HYBRID COMPOSITES

DAZHI YANG
ZHONGGUO WEI
Dalian University of Technology
Dalian, China

INTRODUCTION

Smart materials, or intelligent materials systems, are concepts developed in the late 1980s. Technologically, smart materials could be said to integrate actuators, sensors, and controls with a material or structural component. Scientifically, they can be defined as material systems with intelligence and life features that reduce mass and energy and produce adaptive functionality. The development of smart materials has been inspired by biological structural systems and their basic characteristics of efficiency, functionality, precision, self-repair, and durability. As is well known, few monolithic materials presently available possess these characteristics. Accordingly, smart materials are not singular materials, rather, they are hybrid composites or integrated systems of materials.

Presently, no materials possessing high-level intelligent have been developed. Only some smart hybrid composites that can receive or respond to a stimulus, including temperature, stress, strain, an electric field, a magnetic field, and other forms of stimuli have been developed and studied by materials scientists. These smart hybrid composites are developed by incorporating a variety of advanced functional materials, such as shape memory materials, piezoelectric materials, fiber-optics, magnetostrictive materials, electrostrictive materials, magnetorheological fluids, electrorheological fluids, and some functional polymers. Smart hybrid composites provide tremendous potential for creating new paradigms for material-structural interactions, and they demonstrate varying success in many engineering applications, such as vibration control, sound control, quiet commuter aircraft, artificial organs, artificial limbs, microelectromechanical systems among a variety of others.

Shape-memory materials (SMMs) are one of the major elements of smart hybrid composites because of their unusual properties, such as lie shape-memory effect (SME), pseudoelasticity, or large recoverable stroke (strain), high damping, capacity, and adaptive properties which are due to the reversible phase transitions in the materials. To date, a variety of alloys, ceramics, polymers and gels have been found to exhibit SME behavior. Both the fundamental and engineering aspects of SMMs have been investigated extensively and some of them are presently commercial materials. Particularly, some SMMs can be easily fabricated into thin films, fibers or wires, particles and even porous bulks, enable them feasibly to be incorporated with other materials to form hybrid composites.

SHAPE MEMORY ALLOY FIBER/METAL MATRIX COMPOSITES

The basic design approaches for the SMA fiber/metal matrix composite can be summarized five steps: (1) The SMA fiber/metal matrix composites are prepared and fabricated by using conventional fabrication techniques; (2) the as-fabricated composites will be heated to high temperatures to shape memorize the fibers or to undergo some specific heat treatment for the matrix material, if necessary; (3) since SMAs have much lower stiffness at martensite stage or readily yield at the austenitic stage just above the martensitic transformation start temperature (M_s), the composites are then cooled to lower temperatures, preferably in martensite state; (4) the composites are further subjected to proper deformations at the lower temperature to enable the martensite twinning or the stress-induced martensitic transformation to occur; and (5) the prestrained composites are then heated to higher temperatures, preferably above the austenite finish temperature A_f , wherein martensite detwinning or the reverse transformation from martensite to austenite takes place, and the TiNi fibers will try to recover their original shapes and hence tend to shrink, introducing compressive internal residual stresses in the composites. This design concept can also be applied to polymer matrix composites containing SMA fibers and to the metal matrix composites containing SMA particles.

The internal residual stress in both the fiber and the matrix, and the composite macroscopic strains as a function of external variables such as temperature and applied load or prestrain have been calculated within nonlinear composite models using Eshelby's formulation. As expected, depending on the fiber pretreatment and distribution, as well as the boundary conditions, varying levels of compressive residual stresses can be generated in the matrix of the SMA composite during heating process, resulting in a large negative thermal expansion. For a given SMA fiber reinforcement, the matrix compressive residual stress increases with increasing volume fraction and prestrain of the SMA fibers within a limited range, and optimal prestrain and fiber volume fraction values can be found. In addition, the magnitude of the internal residual stress is limited by the flow strength of both the SMA fibers and the matrix material.

Apart from the dependence on the volume fraction and prestrain, the yield stress of the composite increases with increasing temperature within a limited temperature range. This is because the contributing back stress in the Al matrix induced by stiffness of TiNi fibers and the compressive stress in the matrix originate from the reverse transformation process from the "soft" martensite to the parent phase (austenite) with a several times higher stiffness. For the austenite phase fiber, the critical stress to induce the martensitic transformation shows a strong positive dependence on temperature, as demonstrated in the Clausius-Clapeyron equation and temperature-stress-strain curves of SMAs.

In agreement with modeling predictions, with increasing fiber volume fraction and prestrain, a more significant strengthening effect of the composite by the SMA fibers was observed. It was found that the Young's modulus and tensile yield stress increase with increasing volume fraction of fibers. The crack propagation rate as a function of the apparent stress intensity factor in the composites was measured and a drastic drop of the propagation rate (i.e., crack-closure effect) was observed after the composite was heated to higher temperatures ($>A_f$). The enhancement of the resistance to fatigue crack propagation was suggested to be ascribed to the combination of compressive residual stress, higher stiffness of the composite, the stress-induced martensitic transformation and the dispersion of the mechanical strain energy at the crack-tip.

The SMA fiber-reinforced MMCs also exhibit other improved properties. For instance, the damping capacity of the TiNi fiber/Al matrix composite was measured and the results indicated that the damping capacity of the composite in the temperature range 270 to 450 K was substantially improved over the unreinforced aluminum. The composite was also expected to show high wear resistance.

SHAPE MEMORY ALLOY FIBER/POLYMER MATRIX COMPOSITES

Depending on the SMA fiber pretreatment, distribution configuration, and host matrix material, a variety of hybrid polymer matrix composites can be designed that may actively or passively control the static and/or dynamic

properties of composite materials. Passively, as in the SMA fiber/Al matrix composites, the shape memory alloy fibers are used to strengthen the polymer matrix composites, to absorb strain energy and alleviate the residual stress and thereby improve the creep or crack resistance by stress-induced martensitic transformations. The embedded SMA fibers are usually activated by electric current heating, and hence they undergo the reverse martensitic transformation, giving rise to a change of stiffness, vibration frequency, and amplitude, acoustic transmission or shape of the composite. As a result structural tuning modal modification or vibration and acoustic control can be accomplished through (1) the change in stiffness (inherent modulus) of the embedded SMA elements or (2) activating the prestrained SMA elements to generate a stress (tension or compression) that will tailor the structural performance and modify the modal response of the whole composite system just like tuning a guitar string. The two techniques are termed active property tuning (APT) and active strain energy tuning (ASET), respectively. In general, APT requires a large volume fraction of SMA fibers that are embedded without prior plastic elongation and do not create any large internal forces. While ASET may be equally effective with an order of magnitude by a smaller volume fraction of SMA fibers that are active, however, and impart large internal stresses throughout the structure.

Usually the embedded or bonded shape memory alloy fibers are plastically elongated and constrained from contracting to their “normal” length before being cured to become an integral part of the material. When the fibers are activated by passing current through them, they will start to contract to their normal length and therefore generate a large, uniformly distributed shear load along the length of the fibers. The shear load then alters the energy balance within the structure and changes its modal response. Shape memory alloy fibers can also be embedded in a material off the neutral axis on both sides of the beam in an antagonist–antiantagonist pair. Alternative interaction configurations include creating “sleeves” within the composite laminates and various surface or edge attachment schemes.

Advanced composites such as graphite/epoxy and glass/epoxy composites offer high strength and stiffness at a low weight and moderate cost. However, they have poor resistance to impact damage because they lack an effective mechanism for dissipating impact strain energy such as plastic yielding in ductile metals. As a result the composite materials dissipate relatively little energy during severe impact loading and fail in a catastrophic manner once stress exceeds the composite’s ultimate strength. Typically damage progresses from matrix cracking and delamination to fiber breakage and eventual material puncture. Various approaches to increase the impact damage resistance, and specifically the perforation resistance, of the brittle composite materials have been attempted. The popular design concept is to form a hybrid that utilizes the tougher fibers to increase the impact resistance and also the stiffer and stronger graphite fibers to carry the majority of the load. The hybrids composed of the graphite/epoxy with Kevlaro[®], Spectrag[®], and S-glass fibers have demonstrated modest improvements in impact resistance. Among

various engineering materials, high strain SMAs have a relatively high ultimate strength. They can absorb and dissipate a large amount of strain energy first through the stress-induced martensitic transformation and then through plastic yielding. Accordingly, the impact resistance of the graphite/epoxy composites may be improved by hybridizing them with SMA fibers. Paine and Rogers have developed the concept and demonstrated that under certain load conditions the impact energy absorbing ability of graphite and glass composites can be effectively improved by hybridizing the composites with TiNi SMA fibers. Hybrid composites with improved impact and puncture resistance are very attractive because of their great potential in military and commercial civil applications.

Generally, the shape memory hybrid composite materials can be manufactured with conventional polymer matrix composite fabrication methods, by laying the SMA fibers into the host composite prepreg between or merging with the reinforcing wires and then using either hot press or autoclave and several different types of cure cycles. Previously the few attempts to incorporate embedded TiNi wires directly into a polymer matrix composite proved unsuccessful due to manufacturing difficulties and problems associated with interfacial bonding. To avoid the interface bonding issue, SMA wires were alternatively incorporated into polymer matrix by using coupling sleeves. Both thermoset and thermoplastic composites have been addressed. Comparatively, fiber-reinforced thermoplastics offer some substantial advantages over fiber-reinforced thermosets because of their excellent specific stiffness, high fracture toughness, low moisture absorption, and possible fast and cost-effective manufacturing processes. However, the high process temperatures can be problematic for the embedding of SMA elements. The thermoplastic processing must be performed at higher temperatures, typically between 423 and 673 K. Whereas the thermoset processing cycle of the composites is in the relatively low temperature range of RT – 443 K. The thermoplastic processing cycle has some effect on the microstructure of the SMA fibers as manifested in the change in transformation temperatures and peak recovery stress: the transformation temperatures of the SMA shift upward while the peak recovery stress drops as a result of the thermoplastic processing. The thermoset processing only mildly affects the transformation characteristics of SMA fibers.

However, some dynamical properties of SMA fibers could be significantly affected. Much of previous research on the SMA hybrid composites utilized the one-way shape memory effect, especially in the applications that require recovery stress of the SMA. Much care should be taken to prevent shape recovery of the prestrained SMA fibers or wires during the composite cure cycle. The complexity of manufacturing the SMA composites can be greatly simplified by using the two-way shape memory effect. That is, the SMA wires will be trained to exhibit the two-way shape memory effect prior to embedding in the matrix.

Void content is one of the pressing issues in manufacturing the SMA hybrid composites. Voids in composite materials significantly affect the material integrity and behavior. Their presence in the SMA composites will not only lead to property degradation of the host composite material, but

the efficiency of activation and the level of interfacial bonding between the SMA fibers and host matrix will also be sacrificed. In the hot press composites with graphite/epoxy laminates and embedded TiNi fibers, the average void content was found to be 10.20%. Voids were shown to be concentrated near the embedded TiNi wire locations. Additionally, the interfacial bonding was quite poor. The void content can be reduced as low as 1.29% by autoclave stage curing.

Another concern is the interfacial bonding. In the SMA hybrid composites, the maximum interfacial adhesion between the SMA wire and the polymer matrix is desired because most applications require maximum load transfers, and a strong interfacial bond also increases the structural integrity of the final composite. To improve the interfacial bonding, various surface treatments of SMA fibers have been examined involving the introduction of a coupling interphase. The pullout test was used to qualitatively compare the interfacial adhesion. Five kinds of TiNi fibers—that is, untreated, nitric acid-etched, handsanded, sand-blasted, and plasma-coated—and two kinds of host matrix materials—that is, graphite/epoxy and PEEK/carbon (APC-2) composites—were examined. The pullout test results indicated that in the TiNi fiber/APC-2 system, a brittle interface failure without friction occurred, resulting in overall lower peak pullout stress levels. In the TiNi fiber/GR/EP system, however, strong mechanical interaction or friction between the TiNi fiber and GR/EP composite occurred. As a result the fiber pull-out stress levels show a dependence on the adhesion between TiNi fibers and host composite, and on the average, the peak pull-out stresses are significantly higher than those in the APC-2 composites. Generally, sand-blasting of TiNi fibers increases the bond strength while handsanding and acid cleaning actually decrease the bond strength. Surprisingly, it was found that plasma coating of the fibers did not significantly alter the adhesion strength. The in-situ displacements of embedded SMA wires were also measured and the resulting stresses were induced in the matrix by using heterodyne interferometry and photoelasticity, respectively. As expected, the constraining effect of the matrix increases with increasing bond strength, causing a decrease in the displacement of SMA wire and a corresponding increase in the interfacial shear stress induced in the matrix.

SMA PARTICULATE / ALUMINUM MATRIX COMPOSITES

Particulate-reinforced metal matrix composites (MMCS) have attracted considerable attention because of their feasibility for mass production, promising mechanical properties, and potential high damping capacity. In applications not requiring extreme loading or thermal conditions, such as automotive components, the discontinuously reinforced MMCs have been shown to offer substantial improvements in mechanical properties. In particular, discontinuously reinforced Al alloy MMCs provide high damping and low density and allow undesirable mechanical vibration and wave propagation to be suppressed. As in the fiber-reinforced

composites, the strengthening of the composites is achieved through the introduction of compressive stresses by the reinforcing phases, due to the mismatch of the thermal expansion coefficient between the matrix and reinforcement. The most frequently used reinforcement materials are SiC, Al₂O₃, and graphite (Gr) particles. Although adding SiC and Al₂O₃ to Al matrix can provide substantial gains in specific stiffness and strength, the resulting changes in damping capacity may be either positive or negative. Graphite particles may produce a remarkable increase in damping capacity, but at the expense of elastic modulus. More recently, Yamada et al. have proposed the concept of strengthening the Al MMCs by the shape-memory effect of dispersed TiNi SMA particles. The strengthening mechanism is similar to that in the SMA fiber reinforced composites: the prestrained SMA particles will try to recover the original shape upon the reverse transformation from martensite to parent (austenite) state by heating and hence will generate compressive stresses in the matrix along the prestrain direction, which in turn enhances the tensile properties of the composite at the austenitic stage. In the light of the well-known transformation toughening concept, some adaptive properties such as self-relaxation of internal stresses can also be approached by incorporating SMA particles in some matrix materials. Since SMAs have a comparatively high loss factor value in the martensite phase state, an improvement in the damping capacity of the SMA particulates-reinforced composites is expected at the martensite stage. Accordingly, SMA particles may be used as stress or vibration wave absorbers in paints, joints, adhesives, polymer composites, and building materials.

Shape-memory particulate-reinforced composites can be fabricated by consolidating aluminum and SMA particulates or prealloyed powders via the powder metallurgical route. SMA particulates may be prepared with conventional processes, such as the atomization method and spray or rapid solidification process which can produce powders with sizes ranging from micrometers to millimeters. However, few reports on the production of SMA particles are recorded in the open literature. Recently, Cui has developed a procedure to prepare Ti–Ni–Cu SMA particulates through hydrogenating-ball milling-dehydrogenating. The ternary TiNiCu alloys, where there is a substitution of Ni by Cu by up to 30 atomic %, are of particular interest for their narrow hysteresis, large transformation plasticity, high shock absorption capacity, and basic shape-memory effect. Owing to their unique properties, the ternary Ti–Ni–Cu alloys have shown some promise as smart materials with actuation, sensing, and adaptive strengthening characteristics. When the content of Cu exceeds 15 at%, the ternary alloys become very brittle and hence more easily broken down into particulates by ball milling. Although it was reported previously that Cu–Zn–Al alloy powders had been prepared from commercial Cu–Zn and Al powders, using the mechanical alloying technique, there was no physical evidence to prove that thermoelastic martensitic transformations occurred in the as-received powders. As a matter of fact, most of the attempts to prepare the TiNi and Cu-based SMA particulates by ball milling were

unsuccessful due to the complex mechanical alloying reactions and contaminations during the process.

In an exploratory attempt, a $\text{Ti}_{50}\text{Ni}_{25}\text{Cu}_{25}$ alloy was prepared in a high-frequency vacuum induction furnace. The ingot was homogenized at 1173 K for 10 h. The bulk specimens were cut from the ingot. Chips with a size of about $0.1 \times 3 \times 30$ mm from the ingot were hydrogenated at 673 K for 5.8 h in a furnace under a hydrogen atmosphere, then the chips were ball milled in a conventional planetary ball mill, the weight ratio of balls to chips being 20 : 1. The vials were filled with ether and the rotational speed of the plate was kept constant during the milling. The milled powders were then dehydrogenated in a vacuum furnace at 1073 K for 10 minutes at a vacuum of 10^{-3} Pa. The X-ray diffraction and transmission electron microscopy (TEM) observations indicated that the as-received Ti–Ni–Cu particulates, similar to the bulk counterpart, possessed a mixture structure of B19 and B19' martensites. Differential scanning calorimetry (DSC) results demonstrated that the particulates exhibit excellent reversible martensitic transformations, though the peak temperatures were slightly altered when compared to the bulk material. The Ti–Ni–Cu/Al composite was prepared from 99.99% Al powders of 2 to 3 μm in size and the Ti–Ni–Cu powders of about 30 μm in size. The volume ratio of the Ti–Ni–Cu powder to Al powder was 3 : 7. The powders were mixed in a mixer rotated at 50 RPM for 10 h. The consolidation of the mixed powders was achieved by hot isostatic pressing (HIP) at 793 K for 10 minutes, the relative density of the compact being 98.5%.

The DSC measurements of the Ti–Ni–Cu/Al composite showed evidence the occurrence of the thermoelastic martensitic transformations, just as demonstrated in the Ti–Ni–Cu bulk and particles. These preliminary results suggest that it is feasible to produce some adaptive characteristics within the composite through the shape-memory alloy particulates.

CERAMIC PARTICULATE / SMA MATRIX COMPOSITES

In a shape-memory alloy matrix, dispersed second-phase particles may precipitate or form during solidification or thermal (mechanical) processing, thereby creating a native composite. The martensitic transformation characteristics and properties of the composites can be modified by control the particles, as demonstrated in Ti–Ni(Nb), CuZn–Al, and Cu–Al–Ni–Mn–Ti alloys. Alternatively, the presence of a ceramic second phase within the SMA matrix may lead to a new composite with decreased density and increased strength, stiffness, hardness, and abrasion resistance. Compared with common ceramic/metal composites, a higher plasticity may be expected for this composite because the stress-induced martensitic transformation may relax the internal stress concentration and hence hinder cracking. Previously, Al_2O_3 particle reinforced CuZnAl composites were prepared with conventional casting method, and this kind of composite was suggested to be suitable for applications requiring both high damping and good wear resistance. Using explosive

pressing of the powder mixture, a TiC/TiNi composite was prepared. In the sintered TiC/TiNi composite it was found that the bend strength, compression strength, and stress intensity factors were significantly higher than those for TiC/Ni and WC/Co composites. With increasing TiC content, the hardness and compressive strength increase, while the ductility and toughness decrease. More recently, Dunand et al. systematically investigated the TiNi matrix composites containing 10 vol% and 20 vol% equiaxed TiC particles, respectively. The composites were prepared from prealloyed TiNi powders with an average size of 70 μm and TiC particles with an average size of about 40 μm , using powder metallurgy technique. The TiC particles modify the internal stress state in the TiNi matrix, and consequently, the transformation behavior of the composite: the B2-R transformation is inhibited; the characteristic temperatures A_s and M_f are lowered, while the M_s temperature remains unchanged; and the enthalpy of the martensitic transformation is reduced. Unlike composites with matrices deforming solely by slip, the alternative deformation mechanisms, namely twinning and stress induced transformation, are expected to be operative in the TiNi composites during both the overall deformation of the matrix and its local deformation near the reinforcement, thereby resulting in the pseudoelasticity and rubberlike effect. Compared to unreinforced TiNi, the range of stress for formation of martensite in the austenitic matrix composite is increased, and the maximum fraction of the martensite is lowered. For both the austenitic and martensitic matrix, a strengthening effect can be observed: the transformation or twinning yield stress is increased in presence of the dispersed TiC particles. However, for the austenitic matrix, the transformation yield stress is higher than predicted by Eshelby's load transfer theory, due to the dislocations created by the relaxation of the mismatch between matrix and particles. In contrast, for the martensitic matrix, the twinning yield stress and the apparent elastic moduli are less than predicted by Eshelby's model because of the twinning relaxation of the elastic mismatch between matrix and reinforcement. Besides the elastic load transfer, the thermal, transformation, and plastic mismatches resulting from the TiC particles are efficiently relaxed mainly by localized matrix twinning, as revealed by neutron diffraction measurements. As a result the shape memory capacity, that is, the extent of strain recovery due to detwinning upon unloading, is scarcely affected by the presence of up to 20 vol% ceramic particles.

MAGNETIC PARTICULATE / SMA MATRIX COMPOSITES

Giant magnetostrictive materials $(\text{Tb}_y\text{Dy}_{1-y})_x\text{Fe}_{1-x}$, (Terfenol–D) provide larger displacements and energy density, and superior manufacturing capabilities, as compared to ferroelectrics. However, their applications have been limited by the poor fracture toughness, eddy current losses at higher frequencies, and bias and pre-stress requirements. More recently, composite materials based on Terfenol–D powders and insulating binders have been developed in Sweden. These composites broaden the

useful range of the Terfenol-D material, with improved tensile strength and fracture toughness, and potential for greater magnetostriction and coupling factor. Most recently, Ullakko has proposed a design concept to embed Terfenol-D particles within a shape memory alloy matrix to create a ferromagnetic shape memory composite with combination of the characteristics of shape memory alloys and magnetostrictive materials. The Terfenol-D particles will be elongated by about 0.1% when applying a magnetic field. The generated force is high enough to induce the martensitic transformations in the matrix at appropriate temperatures. Therefore, the orientation and growth of the martensite plates may be controlled by the magnetic field, and by the distribution and properties of the Terfenol-D particles embedded in the matrix. The magnetic control of the shape-memory effect through the magnetostrictive inclusions may be used independently, or simultaneously with the thermal control to achieve optimal performance. Experimentally, a Terfenol-D/Cu-Zn-Al composite was prepared using Cu-Zn-Al SMA and Terfenol-D (15 wt%) powders with the shock wave compaction method. However, the magneto(visco)elastic response and thermomechanical properties of the composite have not been reported. It is probable that this kind of composite is not suitable as an active actuator material due to some technical limitations.

As an alternative, the high passive damping capacity of the magnetic powders/SMA matrix composites may be utilized. It is known that Cu-Zn-Al SMAs have high damping capacity at large strain amplitudes due to thermoplastic martensitic transformations, but their stiffness is inadequate for some structural applications. The ferromagnetic alloys, including Terfenol-D, Fe-Cr, Fe-Cr-Al, and Fe-Al, are known to have relatively high strength as well as high damping capacity in the range of small strain amplitudes, and low damping capacity in the range of large strain amplitudes. In principle, the combination of Cu-Zn-Al matrix and ferromagnetic alloy inclusions should yield high damping capacity over a wide range of strain amplitudes, and higher stiffness than that of the monolithic Cu-Zn-Al alloys. Accordingly, three kinds of metal matrix composites were fabricated from prealloyed Cu-26.5 wt% Zn-4.0 wt%Al powders (as a matrix) and rapidly solidified Fe-7 wt%Al, Fe-20 wt%Cr, and Fe-12 wt%Cr-3 wt%Al alloy flakes (30 vol%), respectively, by powder metallurgy processing. The interfaces between the Cu-Zn-Al matrix and the flakes in the consolidated composites were delineated and were free of precipitates or reaction products. In all of the three composites, the damping capacity with the strain in the range from 1.0×10^{-4} to 6.0×10^{-4} was found overall to show substantial improvements. In particular, the Fe-Cr flakes/Cu-Zn-Al composite demonstrated the highest overall damping capacity and exhibited an additional damping peak at strain 165×10^{-6} .

SMA/SI HETEROSTRUCTURES

The development of shape-memory alloy thin films for microelectromechanical systems (MEMS) is one of the most

important engineering applications of shape-memory alloys during the past decade. Owing to the extensive use in IC microfabrication technologies, silicon is particularly preferable as the substrate to fabricate and pattern SMA thin films in batches. TiNi, Ti-Ni-Cu, and other kinds of SMA films have been deposited onto both single-crystal silicon and polysilicon substrates.

From a thermodynamical point of view, TiNi is unstable compared to Si. As a result interface diffusion and chemical interactions may occur, and Ti and Ni silicides may be formed on postdeposition annealing, especially at higher temperatures, of the SMA films. A thin buffer layer of Nb or Au can prevent the interdiffusion. In particular, a buffer layer of SiO₂ has been proven an effective diffusion barrier and an excellent transition layer favoring the interface adherence.

The delamination of the deposited SMA films from Si arising from the evolution of the intrinsic residual stress must be prevented. Wolf and Heuer reported that the adherence of TiNi with bare Si wafer can be improved if it has been cleaned and etched with a buffered oxide etchant (H₂O + NH₄F + HF) prior to deposition. Also modest heating of the substrate under vacuum to around 473 K, prior to deposition, can minimize contamination and improve adherence. Krulevitch et al. also reported that in-situ heated Ti-Ni-Cu SMA films adhere well to bare silicon. The adherence of TiNi film with both bulk SiO₂ and thermal oxide coated Si (SiO₂/Si) were reported to be excellent. A 50 to 300 nm thick layer of TiNi with parent B2 phase, which remains untransformed, was observed adjacent to the interface. The untransformed interlayer, which may be due to the effect of the strong (110) B2 texture, contributes to the interface adherence by accommodating the strain through a gradient or by absorbing the elastic energy. In some cases, electrical isolation of the film is needed. Wolf and Heuer reported that deposition of a 0.1 μm polysilicon layer on SiO₂ prior to deposition of TiNi resulted in a well-bonded interface.

The structure of the composite films should be properly designed to achieve optimal performance. Owing to the mechanical constraints via the interface, the substrate stiffness, determined by the film/substrate thickness ratio, has a significant effect on the transformation characteristics of the SMA layer and on the output energy of the composite's multiple layers. The optimum SMA film thickness for maximum cantilever deflection depends on the relative stiffness of the SMA film and the underlying beam. The behavior of the film depends on the film thickness and approaches bulk behavior as the film becomes a few micrometers thick. However, more compliant actuating films must be slightly thicker for maximum tip deflection. Up to now, some novel microdevices using the SMA/Si diaphragm have been patterned and fabricated, such as microvalves and microactuators, the micro robot arm, and the microgripper.

SMA/PIEZOELECTRIC HETEROSTRUCTURES

An ideal actuation material would display a large stroke, high recovery force and superior dynamical response.

Shape-memory alloys exhibit large strokes and forces but suffer from a slow response. Ferroelectric ceramics are very sensitive to applied stresses through the direct piezoelectric effect and generate powerful forces by means of the converse piezoelectric effect. The ceramics are characterized by excellent dynamical response (on the order of microseconds), but their displacements are quite small (on the order of a few micrometers) due to their small strain magnitude ($<10^{-3}$). There are a large number of ferroelectric ceramics, but the most widely investigated and currently applied for thin film technology are the titanate and niobate (with oxygen octahedral structure) types, such as lead titanate (PbTiO_3), lead zirconate titanate (PZT), lead lanthanum zirconate (PLZT), barium titanate (BaTiO_3), and strontium titanate (SrTiO_3). Ferroelastic SMAs combined with ferroelectric piezoelectric ceramics have yielded hybrid heterostructures that have the optimum characteristics of both materials.

Piezoelectric thin films can be fabricated with various techniques such as sputtering, chemical vapor deposition (CVD), and sol-gel processing. The sol-gel process of piezoceramics has had increasing applications because the chemical composition can be controlled precisely. Of particular concern here is whether the amorphous piezoelectric materials can be synthesized on SMA, and vice versa. Chen et al. first successfully deposited thin films of PZT and PLZT with 0.6 and 1.4 μm thickness, respectively, onto TiNi SMA foils by the sol-gel process and multistep spin-on coating techniques ($$). The amorphous films were annealed at temperatures above 773 K to obtain the perovskite phases. The dielectric constant and loss tangent at 100 kHz of the TiNi/PZT composite film were about 700 and 0.03, respectively, comparable to that of the bulk ceramics. The PZT films were found adhere well to the TiNi alloy for strains as large as 0.4%, and their ferroelectric properties remain unchanged during repeated cycling through the shape-memory transformation. However, considerable cracking was observed when the diaphragms subjected to a strain of 0.5%. Jardine et al. and Alam et al. also successfully deposited the thin films of PZT, BaTiO_3 , and SrTiO_3 onto commercially available TiNi SMA bulk and thin films by sol-gel and spin-on techniques or with pulsed laser deposition method, though the quality of the multilayer composites was not that desired.

Both types of the amorphous thin films will be crystallized simultaneously if, deposited on amorphous SMA films. Therefore, the fabrication steps must be minimized so as not to promote degradation of performance due to second phases and chemical interactions via diffusion. The composite's multilayers were annealed at various temperatures ranging from 723 to 973 K and a suitable crystallization temperature was found at about 873 K. Although the heterostructures have good SME and piezoelectric properties, the cracking of the piezoceramic thin film layer remains a critical problem. Generally, a thicker PZT film causes more cracks than a thinner film, whereas a smooth surface roughness and a slow cooling rate after annealing will favor the bond of the PZT film with the TiNi SMA substrate. An effective method to lessen cracking is to deposit a buffer layer of TiO_2 onto TiNi SMA foil and then

deposit the piezoelectric film onto the TiO_2/TiNi substrate. Nevertheless, how to accommodate the stress and dynamical coupling of the dissimilar material layers remains a problem that must be solved before SMA films can be used for actuation applications.

Alternatively, the ferroelastic/ferroelectric heterostructures may be effective for active suppression of high amplitude acoustic waves and shock waves. After coupling TiNi SMA to PZT via a TiO_2 layer, the final composite material was found to sense and actuate to dampen structural vibration without the use of external control. The mechanism of the active damping can be explained by considering an approaching stress wave. The stress wave propagates through the TiNi SMA, producing a stress-induced martensitic transformation where some of the mechanical energy is converted into heat. The wave further produces a voltage across the first ferroelectric layer that can be used to produce an out-of-phase stress wave by the second ferroelastic layer and in turn attenuate the stress wave. A mechanical metallic impedance buffer (e.g., Al, Ti, and TiNi) is used to provide time for the counterstress actuation to occur.

SMA/TERFENOL-D HETEROSTRUCTURES

Magnetostrictive materials with either crystalline or amorphous structure provide higher counterforces and up to 20 times higher strains than piezoelectric ceramics. Of the magnetostrictive materials presently available, the compounds Terfenol-D ($\text{Tb}_x\text{Dy}_{1-x}\text{Fe}_2$) have the largest magnetostriction and magnetization at room temperature, the strain output being up to 0.2% when subjected to up to 2500 oersteds (Oe), and in some cases approaching 1%. The alloys are analogous to electrostrictive materials in that they respond quadratically to an applied field. The optimum performance of Terfenol-D is achieved with the combination of a bias field plus a bias compressive stress. The superior properties of Terfenol-D have attracted increasing attention to the use of this material in both bulk and thin film form, applied for actuation.

Terfenol-D films can also be fabricated with the conventional magnetron sputtering techniques. Su et al., Quandt et al., and other researchers have successfully deposited Terfenol-D films of various thicknesses onto Si/SiO₂ substrates by DC magnetron sputtering. The thin films deposited at room temperature are amorphous and the crystallization temperatures are much high (>903 K). However, it should be reminded that the as-received amorphous Terfenol-D films are excellent ferromagnetic materials. The amorphous films show a sharp increase in the magnetostriction at low magnetic fields and no hysteresis during cycling of the field, whereas the crystalline films exhibit magnetostrictive hysteresis loops and large remanence and coercivity, which limit their application. Since the amorphous Terfenol-D films do not need annealing at elevated temperatures to address undesirable chemical interactions or diffusion, the fabrication of hybrid composite films appears to be easy and simple. For instance, the Terfenol-D films can be grown on crystalline TiNi SMA

bulk or deposited films annealed before the sputtering of the Terfenol-D films. Su et al. proposed the concept of a Terfenol-D/NiTi/Si composite in which the ferroelastic actuation can be triggered by magnetic field. Although the Terfenol-D/SiO₂/Si and TiNi/SiO₂/Si composites have been fabricated and characterized, no further report on

the successful fabrication of Terfenol-D/TiNi hybrid composite films are recorded in the open literature. Surely this is a interesting and exciting subject that needs further investigations, and of course, some technical challenges such as their interface compatibility still remain ahead.

MAGNETS, ORGANIC/POLYMER

JOEL S. MILLER

University of Utah
Salt Lake City, UT

ARTHUR J. EPSTEIN

The Ohio State University
Columbus, OH

INTRODUCTION

Magnetism has enabled the development and exploitation of fundamental science, ranging from quantum mechanics, to probing condensed matter chemistry and physics, to materials science. The control of magnetism has resulted in the availability of low-cost electricity and the use of electric motors, leading to the development of telecommunications devices (microphones, televisions, telephones, etc.), and to magnetic storage for computers. Magnets, due to their myriad properties, are suitable components in sensors and actuators, and hence they must be considered the key components in smart materials and systems of the future.

Magnetic materials known from time immemorial are comprised of either transition or rare-earth metals, or their ions with spins residing in d- or f-orbitals, respectively, such as Fe, CrO_2 , SmCo_5 , $\text{Co}_{17}\text{Sm}_2$, and $\text{Nd}_2\text{Fe}_{14}\text{B}$. These materials are prepared by high-temperature metallurgical methods, and generally, they are brittle. In the late twentieth century many metal and ceramic materials were replaced with lightweight polymeric materials. The polymeric materials were designated primarily for structural materials, but examples also abound for electrically conducting and optical materials. More recently, new examples of magnetic materials (1) have been reported. Undoubtedly, in this millennium there will be commercialization of these organic and polymeric magnets (2).

Magnetism is a direct consequence of the coupling of unpaired electron spins. Independent, uncoupled electron spins, as shown in Fig. 1(a), lead to paramagnetic behavior. Strong coupling of the aligned spins, Fig. 1(b), can lead to a substantial magnetic moment and a ferromagnet, while that of opposed spins, Fig. 1(c), to an antiferromagnet as the net moments cancel. In contrast, the incomplete cancellation of spins can lead to a net magnetic moment and a ferrimagnet, Fig. 1(d), or a canted antiferromagnet also termed a weak ferromagnet, Fig. 1(e).

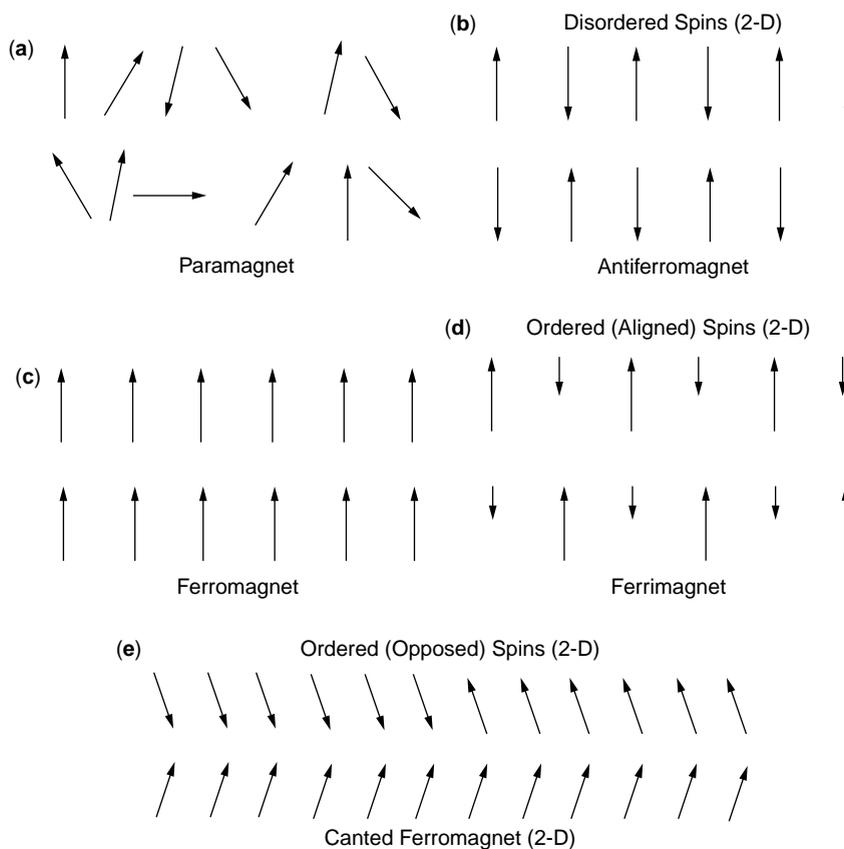


Figure 1. Two-dimensional spin alignment for (a) paramagnet, (b) antiferromagnet, (c) ferromagnet, (d) ferrimagnet, and (e) canted antiferromagnet behavior.

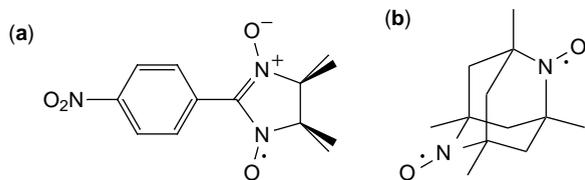


Figure 2. Structure of 4-nitrophenyl nitronyl nitroxide, which orders as a ferromagnet at 0.6 K (a), and a dinitroxide that orders as a ferromagnet at 1.48 K (b).

The recently discovered magnets with spins residing in p-orbitals (organic magnets) have added the following properties to those in the repertoire that already can be attributed to magnets: solubility, modulation of the properties via organic chemistry synthetic methods, and low-temperature (nonmetallurgical) processing, enhancing their technological importance and value for the smart materials or systems.

A few organic nitroxides order ferromagnetically below a T_c of 1.5 K. These include a phase of 4-nitrophenyl nitronyl nitroxide reported by M. Kinoshita et al. (1b) with a T_c of 0.6 K, Fig. 2(a), and a phase of dinitroxide reported by A. Rassat et al. (3) with a T_c of 1.48 K, Fig. 2(b). Each of these molecules can crystallize into more than one polymorph, four for the former and two for the latter (4). However, in each case only one of the possible polymorphs magnetically orders as a ferromagnet. These examples of organic magnets, albeit with very low T_c 's, are crystalline solids, and not polymers.

Organic magnets possessing unpaired electron spins in both p- and d-orbitals also have been reported (1a,c). These include ionic decamethylferrocenium tetracyanoethanide, $[\text{FeCp}_2^*][\text{TCNE}]$ ($T_c = 4.8$ K; $\text{Cp}^* =$ cyclopentadienide, $[\text{C}_5(\text{CH}_3)_5]^-$; $\text{TCNE} =$ tetracyanoethylene), Fig. 3, exhibiting the first evidence for magnetic hysteretic behavior in an organic magnet, as reported by J.S. Miller and A.J. Epstein (5,6). $[\text{Fe}^{\text{III}}\text{Cp}_2^*][\text{TCNE}]^-$ has an alternating $\cdots\text{D}^+\text{A}^-\text{D}^+\text{A}^-\cdots$ ($\text{D} = [\text{FeCp}_2^*]^+$; $\text{A} = [\text{TCNE}]^-$) structure in the solid state, Fig. 4. The observed 16,300 emu Oe/mol saturation magnetization, M_s , is in excellent agreement with the calculated value of 16,700 emu Oe/mol for single crystals aligned parallel to the chain axis. Hysteresis loops with a coercive field of 1 kOe are observed at 2 K (Fig. 5) (1a,5,6).

D and A each have a single spin ($S = \frac{1}{2}$). Above 16 K the magnetic susceptibility behaves as expected for a 1-D ferromagnetically coupled Heisenberg chain with $J/k_B = 27$ K (6). Below that temperature, the susceptibility

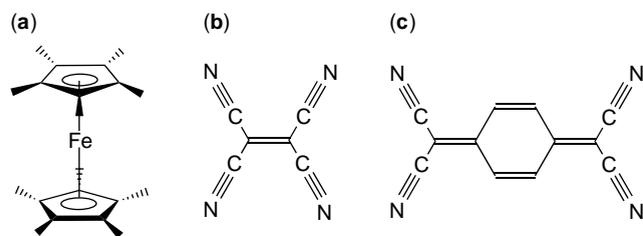


Figure 3. Molecular structures of FeCp_2^* (a), TCNE (b), and TCNQ (c).

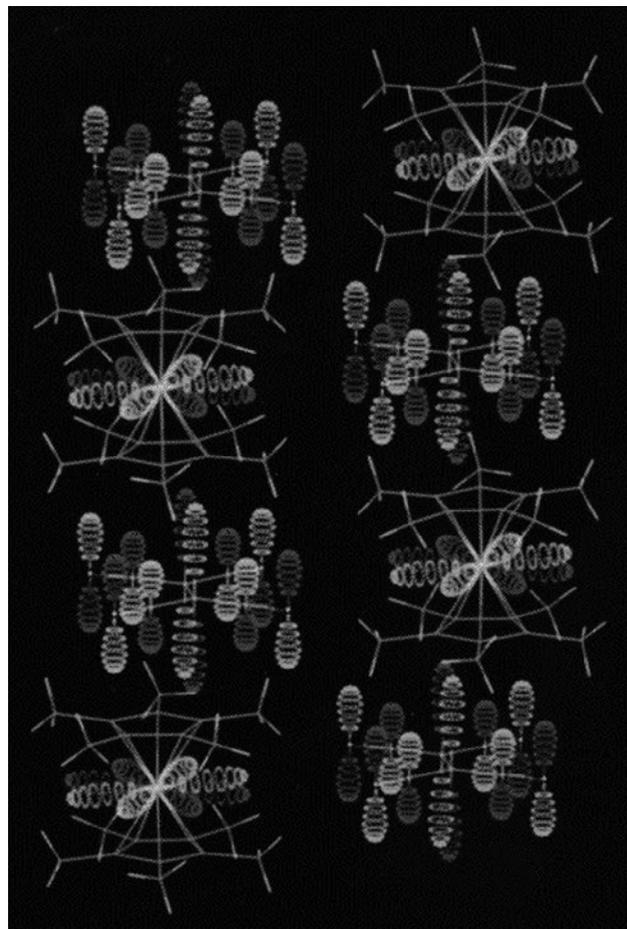


Figure 4. Crystal structure of $[\text{FeCp}_2^*][\text{TCNE}]$ showing the orbitals possessing the largest density of unpaired electrons.

diverges as $(T - T_c)^{-\gamma}$ as anticipated for a 1-D Heisenberg-like system approaching a 3-D magnetically ordered state. Spontaneous magnetization below the 4.8 K ordering temperature follows $(T_c - T)^\beta$ with $\beta \sim 0.5$. Hysteresis loops are well defined with coercive field $H_{cr} = 1$ kG at 2 K (Fig. 5) indicating substantial pinning of the domain walls.

Replacement of Fe by Cr (7) and Mn (8) as well as substitution of TCNE with TCNQ (9–11). [7,7,8,8-tetracyano-*p*-quinodimethane, Fig. 3(c)], leads to ferromagnets with T_c reduced for TCNQ substitution and enhanced when Mn is utilized, Table 1 and Fig. 6. Partial substitution of spinless ($S = 0$) $[\text{CoCp}_2^*]^+$ for ($S = 1/2$) $[\text{FeCp}_2^*]^+$ in $[\text{FeCp}_2^*][\text{TCNE}]^-$ leads to a rapid reduction of T_c as a function of the fraction of spinless sites ($1 - x$) occurs, such as 2.5% substitution of $[\text{FeCp}_2^*]^+$ sites by $[\text{CoCp}_2^*]^+$ decreases T_c by 43% (12).

D. Gatteschi, P. Rey and co-workers (1c) demonstrated and more recently H. Iwamura and co-workers (13) have shown that covalent polymers comprised of bis (hexfluoroacetylacetonate) manganese(II) (Fig. 7) and nitroxides bound to the Mn(II) sites order as ferrimagnets. Using more complex nitroxides Iwamura and co-workers have prepared related systems with T_c 's ~ 46 K. Using TCNE electron-transfer salts of Mn^{II} -(porphyrin)'s, such as $[\text{MnTPP}][\text{TCNE}]$ (TPP = *meso*-tetraphenylporphyrinato),

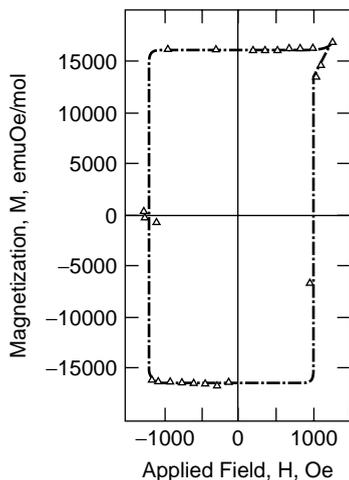


Figure 5. 2 K hysteresis loops with a coercive field, H_{cr} , of 1 kOe for $[\text{FeCp}_2^*][\text{TCNE}]$.

Fig. 8, (14) magnets ($T_c = 13$ K) (15) based on metallo-macrocycles also can be prepared. Both the $[\text{TCNE}]^-$ and the nitroxide each have one spin; however, the Mn(II) in the former 1-D polymeric chain has five spins ($S = 5/2$), while the Mn(III) in the latter polymeric chain has four spins ($S = 2$). In both cases, the Mn and organic spins couple antiferromagnetically, leading to ferrimagnetic ordering. The solid state motif is distinctly different than that for $[\text{MnCp}_2^*][\text{TCNE}]^-$ (Fig. 4) as $[\text{TCNE}]^-$ does not coordinate to the M in the latter system. Thus, the bonding of $[\text{TCNE}]^{\bullet-}$ to Mn is a model for the bonding of $[\text{TCNE}]^{\bullet-}$ to V in the $\text{V}[\text{TCNE}]_x(\text{solvent})$ room temperature magnet (16).

As a consequence of the alternating $S = 2$ and $S = \frac{1}{2}$ chain structure, the $[\text{MnTPP}]^+[\text{TCNE}]^-$ 1-D chain system, which forms a large family of magnets, is an excellent model system for studying a number of unusual magnetic phenomena. This includes the magnetic behavior of mixed quantum/classical spin systems (17a), the effects of disorder (17b), and the role of classical dipolar interaction (in contrast to quantum mechanical exchange) in achieving magnetic ordering (17c). Because of the single ion anisotropy for $[\text{MnTPP}]^+$ and the large difference in orbital overlaps along and between chains,

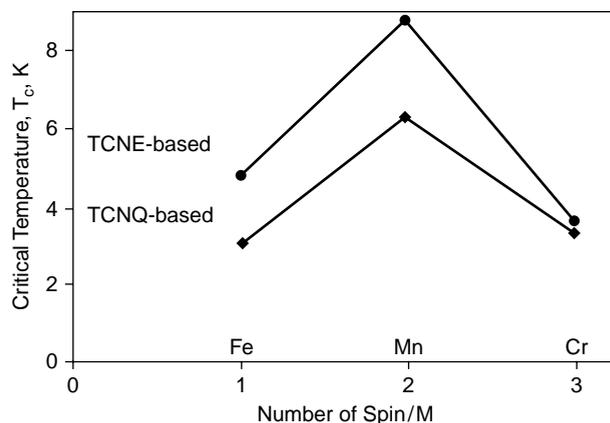


Figure 6. Critical temperature, T_c , for $[\text{MCp}_2^*][\text{A}]$ ($M = \text{Fe}, \text{Mn}, \text{Cr}$; $\text{A} = \text{TCNE}, \text{TCNQ}$).

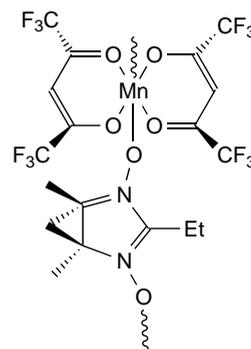


Figure 7. Structure of $\text{Mn}^{\text{II}}(\text{hfac})_2\text{NITet}$ ($\text{hfac} = \text{hexafluoroacetylacetonate}$; $\text{NITet} = \text{ethyl nitronyl nitroxide}$).

this class of materials often undergoes lattice- and spin-dimensionality crossovers as a function of temperature (17d). Metamagnetic-like behavior is noted for many members of this family at low temperature with unusually high critical fields, H_c , and unusually high coercive fields, H_{cr} , as large as ~ 2.7 T (17e). These metamagnetic-like materials are atypical as they exhibit hysteresis with very large coercive fields, H_{cr} , also as high as ~ 2.7 T, and may have substantial remanent magnetizations (Fig. 9).

Table 1. Summary of the Critical Temperatures, T_c , and Coercive Fields, H_{cr} , for $[\text{MCp}_2^*][\text{A}]$

	[TCNE] $^-$			[TCNQ] $^-$		
	$[\text{FeCp}_2^*]^+$	$[\text{MnCp}_2^*]^+$	$[\text{CrCp}_2^*]^+$	$[\text{FeCp}_2^*]^+$	$[\text{MnCp}_2^*]^+$	$[\text{CrCp}_2^*]^+$
S_D	1/2	1	3/2	1/2	1	3/2
S_A	1/2	1/2	1/2	1/2	1/2	1/2
T_c , K	4.8	8.8	3.65	3.0	6.3	3.3
θ , K	+16.9	+22.6	+22.2	+3.8	+10.5	+11.6
H_{cr} , kOe (K)	1.0 (2)	1.2 (4.2)	^a	^b	3.6 (3)	^a
Reference	(5,6)	(7)	(8)	(9)	(10)	(11)

Note: $M = \text{Fe}, \text{Mn}, \text{Cr}$; $\text{A} = \text{TCNE}, \text{TCNQ}$.

^a Not observed.

^b Not reported.

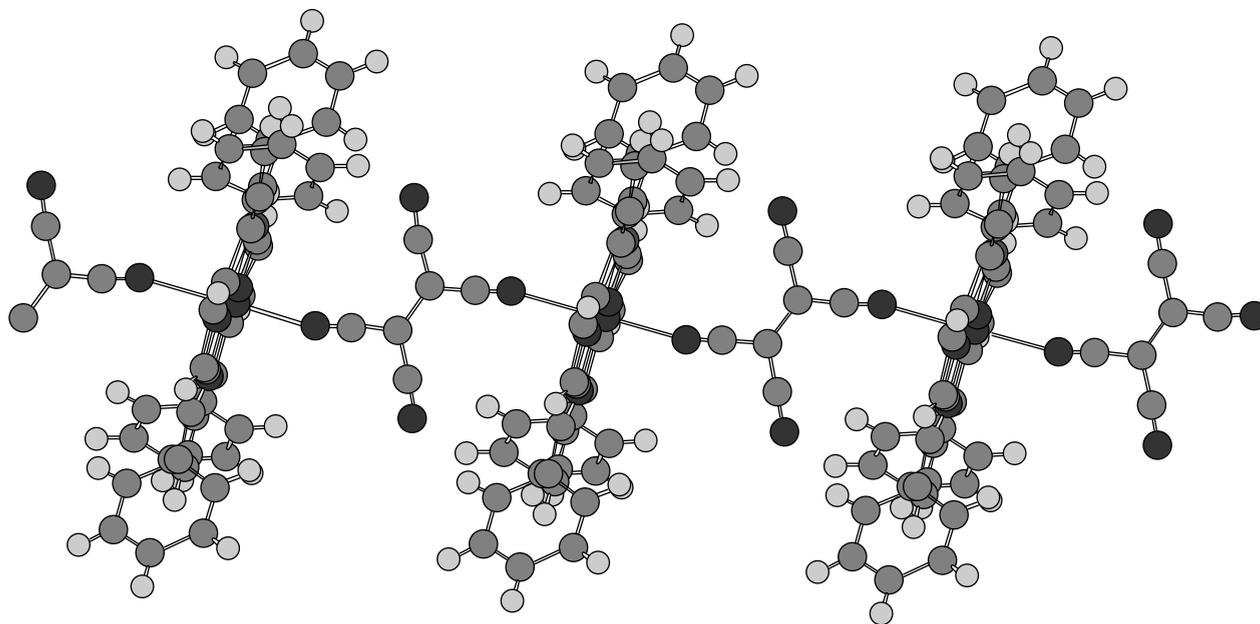


Figure 8. Segment of an 1-D...D⁺A⁻D⁺A⁻... chain of [MnTPP][TCNE]·2C₆H₅CH₃ showing [MnTPP]⁺ *trans*- μ -N- σ -bonding to [TCNE]⁻.

V(TCNE)_xY(SOLVENT) ROOM TEMPERATURE MAGNETS

Reaction of V(C₆H₆)₂ and TCNE in a variety of solvents, such as dichloromethane and tetrahydrofuran, leads to loss of the benzene ligands and immediate formation of V(TCNE)_xyCH₂Cl₂ ($x \sim 2$; $y \sim \frac{1}{2}$). Because of its extreme air and water sensitivities as well as insolubility, compositional inhomogenieties within and between preparations occur. The structure remains elusive. This material, however, is the first ambient temperature organic- or polymer-based magnet ($T_c \sim 400$ K) (16). The proposed structure has each V being octahedrally coordinated with up to 6 ligands (N's from different TCNE's), and each TCNE is reduced and is either planar or twisted and bound to up to four V's.

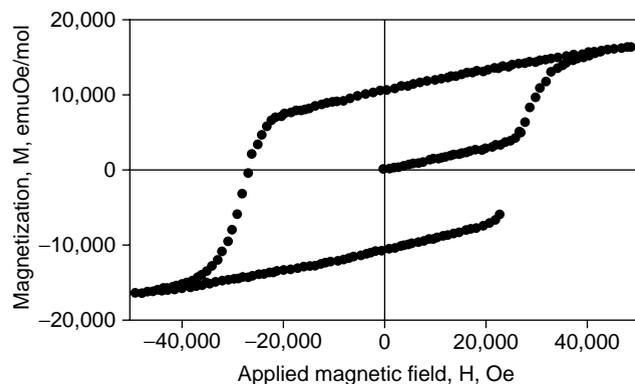


Figure 9. Metamagnetic and hysteresis with 27,000 Oe critical, H_c , and coercive, H_{cr} , fields at 2 K.

Magnets made in dichloromethane have a three-dimensional magnetic ordering temperature, T_c , of ~ 400 K, which exceeds its ~ 350 K decomposition temperature, and hence can be attracted to a SmCo₅ magnet at room temperature. Recently, solvent-free thin magnetic films on a variety of substrates, such as silicon, salt, glass, and aluminum, have been prepared, Fig. 10 (18).

The inhomogeneity in the structure is expected to lead to variations in the magnitude, not sign, of the exchange between the V^{II} ($S = \frac{3}{2}$) and [TCNE]⁻ ($S = \frac{1}{2}$), and the disorder should result in a small anisotropy (16).

V(TCNE)_x prepared in CH₃CN has larger disorder and a lower $T_c \sim 135$ K, which enables the quantitative study of

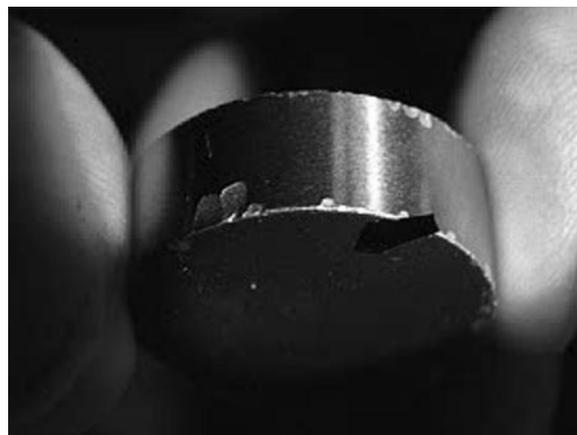


Figure 10. Photograph of ca. 5 μ m coating of the V[TCNE]_x magnet on a glass cover slide being attracted to a Co₅Sm magnet at room temperature in the air (18).

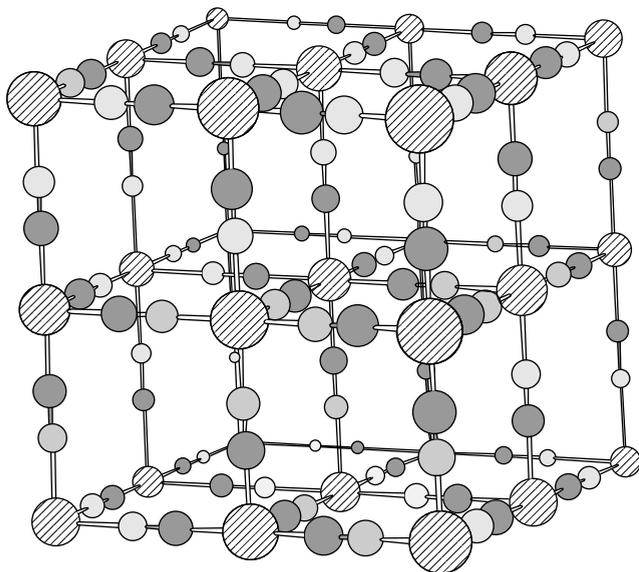


Figure 11. Idealized structure of $\text{Fe}_4^{\text{III}}[\text{Fe}^{\text{II}}(\text{CN})_6]_3 \cdot x\text{H}_2\text{O}$, Prussian blue with $\cdots\text{Fe}^{\text{III}}-\text{N}\equiv\text{C}-\text{Fe}^{\text{II}}-\text{C}\equiv\text{N}-\text{Fe}^{\text{III}}\cdots$ linkages along each of the three unit cell axes.

its critical behavior. Near T_c , a modified equation-of-state approach can be used to obtain the approximate critical exponents for this disordered magnet. A critical isotherm can be determined by plotting $M(H)$ against H at varying temperatures, with M proportional to $H^{1/\delta}$ at T_c . Using this analysis, we determine that $\delta = 4$ and $T_c = 135$ K for the typical $\text{V}(\text{TCNE})_x \cdot y \text{CH}_3\text{CN}$ sample being studied. With T_c and δ determined, exponents β_α and γ_α (for the random anisotropy model) can be determined directly by analyzing isothermal plots of $(H/M)^{1/\gamma_\alpha}$ versus M^{1/β_α} . Given β_α and δ , all of the $M(H, T)$ data in the vicinity of T_c can be collapsed onto a single set of curves for plots of $\ln(M/|t|^{\beta_\alpha})$ versus $\ln(H/|t|^{\beta_\delta})$, where $t = |T - T_c|$; one curve corresponds to $T > T_c$ and the other curve corresponds to $T < T_c$ (19a). Similar results can be obtained for $\text{V}(\text{TCNE})_x \cdot y \text{C}_4\text{H}_8\text{O}$ with $T_c = 205$ K and some different critical exponents (19b). Evident in the analysis of high-temperature magnetization of $\text{V}(\text{TCNE})_x \cdot y \text{CH}_2\text{Cl}_2$ is the important role of random magnet anisotropy (19c).

$\text{M}(\text{TCNE})_2 \cdot x(\text{CH}_2\text{Cl}_2)$ (M = Mn, Fe, Co, Ni) HIGH ROOM TEMPERATURE MAGNETS

New members of the family of high- T_c organic-based magnets $\text{M}(\text{TCNE})_2 \cdot x(\text{CH}_2\text{Cl}_2)$, (M = Mn, Fe, Co, Ni; TCNE = tetracyanoethylene) have been prepared (20). X-ray diffraction studies on powder samples show that these materials are partially crystalline and isomorphous. These materials have T_c 's of 97, 75, 44, and 44 K, respectively. Field-cooled and zero-field-cooled magnetization studies suggest that while the Mn (21) system is a reentrant spin glass, the Fe (22) system is a random anisotropy system. Both systems exhibit complex behavior below T_c . For example, hysteresis curves for the Fe compound, taken at 5 K, are constricted, with a spin-flop shape, indicating

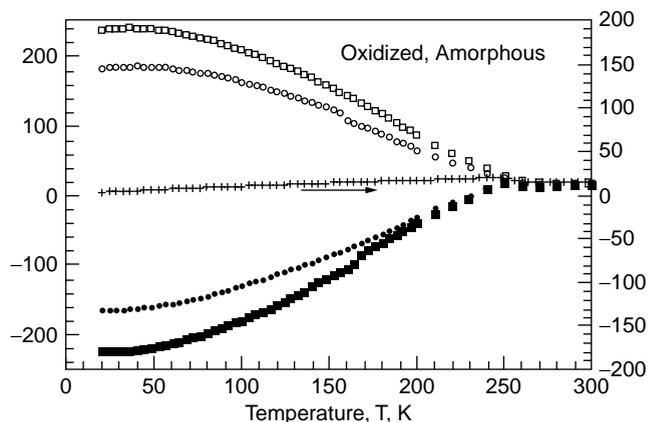


Figure 12. Temperature-dependent magnetization of the amorphous film of $\text{Cr}^{\text{III}}[\text{Cr}^{\text{III}}(\text{CN})_6]_{0.98}[\text{Cr}^{\text{II}}(\text{CN})_6]_{0.02}$, zero field cooled (+) and field cooled in 5 Oe (\circ), -5 Oe (\bullet), 10 Oe (\square), and -10 Oe (\blacksquare) after 2 minutes of oxidation at -0.2 V upon warming in a 5 Oe field (25).

ferrimagnetic behavior, and field- and zero-field-cooled magnetization studies reveal magnetic irreversibilities below T_c for both compounds. Static and dynamic scaling analyses of the dc magnetization and ac susceptibility data for the Mn compound show that this system undergoes a transition to a 3-D ferrimagnetic state at T_c , followed by a reentrant transition, to a spin glass state at $T_g = 2.5$ K. For M = Fe, the results of static scaling analyses are consistent with a high- T transition to a correlated sperimagnet, while below about 20 K, there is a crossover to sperimagnetic behavior.

HEXACYANOMETALLATE MAGNETS

Although rigorously not organic magnets, several materials termed molecule-based magnets are prepared by the same organic chemistry methodologies and have been



Figure 13. Sample of a $\text{V}[\text{Cr}(\text{CN})_6]_y \cdot z\text{H}_2\text{O}$ magnet ($T_c = 372$ K) attracted to a Teflon coated magnet at room temperature in the air (27a).

Table 2. Summary of the Critical Temperatures, T_c , and Coercive Fields, H_{cr} , for Representative Organic-Based Magnets

Magnet	M_s , (emuOe/mol)	T_c (K)	H_{cr} (Oe)	Reference
α -1,3,5,7-Tetramethyl-2,6-diazaadamantane- N,N' -dioxyl (Fig. 2b)	11,000	1.48	<0.1	(3)
β -2-(4'-Nitrophenyl)-4,4,5,5-tetramethyl- 4,5-dihydro-1H-imidazol-1-oxyl-3- N -oxide (Fig. 2a)	5600	0.6	8	(1b)
$[Fe^{III}Cp_2^*][TCNE]$ (Figs. 3a, b; 4)	16,300	4.8	1000	(1a,5,6)
$Mn^{II}(hfac)_2NITEt$ (Fig. 7)	20,000	7.8	320	(1c)
$[MnTPP][TCNE] \cdot 2C_6H_5CH_3$ (Fig. 8)	17,000	13	24,000	(15,17)
$V[TCNE]_x \cdot yCH_2Cl_2$ ($x \sim 2; y \sim 0.5$)	10,000	~ 400	15	(16b)

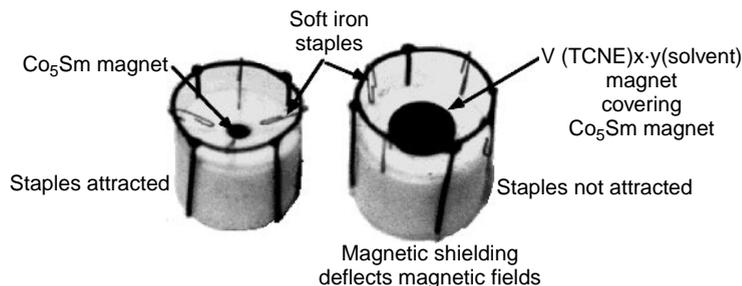


Figure 14. Folded soft iron rods (staples) are shown attracted to a $SmCo_5$ permanent magnet (*left*). The soft iron rods hang freely when a pellet of $V(TCNE)_x \cdot y(CH_2Cl_2)$ at room temperature shields the magnetic field.

reported to magnetically order in many cases. Prussian blue, $Fe^{III}[Fe^{II}(CN)_6]_3 \cdot xH_2O$, possesses a 3-D network structure, with $\dots Fe^{III}-N \equiv C-Fe^{II}-C \equiv N-Fe^{III} \dots$ linkages along each of the three unit cell axes, Fig. 11. It is a prototype structure that stabilizes both the ferro- and ferrimagnetic orders. Replacement of iron with other spin-bearing metal-ions leads to strong magnetic coupling and magnetic ordering with high T_c 's. Examples include ferromagnetic $CsNi^{II}[Cr^{III}(CN)_6] \cdot 2H_2O$ ($T_c = 90$ K) (23) and ferrimagnetic $CsMn^{II}[Cr^{III}(CN)_6] \cdot H_2O$ ($T_c = 90$ K) (24) as well as thin films ($\leq 2\mu m$), both oxidized and reduced, of $Cr^{III}[Cr^{III}(CN)_6]_{0.93}[Cr^{II}(CN)_6]_{0.05}$ ($T_c = 260$ K) (25) that exhibit negative magnetization, Fig. 12. Verdagner and co-workers (48) report that ferrimagnetic $V_{0.42}^{II}V_{0.58}^{III}[Cr^{III}(CN)_6]_{0.86} \cdot 2.8H_2O$ magnetically ordered above room temperature ($T_c = 315$ K). Further studies of this class of materials (26) have led to an enhancement of the T_c to about $100^\circ C$ (373 K), Fig. 13 (27).

USES OF ORGANIC/POLYMERIC MAGNETS

The magnetic (Table 2) as well as chemical/physical properties of organic/polymer magnets, especially in conjunction with other physical properties, may well lead to their use in smart materials in the future (2). Applications include the next generation of electronic, magnetic, and/or photonic devices ranging from magnetic imaging to data storage and to static and low-frequency magnetic shielding and magnetic induction. Particularly promising is the applications, in static and low-frequency magnetic shielding and magnetic induction (2b), as relatively high initial permeabilities have been reported for the $V(TCNE)_x \cdot y(\text{solvent})$ materials. Combined with their low density (~ 1 g/cm³),

relatively low resistivity ($\sim 10^4$ ohm-cm), and low power loss (as low as ~ 2 erg/(cm³ cycle), these properties suggest that magnetic shielding will have future practical applications especially for devices requiring low weight. The feasibility of using $V(TCNE)_x \cdot y(CH_2Cl_2)$ in magnetic shielding applications is demonstrated in Fig. 14. Other potential applications include photoinduced magnetism based on recent studies of Prussian blue-like materials (28,29). Studies of the effects of light on the zero-field cooled and field-cooled behavior show the importance of disorder and defects in stabilizing the photoinduced magnetic phenomena (29).

ACKNOWLEDGMENT

The authors gratefully acknowledge the extensive contributions of their collaborators, students, and post-doctoral associates in the studies discussed herein. The authors also gratefully acknowledge the continued partial support by the Department of Energy Division of Materials Science (Grant Nos. DE-FG02-86ER45271.A000, DE-FG03-93ER45504, and DEFG0296ER12198) as well as the National Science Foundation (Grant No. CHE9320478).

BIBLIOGRAPHY

- (a) J.S. Miller and A.J. Epstein. *Angew. Chem. Int. Ed.* **33**: 385 (1994). J.S. Miller and A.J. Epstein. *Chem. Eng. News* **73**(40): 30 (1995). (b) M. Kinoshita. *Jpn. J. Appl. Phys.* **33**: 5718 (1994). (c) D. Gatteschi. *Adv. Mat.* **6**: 635 (1994).
- (a) C.P. Landee, D. Melville, and J.S. Miller. In O. Kahn, D. Gatteschi, J.S. Miller, and F. Palacio, eds., *NATO ARW*

- Molecular Magnetic Materials*, E198 p. 395 (1991). (b) B.G. Morin, C. Hahn, A.J. Epstein, and J.S. Miller, *J. Appl. Phys.* **75**: 5782 (1994). (c) J.S. Miller and A.J. Epstein. *Chemtech.* **21**: 168 (1991). (d) J.S. Miller. *Adv. Mater.* **6**: 322 (1994).
3. R. Chiarelli, A. Rassat, Y. Dromzee, Y. Jeannin, M.A. Novak, and J.L. Tholence. *Phys. Script.* **T49**: 706 (1993).
 4. J.S. Miller. *Adv. Mater.* **10**: 1553 (1998).
 5. (a) J.S. Miller, J.C. Calabrese, A.J. Epstein, R.W. Bigelow, J.H. Zhang, and W.M. Reiff. *J. Chem. Soc. Chem. Commun.* 1026 (1986). (b) J.S. Miller, J.C. Calabrese, H. Rommelmann, S.R. Chittipeddi, J.H. Zhang, W.M. Reiff, and A.J. Epstein. *J. Am. Chem. Soc.* **109**: 769 (1987).
 6. S. Chittipeddi, K.R. Cromack, J.S. Miller, and A.J. Epstein. *Phys. Rev. Lett.* **58**: 2695 (1987).
 7. G.T. Yee, J.M. Manriquez, D.A. Dixon, R.S. McLean, D.M. Groski, R.B. Flippen, K.S. Narayan, A.J. Epstein, and J.S. Miller. *Adv. Mater.* **3**: 309 (1991).
 8. J.S. Miller, R.S. McLean, C. Vazquez, J.C. Calabrese, F. Zuo, and A.J. Epstein. *J. Mater. Chem.* **3**: 215 (1993).
 9. W.E. Broderick, D.M. Eichorn, X. Lu, P.J. Toscano, S.M. Owens, and B.M. Hoffman. *J. Am. Chem. Soc.* **117**: 3641 (1995).
 10. W.E. Broderick, J.A. Thompson, E.P. Day, and B.M. Hoffman. *Science* **249**: 410 (1990).
 11. W.E. Broderick and B.M. Hoffman. *J. Am. Chem. Soc.* **113**: 6334 (1991).
 12. K.S. Narayan, K.M. Kai, A.J. Epstein, and J.S. Miller. *J. Appl. Phys.* **69**: 5953 (1991). K.S. Narayan, B.G. Morin, J.S. Miller, and A.J. Epstein. *Phys. Rev.* **B 46**: 6195 (1992).
 13. K. Inoue, T. Hayamizu, and H. Iwamura. *Mol. Cryst. Liq. Cryst.* **273**: 67 (1995). A. Izoka, S. Murata, T. Sugawara, and H. Iwamura. *J. Am. Chem. Soc.* **107**: 1786 (1985). *J. Am. Chem. Soc.* **109**: 2631 (1987).
 14. J.S. Miller and A.J. Epstein. *J. Chem. Soc., Chem. Commun.* 1319 (1998). E.J. Brandon, A.M. Arif, J.S. Miller, K.-I. Sugiura, and B.M. Burkhart. *Cryst. Eng.* **1**: 97 (1998).
 15. J.S. Miller, J.C. Calabrese, R.S. McLean, and A.J. Epstein. *Adv. Mater.* **4**: 498 (1992).
 16. (a) J.M. Manriquez, G.T. Yee, R.S. McLean, A.J. Epstein, and J.S. Miller. *Science* **252**: 1415 (1991). J.S. Miller, G.T. Yee, J.M. Manriquez, and A.J. Epstein. In *Proc. Nobel Symp. Conjugated Polymers and Related Materials: The Interconnection of Chemical and Electronic Structure*, Oxford University Press, Oxford, 1993, p. 461. *La Chim. La Ind.* **74**: 845 (1992). A.J. Epstein and J.S. Miller. In *Proc. Nobel Symp. Conjugated Polymers and Related Materials: The Interconnection of Chemical and Electronic Structure*, Oxford University Press, Oxford, 1993, p. 475. *La Chim. La Ind.* **75**: 185 (1993). (b) J. Zhang, P. Zhou, W.B. Brinckerhoff, A.J. Epstein, C. Vazquez, R.S. McLean, and J.S. Miller. *A.C.S. Sym. Ser.* **644**: 311 (1996).
 17. (a) J.S. Miller, J.C. Calabrese, R.S. McLean, and A.J. Epstein. *Adv. Mater.* **4**: 498 (1992). (b) W.B. Brinckerhoff, B.G. Morin, E.J. Brandon, J.S. Miller, and A.J. Epstein. *J. Appl. Phys.* **79**: 6147 (1996). (c) C.M. Wynn, M.A. Girtu, W.B. Brinckerhoff, K.-I. Sugiura, J.S. Miller, and A.J. Epstein. *Chem. Mater.* **9**: 2156 (1997). (d) C.M. Wynn, M.A. Girtu, J.S. Miller, and A.J. Epstein. *Phys. Rev.* **B 56**: 315 (1997). (e) D.K. Rittenberg, K.-I. Sugiura, Y. Sakata, S. Mikami, A.J. Epstein, and J.S. Miller. *Adv. Mater.* **12**: 126 (2000).
 18. K.-I. Pokhodnya, A.J. Epstein, and J.S. Miller. *Adv. Mater.* **12**: 410 (2000).
 19. (a) P. Zhou, B.G. Morin, J.S. Miller, and A.J. Epstein. *Phys. Rev.* **B48**: 1325 (1993). (b) P. Zhou, S.M. Long, J.S. Miller, and A.J. Epstein, *Phys. Lett.* **A 181**: 71 (1993). (c) W.B. Brinckerhoff, J. Zhang, J.S. Miller, and A.J. Epstein. *Mol. Cryst. Liq. Cryst.* **272**: 195 (1995).
 20. J. Zhang, J. Enslin, V. Ksenofontov, P. Gülich, A.J. Epstein, and J.S. Miller. *Angew. Chem. Int. Ed.* **37**: 657 (1998).
 21. C.M. Wynn, M.A. Girtu, J. Zhang, J.S. Miller, and A.J. Epstein. *Phys. Rev.* **B 58**: 8508 (1998).
 22. M.A. Girtu, C.M. Wynn, J. Zhang, J.S. Miller, and A.J. Epstein. *Phys. Rev.* **B 61**: 492 (2000).
 23. V. Gadet, T. Mallah, I. Castro, and M. Verdaguer. *J. Am. Chem. Soc.* **114**: 9213 (1992).
 24. W.D. Greibler and D. Babel. *Z. Naturforsch.* **87b**: 832 (1982).
 25. W.E. Buschmann, S.C. Paulson, C.M. Wynn, M. Girtu, A.J. Epstein, H.S. White, and J.S. Miller. *Adv. Mater.* **9**: 645 (1997). *Chem. Mater.* **10**: 1386 (1998).
 26. S. Ferlay, T. Mallah, R. Ouahes, P. Veillet, and M. Verdaguer. *Nature* **378**: 701 (1995).
 27. (a) O. Hatlevik, W.E. Buschmann, J. Zhang, J.L. Manson, and J.S. Miller. *Adv. Mater.* **11**: 914 (1999). (b) Holmes and S.D. Girolami. *G. J. Am. Chem. Soc.* **121**: 5593 (1999).
 28. O. Sato, T. Iyoda, A. Fujishima, and K. Hashimoto. *Science* **272**: 704 (1996).
 29. D.A. Pejakovic, J.L. Manson, J.S. Miller, and A.J. Epstein. *J. Appl. Phys.* **87**: 6028 (2000).

MAGNETORHEOLOGICAL FLUIDS

ALISA J. MILLAR HENRIE

Brigham Young University
Provo, UT

J. DAVID CARLSON

Lord Corporation
Cary, NC

INTRODUCTION

Magnetorheological (MR) materials are a class of materials whose rheological properties may be rapidly varied by applying a magnetic field. This change is in proportion to the magnitude of the magnetic field applied and is immediately reversible. MR material behavior is often modeled by the Bingham plastic model. Advances in the application of MR materials are parallel to the development of advanced MR materials. These applications include brakes, dampers, and shock absorbers.

DEFINITION

Magnetorheological materials are a class of material whose rheological properties may be rapidly varied by applying a magnetic field. Most commonly, these materials are fluids that consist of micron-sized, magnetically polarizable ferrous particles suspended in a carrier liquid. When exposed to a magnetic field, the suspended particles polarize and interact to form a structure aligned with the magnetic field that resists shear deformation or flow. This change in the material appears as a dramatic increase in apparent viscosity, or the fluid develops the characteristics of a semisolid state. The magnitude of this change is controlled

by the strength of the magnetic field and is immediately reversed upon removing the magnetic field. In addition to fluids, magnetorheological materials also include magnetic field responsive gels, foams, powders, greases, and elastomers.

HISTORY

Jacob Rabinow made advances in early magnetorheological research while at the National Bureau of Standards in the 1940s and early 1950s (1,2). Rabinow's early work on MR fluids eventually led to a host of devices and products based on dry magnetic powders, for example, the magnetic, powder brake. Interest in MR fluids languished until the early 1990s when advancements in MR material composition and the advent of readily available digital control electronics fostered renewed interest. Magnetorheological materials share much in common with their counterpart, electrorheological (ER) materials. These are materials that change their behavior in an electric field. This phenomenon was first documented by Duff in the 1800s, (3), but its potential was not realized until Willis Winslow began working with electrorheological fluids in 1938 (4). Winslow also contributed to early research on MR materials and combined electromagnetorheological materials.

Research in magnetorheology has gone forward in three main areas; developing advanced materials, characterizing the mechanism that causes the MR effect, and applications of MR materials. Advancement in these areas is necessary for the development of mechanical systems that will optimize MR materials. Descriptions of MR compositions, basic models of the MR phenomenon, characteristics of common MR materials, and popular MR applications follow.

COMMON NOMENCLATURE

Some terms commonly used with magnetorheological materials include

Rheology—the study of the deformation and flow of matter

H = Magnetic field strength (a/m) (1 a/m = $4\pi \times 10^{-3}$ oersted)

B = Magnetic flux density (tesla) (1 tesla = 10^4 gauss)

$\dot{\gamma}$ = Shear rate (s^{-1})

τ = Shear stress (Pa)

τ_y = Yield stress (Pa)

η_a = Apparent viscosity (Pa · s) = $\tau / \dot{\gamma}$

η_p = Plastic viscosity (Pa · s) = $d\tau / d\dot{\gamma}$

MATERIAL CHOICE

Common Materials

Magnetorheological materials usually consist of micron-sized (3–8 μm) magnetizable particles suspended in a

liquid. A typical MR fluid consists of 20–40% by volume of relatively pure iron particles suspended in a carrier liquid such as mineral oil, synthetic oil, water, or glycol. A variety of proprietary additives similar to those found in commercial lubricants are used to discourage gravitational settling and promote particle suspension, enhance lubricity, change initial viscosity, and inhibit wear.

Some materials exhibit both MR and ER behavior and react in both magnetic and electric fields. In fact, almost all MR fluids exhibit some ER response due to moisture and mobile ions in the native oxide surface layer on virtually all ferrous particles. For more information on commercially manufactured MR fluids, see (5).

BASIC COMPOSITION

A simple MR fluid may be constructed by mixing fine iron powder (3–8 μm) into mineral oil at approximately one part filings to three parts oil by volume. Stir the mixture with a plastic or wooden implement until a consistent mixture is obtained. Place the mixture in a plastic or glass container (a 50–100 ml beaker works well). A horseshoe magnet or two bar magnets can be used to observe the MR effect. Stir the mixture to observe its viscosity and the way it feels. Now, place the container between the two poles of the magnet, and stir it again. The mixture should feel stiff. If you look closely, you can see the particle structure form between the two poles of the magnet (6).

THEORY

Currently, there is no universally accepted theory for the causes of the MR effect. However, it is generally agreed that the change in properties is due to the realignment of particles in the fluid to form fibrils, or long strands of suspended particles, that resist shear (Fig. 1). It is believed that the observed alignment of particles is related to the displacements and torque produced in the medium by the field and the translational motion and relocation of particles to positions that have local minimum potential energy. This results in an increase in viscosity and an increase in the shear strength of the material. A Bingham plastic model is often considered sufficient to model MR devices:

$$\tau_{\text{total}} = \tau_y(H) + \eta_p \dot{\gamma}, \quad (1)$$

where τ_{total} is the total shear stress of the material, τ_y is the yield stress as a function of the magnetic field, H is the magnetic field strength, η_p is the plastic viscosity or post yield viscosity, and $\dot{\gamma}$ is the shear strain rate in the fluid. Apparent viscosity η_a is defined as the total shear stress divided by the shear rate:

$$\eta_a = \frac{\tau_{\text{total}}}{\dot{\gamma}}. \quad (2)$$

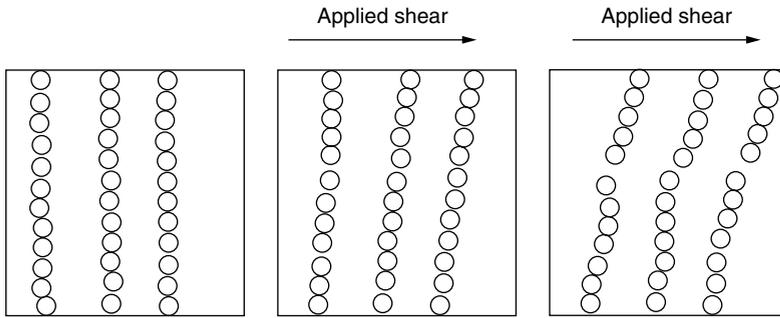


Figure 1. Suspended magnetizable particles align in a magnetic field. The MR fluid will then resist a certain amount of shear before the chains begin to break. This yield point increases as magnetic field strength increases until the magnetic saturation point is reached. This results in the solid feel of MR materials.

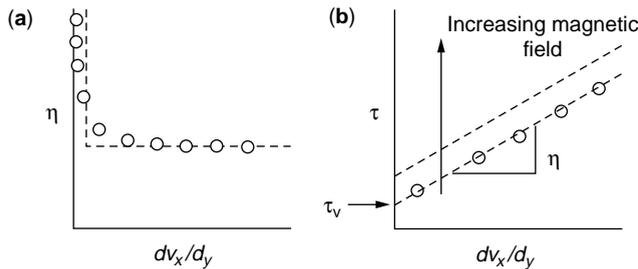


Figure 2. Bingham plastic behavior. Graph (a) shows that the apparent viscosity η is very large until a certain shear rate, dv_x/d_y , is obtained. This shows that the fluid acts as a solid and does not flow until a certain threshold stress, the yield stress, is exceeded. Graph (b) shows that yield stress τ is a function of shear rate and that Bingham plastics experience shear thinning. The yield stress of the material τ_y is the y intercept of the linear regression curve fit of shear stress vs. shear rate. Yield stress τ_y increases as magnetic field increases.

The apparent viscosity of a Bingham plastic material increases dramatically at a low shear rate. At a very high shear rate, the magnitude of the apparent viscosity approaches that of the plastic viscosity, as shown in Fig. 2. Although plastic viscosity is sometimes observed to be a weak function of magnetic field strength, it is often assumed that it is constant for constant magnetic field strength (typically 0.10 to 0.70 Pa · s) (7).

The Bingham plastic model indicates that the material property that changes as magnetic field strength increases is the yield stress τ_y , where flow of the material is initiated. Ideally, below a certain threshold of stress, a Bingham plastic does not flow at all. Beyond this threshold, the flow rate increases in proportion to the applied stress minus the threshold stress (Fig. 2a). Plastic viscosity η_p is the slope of a linear regression curve fit to a measured stress versus shear strain data set, and the yield stress τ_y of an MR fluid is the y intercept of that same linear regression curve (Fig. 2b).

MR PROPERTIES

Magnetorheological materials are useful because the change in their material properties is so large (Fig. 3). Typical magnetorheological materials can achieve yield

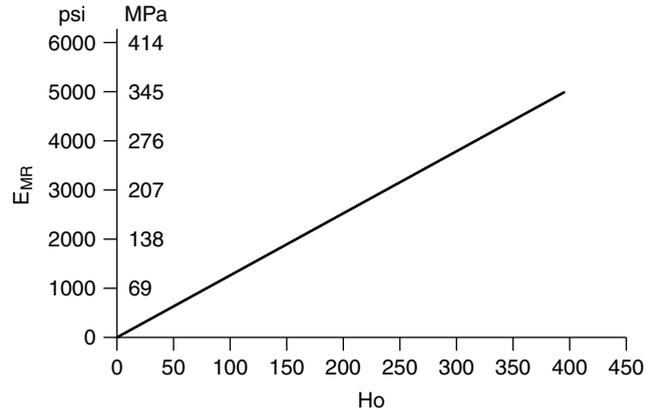


Figure 3. One example of a change in the effective modulus of elasticity (E_{mr} vs. magnetic field strength (H_0)) for a specific MR material.

strengths up to 50–100 kPa at magnetic field strengths of about 150–250 kA/m. For low strains prior to yield, the shear modulus of a MR fluid also shows a very large increase in an applied magnetic field. MR materials eventually reach a saturation point where increases of magnetic field strength do not increase the yield strength of the MR material. This typically occurs around 300 kA/m. MR materials have been developed that are stable in temperature ranges from -40 to 150°C . There are slight changes in the volume fraction and hence slight reductions in the yield strength at these temperatures, but they are small. The MR effect is immediately reversible if the magnetic field is reduced or removed. Response times of 6.5 ms have been recorded (8).

Magnetorheological materials exhibit some advantages over typical ER materials. Unlike ER materials, they are not sensitive to impurities, and thus MR materials are candidates for use in dirty or contaminated environments. They are also unaffected by the surface chemistry of surfactants as ER materials are. The power (1–2 amps or 50 watts) and voltage (12–24V) requirements for MR material activation are relatively small compared with ER materials (8).

Other MR properties that have also been explored include the following. Ginder and Davis (7) investigated the effect of magnetic saturation on the strength of MR

materials by using finite element analysis. It was found that wall roughness on contact with the fluid is important for yield strengths, especially in low magnetic fields (9,10). Size and size distribution of the suspended particles also affect the change in properties of the MR fluid when placed in a magnetic field (11,12). The combined effect of magnetic and electric fields on magnetoelectrorheological materials has also been explored (13). Tang and Conrad conducted a series of systematic experiments to characterize the rheology of MR fluids from technological and scientific viewpoints (14).

APPLICATIONS

Because the state of MR materials can be controlled by the strength of an applied magnetic field, it is useful in applications where variable performance is desired. Many smart systems and structures would benefit from the change in viscosity or other material properties of MR materials. Beginning with the commercialization of MR fluid rotary brakes for use in aerobic exercise equipment in 1995, application of magnetorheological material technology in real-world systems has grown steadily. The past few years have witnessed a blossoming commercialization of MR fluid technology. MR fluid technology has been embraced by a number of manufacturers for inclusion in a diverse spectrum of products that are now commercially available (15):

- linear MR dampers for real-time active vibrational control systems in heavy duty trucks
- linear and rotary brakes for low-cost, accurate, positional and velocity control of pneumatic actuator systems
- rotary brakes to provide tactile force-feedback in steer-by-wire systems
- linear dampers for real-time gait control in advanced prosthetic devices
- adjustable shock absorbers for NASCAR oval and dirt track racing

Other applications that are near commercialization include

- low-cost MR sponge dampers for washing machines (16,17)
- very large MR fluid dampers for seismic damage mitigation in civil engineering structures (17)
- large MR fluid dampers to control wind-induced vibrations in cable-stayed bridges (18)
- real-time controlled shock absorbers and struts for domestic automobiles

MR materials are also being considered for use in compliant mechanism design to change the compliance of flexible members. This behavior can be modified as a function of time resulting in a variably compliant mechanism (19). More applications are being considered and it is likely that,

as MR materials continue to improve, many more applications will be forthcoming.

BIBLIOGRAPHY

1. J. Rabinow, *AIEE Trans.* **67**: 1308–1315 (1948).
2. Magnetic Fluid Torque and Force Transmitting Device. U.S. Pat. 2,575,360, 1951, J. Rabinow.
3. A. Duff, *Phys. Rev.* **4**: 23 (1986).
4. W.M. Winslow, *J. Appl. Phys.* **20**: 1137–1140 (1949).
5. <http://www.mrfluid.com>, 2001.
6. D.J. Klingenberg, *Sci. Am.* pp. 112–113 (Oct. 1993).
7. J.M. Ginder and L.C. Davis, *Appl. Phys. Lett.* **65**: 3410–3412 (1994).
8. K.D. Weiss, T.G. Duclos, J.D. Carlson, M.J. Chrzan, and A.J. Margida, 1993 Int. Off-Highway & Powerplant Congr. Exposition, SAE Technical Paper Series, #932451.
9. T. Miyamoto and M. Ota, *Appl. Phys. Lett.* **64**: 1165–1167 (1994).
10. E. Lemarie, and G. Bossis, *J. Phys. D* **24**: 1473–1477 (1991).
11. E. Lemaire, A. Mennier, G. Bossis, J. Liu, D. Felt, P. Bashtovioi., and N. Matousseritch, *J. Rheol.* **39**: 1011–1020 (1995).
12. Method and Magnetorheological Fluid Formulations for Increasing the Output of a Magnetorheological Fluid, US Pat. 6,027,664, (2000), and US Pat. 5,900,184, (1999), K.D. Weiss, J.D. Carlson, and D.A. Nixon.
13. K. Koyama, *Proc. 5th Int. Conf. Electro-Rheological, Magneto-Rheological Suspensions Associated Technol.*, Sheffield, U.K., July 1995.
14. X. Tang and H. Conrad, *J. Rheol.* **40**(6): 1167–1178 (1996).
15. J.D. Carlson and J.L. Sproston, *Proc. 7th Int. Conf. New Actuators*, Messe Bremen, Bremen, (2000), pp. 126–130.
16. J.D. Carlson, *Mach. Design*, pp. 73–76, Feb. 22, (2001).
17. J.D. Carlson, *Motion Control*, pp. 25–28, March 2001.
18. B.F. Spencer, Jr., G. Yang, J.D. Carlson, and M.K. Sain, *Proc. 2nd World Conf. Struct. Control*, Kyoto, Japan, 1998.
19. A.J.M. Henrie and L.L. Howell, *Proc. 43rd Int. SAMPE Symp./Exhibition*, Irvine, CA, 1998.

MAGNETOSTRICTIVE MATERIALS

MARCELO J. DAPINO
Ohio State University
Columbus, OH

INTRODUCTION

Magnetostrictive materials are a class of smart materials that can convert energy between the magnetic and elastic states. For this reason, magnetostrictive materials and devices based on these materials are often referred to as *transducers*. Due to the bidirectional nature of this energy exchange, magnetostrictive materials can be employed for both actuation and sensing. Alloys based on the transition metals iron, nickel, and cobalt in combination with certain rare-earth elements are currently employed in actuator and sensor systems in a broad range of industrial,

biomedical, and defense applications. Because magnetostriction is an inherent property of ferromagnetic materials, it does not degrade over time as do some poled piezoelectric substances. In addition, newer magnetostrictive materials provide strains, forces, energy densities, and coupling coefficients that compete favorably with more established technologies such as those based on piezoelectricity. As evidenced by the ever-increasing number of patented magnetostrictive systems, transducer designers are finding new opportunities to employ magnetostrictive materials in a wide variety of applications ranging from stand-alone transducers to complex smart structure systems.

A number of design and modeling issues, however, complicate the implementation of magnetostrictive materials in certain applications in which other smart material technologies are currently favored. For instance, due to the required solenoid and related magnetic circuit components, magnetostrictive transducers are usually larger and bulkier than their piezoelectric or electrostrictive counterparts. Hence, they are employed primarily in applications that require high strains and forces but where weight is not an issue. One additional consideration is that the most technologically advanced magnetostrictive materials are costly to manufacture. Advanced crystalline materials are manufactured by employing sophisticated crystal growth techniques that produce directional solidification along the drive axis of the transducer material. The manufacturing process also includes precision machining of laminations, final diameters, and parallel ends of cut-to-length drivers, as well as thorough quality assurance and performance evaluation throughout the process. These technological and cost-related problems have been mitigated to some extent through the advent of new manufacturing techniques that have enabled more capable magnetostrictive materials in various forms, including amorphous or crystalline thin films, magnetostrictive particle-aligned polymer matrix composite structures, and sintered powder compacts suitable for mass production of small irregular shapes. From the perspective of modeling and control, magnetostrictive materials exhibit nonlinear effects and hysteretic phenomena to a degree which other smart materials, for instance electrostrictive compounds, do not. These effects are particularly exacerbated at the moderate to high drive level regimes in which magnetostrictive materials are most attractive. These issues have been addressed through recent modeling techniques, and as new applications are developed, model accuracy and completeness will almost surely follow.

The term magnetostriction is a synonym for magnetically induced strain, and it refers to the change in physical dimensions exhibited by most magnetic materials when their magnetization changes. Magnetization, defined as the volume density of atomic magnetic moments, changes as a result of the reorientation of magnetic moments in a material. This reorientation can be brought about by applying either magnetic fields, heat or stresses. The linear magnetostriction $\Delta L/L$ that results from applying a longitudinal magnetic field on a sample of length L , illustrated in Fig. 1, is the most commonly employed attribute of the magnetostrictive principle in actuator applications.

Though most ferromagnetic materials exhibit linear magnetostriction, only a small number of compounds that contain rare-earth elements provide "giant" magnetostrictions in excess of 1000×10^{-6} . These large magnetostrictions are a direct consequence of the strong magnetomechanical coupling that arises from the dependence of magnetic moment orientation on interatomic spacing. When a magnetic field is applied to a magnetostrictive material, the magnetic moments rotate to the direction of the field and produce deformations in the crystal lattice and strains in the bulk material. Referring again to Fig. 1 which pertains to a material that has positive magnetostriction, note

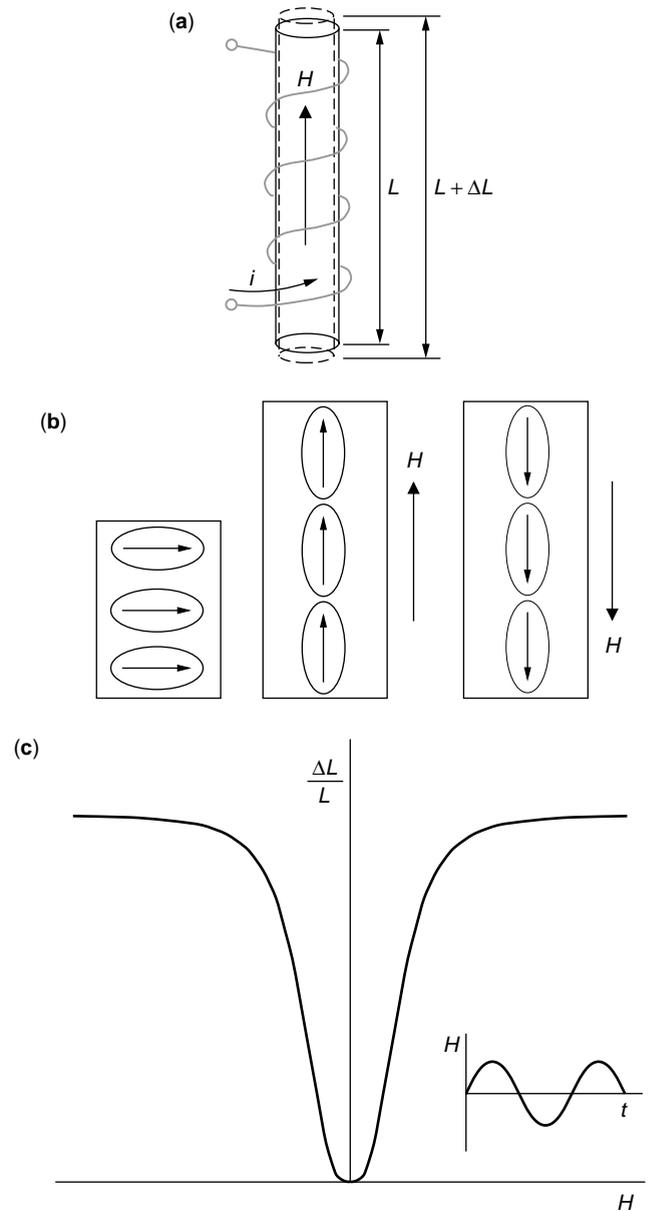


Figure 1. Joule magnetostriction produced by a magnetic field H . (a) H is approximately proportional to the current i that passes through the solenoid when a voltage is applied to it. (b) The rotation of magnetic dipoles changes the length of the sample, and (c) curve $\Delta L/L$ vs. H obtained by varying the field sinusoidally.

that the sample elongates irrespective of the direction of rotation of the magnetic moments. A symmetrical magnetostriction curve is obtained when the field is cycled, as depicted in Fig. 1b,c. If stress is applied instead, the material deformations lead to magnetic moment reorientation and a subsequent change in the magnetization M . This magnetization change can be detected through the voltage induced in a sensing solenoid wrapped around the sample, which provides a mechanism for sensing. In applications, either one sensing solenoid or a plurality of them arranged along the length of the driver can be placed inside of the driving coil. Alternatively, it is possible to employ only one solenoid to drive the magnetostrictive material and sense its magnetization changes. This configuration exhibits the major disadvantage that additional signal processing hardware is required to extract the sensing voltage from the driving signal.

This article provides an overview of magnetostrictive materials in the next section, followed by a description of the physical origin of magnetostriction and a discussion of material behavior. The subsequent section is devoted to linear magnetostriction, and other magnetostrictive effects are discussed in the following section. Finally, a discussion of current transducer designs and modeling techniques is presented.

MATERIALS OVERVIEW

The study of magnetostriction began in 1842 when James P. Joule first observed that a sample of iron changes in length when magnetized by a magnetic field. Subsequent work using other materials such as nickel, cobalt, and their alloys led to numerous applications, including telephone receivers, hydrophones, scanning sonar, fog horns, oscillators, and torque sensors. During World War II, sonar transducer were driven primarily by nickel, which exhibits saturation magnetostrictions of about 50×10^{-6} . A breakthrough in magnetostrictive materials occurred in 1963 when the largest-known magnetostrictions in the rare-earth elements terbium and dysprosium were discovered. The strains in these elements are of the order of $10,000 \times 10^{-6}$, or three orders of magnitude larger than those of nickel, but they are achieved exclusively at cryogenic temperatures. The temperature limitation and the fact that the field of piezoelectricity was gaining technical maturity hindered the development of magnetostrictive materials and led, in the early 1970s, to a search for a new class of transducer materials capable of high room temperature strains. Highly magnetostrictive rare earths (R), principally samarium (Sm), terbium (Tb), and dysprosium (Dy), were combined with the magnetic transition metals nickel, cobalt, and iron by direct compound synthesis and by rapid sputtering into amorphous alloys. In contrast to the normal Curie temperature behavior of the R-Ni and R-Co compounds, R-Fe compounds exhibit an *increase* in the Curie temperature as rare-earth concentration increases. Consequently, huge room temperature magnetostrictions up to $3,000 \times 10^{-6}$ were achieved, particularly in the TbFe₂ compound. However, because the

magnetostriction originates in the strain dependence of magnetic anisotropy, the large magnetostriction in these compounds is obtained at the expense of large anisotropies. This poses a technological limitation since impractically large fields of more than 2 MA/m are needed to bring these compounds to technical saturation.

Partial substitution of dysprosium for terbium in the TbFe₂ system resulted in improved magnetostriction and anisotropy properties. The resulting pseudobinary compound Tb_{0.3}Dy_{0.7}Fe_{1.9-1.95} has been available commercially since the 1980s under the name Terfenol-D (Ter = terbium, Fe = iron, N = Naval, O = Ordnance, L = Laboratory, D = dysprosium). The highest room temperature magnetostriction for Terfenol-D is $1,600 \times 10^{-6}$ at a moderate saturation field of 0.16 MA/m, but even larger magnetostrictions up to $3,600 \times 10^{-6}$ are possible when the material is employed in transducers driven at resonance. The utility of Terfenol-D as a rugged, high-power transducer driver has been increasingly recognized in recent years. At present, Terfenol-D is used in active noise and vibration control systems; low-frequency underwater communications (sonar); linear and rotational motors; ultrasonic cleaning; machining and welding; micropositioning; and detecting motion, force, and magnetic fields. Terfenol-D is currently available in a variety of forms, including monolithic rods (1,2), particle-aligned polymer matrix composites (3-5), and thin films (6,7). Because of the large magnetostrictive anisotropy and strong magnetoelasticity, Terfenol-D and other pseudobinary rare-earth-iron compounds can be synthesized to exhibit a broad range of properties (8).

A second new magnetostrictive material based on amorphous metal was introduced in 1978, produced by rapid cooling of iron, nickel, and cobalt alloys together with one or more of the elements silicon, boron, and phosphorus. These alloys are known commercially as Metglas (metallic glass) and are commonly produced in thin-ribbon geometries. Because of the extremely high coupling coefficients ($k > 0.92$), Metglas is a prime candidate for sensing applications in which a mechanical motion is converted into an electrical current or voltage (2).

The latest materials science research on magnetostrictive materials includes the development of new compounds to minimize magnetic anisotropy and hysteresis and new manufacturing techniques to produce Terfenol-D thin films efficiently (9). Substantial advances have been achieved in the quaternary compounds Terfenol-DH, which are produced by substituting holmium for terbium and dysprosium (10). In addition, new manufacturing techniques are enabling the production of multilayered driver rods which will lead to reduced dynamic losses, thus facilitating operation over a broad frequency spectrum into the megahertz range. Ferromagnetic shape-memory alloys are another class of smart materials which hold much promise due to the large strains that they can provide. The nickel-titanium alloy commercially known as Nitinol features large recoverable strains of the order of $60,000 \times 10^{-6}$, but it suffers from inferior dynamic response. The possibility of combining the desirable aspects of shape memory with magnetostriction through actuating an SMA in a magnetic

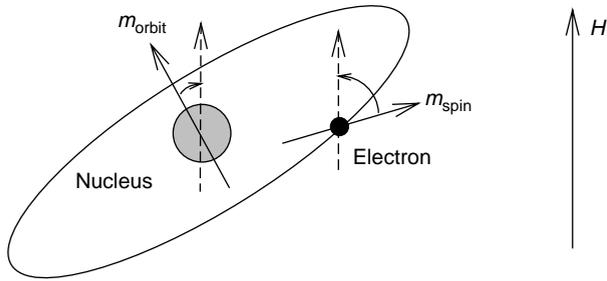


Figure 2. In highly magnetostrictive materials, the spin moment m_{spin} and orbital moment m_{orbit} are strongly coupled. When the spin moment rotates to align with the external field H , the orbital moment rotates along with it and produces considerable lattice deformation.

field is currently being investigated. Promising candidate materials are the Ni_2MnGa system and the Fe-based Invars, which exhibit, in principle, the desired characteristics. Further details on the Ni–Mn–Ga alloys can be found in (11,12).

PHYSICAL ORIGIN OF MAGNETOSTRICTION

Magnetic coupling within atoms can be of two forms, spin–spin and spin–orbit interactions. In ferromagnetic materials, the spin–spin coupling that keeps neighboring spins parallel or antiparallel to one another within domains can be very strong. However, this exchange energy is isotropic because it depends only on the angle between adjacent spins, not on the direction of the spins relative to the crystal lattice.

Magnetostriction is due mainly to spin–orbit coupling, which refers to a kind of interaction between the spin and orbital motion of each electron. This type of coupling is also responsible for crystal anisotropy. Referring to Fig. 2, when a magnetic field is applied and an electron spin tries to align with it, the orbit of that electron also tends to be reoriented. But because the orbit is strongly coupled to the crystal lattice, the orbit resists the rotation of the spin axis. Thus, the energy required to rotate the spin system of a domain away from the preferred orientations is the energy required to overcome spin–orbit coupling. Spin–orbit coupling is weak in most ferromagnetic materials, as evidenced by the fact that a moderate field of a few thousand kiloamperes per meter suffices to rotate the spins. Spin–orbit coupling in rare-earth metals is much stronger by about an order of magnitude. When a magnetic field rotates the spins, the orbital moments rotate, and considerable distortion, and hence magnetostriction, results (13).

MATERIAL BEHAVIOR

Magnetic Anisotropy

Magnetic anisotropy refers to the dependence of magnetic properties on the direction in which they are measured. It can be of several kinds, including crystal, stress, shape,

and exchange anisotropy. Of these, however, only crystal anisotropy is intrinsic to the material, whereas the other types are externally induced.

In crystalline materials, the magnetic moments do not rotate freely in response to applied fields, but rather they tend to point in preferred crystallographic directions. This phenomenon is called magnetocrystalline (or crystal) anisotropy, and the associated anisotropy energy is that required to rotate the magnetic moments away from their preferred direction. Crystal anisotropy energy and linear magnetostriction are closely related. If anisotropy is independent of the state of strain, there will be no linear magnetostriction (14). In rare-earth elements, for instance, large strains are a direct consequence of the huge strain dependence of magnetic anisotropy (8). Under the action of a magnetic field, measurable strains result from the deformations that the crystal lattice undergoes to minimize the energy state of the material.

When a sinusoidal magnetic field is applied to a material that has sufficiently large anisotropy, the resulting magnetization curve is not smooth due to the presence of magnetic moment “jumping.” For example, Fig. 3a,b shows the magnetization and magnetostriction of the alloy $\text{Tb}_{0.67}\text{Dy}_{0.33}$, which has substantial magnetic anisotropy. The discontinuity in both curves near a field value of 40 kA/m occurs because the magnetic moments abruptly enter or leave low energy directions. Elements that have

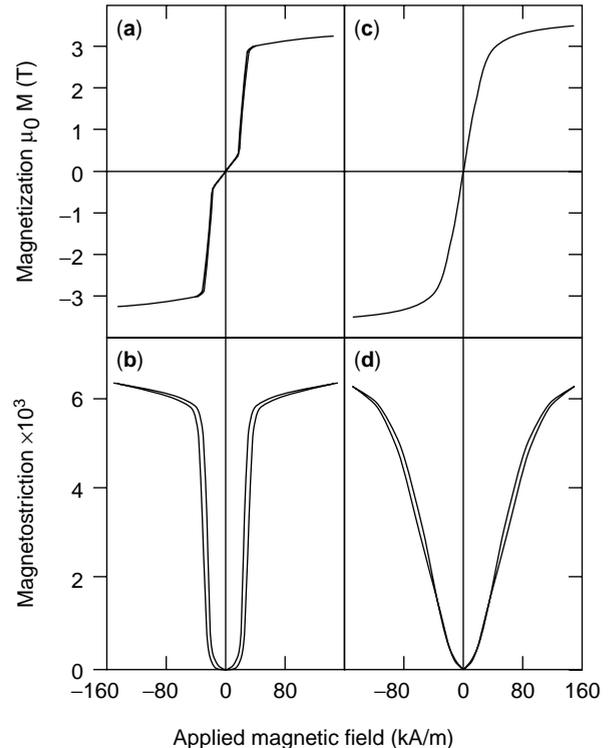


Figure 3. (a,b) Magnetization and magnetostriction jumps in $\text{Tb}_{0.67}\text{Dy}_{0.33}$, a material that has large anisotropy. (c), (d) The same measurements are much smoother in a material that has near zero anisotropy such as $\text{Tb}_{0.6}\text{Dy}_{0.4}$ (36).

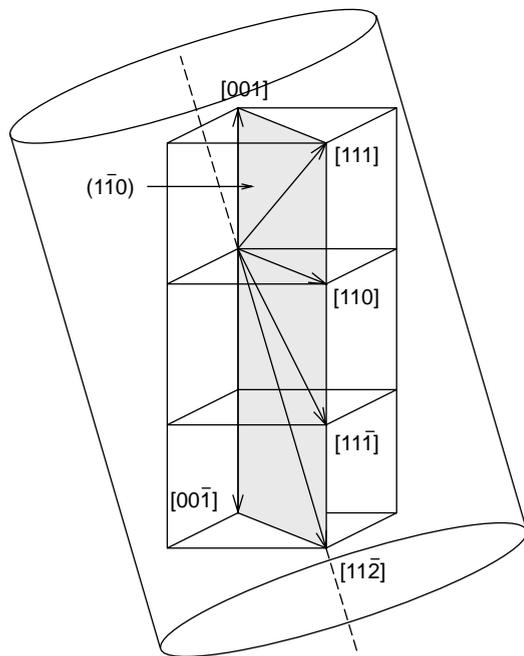


Figure 4. Crystallographic orientations in monolithic Terfenol-D. The square brackets represent the indexes of particular directions such as the edges of a cube: $[100]$, $[010]$, $[001]$, $[\bar{1}00]$, $[0\bar{1}0]$, and $[00\bar{1}]$, in which $\bar{1}$ denotes -1 . The entire set of directions is designated by any one direction in angular brackets, for instance, $\langle 100 \rangle$. Finally, planes of a form are designated by rounded brackets, such as the six faces of a cube: (100) , (010) , (001) , $(\bar{1}00)$, $(0\bar{1}0)$, and $(00\bar{1})$.

opposite anisotropies are often alloyed together to reduce the jumping and associated nonlinearities in material behavior. Such a material is shown in Fig. 3c,d. In this case, the same system but in a different composition ($\text{Tb}_{0.6}\text{Dy}_{0.4}$) exhibits a much smoother response to the applied field.

Accurate models for crystal anisotropy and its relationship to the magnetization process exist for simple cases of cubic and hexagonal crystals (15,16), but models of complex crystal structures often rely on simplifying assumptions that reduce the problem to the simpler cases. For instance, it is often useful to assume operating regimes in which stress anisotropy dominates crystal anisotropy, thus enabling the modeling of highly complex cases such as that of Terfenol-D whose crystals are grown in dendritic twin sheets oriented in the $[11\bar{2}]$ direction, as shown in Fig. 4 and discussed in (1). Terfenol-D has a large and positive magnetostrictive coefficient of $\lambda_{111} = +1,600 \times 10^{-6}$, so a compressive stress applied along the $(11\bar{2})$ direction produces a significant decrease in the internal energy of the crystal at *right angles* to the applied stress. On the other hand, although the anisotropy coefficient of Terfenol-D varies significantly, depending on temperature and stoichiometry ($K_1 = -4$ to -50 kJ/m^3 [3,17]), it is sufficiently large to resist such energy changes by favoring alignment in the $\langle 111 \rangle$ directions. Then, it is inferred that a sufficiently large compressive stress will raise the elastic energy above that of the crystal anisotropy, shifting the

preferred orientation of domains to the $\langle 111 \rangle$ magnetic easy axes that are perpendicular to the $[11\bar{2}]$ direction. Under such compressive stress, the population density in these two orientations increases, and a magnetic field applied in the $[11\bar{2}]$ direction produces nearly isotropic 90° rotations because the energy wells of crystal anisotropy have been effectively removed from the path of the rotations. In addition, under compression, as large populations of magnetic moments align normally to the stress direction, the demagnetized length decreases to a minimum, and the saturation magnetostrictive potential increases to a maximum. The 90° rotations subsequently provide the maximum possible magnetostrictions. To summarize, in materials that have positive magnetostrictions like Terfenol-D, the stress anisotropy generated by compression effectively improves the magnetoelastic state that leads to enhanced magnetostrictions. The manner in which this is implemented in transducer applications is discussed later. In nickel, which has a negative magnetostrictive coefficient, the effect is reversed, and enhanced magnetoelasticity is obtained from tensile stresses. Further details regarding crystal anisotropy can be found in (18–20).

Domain Processes and Hysteresis

The changes in magnetization that result from an applied magnetic field can be either reversible or irreversible. Reversible changes in magnetization are energetically conservative and occur for small field increments in which the material can return to the original magnetic state upon removing the field. Irreversible magnetizations are dissipative because external restoring forces are needed to return the magnetism to its original state, for example, when large fields are applied. In applications, both types of mechanisms contribute to the magnetization process. Magnetization, either reversible or irreversible, can be explained by considering two related mechanisms: the rotation of magnetic moments and the movement of domain walls. The presence of domain walls lies in the domain structure characteristic of ferromagnetic materials below their magnetic phase transition temperature or Curie temperature, T_c (see Table 1 for values of T_c for several magnetostrictive materials). When a ferromagnetic material is cooled below its Curie temperature, the magnetic moments become ordered across volumes, called domains, that contain large numbers of atoms. The domain structure can be observed under a microscope, and it typically consists of 10^{12} – 10^{15} atoms per domain. The transition regions between neighboring domains are called domain walls. All of the moments of each domain are aligned parallel, producing a spontaneous magnetization M_s , but without a field, the direction of M_s varies from domain to domain, so that the bulk magnetization in the material averages zero. This is illustrated by the randomly oriented regions of Fig. 5a.

When a small magnetic field H is applied, as depicted in Fig. 5b, domains oriented favorably to the field grow at the expense of the remaining domains, and the main magnetization mechanism is domain wall motion. As the field is increased (see Fig. 5c), entire domains rotate to align

Table 1. Magnetoelastic Properties of Some Magnetostrictive Materials

Material	$\frac{3}{2}\lambda_s (\times 10^{-6})$	ρ (g/cm ³)	B_s (T)	T_c (°C)	E (GPa)	k
Fe	-14 (8)	7.88 (14)	2.15 (14)	770 (14)	285 (14)	
Ni	-50 (14)	8.9 (14)	0.61 (14)	358 (14)	210 (1)	0.31 (8)
Co	-93 (14)	8.9 (14)	1.79 (14)	1120 (14)	210 (1)	
50% Co-50% Fe	87 (2)	8.25 (8)	2.45 (76)	500 (14)		0.35 (8)
50% Ni-50% Fe	19 (2)		1.60 (76)	500 (14)		
TbFe ₂	2630 (8)	9.1 (14)	1.1 (2)	423 (8)		0.35 (8)
Tb	3000 (-196°C) (36)	8.33 (14)		-48 (13)	55.7 (1)	
Dy	6000 (-196°C) (36)	8.56 (14)		-184 (1)	61.4 (1)	
Terfenol-D	1620 (8)	9.25	1.0	380 (76)	110 (77)	0.77 (78)
Tb _{0.6} Dy _{0.4}	6000 (-196°C) (36)					
Metglas 2605SC	60 (36)	7.32 (2)	1.65 (76)	370 (2)	25-200 (2)	0.92 (1)

^aUnless otherwise specified, all measurements were performed at room temperature.

with the easy $[11\bar{1}]$ axis. This produces a burst region in the magnetization versus field ($M-H$) and strain versus field ($\epsilon-H$) curves by virtue of which small field changes produce large magnetization or strain changes. In the final stage shown in Fig. 5d, the material acts as a single domain as magnetic moments rotate coherently from the easy axis to the direction of the field. This produces saturation of the magnetization. Typical magnetization and strain loops shown in Fig. 6 illustrate the burst region and saturation effects. From a design perspective, magnetic biasing described later is used to center operation in the burst region for optimum performance.

For low magnetic field levels, partial excursions in the $M-H$ or $\epsilon-H$ curve are observed that are approximately linear. However, hysteresis is always present, particularly when the materials are employed at high field levels such

as those of Fig. 6. The hysteresis can be attributed to the irreversible impediment to domain motion by pinning sites, such as when domain walls move across twin boundaries in Terfenol-D. Modeling hysteresis and nonlinear behavior is currently a focal point in designing and controlling magnetostrictive materials. Extensive details on the topic of ferromagnetic hysteresis can be found in (14,21,22).

Material Properties

Strains are generated by magnetostrictive materials when magnetic moments rotate to align with an applied field. This phenomenon is governed by an energy transduction process known as magnetomechanical coupling that is intrinsically bidirectional and that facilitates both actuating and sensing mechanisms in a material. From a design

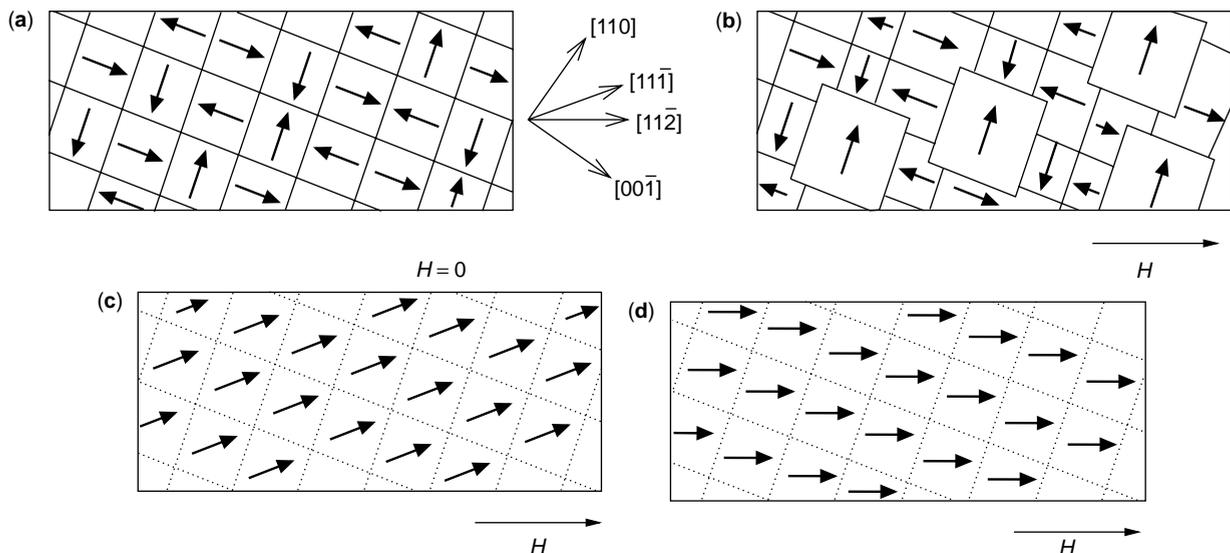


Figure 5. Domain processes in the $(1\bar{1}0)$ plane of single crystal Terfenol-D under the application of a field H along the $[11\bar{2}]$ axis: (a) demagnetized specimen, (b) partial magnetization by domain-wall movement, (c) from partial magnetization to the knee of the magnetization curve by irreversible domain magnetization rotation into the $[11\bar{1}]$ axis, and (d) from the knee of the magnetization curve to technical saturation by reversible (coherent) rotation to the $[11\bar{2}]$ axis (21).

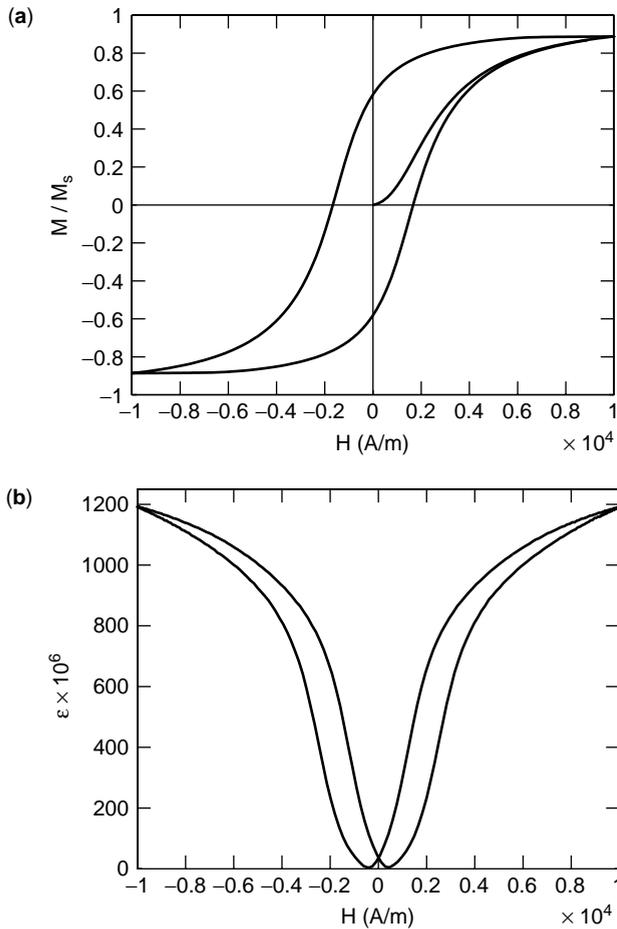


Figure 6. Relative magnetization M/M_s and total strain ε as a function of magnetic field H in a magnetostrictive material.

standpoint, the linear coupling coefficient k that quantifies the conversion efficiency between mechanical and elastic energies must be close to unity. Significant magnetic moment rotation occurs only when a domain structure is present, that is to say, in the ferromagnetic state below the Curie temperature T_c . Hence, the material must be designed so that its Curie temperature is well above the operating temperature range. In smart structure systems, large forces are often involved that the magnetostrictive material must support. The stiffness of the material is quantified by the elastic modulus E . The material must have a large value of E to support large forces. Finally, the magnetostrictive material must feature a large saturation magnetization M_s (or, equivalently, a large saturation induction B_s), and the magnetic anisotropy must be small.

Shown in Table 1 is a list of nominal properties for several magnetostrictive materials of interest. Note that the magnetomechanical coupling responsible for diverse material properties is a highly complex function that depends on quantities such as magnetic field, stress, temperature, and frequency. These quantities are collectively known as “operating conditions” and typically vary during device operation. It has been demonstrated that small variations in operating conditions often produce large changes

in material properties (23). Efficient transducer design requires accurately assessing how material properties behave under varying operating conditions.

LINEAR MAGNETOSTRICTION

Linear or Joule magnetostriction pertains to the strain produced in the field direction and is the most commonly used magnetostrictive effect. Because linear magnetostriction occurs at constant volume, there must be a transverse strain of sign opposite to that of the linear magnetostriction,

$$\lambda_{\perp} = -\frac{\lambda}{2}.$$

Isotropic Spontaneous Magnetostriction

It was mentioned earlier that when a ferromagnetic material is cooled through its Curie temperature, a transition from paramagnetism to ferromagnetism takes place, and magnetic moments become ordered giving rise to spontaneous magnetization M_s within domains. This process is also accompanied by a strain which is known as *spontaneous* magnetostriction λ_0 . It is possible to derive a useful relationship between λ_0 and *saturation* magnetostriction λ_s . To that end, we consider an isotropic material in the disordered state above T_c , which is therefore modeled by spherical volumes, as shown in Fig. 7a. As the material is cooled below T_c , spontaneous magnetization M_s is generated within magnetic domains along with the corresponding spontaneous magnetostriction λ_0 . The domains are represented in Fig. 7b by ellipsoids that have spontaneous strain e . Because the material is isotropic, the magnetic domains are oriented randomly; each bears an angle θ with respect to the direction of measurement. Net magnetization is consequently zero, and the length in the

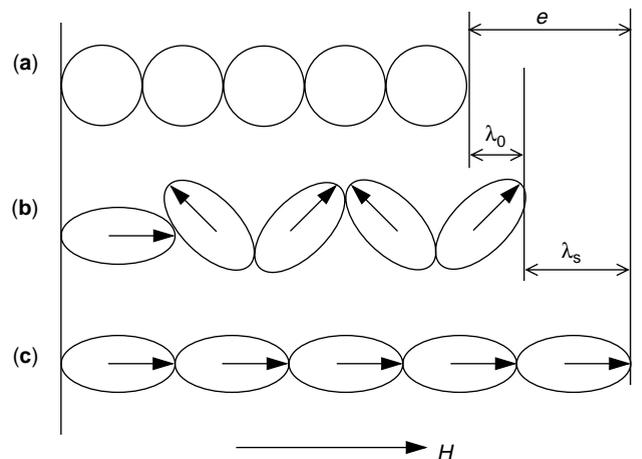


Figure 7. Schematic diagram illustrating the magnetostriction of a ferromagnetic material: (a) paramagnetic state above T_c ; (b) after it has been cooled through T_c ; and (c) after it has been brought to saturation by a field H .

direction of interest is given by (13)

$$e(\theta) = e \cos^2 \theta. \quad (1)$$

Then, the average domain deformation at the onset of spontaneous magnetostriction can be obtained by integration in all possible directions,

$$\lambda_0 = \int_{-\pi/2}^{\pi/2} e \cos^2 \theta \sin \theta d\theta = \frac{e}{3}.$$

Spontaneous magnetostriction λ_0 is homogeneous in all directions, so that the material has changed its dimensions but not its shape. When a magnetic field is applied, the domains rotate and become aligned either parallel to the field or perpendicular to it, depending on whether the material exhibits positive or negative magnetostriction. Assuming positive magnetostriction, the domains will rotate into the field direction, as depicted in Fig. 7c. Near saturation, the material will be a single domain, and the total strain will be e . Then, the total available saturation magnetostriction is given by the difference between e and λ_0 ,

$$\lambda_s = e - \lambda_0 = \frac{2}{3} e = 2 \lambda_0. \quad (2)$$

This expression provides a method of measuring the spontaneous strain λ_0 by measuring λ_s . Methods to determine λ_s are discussed next.

Saturation Magnetostriction

Assuming again for simplicity that the medium is isotropic, saturation magnetostriction at an angle θ to the direction of the field is given by (13)

$$\lambda_s(\theta) = \frac{3}{2} \lambda_s \left(\cos^2 \theta - \frac{1}{3} \right), \quad (3)$$

where $\lambda_s(\theta)$ is the saturation magnetostriction at an angle θ to the field and λ_s is the saturation magnetostriction along the direction of magnetization.

The saturation magnetostriction is then calculated from the difference between the maximum magnetostriction when the field is parallel to a given direction ($\lambda_{s\parallel}$) and that when the field is perpendicular to the given direction ($\lambda_{s\perp}$). Substituting $\theta = 0^\circ$ and $\theta = 90^\circ$ in Eq. (3) gives

$$\lambda_{s\parallel} - \lambda_{s\perp} = \lambda_s + \frac{1}{2} \lambda_s = \frac{3}{2} \lambda_s, \quad (4)$$

which defines λ_s independently of the demagnetized state.

Magnetostriction data from Clark (8) taken from polycrystalline $\text{Tb}_x\text{Dy}_{1-x}\text{Fe}_2$ samples are reproduced in Fig. 8. The data points correspond to $\lambda_{s\parallel} - \lambda_{s\perp}$ at room temperature and field values of $H = 10$ kOe (0.8 MA/m) and $H = 25$ kOe (2 MA/m). Near $x = 0.3$, the magnetostrictive curve shows a peak in accordance with the near zero magnetic anisotropy observed at this composition. From the magnetostrictive value at the peak, about 1600×10^{-6} ,

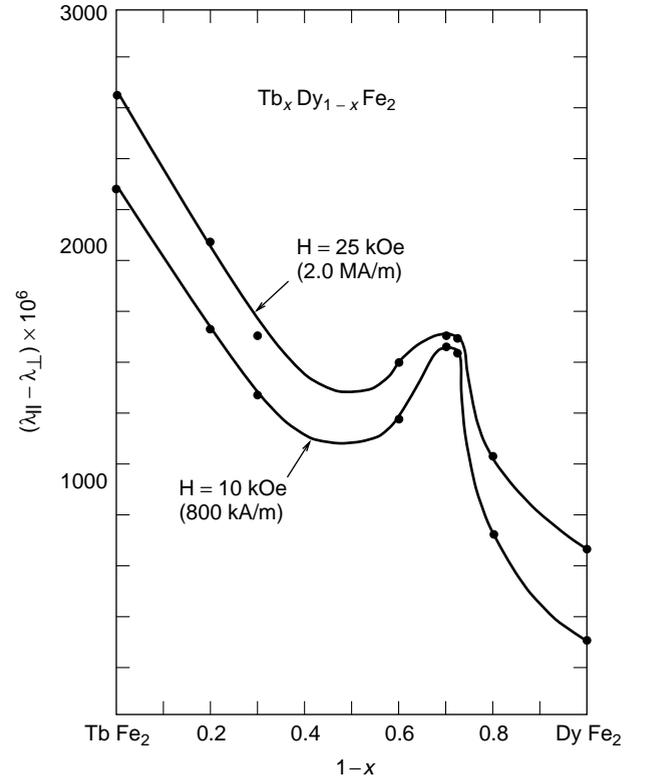


Figure 8. Magnetostriction of polycrystalline $\text{Tb}_x\text{Dy}_{1-x}\text{Fe}_2$ at room temperature (8).

Eq. (4) gives $\lambda_s = 1000 \times 10^{-6}$ which is a widely employed value for the saturation magnetostriction of Terfenol-D.

Anisotropy is present to some degree in all magnetic materials, and therefore, the saturation magnetostriction needs to be defined relative to the axis along which the magnetization lies. One exception is nickel, whose magnetostriction is almost isotropic (see Table 2). Recognizing that there are two independent magnetostriction constants λ_{100} and λ_{111} for cubic materials, the saturation magnetostriction, assuming a single crystal, single domain material is given by a generalization of Eq. (3) for isotropic materials:

$$\lambda_s = \frac{3}{2} \lambda_{100} \left(\alpha_1^2 \beta_1^2 + \alpha_2^2 \beta_2^2 + \alpha_3^2 \beta_3^2 - \frac{1}{3} \right) + 3 \lambda_{111} (\alpha_1 \alpha_2 \beta_1 \beta_2 + \alpha_2 \alpha_3 \beta_2 \beta_3 + \alpha_3 \alpha_1 \beta_3 \beta_1), \quad (5)$$

where λ_{100} and λ_{111} are the saturation magnetostrictions along the $\langle 100 \rangle$ and $\langle 111 \rangle$ axes of the crystal. Cosines α_i

Table 2. Magnetostrictive Coefficients of Cubic Crystal Materials

Material	$\lambda_{100} (10^{-6})$	$\lambda_{111} (10^{-6})$
Nickel	-46	-24
Iron	21	-21
Terfenol-D	90	1600

($i = 1, 2, 3$) define the direction along which the magnetic moments are saturated, and cosines β_i define the direction in which the saturation magnetization is measured. The saturation magnetostriction in the field direction is obtained by using $\alpha_i = \beta_i$ in Eq. (5), which leads to

$$\lambda_s = \lambda_{100} + 3(\lambda_{111} - \lambda_{100})\left(\alpha_1^2\alpha_2^2 + \alpha_2^2\alpha_3^2 + \alpha_3^2\alpha_1^2\right). \quad (6)$$

Note that Eqs. (5) and (6) apply only to single domain materials. In the saturated state, the whole specimen consists of a single domain whose magnetization M_s is aligned parallel to the applied field. However, when a domain structure is present such as in polycrystals, the magnetostriction can be calculated only by averaging the effects, unless the domain structure is known specifically. Note that different domain configurations can give the same bulk magnetization and different magnetostrictions (see Fig. 9). So, assuming that there is no preferred grain orientation, Eq. (6) simplifies further and becomes

$$\overline{\lambda_s} = \frac{2}{5}\lambda_{100} + \frac{3}{5}\lambda_{111}.$$

Extensive magnetostriction data on the R-Fe₂ compounds can be found in (8), and calculations of λ_s in different crystallographic structures such as cubic, hexagonal, and polycrystalline can be found in (2,13,16,21).

Magnetostriction below Saturation

Although saturation magnetostriction λ_s can be determined by employing the methods previously discussed, magnetostriction between the demagnetized state and saturation is structure sensitive, so general constitutive

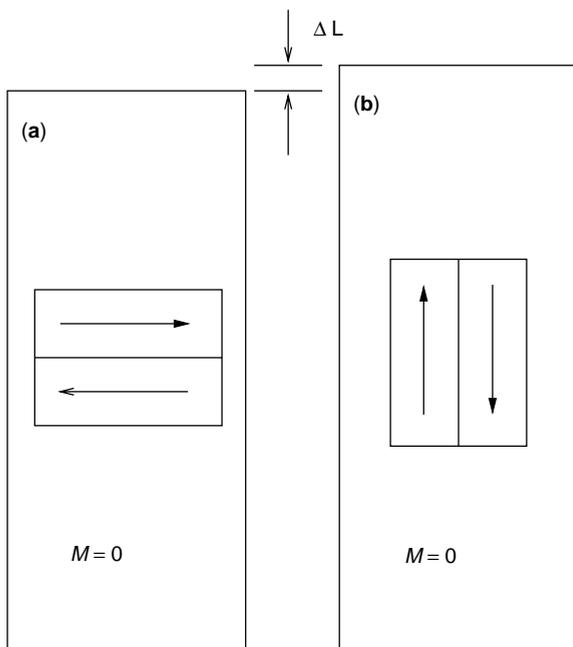


Figure 9. Demagnetized specimen featuring a 180° domain wall in the (a) horizontal or (b) vertical direction. The length of the specimen is different in either case, even though the magnetization is the same.

relationships for magnetostriction are not feasible. However, an explicit solution exists when strains are due primarily to 90° domain rotations. In practice, these rotations occur in (1) a single crystal that has uniaxial anisotropy in which the field is applied in a direction perpendicular to the easy axis or (2) a polycrystalline material in which the magnetic moments have been completely aligned in a direction perpendicular to the applied field, such as Terfenol-D under extreme compression or nickel under tension. The latter implies that perpendicular stress energy is sufficient to dominate crystal anisotropy, as discussed earlier. For that regime, combining Eqs. (1) and Eq. (2) gives

$$\lambda = \frac{3}{2}\lambda_s \cos^2 \theta, \quad (7)$$

where θ is the angle between the M_s vectors and the field direction. Recognizing that the bulk magnetization in the field direction is given by $M = M_s \cos \theta$, Eq. (7) becomes

$$\lambda = \frac{3}{2}\lambda_s \left(\frac{M}{M_s}\right)^2, \quad (8)$$

which provides a quadratic relationship between magnetization and magnetostriction. It has been shown that this expression is sufficiently accurate in a broad range of transducer regimes in which high mechanical preloads are employed to optimize transducer performance (24). A generalized version of this equation has been given in (25), and more elaborate models of magnetostrictive hysteresis have been presented in (19,26–28). Additional effects such as stress dependences have been also considered (25,29). Finally, the dependence of the magnetostriction of the R-Fe₂ compounds on temperature has been discussed in (8).

OTHER MAGNETOSTRICTIVE EFFECTS

Linear magnetostriction is just one of several manifestations of a more general phenomenon, the coupling between the magnetic and elastic states in a material. These effects are briefly discussed following and are summarized in Table 3.

Villari Effect

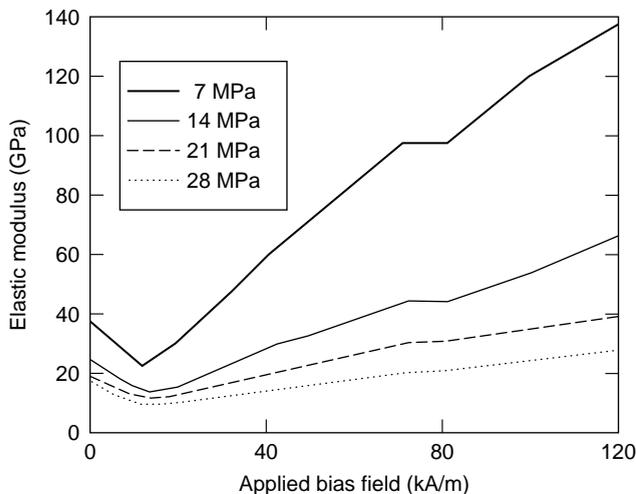
The Villari effect, also known as the magnetomechanical effect, are the changes in magnetization that a magnetostrictive material undergoes when subjected to an applied uniaxial stress. This effect pertains to the transduction of energy from the elastic to the magnetic state and is inverse of Joule magnetostriction. Furthermore, the Villari effect exhibits many of the attributes of the direct magnetostrictive effect inasmuch as its physical origin also lies in magnetoelastic coupling. The Villari effect has been the object of much study, given its relevance in applications such as nondestructive evaluation and sensing. Extensive theoretical and experimental details can be found in (25). The effect of stress on magnetostrictive materials in particular has been discussed in (30).

Table 3. Magnetostrictive Effects

Direct Effects	Inverse Effects
Joule magnetostriction Change in sample dimensions in the direction of the applied field	Villari effect Change in magnetization due to applied stress
ΔE effect Magnetoelastic contribution to magnetocrystalline anisotropy	Magnetically induced changes in elasticity
Wiedemann effect Torque induced by helical anisotropy	Matteucci effect Helical anisotropy and emf induced by torque
Magnetovolume effect Volume change due to magnetization (most evident near the Curie temperature)	Nagaoka–Honda effect Change in the magnetic state due to a change in volume

ΔE Effect

The elasticity of magnetostrictive materials is composed of two separate but related attributes, the conventional stress–strain elasticity that arises from interatomic forces and the magnetoelastic contribution due to the rotation of magnetic moments and the ensuing strain that occur when a stress is applied. The latter contribution, known as the ΔE effect, is quantified by $\Delta E = (E_s - E_0)/E_0$, where E_0 is the minimum elastic modulus and E_s is the elastic modulus at magnetic saturation. Because the strain produced by magnetic moment rotation adds to the non-magnetic strain [see Eq. (9)], the material becomes softer when the moments are free to rotate. This is illustrated in Fig. 10. Note that the material becomes increasingly stiff as saturation is approached and magnetic moment mobility decreases. The ΔE effect is small in nickel ($\Delta E = 0.06$) but is quite large in Terfenol-D (ΔE up to 5) and certain transverse-field annealed $\text{Fe}_{81}\text{B}_{13.5}\text{Si}_{3.5}\text{C}_2$ (Metglas 2605SC) amorphous ribbons ($\Delta E = 10$). The ΔE effect of Terfenol-D can be advantageously employed in tunable vibration absorbers and broadband sonar systems (31).

**Figure 10.** Magnetoelastic modulus of $\text{Tb}_{0.3}\text{Dy}_{0.7}\text{Fe}_2$ under various stresses (79).

Wiedemann Effect

A current-carrying ferromagnetic or amorphous wire produces a circular magnetic field in a plane perpendicular to the wire, and the moments align predominantly in the circumferential direction. When an axial magnetic field is applied, some of the moments align helically and create a helical magnetic field. The twist observed in the wire is called the Wiedemann effect. The inverse Wiedemann effect, known as the Matteucci effect, is the change in axial magnetization of a current-carrying wire when it is twisted. Further details can be found in (2).

Magnetovolume Effect

The volume of a magnetostrictive material remains virtually unchanged during normal operation, but in certain extreme regimes, the volume of the material may change in response to magnetic fields. This anomalous volume change is called the volume magnetostriction or Barret effect. The effect has little applicability in smart structure systems. For instance, the magnetostriction curve of nickel rapidly reaches -35×10^{-6} at only 10 kA/m, but the fractional volume change is only 0.1×10^{-6} in a much larger field of 80 kA/m. In the alloy Invar (36% nickel–64% iron), the fractional volume change at the Curie temperature, which is slightly above room temperature, compensates for the intrinsic thermal expansion and gives a compound that has nearly zero thermal expansion at room temperature. The inverse of the Barret effect, the Nagaoka–Honda effect, is the change in magnetic state caused by a volume change (2,14).

MAGNETOSTRICTIVE TRANSDUCERS

One advantage of magnetostrictive transducers over other types of transducers is that they can be driven by conventional low impedance amplifiers, particularly at frequencies well below resonance; in this case, the low impedance of a magnetostrictive transducer means that driving voltages can be low. This can prove useful in medical applications and in general can greatly simplify amplifier design. Figure 11 shows the measured complex electrical impedance frequency response function $Z_{ee} = V/I$ of a Terfenol-D transducer designed following the generic

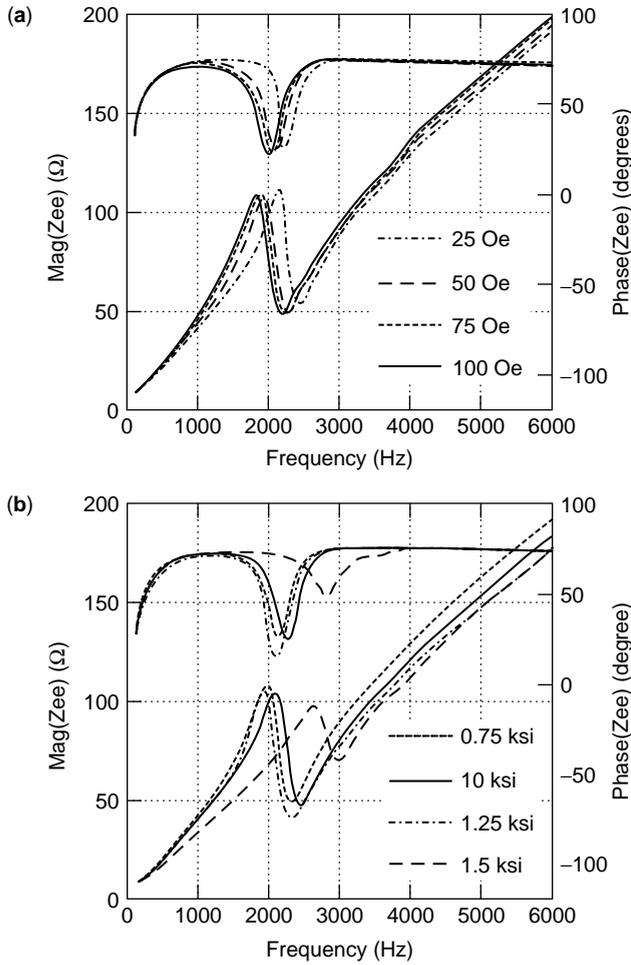


Figure 11. Total electrical impedance Z_{ee} versus frequency, expressed as magnitude and phase: (a) bias condition of 5.2 MPa, 24 kA/m (0.75 ksi, 300 Oe), and varied ac drive levels; (b) constant ac drive level of 8 kA/m (100 Oe) and varied bias conditions (79).

configuration indicated in Fig. 12 (32,33). This transducer consists of a cylindrical magnetostrictive rod, a surrounding copper-wire solenoid, a preload mechanism that consists of a bolt and a spring washer, magnetic couplers, and a barrel-like permanent magnet that provides bias magnetization. Specific design details depend on the particular smart structure application; however, this configuration depicts the basic components needed to extract maximum performance from a magnetostrictive material.

Because magnetostriction is produced by the rotation of magnetic moments, a magnetostrictive transducer driven by an ac magnetic field vibrates at twice the drive frequency, and the motion occurs in only one direction. This is illustrated in Fig. 13, where the solid lines represent the unbiased input and corresponding strain output. The dashed lines demonstrate the performance improvements achieved by applying a magnetic bias. In the biased regime, the frequency of the input is preserved, the output is bidirectional, and the ratio of output per input is substantially larger. To center operation accurately around the

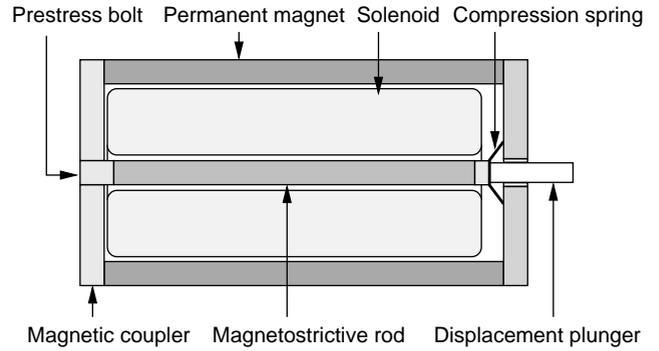


Figure 12. Cross section of a typical magnetostrictive transducer.

desired bias point, a permanent magnet is often employed in combination with a static field generated by passing a dc current through the solenoid. Note that exclusive permanent magnet biasing has the advantage of substantial power savings, but it has the disadvantage of added bulk and weight. Conversely, dc currents produce considerable power losses through ohmic heating but facilitate savings in bulk and weight. Magnetic biasing can be alternatively provided by magnets located in series with the rod or rods; this design is known as a stacked-magnet configuration. The stacked-magnet configuration can improve the magneto-mechanical coupling up to 5% for large rods ($L > 20$ cm, $D > 2.5$ cm) compared to the barrel-magnet configuration. However, collateral problems such as saturation effects and resonance frequency shifts are common in stacked-magnet designs. Carefully designed transducers must provide efficient magnetic flux closure within the circuit formed by the rod itself, the couplers, and the permanent magnets.

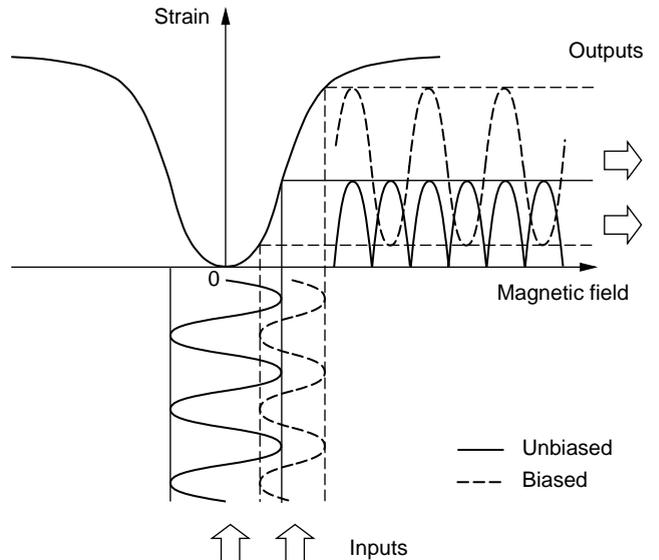


Figure 13. Effect of magnetic bias on the strain produced by a magnetostrictive transducer.

Finally, although modern magnetostrictive materials such as Terfenol-D are manufactured so that their magnetic moments are nearly perpendicular to the rod axis, a static stress (mechanical preload) is nevertheless, required to achieve full alignment of all of the moments. The moments of a mechanically free rod are aligned randomly, and the rod produces only about half of its maximum magnetostriction because the moments initially aligned with the rod axis do not contribute to the magnetostriction. Furthermore, the stress anisotropy generated by the static compression (or tension for materials of negative magnetostriction) enhances the overall magnetoelastic state, as described previously. Note that in designs that employ linear washers for preloading, the stress in the magnetostrictive rod can vary significantly relative to the nominal preload during dynamic transducer operation. By virtue of the magnetomechanical coupling, this can have a profound impact on the performance of the magnetostrictive transducer and driving electronics by affecting the magnetic state and, through it, the electrical regime (see Fig. 11). The effects of mechanical preload and magnetic bias on the performance of a Terfenol-D transducer have been studied in (34). A second reason for employing a mechanical preload is to avoid operating the rod in tension, particularly when driving brittle materials such as Terfenol-D ($\sigma_t = 28 \text{ MPa}$, $\sigma_c = 700 \text{ MPa}$) at or near mechanical resonance.

Actuator Applications

The number of actuator applications based on magnetostrictive materials, mainly Terfenol-D, is continuously increasing as a consequence of the high energy density, high force, broad frequency bandwidth, and fast response that these materials provide. Even though the cost of Terfenol-D is high at present, the range of applications is likely to continue increasing as manufacturing techniques are perfected and prices decline. Actuators designed according to the configuration shown in Fig. 12 have been employed in these applications: sonar, chatter control for boring tools, high-precision micropositioning, borehole seismic sources, geological tomography, hydraulic valves for fuel injection systems, deformable mirrors, hydraulic pumps, bone-conduction hearing aids, exoskeletal telemanipulators, self-sensing actuators, degassing in manufacturing processes such as rubber vulcanization, and industrial ultrasonic cleaning. Four main application subgroups of current transducer designs are discussed here: sonar transducers, linear motors, rotational motors, and hybrid smart material transducers. The reader is directed to (1,2) for more complete details.

Sonar Transducers. Efficient sonar transducers must produce high mechanical power at low frequencies and often have the additional constraint that a broad frequency bandwidth or equivalently, a low quality factor Q , must be attained. Although nickel was widely employed in sonar applications during World War II, it has a low magneto-mechanical coupling coefficient of $k = 0.30$ which typically demands a high Q to achieve good efficiencies. In

contrast, the newer giantmagnetostrictive materials have much higher coupling coefficients of more than $k = 0.70$ which makes it possible to operate the transducer at low Q and attain high power output simultaneously. For example, a Terfenol-D Tonpitz transducer similar to that depicted in Fig. 14a can produce a bandwidth of 200 Hz at a resonance frequency of 2 kHz ($Q = 10$) and a source level of 200 dB ref. $1 \mu\text{Pa}$ at 1 m (35). Another Terfenol-D transducer reportedly produces a maximum output of 206 dB

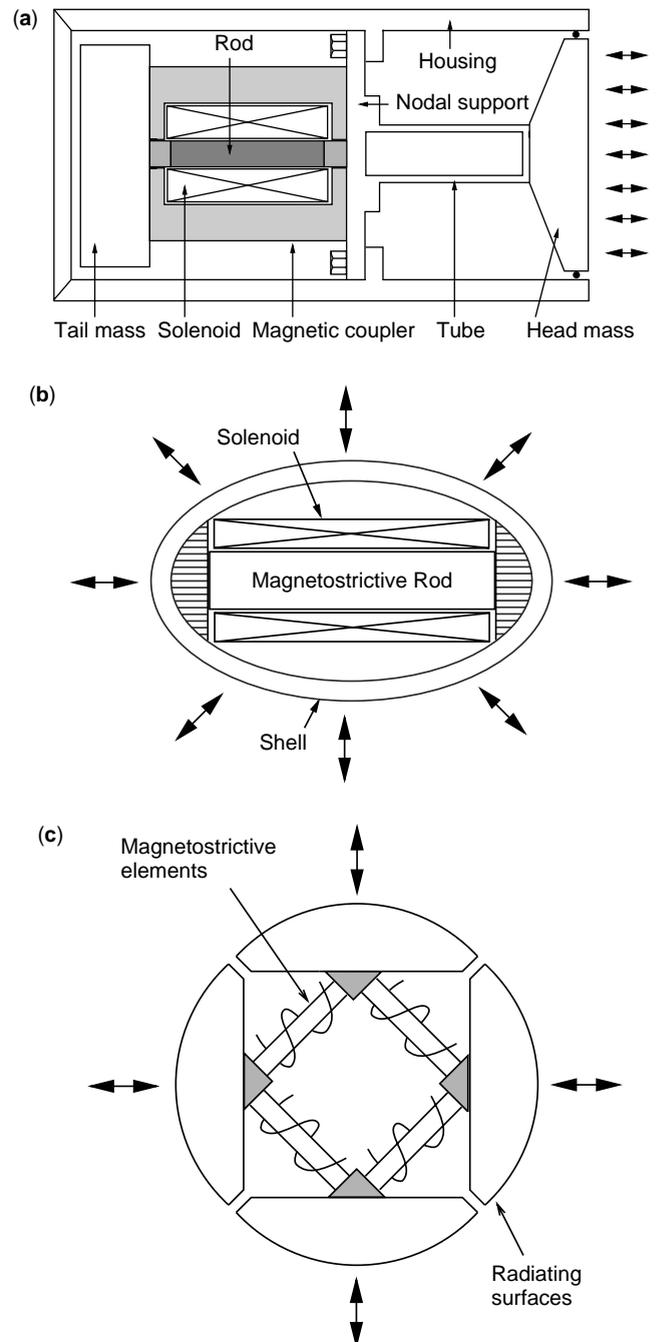


Figure 14. Magnetostrictive sonar transducers: (a) Tonpitz, (b) flextensional, and (c) square ring.

ref. $1 \mu\text{Pa}$ at 1 m and operates across a broad usable bandwidth of 5–50 kHz (36). Other designs employ the linear motion of cylindrical magnetostrictive rods to flex a surrounding shell or to induce radial vibrations in a tube or ring, as shown in Fig. 14b,c. Further details and references are provided in (37).

Linear motors. Direct coupling between the load and the magnetostrictive element in Fig. 12 implies that the net load displacement is limited by magnetostriction. For instance, an 11.4-cm Terfenol-D actuator provides maximum displacements of about 0.2 mm. This displacement level is sufficient for many vibration control applications, but certain applications such as flow control valves or aircraft flap positioners typically require much larger strokes.

The fact that Joule magnetostriction takes place at constant volume is employed in the Kiesewetter motor to displace loads beyond the maximum strain normally achievable by a Terfenol-D rod. This motor (38) consists of a cylindrical Terfenol-D rod that fits snugly inside a stiff stator tube when no magnetic field is applied. Several short coils surround the stator to produce a magnetic field profile that sweeps along the Terfenol-D rod. When one of the coils is energized, for instance, coil 1 in Fig. 15a, the section of rod directly exposed to the magnetic field elongates and shrinks. As the field is removed, the rod clamps itself again inside the stator but at a distance d to the left of the original position. As the remaining coils are energized sequentially and the magnetic field profile is swept, the rod moves in a direction opposite to the sweeping field. The direction of motion is changed by inverting the sequence in which the coils are energized. From a design perspective, the total displacement is limited only by the length of the Terfenol-D rod, whereas the speed of motion is proportional to the sweeping frequency and the magnetostriction of the rod. Other factors that affect the smoothness and speed of the motor are the number of traveling pulses, the spacing between excitatory coils, the stiffness of the Terfenol-D material, and skin effect degradation due to eddy currents. The Kiesewetter motor is self-locking when unpowered, which is an important attribute for many robotic applications.

A proof-of-concept Kiesewetter motor presented in (39) produces 1000 N of force and 200 mm of useful stroke at a speed of 20 mm/s; it is intended for uses such as control of coat weight and fiber distribution in the paper industry and valve operation and precision positioners for the machine tool industry. An improved design presented in (40) addresses some of the technical problems of the Kiesewetter motor, particularly the degradation of fit between the stator and the rod caused by wear and thermal expansion. Furthermore, this revised design enables rotary motion in a way that is otherwise impossible to achieve by using the original Kiesewetter design.

Another variant of the inchworm principle is shown in Fig. 15b. This motor consists of translating clamps, fixed position clamps, pusher transducers, and a load shaft. By coordinating the clamping and unclamping actions of the clamps with the action of the pushing transducers, it is possible to induce bidirectional motion of the load shaft. The load rating is limited by the frictional force between the clamps and the load shaft. Note that the inchworm principle can

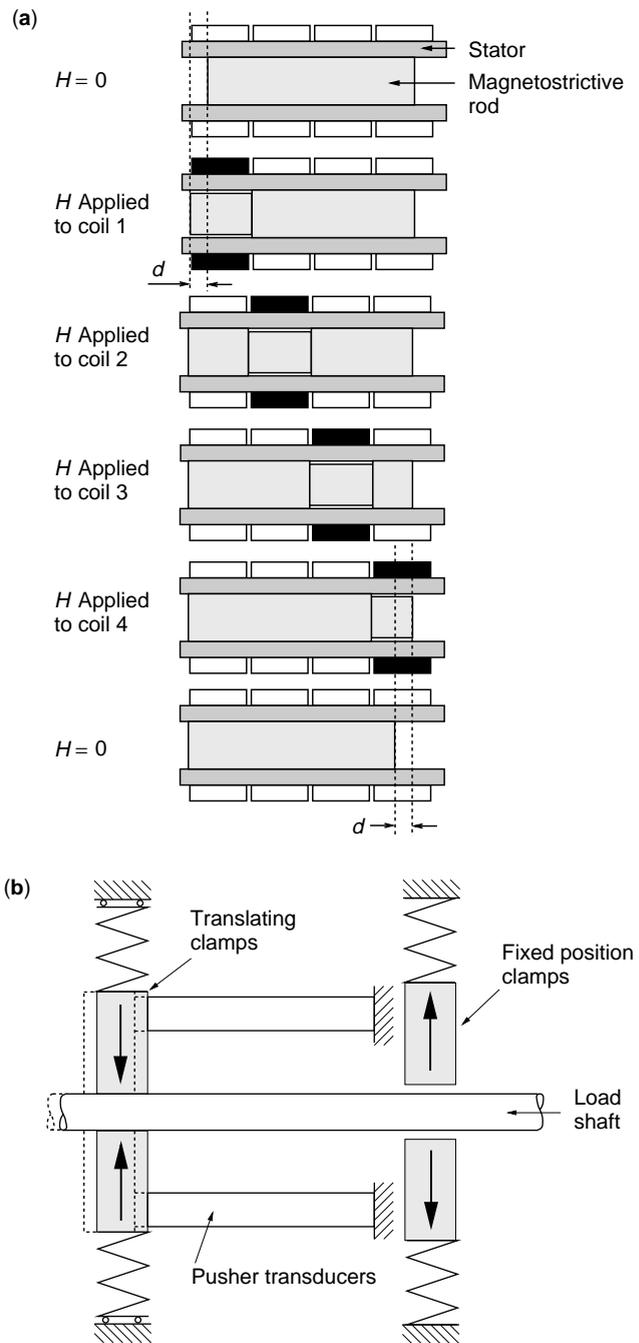


Figure 15. (a) The Kiesewetter inchworm motor. Black rectangles indicate energized coils; white rectangles indicate inactive coils. (b) Inchworm linear motor.

also be implemented by using other smart materials such as piezoelectric stacks (41) or a combination of piezoelectric and magnetostrictive elements, as shown later.

Piezoelectric transducers are often preferred for ultrasonic power generation in the megahertz range, but certain applications in the low-ultrasonic range benefit from the ruggedness and lack of depoling mechanisms of magnetostrictive materials. For instance, nickel is extensively used in applications such as degassing liquids (20–50 kHz)

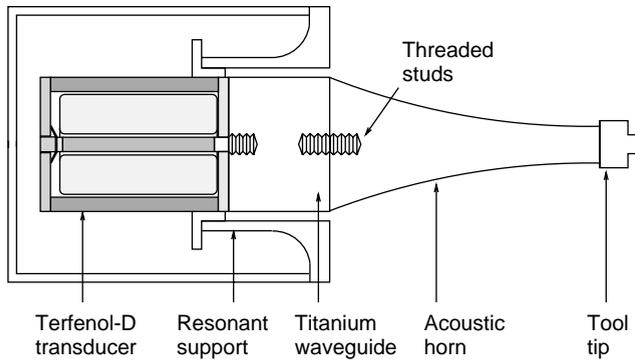


Figure 16. Schematic diagram of an ultrasonic Terfenol-D device consisting of a quarter-wave transducer coupled to a quarter-wave titanium waveguide and a half-wave acoustic horn. Different tool tips can be used as needed.

and cleaning dental or jewelry pieces (more than 50 kHz). A surgical ultrasonic tool based on Terfenol-D was developed recently that reportedly provides enhanced power and displacement output compared to existing piezoelectric tools; it is also lighter, more compact, and can deliver a 600-V, 1-MHz signal to cauterize bleeds without affecting its surgical function. In this device, illustrated in Fig. 16, a laminated quarter-wavelength Terfenol-D rod is coupled to a quarter-wavelength titanium waveguide that provides the resonant subassembly to which a half-wavelength acoustic horn is attached. The acoustic horn has an amplification factor from 15–30; thus, it provides extreme accelerations and energy concentration at the tip of the tool (42,43). Other current or potential uses for this transducer design include industrial cleaning, sonic cell disruption and sterilization, friction welding, and treatment of diverse chemical and biological processes (1).

Rotational Motors. Smart material motors based on the magnetostrictive principle are potentially simpler and more reliable than conventional hydraulic or electromagnetic systems. The inchworm technique has been employed in a rotational motor that produces a torque of 3 Nm at a speed of 0.5 rpm (44). Another inchworm type device also provides a speed of 0.5 rpm but produces a very high torque of 12 Nm and precision microsteps of $800 \mu\text{rad}$ (45). Despite the great positional accuracy and high holding torques, the current inchworm-type rotational motors tend to lack efficiency. Much of the efficiency limitation has been overcome in the resonant rotational motor proposed by Claeysen et al. (46). Two linear Terfenol-D actuators are used to induce elliptical vibrations in a circular ring that acts as a stator and transmits the vibrations to rotational rotors pressed against the ring. The prototype reportedly provides a maximum torque of 2 Nm at a maximum speed of 17 rpm.

The field of ultrasonic rotational motors has aroused much research and commercial interest. These motors are employed in a wide range of applications from autofocusing camera lenses to robotic manipulators. A rotational actuator developed by Akuta (44) employs Terfenol-D to achieve a relatively high speed of 13.1 rpm and a maximum torque

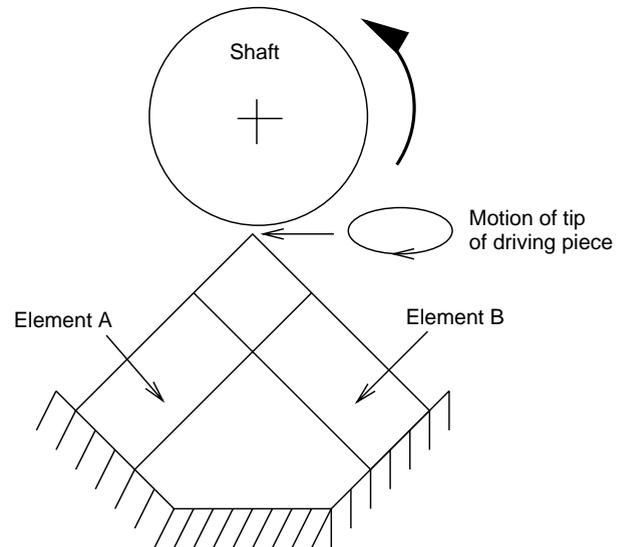


Figure 17. Rotational ultrasonic motor (44).

of 0.29 Nm. As depicted in Fig. 17, this motor employs two Terfenol-D exciter rods to induce rotations in the shaft.

Hybrid Magnetostrictive/Piezoelectric Devices. Given their technological interest, hybrid smart material actuators can be considered a separate class independently of whether they are intended for sonar, linear, or rotational applications. Because magnetostrictive materials are inductive and piezoelectric elements are capacitive, it is advantageous to combine both types of materials in the same device, so that a resonant electric circuit is formed. When driven at resonance, such a device behaves like a purely resistive load and only the energy that is effectively converted to mechanical motion or lost to inner losses needs to be supplied externally. This greatly simplifies amplifier design and helps to attain high efficiencies.

The hybrid device has been demonstrated that overcomes the difficulties involved in achieving motion at only one end of a Tonpitz piston-type sonar transducer; it consists of a quarter-wavelength stack of piezoelectric Navy type I ceramic rings joined to a quarter-wavelength Terfenol-D composite tube [see Fig. 18a and (47)]. The inherent 90° phase shift between the magnetostrictive and piezoelectric velocities in combination with the quarter-wavelength design of the elements ensures addition at one end and cancellation at the other. The device is *mechanically* unidirectional, but it becomes *acoustically* unidirectional only under array-baffled operation. The measured front-to-back pressure ratio is 5 dB for the device alone and 15 dB under array-loaded conditions. The concept of hybrid piezoelectric/magnetostrictive transduction has also been implemented for linear inchworm motors (48,49) and rotational motors (50). For example, the prototype presented in (48) has the configuration shown in Fig. 15b, but the clamping is done by piezoelectric stacks, and the translation is provided by Terfenol-D rods. The intrinsic 90° phase lag between the two types of elements provides natural drive timing for the inchworm, and the direction of motion

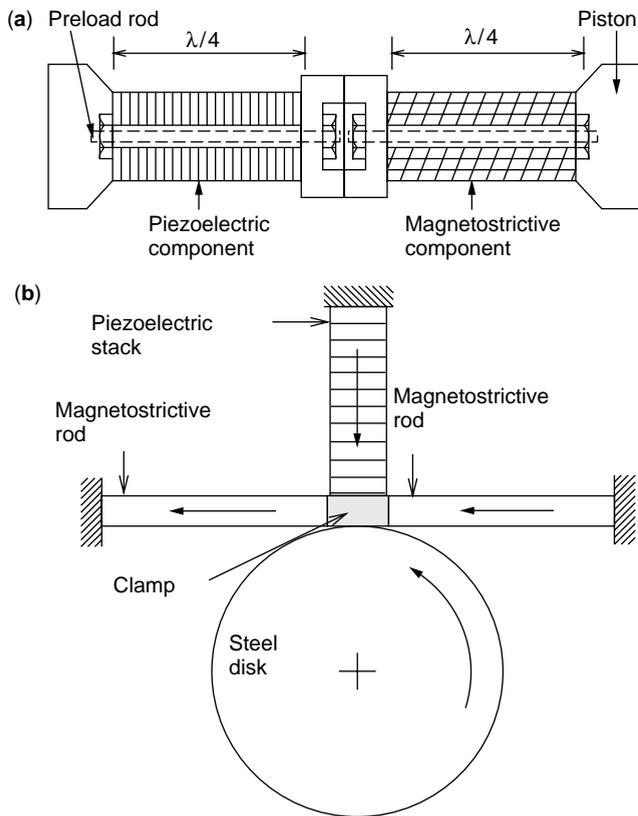


Figure 18. Hybrid magnetostrictive/piezoelectric transducers: (a) sonar projector (47); (b) rotational motor (50).

is easily reversed by changing the magnetic bias on the Terfenol-D elements. This motor achieves a zero-load speed of 25.4 mm/s and a stall load of 115 N.

The hybrid magnetostrictive/piezoelectric rotational motor illustrated in Fig. 18b follows the proof-of-concept transducer presented in (50). A piezoelectric stack clamps a piece of friction material onto the rotating disk, and two magnetostrictive rods move the clamp tangentially to the disk to produce rotational motion. As indicated before, the sequence of the motion is determined by the natural timing of the piezoelectric and magnetostrictive responses. The device produces a speed of 4 rpm at excitatory voltages from 30–40 V and frequencies from 650–750 Hz.

Sensor applications

Magnetostrictive materials are being employed in a wide variety of sensor designs as evidenced by the growing number of publications and patents. In this overview, the term sensor is used in a broad sense to indicate the attributes of magnetostrictive materials that facilitate generation of electrical signals in response to mechanical excitations, such as force, strain, and torque, or magnetic excitations such as magnetic fields. By virtue of the magnetoelastic coupling, changes in the magnetoelastic state through these parameters (or a combination of them) produce measurable change in magnetization anisotropy. To complete the sensing mechanism, a pick-up coil is often wrapped

around the magnetostrictive material to detect magnetization changes; this effectively provides a mechanism for conversion of energy from magnetic to electrical regimes. The principle that links the magnetization with the voltage V generated across a pick-up coil is the Faraday–Lenz law of electromagnetic induction:

$$V = -NA \frac{dB}{dt},$$

in which N and A are, respectively, the number of turns and constant cross-sectional area of the coil and B is the magnetic induction that quantifies the magnetization state through the relationship $B = \mu_0(H + M)$ (21). Alternatively, interferometric techniques can be employed to detect the changes in wave speed that occur when the magnetostrictive material changes its properties in the presence of external excitations, for instance, the stiffness changes associated with the ΔE effect described earlier.

An overview of sensor designs is presented next. The main principles that enable operation of the sensors are emphasized. The list is not comprehensive, but it shows that a huge number of alternative designs can be devised based on the fundamental operation principles presented here. Further details can be found in the references provided.

Torque sensors. Magnetostrictive noncontact torque meters have been devised based on the principle that the torque applied to a shaft generates stresses of opposite sign, $+\tau$ and $-\tau$, oriented $\pm 45^\circ$ from the shaft axis. If the shaft is magnetostrictive or has a magnetostrictive amorphous ribbon bonded to it, the magnetic properties along the directions of $+\tau$ and $-\tau$ change as discussed previously. These properties can be measured either differentially by a set of perpendicular coils, as shown in Fig. 19a, or through a single Hall effect or similar magnetic field intensity sensor (51). This kind of sensor can be employed, for instance, in fly-by-wire steering systems for the automotive and aerospace industries. Additional details and references can be found in (2).

Another class of noncontact torque meters relies on the changes in permeability exhibited by a magnetostrictive material subjected to torsional stress. In particular, applications that require less sensitivity can benefit from the elevated mechanical strength that magnetic steels or alloys provide. One example is shown in Fig. 19b, where the working torque on a drill bit is detected by two sensing coils connected in series, one located over the flutes and the other over the shank (the permeability of the shank is less sensitive to changes in torque than the flutes.) A coil provides the ac magnetic field excitation, and the sensor's proportional output is the differential voltage generated by the sensing coils as the permeability of the bit changes due to the applied torque (52).

Deformational and Position Sensors. Transverse-field annealed magnetostrictive ribbons or wires make very sensitive strain gauges. A sensor of this kind has been made from strips of Metglas 2605SC transverse annealed for 10 min in a 208 kA/m (2.6 kOe) magnetic field at 390°C

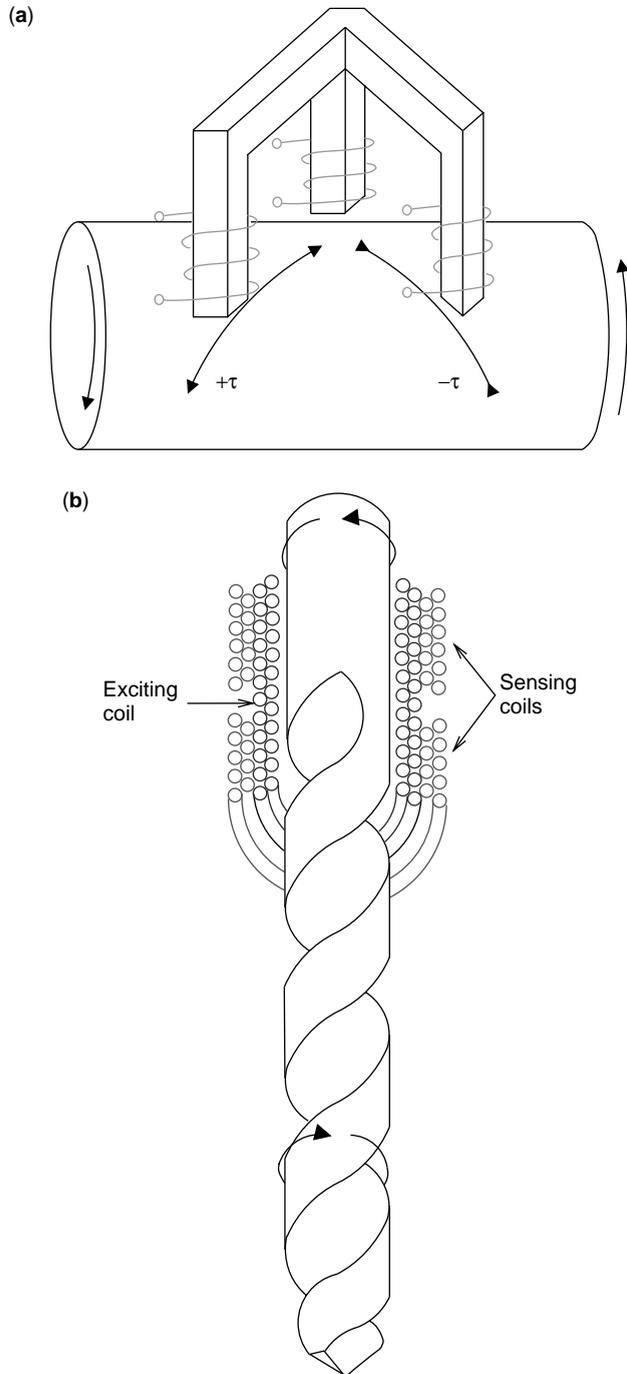


Figure 19. Magnetostrictive noncontact torque sensors: (a) Differential reading in directions oriented $\pm 45^\circ$ from the shaft axis and (b) differential reading of the permeability changes experienced by a drill bit subjected to a torque.

and rapidly cooled in a saturation field (53). The sensor responds to the changes in the permeability of the ribbon, which by virtue of the magnetomechanical coupling depends in turn on the state of strain in the material. Defining a dimensionless gauge factor as the fractional change in the measured parameter (in this case permeability) divided by the change in strain, $F = (\partial\mu/\partial S)/\mu$,

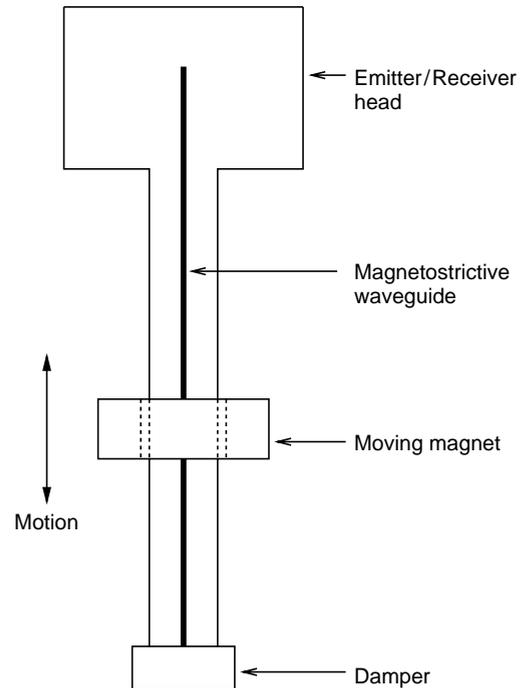


Figure 20. Magnetostrictive waveguide position sensor.

this sensor has an F value equal to about 250,000, which compares extraordinarily well with resistive strain gauges ($F = 2$) and semiconductor gauges ($F = 250$). One problem encountered in using this device is that normal thermal expansion can saturate the sensor. This problem can be overcome by bonding the material using a highly viscous liquid, although this limits operation to ac regimes.

A position detector can be made by using a magnetostrictive material as an acoustic waveguide. This device, shown in Fig. 20, consists of a permanent magnet that is connected to the target and rides along the length of the waveguide, an emitter/receiver head that sends and receives either an acoustic or current pulse down the waveguide, and a damper that prevents unwanted wave reflections. The sensor's operating principle is rather simple; the magnet interacts with the magnetostrictive waveguide and locally changes its material properties. These material property changes can be detected in different ways. In one version, the stiffness discontinuity produced by the magnet (ΔE effect, see earlier) partially reflects an acoustic pulse sent by the emitter. In a second version, the emitter sends a continuous current pulse down the waveguide that produces a circumferential magnetic field that interacts with the axial field from the magnet. The resulting helical field produces a twist in the wire (Wiedemann effect) that travels back to the receiver head. In both versions, the transit times of the original and reflected pulses provide a measure of the location of the magnet along the waveguide. This sensor can be used for measuring fluid levels by connecting the magnet to a float or for generic position sensing of up to 50 m at ± 1 mm accuracy (2).

Magnetometers. If the magnetostriction of a given material is known as a function of magnetic field, the problem

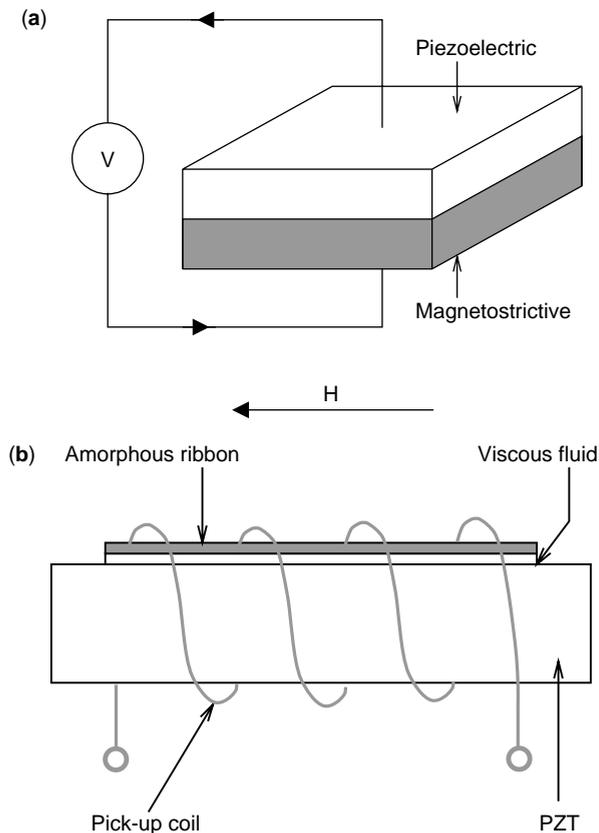


Figure 21. Hybrid magnetostrictive/piezoelectric magnetic field sensors. When a magnetic field is applied to one of these sensors, the magnetostrictive material strains, which either (a) generates a voltage across the piezoelectric plate or (b) induces an emf in a surrounding pick-up coil that can be extracted from the alternating carrier emf produced as the piezoelectric plate resonates.

of measuring the magnetic field reduces to one of measuring length. The length can be measured by a laser interferometer, optic fiber, strain gauge, capacitor, or another calibrated material such as a piezoelectric compound. For example, a very simple design consists of two slabs of magnetostrictive and piezoelectric materials bonded together (Fig. 21a). When a magnetic field is applied to the magnetostrictive material, it strains and induces a proportional voltage in the piezoelectric material. In another version, shown in Fig. 21b, a magnetometer is realized by bonding a field-annealed metallic glass ribbon onto a resonating PZT plate by using a viscous fluid. An alternating voltage is applied to the PZT plate, which generates a longitudinal stress field. By using proper bonding techniques, the dynamic stress in the metallic ribbon is congruent with that in the PZT, and the static component is filtered out by the viscous fluid. By virtue of the Villari effect, these dynamic stresses create an oscillating electromotive force (emf) in the surrounding pick-up coil. When exposed to low-frequency magnetic fields, a low-frequency emf is generated in the coil that is extracted from the carrier emf by conventional phase sensitive detection techniques. The measured detection limit can reach 6.9×10^{-6} A/m at 1 Hz (54), which compares with that of fluxgate magnetometers.

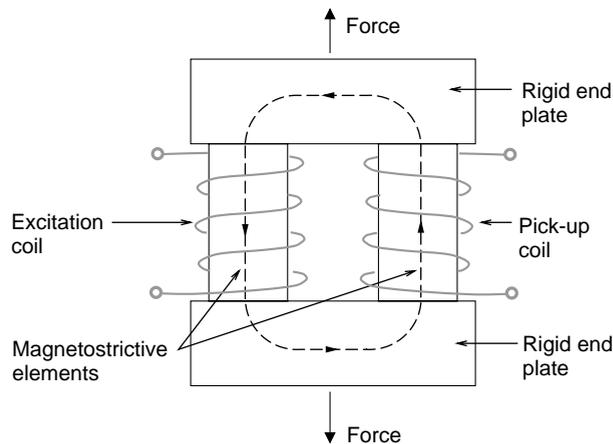


Figure 22. Magnetostrictive force sensor based on the Villari effect.

Another type of magnetometer consists of a magnetostrictive film bonded to an optic fiber. When the sensor is exposed to magnetic fields, the magnetostrictive material deforms and induces a deformation in the optic fiber. This causes changes in the optical path length of laser beams that pass through the optic fiber and can be detected by an interferometer (55). Highly sensitive metallic glass ribbons have been employed in devices so designed that yield quasi-static resolutions from 1.6×10^{-3} – 8.0×10^{-3} A/m (56). Finally, a diode laser interferometer has been used to detect changes in the length of a Terfenol-D rod that are produced when a magnetic field is applied (57). A maximum sensitivity of 160×10^{-6} A/m was achieved, although certain non-linear dependences were observed that make it critical to operate the sensor within its optimum mechanical preload range.

Force Sensors. By employing the Villari effect, it is possible to realize a simple and rugged force sensor from either crystal or amorphous magnetostrictive materials. The magnetostrictive attribute that provides the operating principle for such a sensor is the dependence of magnetization on the state of stress in a material. To illustrate, the design in Fig. 22 consists of two magnetostrictive elements, one surrounded by an excitation coil and the other surrounded by a pick-up coil, and two rigid end plates. In one mode of operation, an ac voltage is applied to the excitation coil that generates a magnetic flux in the sensor and a corresponding voltage in the sensing coil. As a force is applied, the magnetostriction in the elements produces a change in the magnetic flux that is detected as a proportional voltage change in the pick-up coil. In a second mode of operation at constant flux, the excitation voltage is allowed to change to maintain a constant pick-up coil output voltage. The change in excitation voltage is then related to the change in the applied force. Compared to conventional force sensors such as those based on strain gauges, this sensor is simpler, more rugged, and requires simpler electronics. A similar Villari effect sensor based on amorphous ribbons has been discussed in (58). Numerous other designs have been discussed or patented, including

percussion sensors, pressure sensors, and force sensors based on magnetoelastic strain gauges. The reader is directed to (2) for further details and references.

Transducer Models

Fundamental models of the performance of magnetostrictive materials used in smart structure systems are particularly difficult to develop. The difficulty lies in the coupling between regimes and the strong nonlinear dependence of material behavior on operating conditions and the hysteresis inherent in the materials. Notwithstanding, models based on physical principles are necessary to optimize transducer design and control, particularly considering that first-principles models provide the ability to *scale the results* in a sense that empirical models typically do not. These include magnetization models based on Preisach, Stoner–Wohlfarth, Jiles–Atherton, and micromagnetic theories, as well as domain rotation and higher order magnetostrictive models. However, comprehensive models for designing and controlling integrated smart structure systems must address the interaction between the magnetostrictive material’s elastic, magnetic, thermal, and electrical regimes and the dynamics of the underlying structure. This can include elastic dynamics and also acoustics attributes. Following is an overview of models that have been employed to design and control magnetostrictive transducers. The reader is directed to the references for more comprehensive details on individual modeling techniques.

Piezomagnetic Constitutive Equations. Piezomagnetic models analogous to those classically employed for piezoceramics can prove useful for low drive level applications where hysteresis is minimal and behaviors are quasi-linear. Neglecting thermal effects, the total strain ε of a magnetostrictive material includes two contributions: (1) the magnetostriction λ produced by the rotation of magnetic moments as they align with externally applied magnetic fields and (2) a purely elastic component $\varepsilon = s\sigma$ of the kind found in conventional nonmagnetic materials. Analogously, the magnetic induction B consists of two contributions: (1) a magnetomechanical component dependent on the stress and (2) the constant-stress magnetic constitutive law $B = \mu H$. From these considerations, the constitutive linearized equations that describe magnetostriction or piezomagnetism can be written in differential form, as follows:

$$\begin{aligned}\Delta\varepsilon &= s^H \Delta\sigma + d \Delta H, \\ \Delta B &= d^* \Delta\sigma + \mu^\sigma \Delta H,\end{aligned}\quad (9)$$

where ε is the total strain in the magnetostrictive material, s^H is the compliance at constant field H , d and d^* are the piezomagnetic coefficients, σ is the stress, B is the magnetic induction, and μ^σ is the magnetic permeability at constant stress.

The ratio of fundamental constants $k^2 = d^2 / (s^H \mu^\sigma)$ for this one-dimensional linearized case provides a formulation which has been of value in describing the magnetomechanical coupling coefficient k of magnetostrictive

materials. This coupling coefficient provides a measure of the conversion efficiency between magnetic and mechanical energies. Specifically, the magnetomechanical coupling coefficient squared represents the fraction of maximum stored magnetic energy that can be converted into elastic energy or conversely, the fraction of maximum stored elastic energy that can be converted into magnetic energy. It has been shown (59) that k can be expressed in terms of suitable internal energies,

$$k = \frac{U_{me}}{\sqrt{U_e U_m}},$$

in which $U_{me} = (1/2)Hd^*\sigma$ is the magnetic energy due to elastic energy, $U_e = (1/2)s\sigma^2$ is the elastic internal energy and $U_m = (1/2)\mu H^2$ is the magnetic internal energy. Note that these expressions for k are based on the intrinsic assumption of piezomagnetic reciprocity, that is, $d = d^*$ and $U_{me} = U_{em} = (1/2)dH\sigma$. In that sense, these expressions must be employed strictly within the linear operating regimes in which the reciprocity assumption is sufficiently accurate. Typical values of the coupling constant are provided in Table 1. Further details can be found in (1,8).

The piezomagnetic model given by Eq. (9) has been augmented by adding saturation phenomena, temperature dependences, and the quadratic relationship between the magnetization and free strains given by Eq. (8). This higher order anhysteretic model has been described elsewhere (4). Like the original formulation from which it derives, this augmented model has the advantage of simplicity in addition to adding certain nonlinear features. In this model, however, complete nonlinear phenomena or hysteretic effects can be addressed only by using complex model parameters. Furthermore, this model addresses material attributes primarily but disregards important transducer effects such as the electrical and dynamic regimes. These effects have been addressed to some degree by coupling the piezomagnetic equations (9) with transducer models such as the linearized canonical equations for electromechanical transducers,

$$\begin{aligned}V &= Z_e I + T_{em} v, \\ F &= T_{me} I + z_m v,\end{aligned}\quad (10)$$

that describe the behavior of a “black box” analog of a transducer system that consists of two coupled electrical and mechanical regimes (60,61). In these equations and Fig. 23,

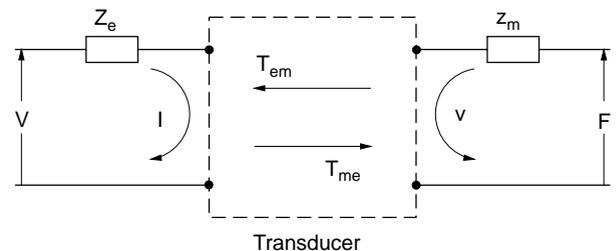


Figure 23. Schematic representation of an electromechanical transducer.

V and I are the voltage and current in the coil, Z_e is the blocked electrical impedance (electrical impedance that is observed when the mechanical system is prevented from moving), F and v are force and velocity on the mechanical side, z_m is the mechanical impedance, and T_{em} and T_{me} represent transduction coefficients that describe the electromechanical coupling. Although this model is phenomenological, certain loss mechanisms based on physical effects such as eddy currents have also been incorporated (1).

Magnetization Models. Magnetization plays a fundamental role in both the magnetostriction that arises from the rotation of magnetic moments (actuator mode) and the change in magnetic anisotropy that occurs when the material strains (sensor mode). From a modeling perspective, the hysteretic phenomena that arise as magnetic moments rotate in response to magnetic fields, stresses, or thermal energy are often the primary magnetization attribute that needs to be quantified. Models for magnetization hysteresis in ferromagnetic materials range from micromechanical models to phenomenological characterizations. Micromechanical models (28,62) describe the coupled magnetoelastic interactions through first-principles formulations that involve the elasticity, thermodynamics, and the electromagnetic energy states. This can lead to highly accurate results at the magnetic domain level, but the large number of parameters required currently prohibits implementing micromechanical models at the system level. The Preisach model has been extensively used to characterize the magnetization of ferromagnetic materials, and more recently of magnetostrictive materials as well (63–66). This model approximates the multivalued hysteretic map by employing a parallel collection of operators. This method has the advantage of intrinsic generality, but typically it requires identifying a large number of nonphysical parameters, and it does not employ or provide insights regarding the dynamics of the system. One advantage of this method is that the Preisach operator can be inverted to facilitate linear control design. Domain rotation models include the classical anisotropic formulation of Stoner and Wohlfarth (18) and later extensions to include cubic anisotropies (67) and compressive loading effects (19). A more complete generalization includes anisotropy, magnetoelastic, and field energies in all three spatial directions (20). Although model results agree well with measured data, identifying the fractional occupancies that define the participation of different easy axes in the total magnetization is rather complex. The domain wall theory of Jiles and Atherton has provided accurate results in ferromagnetic (68), magnetostrictive (30,69), and ferroelectric (70–72) materials. This model is constructed from a thermodynamic difference between the magnetic energy available to magnetize a material and the energy lost as domain walls attach to and detach from inclusions or pinning sites in the material. This yields a set of differential equations that depend on five physical parameters which quantify reversible, irreversible, and total magnetization changes. In the limit when domain wall pinning goes to zero, the Jiles–Atherton model reduces to Langevin’s anhysteretic function, although alternative anhysteretic models that employ

the Isin spin relation have been studied (72). Finally, the Jiles–Atherton model has been employed in combination with Eq. (8) to map the H – λ constitutive relationship (24), and later with dynamical transducer effects to provide a set of reciprocal constitutive relationships among H , M , ε , and σ (30,69). This modeling approach has proven useful in inverse compensation designs (73), characterization of collocated actuation and sensing in transducers (74), and quantification of the ΔE effect in magnetostrictive materials (75).

CONCLUDING REMARKS

Magnetostrictive materials are a class of smart materials that can convert energy between the magnetic and elastic states. The phenomenon of magnetostriction is ultimately due to the coupling between magnetic moment orientation and interatomic spacing, or magnetomechanical coupling. This type of coupling provides a mechanism for bidirectional conversion of energy between magnetic and elastic regimes which is attractive for actuator and sensor systems because it does not degrade in time and it provides almost immediate response. Newer materials such as Terfenol-D or amorphous metallic ribbons exhibit a unique combination of high forces, strains, energy densities, operating bandwidths, and coupling coefficients that justify their use in an ever-increasing number of applications ranging from micropositioners to vibration control systems for heavy machinery. The excellent performance of magnetostrictive materials is sometimes obscured by the large magnetic hysteresis losses and substantial nonlinear effects exhibited by these materials when operated at the high regimes in which their attributes become more prominent. In this sense, realizing the full potential of magnetostrictive materials presents rigorous engineering challenges in a way that other less capable smart materials do not. However, as evidenced by the increasing number of patented devices based on magnetostrictive principles, designers continue to overcome these challenges and make advances in designing, modeling, and controlling magnetostrictive transducers. Furthermore, clever transducer designs are possible solely due to the rich performance space that arises from the otherwise undesirable nonlinear characteristics of these materials. As material advances continue, it is expected that magnetostrictive device designers will find new magnetostrictive solutions for an ever-growing variety of transducer applications.

BIBLIOGRAPHY

1. G. Engdahl ed., *Handbook of Giant Magnetostrictive Materials*. Academic Press, San Diego, 2000.
2. E. du Trémolet de Lacheisserie, *Magnetostriction Theory and Applications of Magnetoelasticity*. CRC, Boca Raton, FL, 1993.
3. T. Cedell, Ph.D. thesis, Lund University, Lund, Sweden, 1995. LUTMDN/(TMMV-1021)/1-222/(1995).
4. T.A. Duenas, L. Hsu and G.P. Carman, in *Advances in Smart Materials Fundamentals and Applications*. Boston, 1996.

5. M. Anjanappa and Y. Wu, *Smart Mater. Struct.* **6**:393–402 (1997).
6. H. Uchida, M. Wada, A. Ichikawa, Y. Matsumara, and H.H. Uchida, *Proc. Actuator 96, 5th Int. Conf. New Actuators*, Bremen, Germany, 1996, pp. 275–278.
7. C. Body, G. Reyne, and G. Meunier, *IEEE Trans. Magn.* **33**(2):1620–1623 (1997).
8. A. E. Clark, in *Ferromagnetic Materials*, E.P. Wohlfarth, ed., North Holland, Amsterdam, 1980, Vol. 1, Chap. 7, pp. 531–589.
9. E.A. Lindgren, J.C. Poret, J.J. Whalen, L.P. Martin, M. Rosen, M. Wun-Fogle, J.B. Restorff, A.E. Clark, and J.F. Lindberg, *U.S. Navy Workshop Acoust. Transduction Mater. Devices*, State College, PA, April 13–15, 1999.
10. J.B. Restorff, M. Wun-Fogle, and A.E. Clark, *U.S. Navy Workshop Acoust. Transduction Mater. Devices*, State College, PA, April 13–15, 1999.
11. R. Tickle, R.D. James, T. Shield, M. Wuttig, and V.V. Kokorin, *IEEE Trans. Magn.* **35**(5):4301–4310 (1999).
12. R.C. O'Handley, *J. Appl. Phys.* **83**(6):3263–3270 (1998).
13. B.D. Cullity, *Introduction to Magnetic Materials*, Addison-Wesley, Reading, MA, 1972.
14. R.M. Bozorth, *Ferromagnetism*. Van Nostrand, New York, 1968.
15. C. Kittel, *Rev. Mod. Phys.* **21**:541–583 (1949).
16. E.W. Lee, *Rep. Prog. Phys.* **18**:184–220 (1955).
17. J.P. Teter, A.E. Clark, and O.D. McMasters, *J. Appl. Phys.* **61**:3787–3789 (1987).
18. E.C. Stoner and E.P. Wohlfarth, *Philos. Trans. R. Soc.* **A240**:599–642 (1948).
19. A.E. Clark, H.T. Savage, and M.L. Spano, *IEEE Trans. Magn.* **MAG-20**(5): pp. 1443–1445 (1984).
20. D.C. Jiles and J.B. Thoeleke, *J. Magn. Magn. Mater.* **134**:143–160 (1994).
21. D.C. Jiles, *Introduction to Magnetism and Magnetic Materials*. 2e, Chapman & Hall, London, 1998.
22. S. Chikazumi, *Physics of Magnetism*. Krieger, Malabar, FL, 1984.
23. M.J. Dapino, A.B. Flatau, and F.T. Calkins, *Proc. SPIE Smart Struct. Mater. 1997*, San Diego, March 1997, Vol. **3041**, pp. 256–267.
24. F.T. Calkins, R.C. Smith, and A.B. Flatau, *IEEE Trans. Magn.* **36**(2):429–439 (2000).
25. D.C. Jiles, *J. Phys. D Appl. Phys.* **28**:1537–1546 (1995).
26. M.J. Sablik and D.C. Jiles, *J. Appl. Phys.* **64**(10):5402–5404 (1988).
27. M.J. Sablik and D.C. Jiles, *IEEE Trans. Magn.* **29**(4): pp. 2113–2123 (1993).
28. R.D. James and D. Kinderlehrer, *Philos. Mag. B* **68**(2):237–274 (1993).
29. V. Agayan, *Physica Scripta* **54**:514–521 (1996).
30. M.J. Dapino, R.C. Smith, L.E. Faidley, and A.B. Flatau, *J. Intelligent Mater. Syst. Struct.* **11**(2): 135–152 (2000).
31. A.B. Flatau, M.J. Dapino, and F.T. Calkins, *Proc. SPIE Smart Struct. Mater.* San Diego, March 1998, Vol. **3327**, pp. 463–473.
32. F.T. Calkins and A.B. Flatau, *Proc. SPIE Smart Struct. Mater. 1996*, San Diego, March 1996, Vol. **2717**, pp. 709–719.
33. A.B. Flatau, F. Pascual, M.J. Dapino, and F.T. Calkins, CATD-IPIRT Contract #95-05, October 1996.
34. F.T. Calkins, M.J. Dapino, and A.B. Flatau, *Proc. SPIE Smart Struct. Mater. 1997*, San Diego, CA, Vol. **3041**, pp. 293–304.
35. G.A. Steel, in *Transducers for Sonics and Ultrasonics*, Technomic, Lancaster, PA, 1993, pp. 250–258.
36. J.B. Restorff, in *Encyclopedia of Applied Physics*, VCH, New York, 1994, Vol. **9**, pp. 229–244.
37. M.J. Dapino, F.T. Calkins, and A.B. Flatau, in *22nd Encyclopedia of Electrical and Electronics Engineering*, J.G. Webster, ed., Wiley, NY, 1999, Vol. **12**, pp. 278–305.
38. L. Kiesewetter, *Proc. 2nd. Int. Conf. Giant Magnetostrictive Alloys*, Marbella, Spain, October 12–14, 1988.
39. R.C. Roth, *Proc. 3rd Int. Conf. New Actuators*, Bremen, Germany, 1992, pp. 138–141.
40. J.H. Goldie, M.J. Gerver, J. Kiley, and J.R. Swenbeck, *Proc. SPIE Smart Struct. Mater. 1998*, San Diego, March 1998, Vol. **3329**, pp. 780–785.
41. W. Chen, J. Frank, G.H. Koopmann, and G.A. Lesieutre, *Proc. SPIE Smart Struct. Mater. 1999*, Newport Beach, CA, March 1999.
42. T.T. Hansen, Technical report, Chemtech, American Chemical Society, 1996, pp. 56–59.
43. J.R. Frederick, *Ultrasonic Engineering*. Wiley, NY, 1965.
44. T. Akuta, *Proc. 3rd. Int. Conf. New Actuators*, Bremen, Germany, 1992, pp. 244–248.
45. J.M. Vranish, D.P. Naik, J.B. Restorff, and J.P. Teter, *IEEE Trans. Magn.* **27**:5355–5357 (1991).
46. F. Claeysen, N. Lhermet, and R.L. Letty, *IEEE Trans. Magn.* **32**(5):4749–4751 (1996).
47. J.L. Butler, S.C. Butler, and A.L. Butler, *J. Acoust. Soc. Am.* **94**:636–641 (1993).
48. J.E. Miesner and J.P. Teter, *Proc. SPIE Smart Struct. Mater. 1994*, Orlando, FL, 1994, Vol. **2190**, pp. 520–527.
49. B. Clephas and H. Janocha, *Proc. SPIE Smart Struct. Mater. 1997*, San Diego, March 1997, Vol. **3041**, pp. 316–327.
50. R. Venkataraman, W.P. Dayawansa, and P.S. Krishnaprasad, Technical report, CDCSS, University of Maryland, College Park, MD, 1998.
51. I.J. Garshelis, *IEEE Trans. Magn.* **28**(5):2202–2204 (1992).
52. I. Sasada, N. Suzuki, T. Sasaoka, and K. Toda, *IEEE Trans. Magn.* **30**(6):4632–4635 (1994).
53. M. Wun-Fogle, H.T. Savage, and M.L. Spano, *J. Mater. Eng.* **11**(1):103–107 (1989).
54. M.D. Mermelstein and A. Dandridge, *Appl. Phys. Lett.* **51**(7):545–547 (1987).
55. A. Yariv and H. Windsor, *Opt. Lett.* **5**:87 (1980).
56. A. Dandridge, K.P. Koo, F. Bucjolls, and A.B. Tveten, *IEEE Trans. Magn.* **MAG-22**:141 (1986).
57. R. Chung, R. Weber, and D.C. Jiles, *IEEE Trans. Magn.* **27**(6):5358–5360 (1991).
58. J. Seekercher and B. Hoffmann, *Sensors and Actuators* **A21–A23**:401–405 (1990).
59. D.A. Berlincourt, D.R. Curran, and H. Jaffe, in *Physical Acoustics, Principles and Methods*, W.P. Mason ed., Academic Press, NY, 1964, Vol. **1**, Part A.
60. F.V. Hunt, *Electroacoustics: The Analysis of Transduction and Its Historical Background*. American Institute of Physics for the Acoustical Society of America, 1982.
61. D.L. Hall, Ph.D. Dissertation, Iowa State University, Ames, Iowa, 1994.
62. W.F. Brown, *Magnetoelastic Interactions*. Springer-Verlag, Berlin, 1966.
63. I.D. Mayergoyz, *Mathematical Models of Hysteresis*. Springer-Verlag, NY, 1991.

64. J.B. Restorff, H.T. Savage, A.E. Clark, and M.Wun-Fogle, *J. Appl. Phys.* **67**(9):5016–5018 (1990).
65. A. Reimers and E. Della Torre, *IEEE Trans. Magn.* **35**:1239–1242 (1999).
66. R.C. Smith, *J. Math. Syst. Estimation Control* **8**(2):249–252 (1998).
67. E.W. Lee and J.E. Bishop, *Proc. Phys. Soc. London* **89**:661 (1966).
68. D.C. Jiles and D.L. Atherton, *J. Magn. Magn. Mater.* **61**:48–60 (1986).
69. M.J. Dapino, R.C. Smith, and A.B. Flatau, *IEEE Trans. Magn.* **36**(3):545–556 (2000).
70. R.C. Smith and C.L. Hom, *J. Intelligent Mater. Syst. Struct.* **10**(3):195–213 (1999).
71. R.C. Smith and Z. Ounaies, *J. Intelligent Mater. Syst. Struct.* **11**(1):62–79 (2000).
72. R.C. Smith, Smart structures: model development and control applications. Technical Report CRSC-TR01-01, Dept. of Mathematics, North Carolina State Univ., Raleigh, NC, 2001.
73. R.C. Smith and R.L. Zrostlik, *Proc. 1999 IEEE Conf. Decision and Control*, Phoenix, AZ, December 7–10, 1999.
74. M.J. Dapino, F.T. Calkins, R.C. Smith, and A.B. Flatau, *Proc. ACTIVE 99*, Ft. Lauderdale, FL, December 2–4, 1999, Vol. 2, pp. 1193–1204.
75. R.C. Smith, M.J. Dapino, and A.B. Flatau, *Proc. SPIE Smart Struct. Mater. 2000*, Newport Beach, CA, March 6–9, 2000, Vol. **3985**, pp. 174–185.
76. D.C. Jiles, *Introduction to the Electronic Properties of Materials*. Chapman & Hall, London, 1994.
77. R.A. Kellogg and A.B. Flatau, *Proc. SPIE Smart Struct. Mater. 1999*, Newport Beach, CA, March 1999, Vol. **3668**.
78. A.E. Clark, J.P. Teter, M. Wun-Fogle, M. Moffett, and J. Lindberg, *J. Appl. Phys.* **67**(9): pp. 5007–5009 (1990).
79. F.T. Calkins, Ph.D. Dissertation, Iowa State University, Ames, Iowa, 1997.

MICROROBOTICS, MICRODEVICES BASED ON SHAPE-MEMORY ALLOYS

YVES BELLOUARD

Institut de Systèmes Robotiques
Ecole Polytechnique Fédérale de Lausanne
Switzerland

INTRODUCTION

For the past few years, consumer products such as computers, mobile phones, and cameras have drastically decreased in size and their functionalities have increased. In addition, mini-invasive techniques in surgery have led to a growing need for small, highly reliable components that can go through arteries or veins. In general, there are nearly no highly technological products that do not benefit from miniaturization. However, one cannot simply scale down electrical or mechanical components from “macroscale” because physical forces do not scale down at the same amplitude. Therefore, the efficiency of actuators depends on their size. Consequently, there is a need for new actuating technologies adapted to the microworld

that can replace usual actuators such as electromagnetic motors. Among these technologies, smart materials such as piezoceramics, magnetostrictive, electrostrictive, and shape-memory materials are of particular interest. This article describes the use of shape-memory alloys (SMA) for microengineering applications.

In the first part, a brief description of SMA properties is given. The second part addresses some usual design principles. Finally, a literature survey and a concept of monolithic microdevices concludes the article.

SHAPE-MEMORY ALLOYS (SMA): A SUMMARY OF THEIR PROPERTIES

Properties of SMA: A Brief Description

Three main functionalities can be associated with the martensitic transformation: the shape-memory effect, superelasticity, and high damping capability. Special care is taken to describe the shape-memory effect, which is the essence of the SMA actuator. For a thorough understanding, the reader may refer to other sections of this encyclopedia. General books about SMA material may also be consulted (1–3).

The shape-memory effect. This happens when a material, previously deformed in martensite—the low-temperature phase—recovers its original shape when heated up to the austenite—the high-temperature phase. Figure 1 illustrates the shape-memory effect. The graph on the left side represents the stability zone for the two phases in a stress–temperature representation. According to the Clausius–Clapeyron relationship between stress and temperature, the oblique lines indicate the boundary between phases and the transition period. The martensitic transformation occurs across a given range of temperature (M_s to M_f , from austenite to martensite and A_s to A_f , from martensite to austenite).

Let us consider an example: a strip that is flat in the austenite phase, that is, the “memorized shape” (step 1 in Fig. 1). If a stress is applied below the martensite finish temperature (M_f) and if this stress is higher than the critical stress to detwin martensite, the variant reorientation occurs, and the strip is deformed in a plasticlike way (step 2). This means that when the stress is released, almost all of the deformation remains. When the material is heated above A_f (step 3), the material transforms to austenite and recovers its original shape. Repetitive cooling (step 4) does not cause any shape change.

The term “one-way, shape-memory-effect” is often used to address this effect, because it is a one-time occurrence; the material keeps its original austenitic shape upon further cooling cycles below M_f . Because nonoriented martensite is created upon cooling, martensitic structures have the same specific macroscopic volume as the austenitic crystal. A plasticlike deformation of martensite (“detwinning”) is required to observe a shape change while heating up to the austenite. This phenomenon can be well understood using a one-dimensional phenomenological model

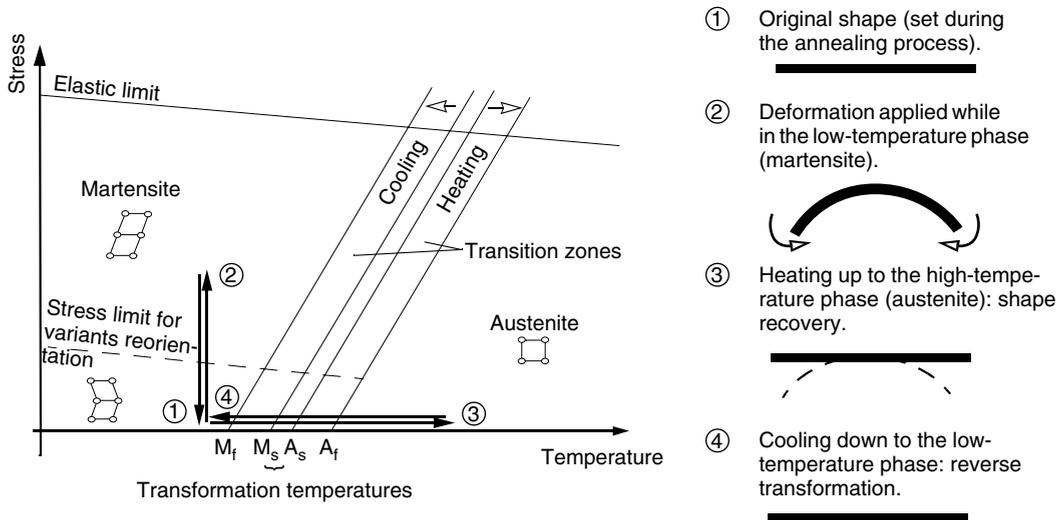


Figure 1. An example of the one-way shape-memory effect: step-by-step description and representation in a stress–temperature diagram.

(Fig. 2). This description is considerably simplified, but it conveys some of the fundamental aspects of the shape-memory effect.

Two martensitic variants are symbolically represented by vertical and horizontal rectangles (M1, M2). The material is assumed to be free of any internal stress. When there is no stress (set 1), the martensitic variants are randomly

distributed: each variant has the same probability. When applying a force (set 2), the variants are reoriented to minimize the stress within the structure. Therefore, the distribution between the two variants is modified. When heated because the austenite that is symbolically represented by a square has higher symmetry than martensite, the two martensitic variants are transformed into the same

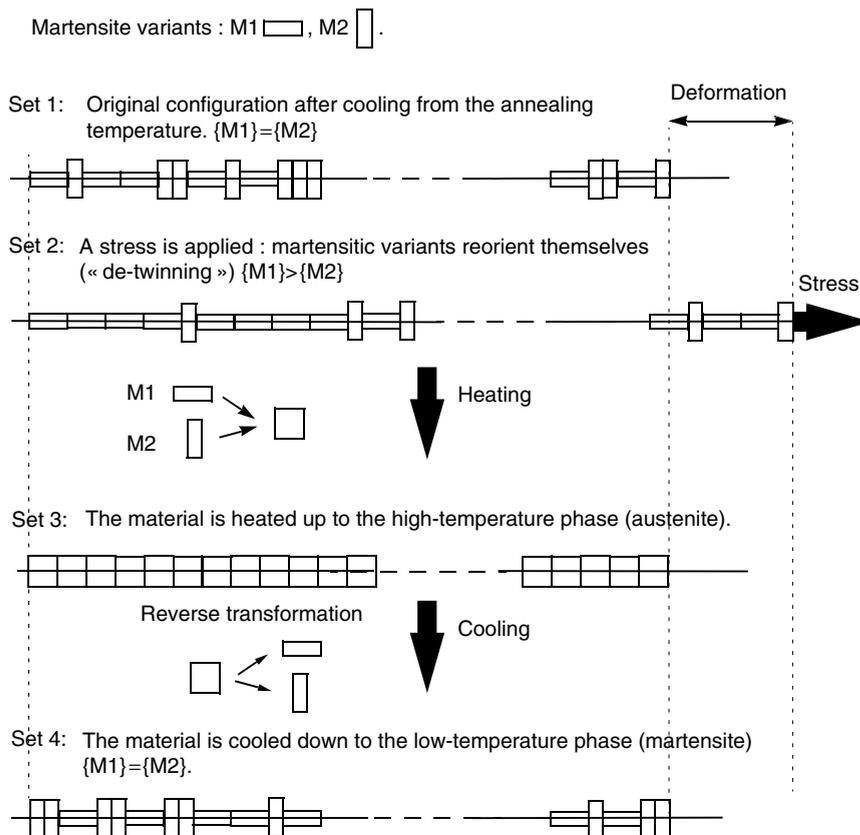


Figure 2. A one-dimensional phenomenological description of the one-way shape-memory effect. The twinned martensitic variant occupies the same specific volume as the austenitic.

“crystallographic” structure, the so-called parent phase: whatever the set of martensite variants is (set 1 or set 2, for instance), the parent shape will be the same. On cooling (set 4), if no stress is applied, the two martensitic variants are equally distributed: the macroscopic shape will remain the same as the austenitic shape. This nonreversibility, from the point of view of change of shape, is the basis for understanding the design of shape-memory actuators. As we see later, all of the design concepts deal with methods for obtaining a reversible effect.

The Superelastic Effect. This occurs when the martensitic transformation is induced by applying a stress at a constant temperature. Figure 3 shows a schematic representation of the superelastic transformation on stress–temperature and stress–strain graphs. The material in its austenitic state is loaded at a constant temperature. The mechanical transformation is purely elastic until the stress reaches a critical level where the transformation starts. At this point, the stress remains constant, and the strain is still increasing. This region, usually called the stress “plateau” extends until the strain has reached typically 8% and even 15% or more for single crystals. When all of the austenite has been transformed into martensite, another elastic domain can be observed that corresponds to the elasticity of the martensite. When unloading, the material displays hysteresis. It is very important to note that the stress level of the plateau is temperature-dependent.

Dissipation Mechanism and Damping

Due to a large density of mobile interfaces between martensitic variants and the austenitic matrix during martensitic transformation, the material displays a very efficient energy dissipation mechanism. Moreover, in martensite, these alloys have very high damping capacity due to internal friction of martensitic variants. Unlike other high-damping metals, the dissipation mechanism in SMAs

does not significantly affect other important mechanical parameters of the material. The high damping capacity of SMA has been successfully used in composite structures that have embedded shape-memory fibers (4) as well as in developing a competition ski (5):

Material Selection (see “Shape Memory Alloy, Types and Functionalities” in this Encyclopedia for an in-depth Description)

Among all of the shape-memory alloys, three main types are commercially available: Ni–Ti–X (where X is an additional component such as Cu, Pd, Hf), Cu–Zn–Al, and Cu–Al–Ni. Ni–Ti alloys are the most interesting for many reasons: they are ductile (about 30 to 40%), biocompatible (binary alloys), and they have high electrical resistivity, which allows efficient actuation by Joule heating.

To appreciate the mechanical properties of Ni–Ti alloys, Table 1 compares Ni–Ti alloys to stainless steels. Ni–Ti alloys have other useful properties such as high wear resistance (6) and good corrosion resistance. Moreover, some Ni–Ti alloys have an additional martensitelike transition: the so-called R-phase transition (1,2). The R-phase transition appears upon cooling before the martensitic transformation under certain conditions of material fabrication and processing. This phase has a narrow hysteresis (nearly 1.5°C) and very good thermal cycling stability. For these reasons, the R-phase has been used in many actuator applications. However, the strain is fairly small (about 1%).

SMA Compared to Other Smart Materials and Bimorph Structures

Criteria for comparison between smart materials obviously depend on the application specifications, functionalities, and technological aspects such as machinability, miniaturization capability and control properties.

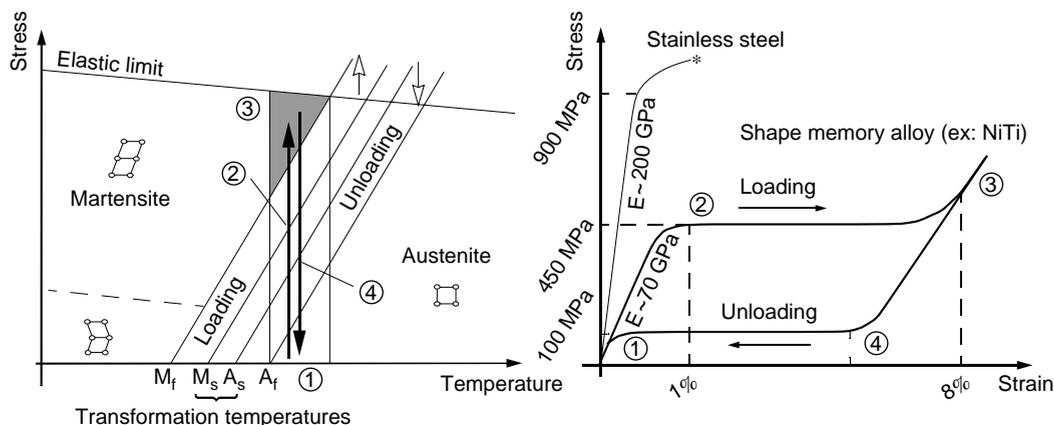


Figure 3. Superelastic behavior: representation on stress–temperature (left) and stress–strain graphs (right). The hatched zone represents the superelastic window, that is the temperature window where the material is fully transformed to martensite when loading and where no residual strain due to incomplete transformation is observed when unloading. Datas displayed on the graph are for a binary NiTi and a stainless steel.

Table 1. Mechanical Properties of Ni-Ti Compared to Stainless Steel^a

	Ni-Ti		Typical Data for a Stainless Steel	
Max. reversible elastic deformation	Typ. 8%		0.8%	
Mass density	6450 kg.m ⁻³	12.52 slugs.ft ⁻³	7850 kg.m ⁻³	15.2 slugs.ft ⁻³
Young's modulus (<i>E</i>)	M: 28–41 Gpa A: 83 GPa	M: 4–6 × 10 ³ ksi A: 12 × 10 ³ ksi	190–210 GPa	28–30 10 ³ ksi
Shear modulus (<i>G</i>)	M: 10–15.5 GPa A: 31 GPa	M: 1.5–2.2 × 10 ³ ksi A: 4.5 × 10 ³ ksi	75–80 GPa	11–12 10 ³ ksi
Poisson's ratio (<i>ν</i>)	0.33		0.27–0.30	
Yield stress ^b	A: 195–690 MPa M: 70–140 MPa	A: 28–100 ksi M: 10–20 ksi	400–1600 ^c MPa	60–240 ^c ksi
Ultimate stress	895–1900 MPa	130–275 ksi	700–1900** MPa	100–270** ksi
Coefficient of thermal expansion	A: 11 × 10 ⁻⁶ /°C M: 6.6 × 10 ⁻⁶ /°C	A: 6.11 × 10 ⁻⁶ /°F M: 3.67 × 10 ⁻⁶ /°F	8–10 10 ⁻⁶ /°C	4.5–5.5 10 ⁻⁶ /°F

^aThe data on Ni-Ti, are taken from (2,7). The data for stainless steel are taken from (8). "A" refers to austenite and "M" refers to martensite.

^bThe yield stress for the SMA is not really a yield stress but rather a critical stress to induce martensite when in the austenitic state and a critical stress to reorient martensitic variants when in the martensitic state.

^cThese values apply typically to a spring steel.

However, comparing several physical properties of different smart materials allows a quick assessment of some of the major advantages and drawbacks of shape-memory materials (Table 2). The energy density of SMA is the highest. However, the bandwidth is rather small compared to magnetostrictive and piezoelectric material, for instance. Considering the working mode, SMA can be loaded in bending, torsion, tension, and compression. Unlike bimorph actuators or conducting polymers where the bending response is a consequence of the actuator structure, the stress/strain response will depend on the direction of the applied strain. Piezoelectric and giant magnetostrictive also have different working modes that will correlate with polarization and the applied field. Conducting polymers can achieve higher strain than SMA materials. However, Young's modulus of conducting polymers is rather low compared to SMA.

DESIGNING SMA ACTUATORS: GENERAL PRINCIPLES

Introduction

Now, it is clear that SMAs have some very interesting properties that can be used for actuating. Their unique behavior requires some appropriate design principles based on material properties.

Because the shape-memory effect is a one-time occurrence, it is not sufficient to create a reversible actuator. To bypass this intrinsic property, two main strategies can be pointed out:

- The first is to find a way to make the material reversible by itself without any external "help." This requires thermomechanical treatment or specific manufacturing processes.

Table 2. Comparison Between SMA and Other Smart Materials and Bimorph Actuators

	SMA	Piezoelectric	Giant Magnetostrictive	Conducting Polymers (11)	Bimorph Actuators
Physical phenomenon	Martensitic transformation	Piezoelectricity	Magnetostriction	Electrochemical doping	Differential thermal expansion coefficient
Actuation principle	Thermal	Electrical field	Magnetic field	Voltage	Thermal
Energy density (J m ⁻³)	~10 ⁶ to 10 ⁷	~10 ² (PZT) ~10 ³ (PMN) (9)	~10 ⁴ –10 ⁵ (Terfenol D) (9)	~10 ³	~10 ⁵ (Ni / Si) (10)
Bandwidth (order of magnitude)	Low ^a (10 ² Hz)	High (100 kHz)	High (100 kHz)	High (10 kHz)	Low ^a (10 ² Hz)
Working mode	Bending, torsion, tension, compression	Depends on the electrical field direction	Depends on the magnetic field direction	Electrolyte storage scheme	Bending
Typical strain	1–8% ^b	0.12–0.15%	0.58–0.81%	>10%	~5, 23 10 ⁻⁴ %/C

^aStrongly size and shape dependent.

^bDepends on the lifetime specification. For single crystals, the maximum strain can increase up to 15%.

- The second consists of using an additional element, usually called a biasing element, which will be used as a buffer to provide the required energy to reorient the martensite upon cooling and thus “reset” the actuator.

Another important thing to notice is the complexity of the thermomechanical behavior of these alloys: they have non-linear, dissipative, and temperature-dependent mechanical behavior. Therefore, usual design methods based on linear mechanics of structure, for instance, Castigliano's theorem and the superposition principle, are valid only in the elastic domain of the mechanical characteristic. Nevertheless, the isothermal behavior can be simulated reasonably accurately by using an elastoplastic model. In addition, several models, which consider thermomechanical coupling, have also been proposed in the literature [see, for instance, (12–14)].

Study of Some Loading Modes

The shape-memory effect can be used in any direction of space. This means that an SMA actuator can work under tensile–compressive stress, bending load, torsion load, or any combined mode. This kind of isotropy greatly enhances the design capabilities of these materials. The trade-off between one loading mode and another can be done, depending on actuator specification such as stroke, force, volume, efficiency, etc.

In the next section, we examine briefly some usual loading modes. A simple way to convey isothermal mechanical behavior is to consider elastoplasticlike behavior, such as that shown in Fig. 4.

This kind of approximation can be used for a superelastic curve as well as a martensitic curve. In the martensitic case, the critical stress will refer to that required to induce a reorientation of martensitic variants, above A_f , and the critical stress will refer to the plateau's stress.

Tension. This is the simplest loading mode; the material is axially loaded. The normal stress is uniformly distributed within the material. The force is directly

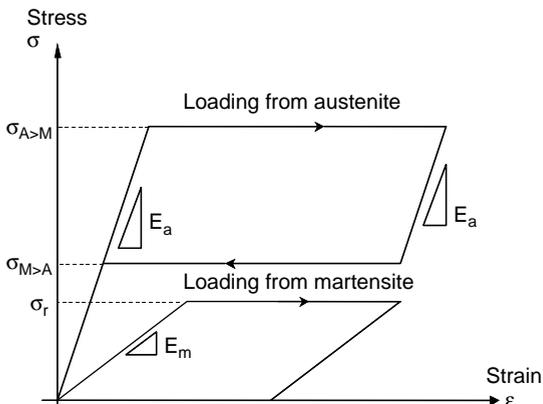


Figure 4. Isothermal elastoplastic model of the SMA mechanical characteristics.

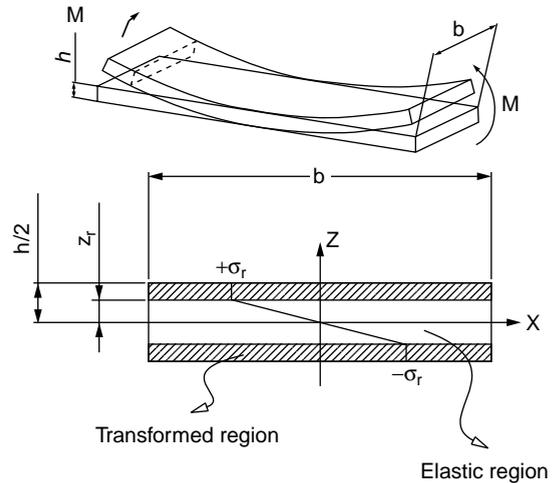


Figure 5. Pure bending loading mode.

proportional to the normal stress. In term of efficiency of energy transformation, this is the optimal mode because all of the material contributes to the transformation. Nonetheless, the motion range is very small.

Bending. Stress distribution in the bending mode is not constant across the material. Figure 5 shows a cantilever subjected to bending deformation. The outer part of the material is in tension and the inner part is in compression. The strain on the neutral fiber is equal to zero, and in a first approximation, the position of the neutral fiber can be mixed with the center of gravity. To characterize the stress distribution, the deformation process during loading can be divided into periods of different deformation mechanisms (8,15) (see Fig. 5). The first period is determined by pure linear-elastic deformation behavior until the stress reaches a critical value corresponding to the beginning of the stress-induced martensitic transformation. The second period is a combination of elastic deformation and ideal-plastic deformation. In a first approximation, the behavior upon loading can be approximated by elastoplasticlike behavior. In this case, the tensile behavior, it is assumed, is the same as compressive behavior. However, Liu et al. (16) showed that this assumption is not true for Ni–Ti alloys: there is no real “plateau” in compression but rather a decrease of the slope of the stress–strain curve. Therefore, Eq. (1) is a rough approximation. The accuracy of this model can be enhanced by adding a strain hardening condition. Let σ_R be the critical stress to induce a reorientation of the martensite, where a cantilever beam of thickness h and width b is considered. Let R be the radius of curvature and z the relative thickness measured from the neutral line which, in a first approximation, can be considered the center of the strip. The strain is a function of z and is equal to z/R . If it is assumed that the cantilever is in the austenitic state, the stress within a section of a cantilever beam can be expressed by

$$\sigma(z) = \begin{cases} E \frac{z}{R} & z \leq z_R \\ \sigma_R & z > z_R \end{cases}$$

where

$$-\frac{h}{2} \leq z \leq \frac{h}{2}, \quad (1)$$

where z_R is the boundary between the elastic and the transformed zones.

Therefore, the bending moment can be written

$$M = \int_A z \sigma(z) dA = \frac{2}{3} \frac{Eb z_R^3}{R} + b \sigma_R \left(\frac{h^2}{4} - z_R^2 \right). \quad (2)$$

The first term of the bending moment expresses the elastic contribution, and the second term denotes the plasticlike mechanical behavior of the transformed region.

Assuming $z_R = \sigma_R R / E$, the bending moment can be written as a function of the radius of curvature:

$$M(R) = \begin{cases} \frac{bh^2}{4} \sigma_R - \frac{b\sigma_R^3 R^2}{3E^2}, & R \leq \frac{Eh}{2\sigma_R} \\ \frac{bh^3}{12} \frac{E}{R}, & R > \frac{Eh}{2\sigma_R} \end{cases}. \quad (3)$$

Torsion. As in bending, the stress distribution is not uniform (see Fig. 6). Consider a wire to which a pure torsion moment is applied: the middle of the wire remains in the elastic region, and the first transformed region is on the edge. Let M_t be the torsional moment, τ the shear stress, G the shear modulus, r the radius along the cross section of the wire, R the wire radius, and θ the twist angle per unit length.

Assuming elastoplastic behavior, the shear stress is

$$\tau = \begin{cases} Gr\theta, & 0 \leq r \leq r_T \\ \tau_R, & r_T < r \leq R \end{cases}, \quad (4)$$

where τ_R is the transition between the elastic region and the transformed region. The torsional moment Γ is:

$$\begin{aligned} \Gamma &= \int_A \tau r dA = 2\pi \int_0^R \tau r^2 dr \\ &= 2\pi \left[\int_0^{r_T} (Gr\theta) r^2 dr + \int_{r_T}^R \tau_R r^2 dr \right]. \end{aligned} \quad (5)$$

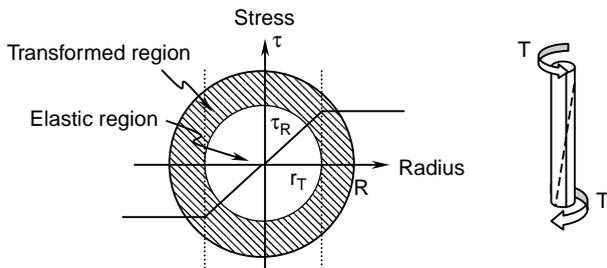


Figure 6. Loading mode in pure torsion.

The first integral expresses the elastic contribution, and the second expresses the contribution of the transformed region. Thus, the torsional moment can be written as a function of the twist angle φ :

$$\Gamma(\varphi) = \begin{cases} \frac{\pi r^4}{2} \frac{G\varphi}{L}, & \varphi \leq \frac{\tau_R L}{GR} \\ 2\pi \left(\tau_R \frac{R^3}{3} - \frac{1}{12} \frac{L^3 \tau_R^4}{G^3 \varphi^3} \right), & \varphi > \frac{\tau_R L}{GR}, \end{cases} \quad (6)$$

where L is the length of wire.

Application Example: Equation of a SMA Helical Spring.

Let us consider a spring whose outer radius is ρ , wire radius is R , and has n coils. Considering the expression of torsion from (Eq. 6), the moment becomes

$$\Gamma_t = P \frac{\rho}{2} = \begin{cases} \frac{\pi r^4}{2} \frac{G\varphi}{L}, & \varphi \leq \frac{\tau_R L}{GR} \\ 2\pi \left(\tau_R \frac{R^3}{3} - \frac{1}{12} \frac{L^3 \tau_R^4}{G^3 \varphi^3} \right), & \varphi > \frac{\tau_R L}{GR}. \end{cases} \quad (7)$$

The deflection can be expressed by

$$\delta = \rho L \varphi = 2\pi n \rho^2 \varphi. \quad (8)$$

Substituting Eq. (8) for φ in Eq. (7), we find the equation for an SMA helical spring:

$$P = \begin{cases} \frac{Gr^4}{4n\rho^3} \delta & \delta < \delta_e \\ \frac{2\pi \tau_p r^3}{3\rho} \left[1 - 2 \left(\frac{\pi n \tau_p \rho^2}{G \delta r} \right)^3 \right] & \delta \geq \delta_e, \end{cases}$$

where

$$\delta_e = \frac{\rho \tau_p}{GR}. \quad (9)$$

Tobushi and Tanaka (17) showed that this method provides reasonably accurate results in analyzing a SMA helical spring.

Creating a Reversible Actuator: The External Way

Several studies, for instance, Hirose et al. (18), Waram (1), and Liang and Rogers (13), have been conducted to give basic principles for SMA actuator design. This section gives an overview of these basic design principles.

Constant Force Loading and Transformation under Constraint. The simplest actuator is one working under constant force loading or deadweight (Fig. 7). The stroke is given by the intersection between austenitic and martensitic mechanical characteristics and the straight line corresponding to the force level. If the actuator is mechanically constrained during the transformation to the high-temperature phase, recovery stresses are generated at a constant stress rate (2) until the stress reaches the

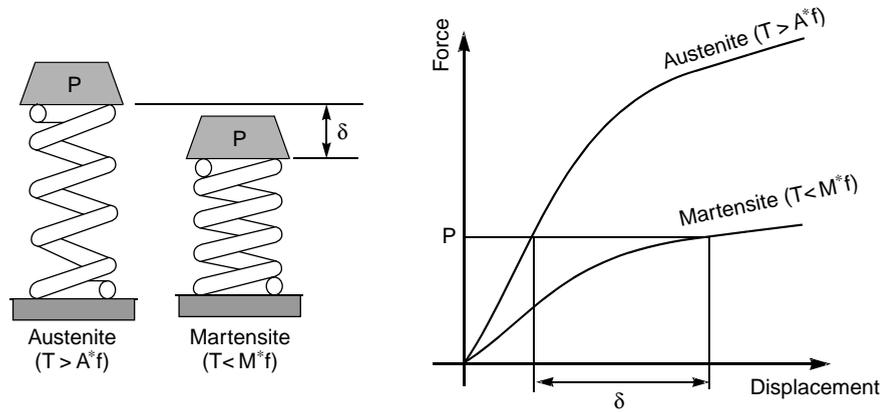


Figure 7. Constant force loading of a SMA actuator. A^*f and M^*f are the end-of-transformation temperatures for a given load P .

yield strength of the material. Using the constrained recovery property, one can design a fixture that has a grasping force nearly independent of the tolerances of the grasped object. In addition, the stress is well defined. A detailed and clever presentation of constrained recovery can be found in Duerig and Proft (1).

Figure 8 illustrates this phenomenon: for a stainless steel fixture, the applied force will be variable depending on the stiffness of the fixture and the tolerances of the object. However for SMA fixtures that have at least 1.5% of unresolved recovery, the force remains nearly constant whatever the object's tolerances are.

Spring-Biased Actuator ("Pullback Spring Design"). In this configuration, a passive mechanical element is used to apply a force on the SMA element. This mechanical element or "bias spring" has a known elastic characteristic.

Figure 9 illustrates the principle of this actuator. The SMA element and the bias spring are symbolically represented by two helical springs attached together. The output of the mechanism is symbolically represented by a black rectangle. It is clear that this concept is general and

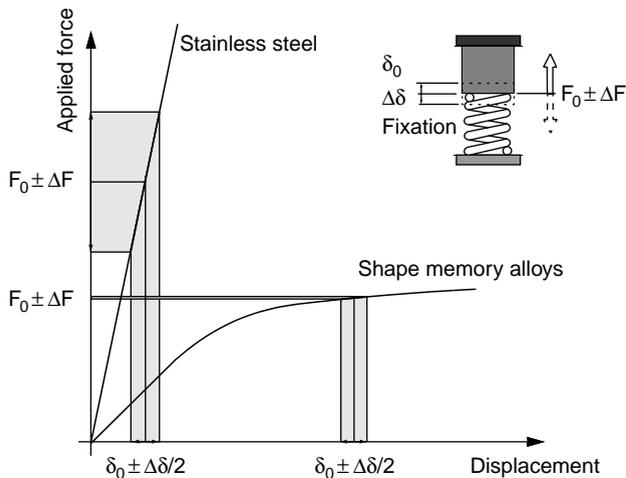


Figure 8. Illustration of the constant clamping force obtained using the constrained-recovery property of an SMA material.

applies for any kind of SMA design and bias spring characteristics. During the assembly process, the SMA is deformed and prestressed in its martensitic phase (distance OO' in Fig. 9). When the SMA element is heated up to the austenitic temperature, the material tries to recover its original shape and thus generates a recovery force. This force pulls the bias spring which stores some elastic energy. When cooling, the reverse transformation occurs resulting in a change in the mechanical characteristics within the SMA (a decrease in Young's modulus and nucleation of martensitic variants). Thus, the elastic energy stored by the bias spring is released and induces the reorientation of martensitic variants resulting in macroscopic deformation of the SMA element. The two equilibrium points (A and B) can be found by using the corresponding isothermal loading curves of the SMA element.

Antagonistic or "Push-Pull Design". The push-pull actuator consists of two SMA elements that move in opposite directions. A design is shown on the left side of Fig. 10. In this example, it is assumed that the two SMA springs are identical. The motion reference is taken between the two SMA actuators. As for the spring-biased actuator, the structure is prestressed to deform each SMA element in martensite. The principle is to heat up one actuator at a time. The heated element pulls the mobile part in its direction and deforms its counterpart. Inverting the actuation between the heated SMA element and the cool element has the reverse motion. The graph on the right side of Fig. 10 represents the mechanical characteristics of the two actuators: the intersection between the martensitic characteristic of one of the actuators and the austenitic characteristic of the other gives the two extreme positions of the linear stage. Between these two points, different paths are available, depending on heating/cooling strategy. For example, one can choose to supply one actuator at a time or to modulate the power applied to the two actuators at the same time: the dynamic behavior will be different.

Creating a Reversible Actuator: The Internal Way

The Two-Way Shape-Memory Effect (Induced by a Training Process). The "two-way" shape-memory effect (TWSME) (1,2,19), in contrast with the "one-way" shape-memory

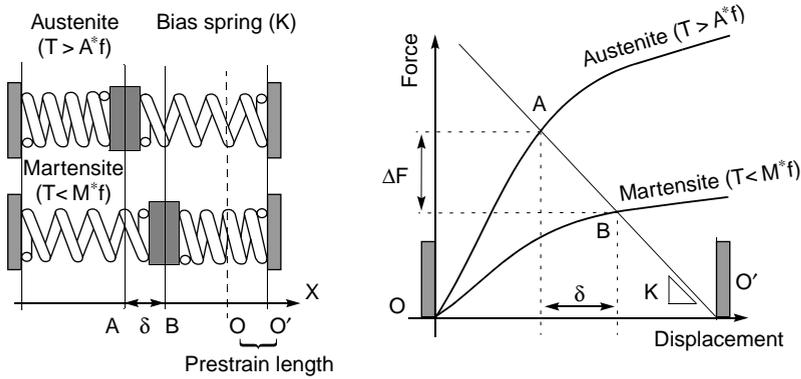


Figure 9. Design characteristics of a SMA actuator working against a bias spring. K denotes the stiffness of the bias spring. A_f^* and M_f^* are the end-of-transformation temperatures for a given prestrain length (OO') and stiffness K .

effect, is reversible. This means that the material “memorizes” a shape when cooling, resulting in a spontaneous shape change without any external stress. However, this effect is not intrinsic to the alloy properties: as mentioned by Perkins and Hodgson in (1), this is “learned” behavior. In other words, the material must be subjected to special thermomechanical treatments to exhibit this strange effect. The term “training process” is often used to describe these processes. The TWSME can be explained by the presence of internal stress that results from the training process and creates a given distribution of martensitic variants upon cooling. In the opposite of the one-way effect, the TWSME, preferred oriented martensitic variants are created. The TWSME usually has strain amplitude between 1 and 2%. However, the force generated when heating is much larger than when cooling. This drawback limits the TWSME to specific applications. In the next sections, other methods, that have a reversible effect, are presented. Nevertheless, for better understanding, the term TWSME is reserved throughout this article to the reversible motion induced by training processes.

All-Round Effect [20]. This effect has been observed in Ni-rich Ti–Ni alloys (Ni content > 50.6 at%) (Fig. 11): a flat strip is annealed under constraint into a round shape at a temperature where formation of precipitates is possible (typically above 673 K). After this heat treatment, the specimen has a ring shape in austenite and an

opposite ring shape in martensite so that the inner part of the ring becomes the outer part and vice versa. Kainuma et al. (20) described the mechanism of this phenomenon: oriented precipitates are created within the structure. Under compression, these precipitates have a given orientation, which is perpendicular to the orientation of the precipitates formed in tension. When the specimen is annealed in a round shape, the inner part of the strip is in compression, and the outer part is in tension, resulting in two different groups of oriented precipitates that are perpendicular to each other. When cooling, due to the internal stress induced by precipitates, preferred oriented martensite variants appear, leading to the opposite round shape.

Irradiation. Neutron irradiation can suppress martensitic transformations in Ti–Ni alloys (21). By controlling this effect, one can create a cantilever beam, for example, half of whose thickness is irradiated, and the other half is unaffected. The lattice in the irradiated zone has been altered and does not exhibit any martensitic transformation. Using the irradiated part as the biasing element and the unaffected zone as the actuator allows creating a reversible motion. In a certain way, a bimorphlike structure is created, and the resulting motion is a bending motion (21).

Special Production Methods. Some fabrication processes can create reversible effects in SMA. For instance, melt

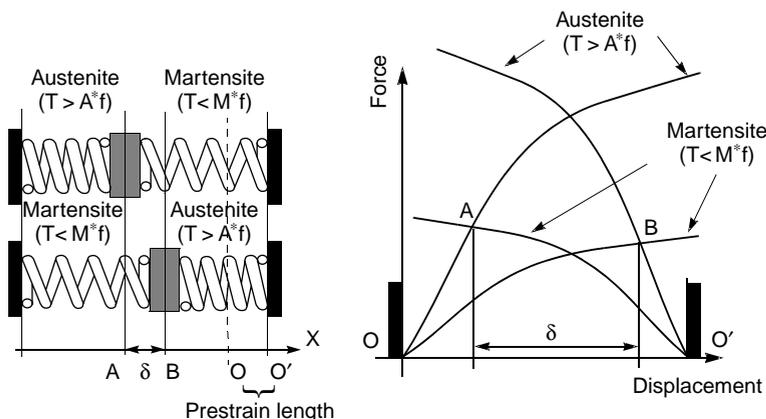


Figure 10. Characteristics of an antagonistic design of SMA actuators. A_f^* and M_f^* are the end-of-transformation temperatures for a given prestrain (OO') and mechanical characteristic of an SMA actuator.

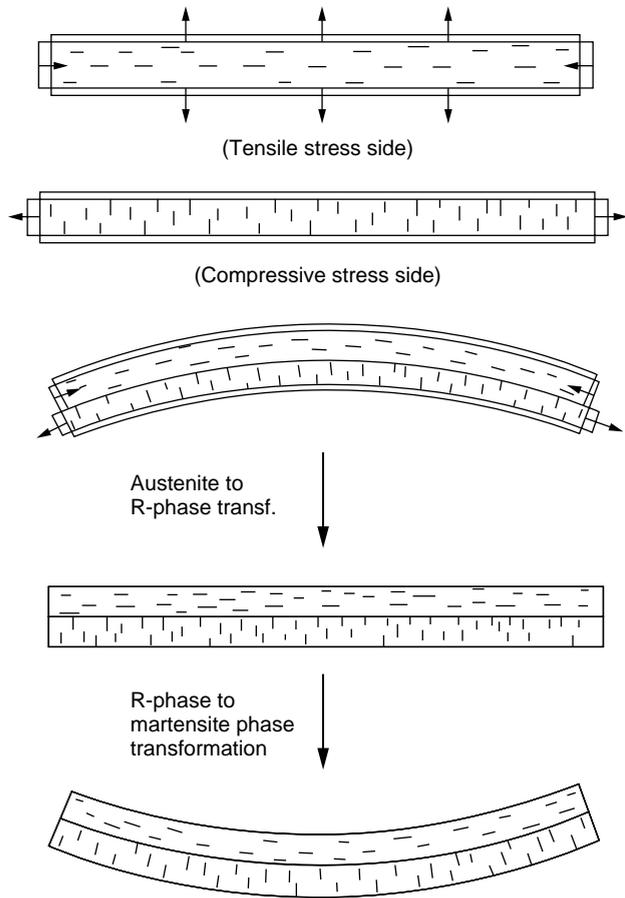


Figure 11. Mechanism of the all-round shape-memory effect [after Kainuma et al. (20)].

spinning produces an intrinsic reversible effect under certain conditions (22). Another example is the fabrication process of Ni–Ti shape-memory films developed by Lehnert et al. (23). The method consists of successively depositing pure layers of Ni and Ti ranging from 10 to 20 nm thick. The heat process allows interdiffusion of Ni and Ti and subsequent crystallization. This process leads to an intrinsic two-way effect resulting in a bending motion.

Outlook for Reversible SMA Actuators

SMA actuators can be divided into two categories, depending on the method used to produce the reversible effect (Fig. 12). The first, called *monolithic*, refers to methods for which no external elements are used to produce the reversible effect, and the second, called *multiparts or mechanism*, refers to a reversible effect obtained by adding a second element. The subclass “smart design” denotes the use of design strategies to obtain a reversible motion within a single piece of material. These design methods are presented later.

Thermal Response of SMA Actuators

Time response is the well-known limitation of shape-memory alloys. Because the actuation method is thermal,

only a limited bandwidth can be achieved. However, as the mass decreases, scaling down SMA actuators increases their time response. Table 3 shows the thermal and related properties of Ni–Ti alloy compared to pure copper.

Modeling heat transfer within an SMA can be a tough job if all effects are considered. However, using a few assumptions, the transient response can be evaluated reasonably accurately. One of the basic assumptions is to use the lumped capacitance method. In this method, the solid, it is assumed, is spatially uniform at any instant during the transient process. In other words, the temperature gradient within the solid is negligible, and the thermal equilibrium equation can be written as follows:

$$\rho Vc \frac{dT}{dt} = Q_h - (Q_{\text{conduction}} + Q_{\text{convection}} + Q_{\text{radiation}}), \quad (10)$$

where ρ is the material density, V the volume, c the specific heat, T the temperature, and Q_h denotes the heat source. Considering the usual working temperature of SMA actuators and an application at ground level, the loss by radiation can be neglected ($Q_{\text{radiation}} = 0$). Convection is simply expressed by $hA(T - T_\infty)$, and the conduction heat loss should be evaluated depending on the application. Then, the transient response can be expressed as

$$T(t) = \begin{cases} T_\infty + \frac{Q}{hA} [1 - e^{-t/\tau}] & (\text{heating}) \\ T_\infty + (T_0 - T_\infty)e^{-t/\tau} & (\text{cooling}) \end{cases} \quad (11)$$

where T_∞ is the temperature of the surrounding medium, T_0 the initial temperature, A the area of the material, V the volume, h the convection coefficient, and Q describes the uniform heat medium. The time constant τ is simply $\tau = \rho Vc/hA$, where ρ is the density and c the specific heat. For electrical heating, which is the most common way to provide heat to an SMA actuator, Q is equal to I^2R , where I is the current through the material and R is the electrical resistance.

Using Eq. (11), the heating and cooling time can be expressed by

$$t = \begin{cases} \frac{\rho Vc}{hA} \ln \left(\frac{T - T_\infty}{T_0 - T_\infty} \right) & (\text{cooling}) \\ \frac{\rho Vc}{hA} \ln \left[\frac{(T - T_\infty) - \frac{Q}{hA}}{(T_0 - T_\infty) - \frac{Q}{hA}} \right] & (\text{heating}) \end{cases} \quad (12)$$

However, this simple approach does not consider the release and absorption of internal latent heat due to the phase transformation. Therefore, using Eq. (12), the time response is always underestimated. To include the effect of latent heat, a method used is based on a temperature-dependent specific heat. For instance, Brailovski et al. (24) considered a polynomial approximation of the specific heat during the transformation, where the polynomial’s coefficients are determined by using differential scanning calorimeter (DSC) measurements. The delay introduced by the phase transformation is often negligible in heating. It is also interesting to note that the latent heat depends on

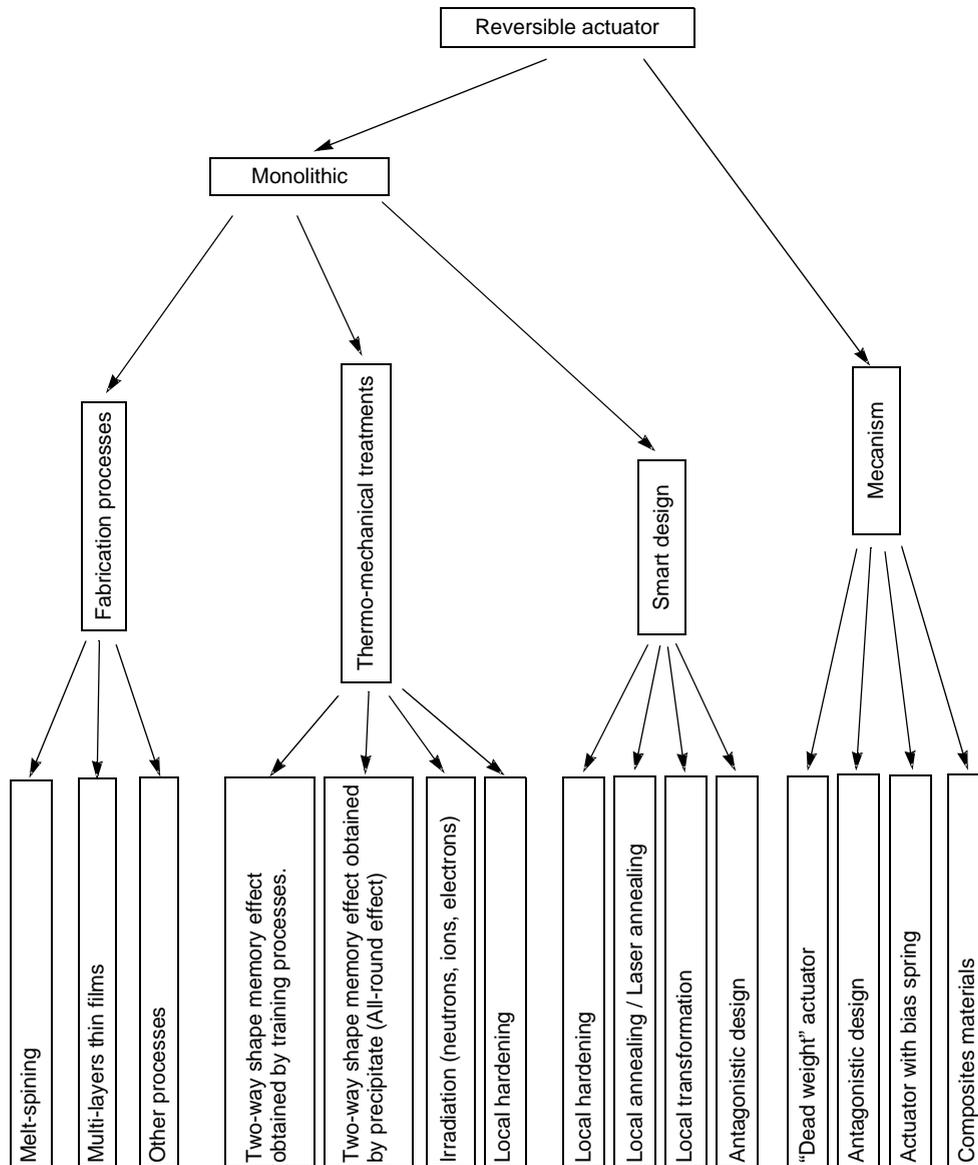


Figure 12. SMA reversible actuators: an outlook on the known methods for producing reversible motion.

Table 3. Thermal Properties of SMA vs. Copper at 300 K

	Ti-Ni	Copper (pure)
Density	6450 kg·m ⁻³ (12.52 slugs/ft ³)	8933 kg·m ⁻³ (17.34 slugs/ft ³)
Thermal conductivity	A: 18 W/m K (10.4 BTU/ft h °F) M: 8.6 W/m K (5.0 BTU/ft h °F)	401 W/mK (231.67 BTU /ft h °F)
Specific Heat (c_p)	836 J/kg K (0.20 BTU/lb °F)	385 J/kg K (0.09 BTU/lb °F)

the applied stress and the material. The higher the stress, the smaller the latent heat. Of course, this stress also affects the fatigue performance of the actuator.

Fatigue

An actuator working at a strain level of 8% obviously does not have the same lifetime as an actuator working at a strain level of 2%. Therefore, fatigue specifications contribute to establishing the design parameters of the actuator.

Many parameters make fatigue analysis complex. Alloy composition, loading speed, heat treatment, amount of cold-work, etc. are some of the relevant parameters that need to be considered. A method based on factorial analysis

Table 4. Fatigue Properties as a Function of Strain and Stress^a

Number of Cycles	Typical Strain (%)	Maximum Stress [MPa (ksi)]
1	8	500 (72.57)
100	4	275 (39.91)
10000	2	140 (20.32)
+100000	1	70 (10.16)

^a Ref. 26.

was proposed by Morgan (25) to determine critical parameters involved in fatigue. Moreover, as Van Humbeeck and Stalmans suggested in (2), three different types of fatigue have to be considered:

- failure by fracture due to stress or strain cycling at constant temperature
- changes in physical, mechanical and functional properties, for instance, the two-way, shape-memory effect.
- degradation of the shape-memory effect due to stress, strain, or temperature cycling in the transformation region

Considering all of these aspects, it is difficult to give some general rules. However, some recommendations are made about the maximum strain versus the number of cycles (Table 4).

For binary Ni-Ti, long cycle lifetimes higher than 1,000,000 have been obtained using the R-phase transformation, which, it is known, is very stable (1,2).

SHAPE-MEMORY ALLOYS FOR MICROAPPLICATIONS

Introduction

In the previous sections, some basic design rules were briefly introduced. These rules fit quite well with conventional size actuators such as devices that can be assembled where frictional and adhesive forces do not interfere with the functionalities of the device. What happens if we decrease the size of components? How reliable are the aforementioned design methods?

As Richard Feynman mentioned about miniaturization in his visionary and famous talk “There’s Plenty of Room at the Bottom” (27): “*the electrical equipment won’t simply be scaled down; it has to be redesigned.*” Let us focus now on the use of shape-memory alloys in microengineering.

Microrobotics and Microdevices

We usually speak of microdevices when the resolution of the motion and the dimensions of the parts are smaller than the precision and dimension usually achievable in a workshop. In a resolution-of-motion versus size-of-components representation, microrobotics is typically located in a region defined by resolutions ranging from 10 microns (about 4×10^{-4} inch) to 1 nanometer (about 4×10^{-8} inch) and dimensions ranging from 10 mm (about 0.4 inch) to 1 micron (about 4×10^{-5} inch). These boundaries are rather a trend than a definition. Figure 13 schematically illustrates this idea and some well-known mechanisms are presented.

Micro-devices, which integrate other functions such as controlling integrated circuits, are usually called “microelectromechanical systems” (MEMS). The MEMS are also often associated with silicon-based technologies and

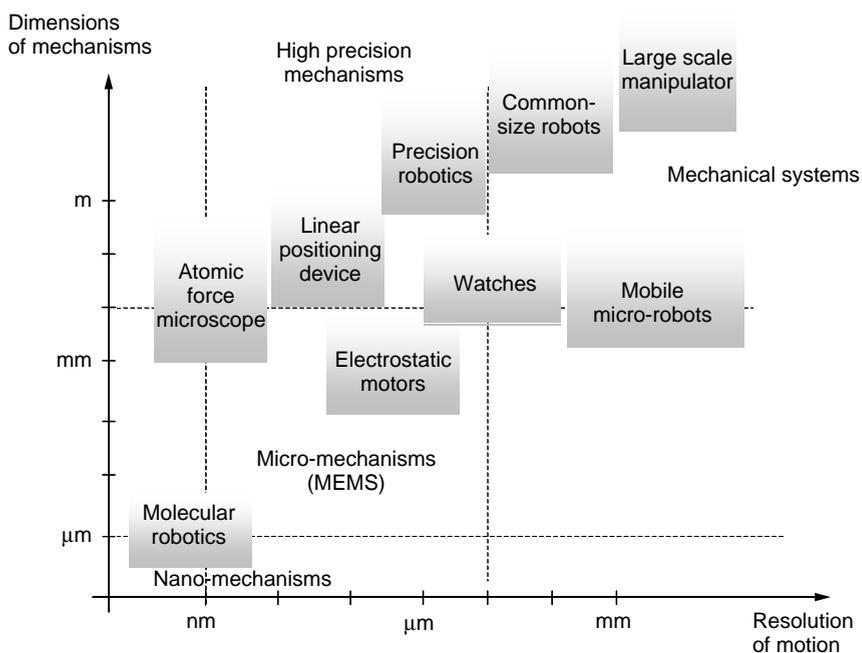


Figure 13. Micromechanisms: a definition as a function of dimensions and resolution of motion.

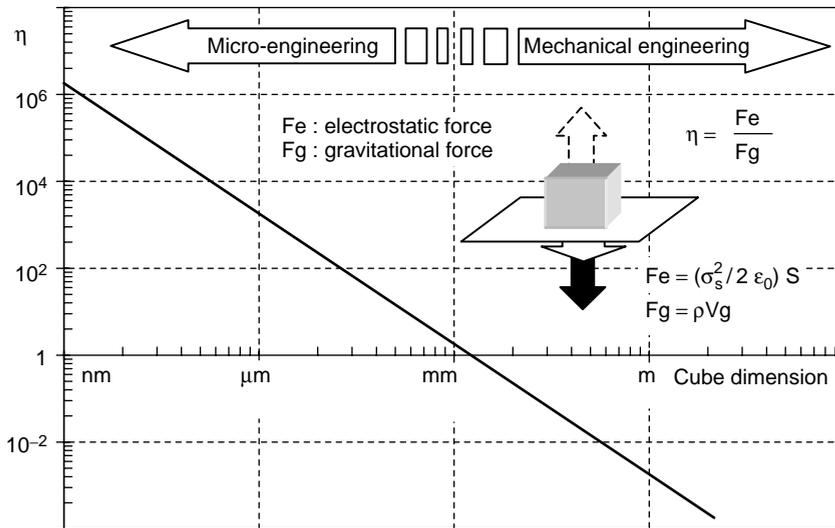


Figure 14. An illustration of the scaling effect on cube: ratio between electrostatic force and gravitational force as a function of cube dimension.

processing. However, according to the definition, MEMS should not be restricted to these special processes. Therefore, as a clever definition, the term “microdevices” is used throughout this article, rather than MEMS, to qualify small mechanisms, integrated or not.

Scaling Effect

We have all tried to manipulate small objects in our childhood and found that they stick very well to our fingers! This means that the relevant force is not gravity but rather adhesive forces. For instance, electrostatic or surface tension forces cannot be neglected when considering manipulation on a millimeter scale. This idea is briefly illustrated in Fig. 14 where the ratio between electrostatic force and gravitational force for a small cube is plotted as a function of the cube’s dimensions.

These scaling effects drastically modify our perception of the microworld and a lot of things have to be redesigned to adapt to “this new world.” New actuators and new mechanical components have to be created. SMA is certainly one of the most interesting candidates in this fascinating field of research as actuators and also as high strain flexible structures, as shown later.

SMA as a Smart Material for Microrobotics

The shape-memory effect in combination with an appropriate design or treatment can be used for microactuators as well as for microfastening devices. For instance, super-elasticity can provide an efficient method for enhancing the stroke of flexural hinges. In the previous section, SMA materials were compared to other candidates for small applications. In addition to some of the advantages already described, several interesting properties, relevant to microengineering applications, can be mentioned:

- As mentioned before, SMAs have the highest power/weight ratio among all known actuators in microengineering. Applications that require force production can be very compact.

- SMAs offer solid-state actuation and thus do not produce any dust. They are suitable for clean room conditions.
- The resistivity of the material (similar to stainless steel) is low, which allows efficient Joule heating. Low voltage and simple electronics can be used to power SMA actuators.
- Ni–Ti is compatible with MEMS processes.

However, a few drawbacks need to be considered carefully, such as their thermal activation, which limits the bandwidth. For fast applications, an SMA cannot compete with electrically or magnetically field driven actuators. The actuator efficiency is also very poor (typically a few %), which often excludes SMA microactuators from applications whose power consumption is very low. For example, the power consumption of a watch motor is typically a few picowatts!

Micromachining and Fabrication of SMA Microdevices

Silicon-Based MEMS: Fabrication Processes (28,29). Silicon technologies and related processes were major breakthrough in microelectronics as well as in sensors, actuators, and microsystems. Silicon based processes provide a unique method for large-scale production and miniaturization in the development of microactuators. These production methods are massively parallel and allow batch processing. However, these fabrication processes have a few limitations:

- Silicon microstructuring technologies are planar.
- The technological investment is very high.
- These methods are usually confined to structures of limited aspect ratios.

Ni–Ti and Ni–Ti–X (where X = Cu, Hf, Pd) alloys can be deposited on various substrates such as silicon, silicon dioxide (SiO₂) or titanium. However, the Si substrate is usually avoided because of the possible formation of silicide (SiNi) during crystallization of Ni–Ti thin films. The most

Table 5. HF/HNO₃/H₂O Etching of a Ni-Ti Thin Film Deposited on a Si Substrate^a

HF (46%): HNO ₃ (61%): H ₂ O Etchant Composition	Si Etching Rate ^a (nm/s)	Ni-Ti Etching Rate ^a (nm/s)	Etch Factor
1:1:0	730	180	0.1
1:1:2	1.3	50	0.45
1:1:4	<1.3	30	0.55

^a Ref. 32.^b(1 nm = 4 × 10⁻⁸ in).

common process for producing Ti-Ni thin films is deposition by dc or RF magnetron sputtering. An in-depth description of this process as well as the properties of thin film produced can be found in (29). Other processes such as deposition by flash evaporation (30), laser ablation (31), and multilayer processing (23) can also be used to produce thin films.

Ni-Ti thin films are usually etched by using HF/HNO₃/H₂O solutions. Table 5 shows experimental data obtained by Makino et al. (32). According to their results, a solution of HF:HNO₃:H₂O = 1:1:4 allows selectively etching of a Ni-Ti film deposited on a Si substrate.

Different fabrication methods can be used that depend on the type of actuator. When the actuator is moving freely, an intermediate layer, the so-called sacrificial layer, is required.

Figure 15 presents a method of producing a freely moving micromechanical element represented by a spring (33). In the first step, a polyimide layer is deposited by spin coating and is subsequently cured. The second step is Ni-Ti thin film deposition on the polyimide layer. In the next step, a photoresist thin film is spin coated over the Ni-Ti thin film. Then, the photoresist is baked, patterned, and developed to expose specific portions of the Ni-Ti layer selectively. Then the Ni-Ti film is etched. Finally, the structure is freed through reactive ion etching of the sacrificial polyimide layer.

In another process proposed by Buchaillot et al. (34), the Ni-Ti film is deposited just before removing the sacrificial layer, as is illustrated in Fig. 16 (29). A sacrificial Cr layer is deposited on a SiO₂ substrate. A polyimide layer is spin-coated on the Cr layer. Then, a second Cr layer is deposited on the polyimide. The upper Cr layer is patterned by photolithography and wet etched. The polyimide layer is vertically etched by oxygen plasma using the previously patterned Cr layer as a mask. Then, the Ni-Ti layer whose thickness is less than that of the polyimide layer is deposited. The final step consists of removing the polyimide layer by wet etching (“liftoff”) and releasing the mobile part by wet etching the Cr layer. Because the Ni-Ti layer is never etched, the aspect ratio of the structure is theoretically not limited.

Several applications such as microvalves use membrane actuators or biomorphs of Ni-Ti on SiO₂. This structure can be fabricated by using the process proposed by Wolf et al. (35) (Fig. 17). A Si wafer that has an oxide layer on both sides (SiO₂) is coated by a photoresist and softbacked. The back side photoresist is patterned and results in direct exposure of the Si in defined parts that creates window-like structures. Then, the Si is etched and almost all of the

volume “behind the window” is removed. The front layer is then removed using a buffered oxide etchant. The next step is depositing the Ni-Ti layer. Finally, the rest of the silicon is removed. When bimorphlike structures are desired, the front SiO₂ layer is not removed, and Ni-Ti is deposited directly on it.

Laser Machining. A laser can be used for milling and cutting SMA elements. Nd-Yag lasers are usually used. The sample is fixed on a two-axis linear stage under a fixed focusing objective. A rotational stage is sometimes added for tube cutting. The laser is focused on the sample, and an additional gas flow is usually used to drain off the molten material. The cutting precision depends on the object size and the material. It is possible to cut sheets whose thicknesses range from a few millimeters down to 0.005 m

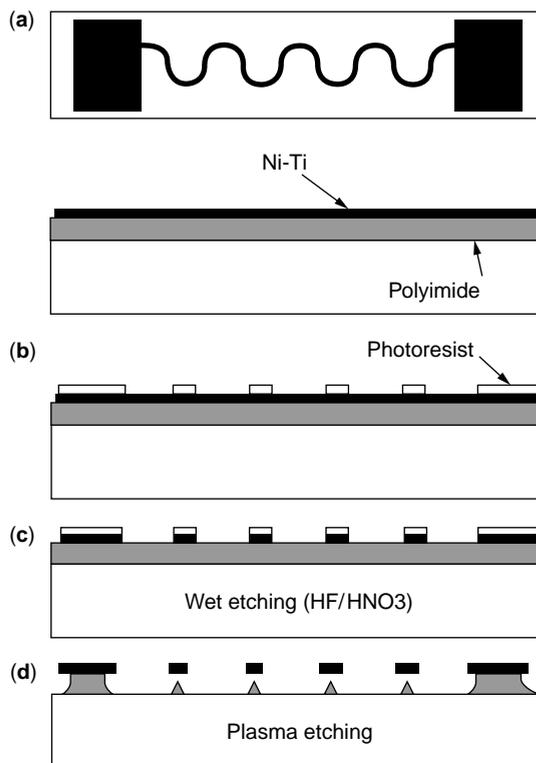


Figure 15. Thin-film microactuator fabrication process: (a) deposition of the layers, (b) photoresist deposition and patterning, (c) wet etching of SMA layer, and (d) plasma etching of the polyimide layer [after Walker et al. (33)].

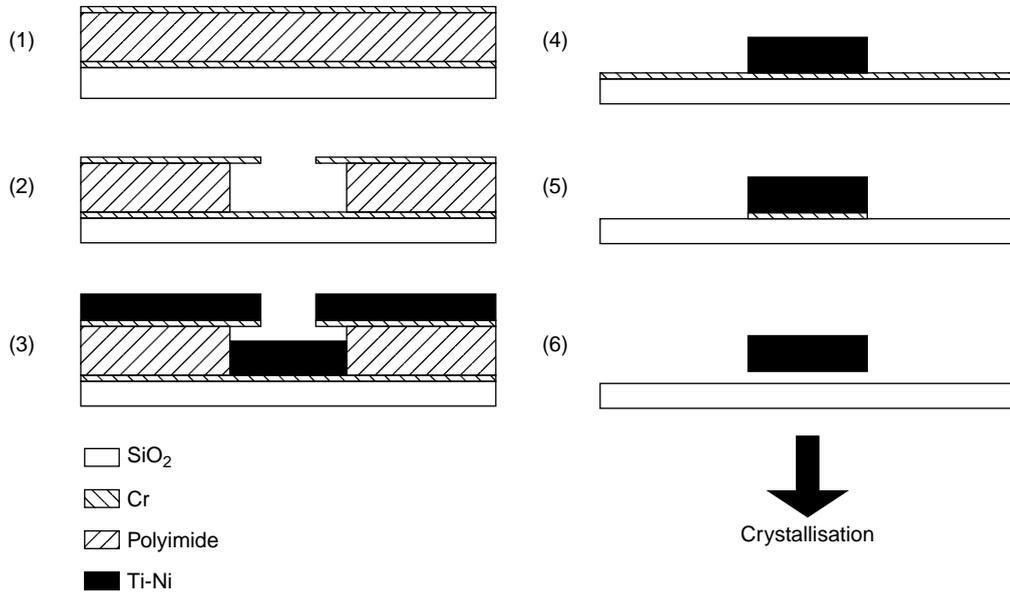


Figure 16. “Liftoff” fabrication process: (1) deposition of the layers (Cr, Polyimide), (2) Cr upper layer patterning and polyimide etching, (3) Ti–Ni sputtering (4) liftoff by removing polyimide using KOH, and (5,6) CR wet etching [after Buchailot et al. (34)].

(2×10^{-4} in). The spot size varies according to the sample’s thickness. Results for Ni–Ti have shown that the minimum spot size ranges from 0.02 mm (8×10^{-4} in) for 0.01-mm sheet thickness to 0.08 mm (4×10^{-3} in) for 1-mm sheet thickness (0.04 in).

Figure 18 shows part of a microgripper machined by a Nd–Yag slab laser working in the fundamental mode (TEM₀₀). An additional treatment has been used to decrease the edge surface roughness. For small to medium production volume and prototyping, laser cutting is an

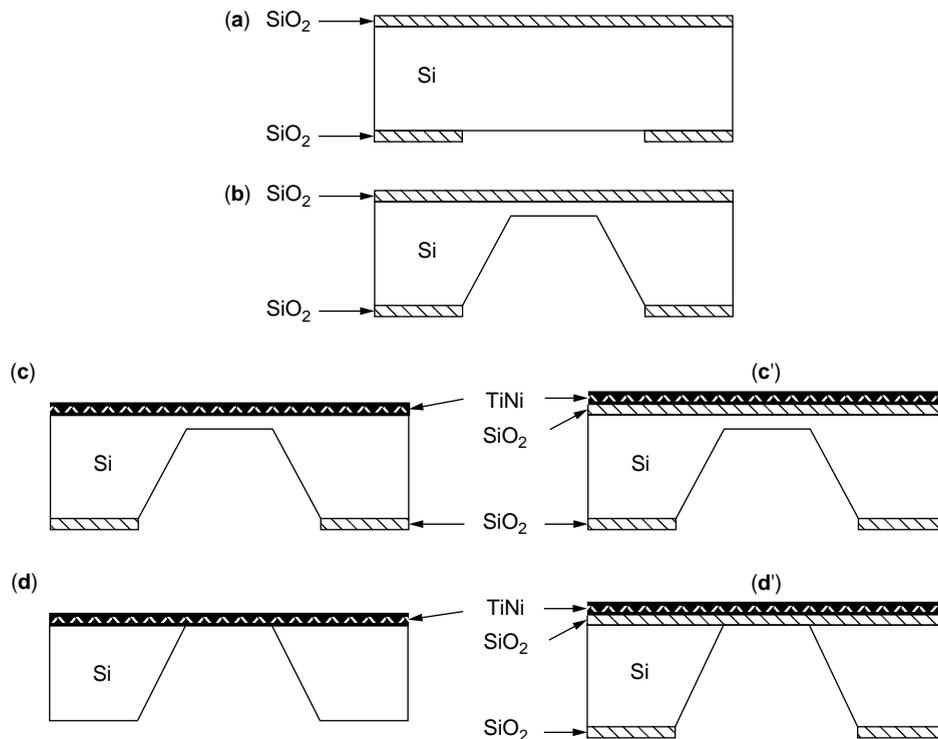


Figure 17. Microfabrication of Ni–Ti diaphragms with and without a SiO₂ layer. (a) Coating SiO₂ layers and patterning of the bottom one, (b) silicon etching, (c,c’) Ni–Ti deposition on Si, (c) (the SiO₂ has been etched) or directly on the SiO₂ layer, and (d,d’) removing the remaining silicon (35).

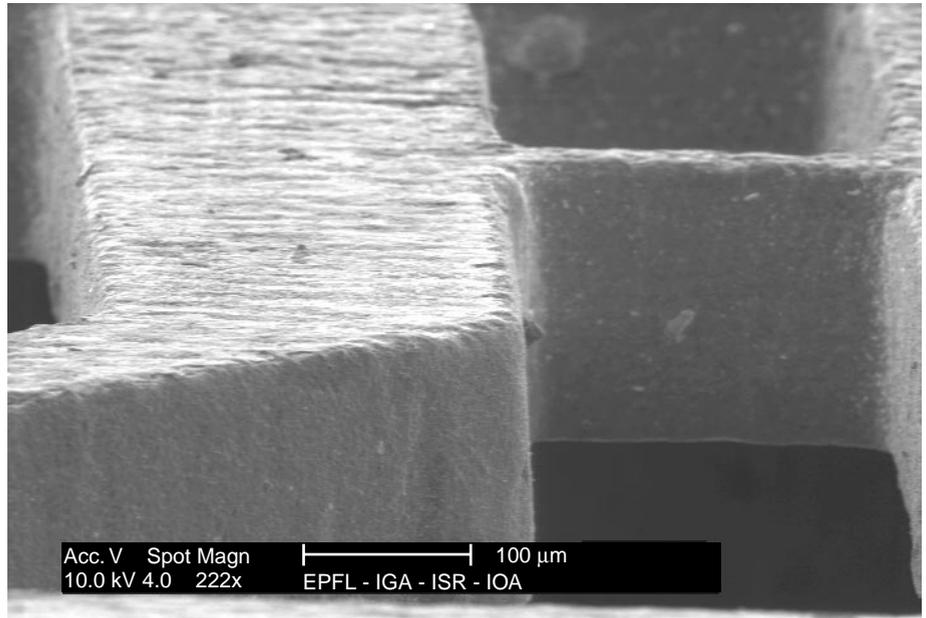


Figure 18. A Ni-Ti structure that was laser-cut by a Nd: Yag slab laser.

efficient method for Ni-Ti. However, due to laser heating, a heat-affected zone exists around the cutting edge. The depth of this heat-affected zone is typically of the order of 10 microns (4×10^{-4} inch).

Electrodischarge Machining. Electrodischarge machining (EDM) consists of using sparks created between an electrode and a sample to machine the material. There are two different processes: diesinking EDM and wire-cut EDM. In die-sinking EDM, the required shape is formed negatively in the metal by using a three-dimensional electrode. Various shapes, indentations, and cavities can be created. In the wire-cut process, the required shape is cut by guiding the wire along the given stretch.

Figure 19 shows a microsurgical tool in which a groove about 40 microns (1.57×10^{-3} inch) wide and about 200 microns (7.9×10^{-3} inch) deep has been machined by using a micro-EDM (36). The groove is 7 millimeters (0.27 inch) long. EDM is the most efficient method for creating structures of high aspect ratios that have low surface roughness. EDM is usually used for micromachining of elastic structures of high aspect ratio.

Recent Developments in Microrobotics and Microdevices: The State of the Art

In this section, some recent applications are presented. This discussion is not extensive but gives some general

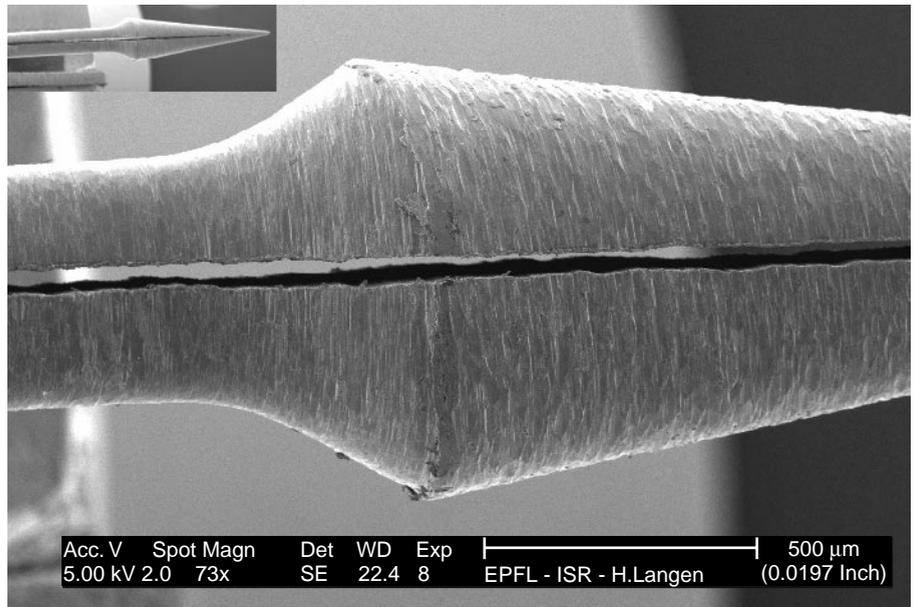


Figure 19. A Ni-Ti micro-surgical tool: the groove was machined by electrodischarge machining. The groove is 40 microns wide (0.0015 inch) by about 200 microns deep (0.0078 inch) (36).

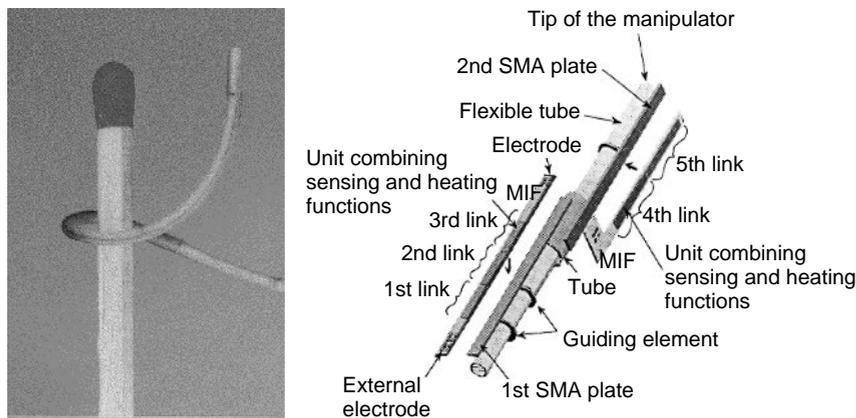


Figure 20. The Olympus Co. microendoscope [39]. (a) The endoscope winding around a match; (b) exploded view of the inside structure.

trends in the field of microengineering. Electronic connectors and other small fasteners are not mentioned in this section. The reader interested may consult the applications section of this encyclopedia or (1,2) for a more detailed description.

Active Catheter and Snakelike Robotic Systems. One of the first microrobotic applications was a prototype of an active miniature endoscope for a gastrointestinal intervention system (18). This endoscope was designed to pass smoothly through the sigmoid colon, which has a very small radius of curvature. The outside diameter was about 13 mm (0.51 in). As one can easily imagine, the heat dissipation of active surgery instruments is a critical issue. To address this problem, authors have proposed a cooling water tube going through the structure. As a result of this pioneering research, many active endoscopes for gastrointestinal surgery have been developed worldwide, for example, by Reynaerts et al. (37). All of these projects are minirobotic rather than microrobotic. Outside diameters range from 8 mm (0.31 in) to 15 mm (0.59 in). In the 1990s, smaller active endoscopes and active catheters were proposed in Japan. Fukuda et al. (38) designed an active catheter whose diameter ranged from about 1.33 mm (0.052 in) to 2 mm (0.079 in). The catheter has also been tested *in vivo*.

A very impressive five-degrees-of-freedom tube-type micromanipulator was recently introduced by Olympus Optical Company in Japan (39). This manipulator, shown in Fig. 20, is dedicated to inspection and maintenance in narrow spaces and for medical applications. The diameter is one mm (0.04 in), which makes this snakelike robot one of the thinnest in the world. The active parts consist of SMA strips working in bending modes. The strips have a two-way reversible effect and are heated by thermal conduction. A “multifunction integrated film” (MIF) is attached to the SMA strip. This circuit combines a heating function and a strain sensor by using a strain gauge measurement principle. The integrated circuit is constructed by successive sputtering of Ti and Pt on a polyimide layer. The authors developed a position controller based on strain sensing feedback, which gives positional accuracy of less than ± 0.25 mm (0.04 in). In another version, tactile sensors have been added that give reflex functions to the catheter, and

when the tube touches something, it automatically bends in the opposite direction.

Mini- to Microgrippers. At the end of the 1980s Ikuta (40) proposed a miniature gripper suitable for clean room conditions. The device consists of two bronze fingers that have two SMA coil springs attached to each finger. The coil springs are mounted opposite one another in an antagonistic or “push-pull” actuating arrangement. These springs were made from a thin film sputter-deposited on a glass substrate. The alloy showed an R-phase transformation around 320 K. The electrical resistivity change was used as an internal variable for monitoring the transformation. In this design, Ikuta proposed a controller scheme that combined a resistance and positional feedback where the positional feedback is given by a photosensor.

In another realization, Hesselbach et al. (41) in Germany developed a compliant microgripper.

In one version (Fig. 21), the compliant structure with flexural hinges is machined by electric discharge machining (EDM) from a superelastic material. Thus, the gripper uses SMA materials for two purposes: for actuation and for the guiding mechanism. The compliant mechanism uses a four-bar linkage that has a transmission ratio of -1 between the input and the output of the mechanism. If the

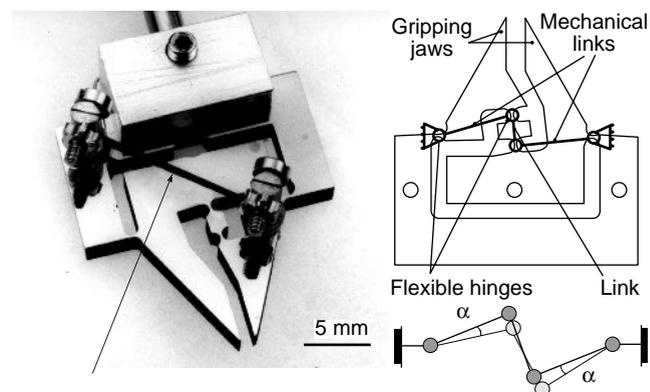


Figure 21. The microgripper developed by Hesselbach et al. (41). (a) Overall view of the microgripper (5 mm = 0.2 in); (b) design and mechanical equivalent of the flexible structure.

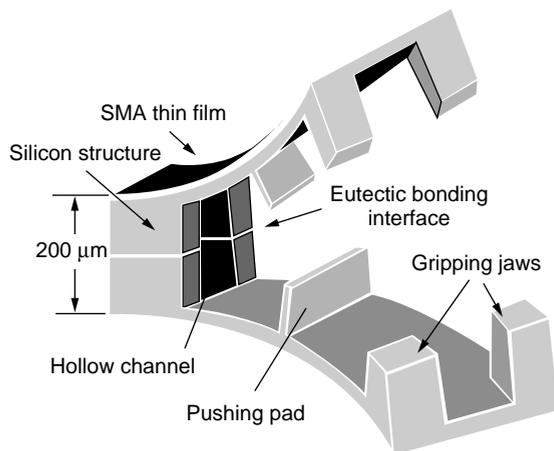


Figure 22. The Lawrence Livermore National Laboratory's microgripper (200 microns = 0.008 inch) (42).

input is one jaw and the output is the other jaw, a movement of the input jaw induces exactly the opposite movement of the output jaw. Hence, the jaws center every grasped object, regardless of the different stiffness of the flexural hinges or disturbance forces. An SMA actuator that has a strip shape deforms the compliant structure. The reversible motion is obtained by using the elasticity of the compliant structure, which acts as a bias spring. A resistance feedback is used to control the position of the finger. The authors have reported a lifetime of more than 450,000 gripping cycles and a close/open time of 0.5 s.

In microrobotics, Lee et al. (42) realized one of the smallest SMA microgrippers (Fig. 22). This gripper's dimensions are $1 \times 0.2 \times 0.38 \text{ mm}^3$ ($0.04 \times 0.008 \times 0.015 \text{ in}$). The design principle is a kind of "bimorph" of SMA/Si materials. The Si layer acts as a bias spring. The mechanism consists of two identical jaws actuated by Ni-Ti-Cu. The thin films are deposited on both external sides of the gripping jaws. The jaws are made of silicon and were shaped by a

combination of precision sawing and bulk micromachining of silicon. The upper part of the gripper is bonded to the lower part by selective eutectic bonding. The silicon gripper cantilevers act as bias springs. When heated, the Ti-Ni-Cu films bend the silicon cantilever, and when cooled, the cantilever deflects back and stretches the shape-memory film. A pushing pad is designed to assist in releasing the gripped object by pushing forward as the gripper is opened. An IC fabricated thin-film resistor heater pad applies heat to the microgripper. Experimental results have shown a gripper opening to 110 microns ($4.3 \times 10^{-3} \text{ inch}$), when fully actuated, and an estimated gripping force of 13 mN. The time response was estimated at 0.5 second.

A microgripper (43) for micro endoscopic (i.e., an endoscope whose diameter is about 1 mm (0.04 inch)) assembly is presented in more detail in the next paragraph.

Linear Actuator and Other Small Actuators. Grant and Hayward (44) proposed a high-speed linear actuator consisting of several thin Ni-Ti fibers woven in a counterrotating helical pattern around supporting disks (Fig. 23). The volume required by the actuator is a 17-mm diameter cylinder $\times 30 \text{ mm}$ ($0.67 \text{ in} \times 1.18 \text{ inch}$) long that belongs to the category of "minirobotics" rather than "microrobotics." Preloaded springs separate the disks, which keep the fibers under tension when unheated. When heated, the fibers shrink and pull the disks, together. The weave pattern of the fibers accomplishes displacement amplification. A rotary actuator has been developed using two of these linear actuators mounted opposite each other. In this configuration, a time response of less than 100 ms was obtained. However, due to the large amount of fibers in parallel, this woven design requires a high level of electrical current (typically 4–8 A).

Using Si-based processes, Buchailot et al. (34) developed an XY linear stage made of Ni-Ti thin films. This linear stage consists of four leaf springs operating in the same plane as the substrate. Each actuator has two parallel leaf springs and is initially stretched during mounting to induce a martensitic reorientation.

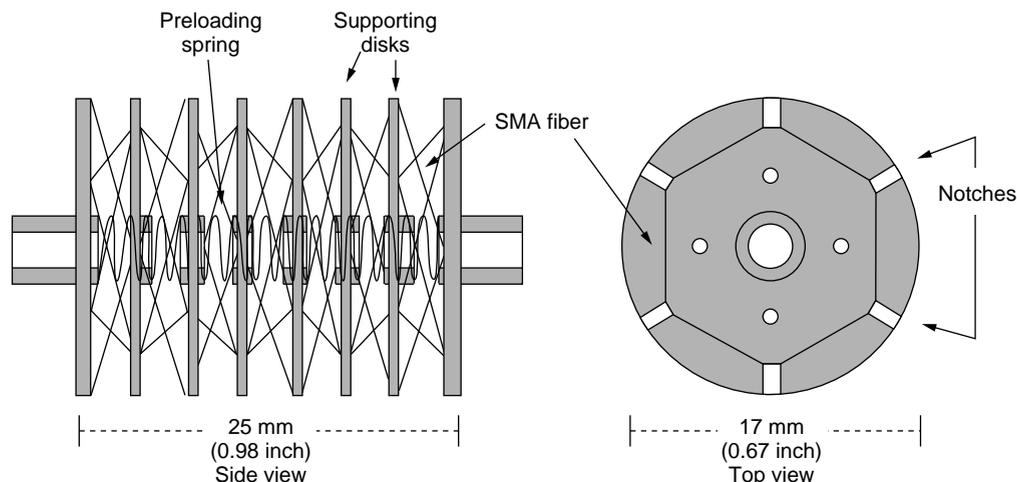


Figure 23. The "woven structure" developed by Grant and Hayward (44).

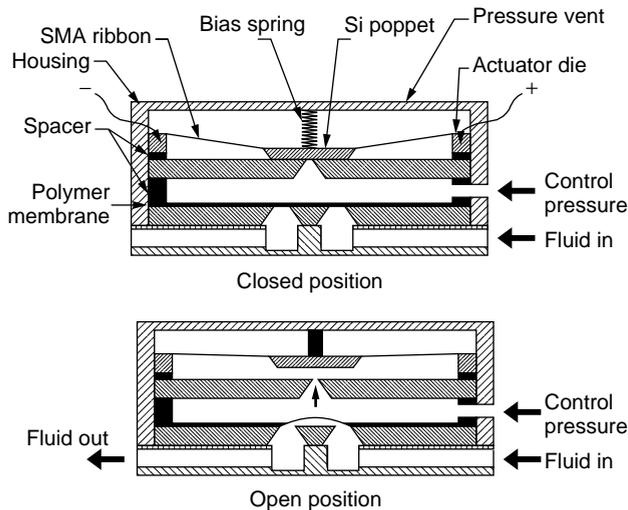


Figure 24. The microvalve developed by Johnson et al. (47,48).

Actuators are mechanically connected to each other and are designed to avoid any coupling between degrees of freedom. Therefore, the basic actuating principle is an antagonistic design, where each actuator deforms its counterpart. The whole structure is a few millimeters square (0.04 inch^2).

Many other reversible actuators have been proposed in the literature. Among these designs, Kuribayashi et al. (45) in 1993 proposed a millimeter-sized robotic arm. The all-round effect (see earlier) was used to produce a reversible effect in a Ni-Ti alloy beam. Three of these microcantilevers were combined to realize a SCARA-type microrobot that has three degrees of freedom. Later, Kuribayashi and Fujii developed a microcantilever made of a SMA thin film prestressed by a polyimide layer (46). The polyimide layer provides reverse motion when cooling.

Fluidic Applications. The pioneers of this field of research were Busch and Johnson (47). In 1989, they patented a micro valve, that uses a SMA thin film that has been processed by MEMS technologies. The valve shown in Fig. 24, consists of a silicon orifice die, an actuator die that has a poppet controlled by a Ni-Ti shape-memory alloy microribbon, and a bias spring. An electrical current heats the actuator (the Ni-Ti microribbon). When no electric current passes through the actuator, the bias spring pushes the poppet against the orifice and closes the valve. When an electric current is applied, the Ni-Ti actuator contracts and lifts the poppet from the orifice, opening the valve. The whole device is less than one-half square centimeter (0.2 in^2). The displacement of the poppet is more than 100 microns (0.004 inch) and produces a force of one-half newton. In a rough approximation, flow through the valve is proportional to the current applied to the actuator. The current and an appropriate feedback may be used to control the flow (48).

In another realization, Kohl et al. (49) designed microvalves by integrating SMA bending actuators fabricated from laser-cut, cold-rolled Ni-Ti sheets (Fig. 25).

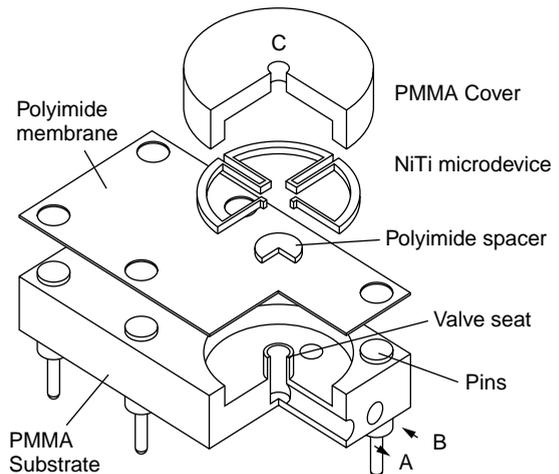
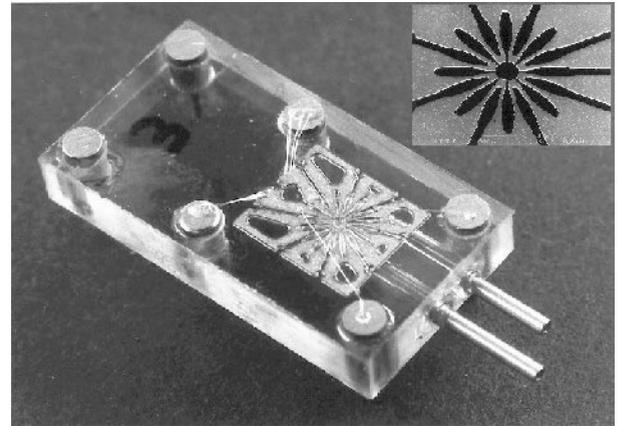


Figure 25. The microvalve developed at the Forschung Zentrum in Karlsruhe (49).

The valve consists mainly of a PMMA housing that has an integrated valve seat, a polyimide membrane, a polyimide spacer, and a SMA microdevice. The SMA microactuator is used for deflection control of the membrane, which opens or closes an opening in the valve seat. The actuator is made of “stress-optimized” cantilever elements whose actuation direction is perpendicular to the plane of the substrate.

A CONCEPT OF SMART SMA MICRODEVICES

Basic Idea

Active mechanical systems generally consist of a force generator, a coupling device, a transmission system, a guiding system, and an output element. The basic idea is to integrate all of these elements and functions within one single piece of material to form “a smart SMA microdevice.” This concept is well adapted to microrobotics because a monolithic device does not require any assembly and is free of wear, friction, and dust. Nevertheless, to be reversible, a monolithic design requires that one region of the material act as a biasing spring. Thus, several methods have been developed to address this issue (50,51).

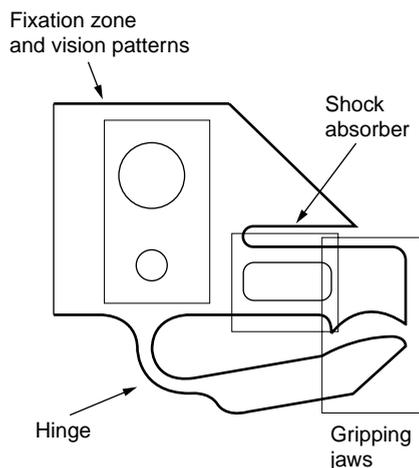
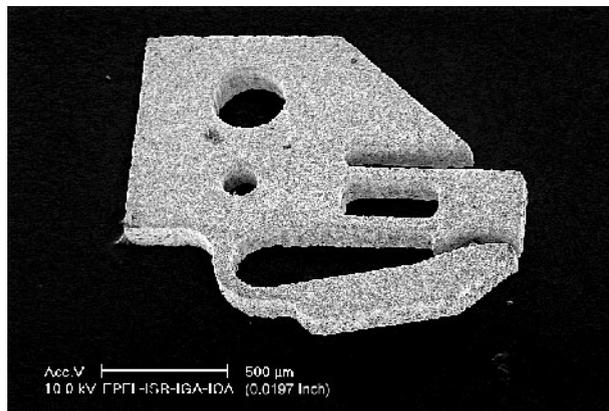


Figure 26. The EPFL's microgripper for submillimeter lens manipulation (43,50).

The Two-Way Shape-Memory Effect (TWSME) Applied to Small Devices

As presented before, thermomechanical treatment of a material can introduce internal stress. This stress will generate preferentially oriented martensitic variants when cooling. Using an appropriate design, the TWSME can also be used to create reversible motion within a device that has a more complex shape than wires, strips, or springs.

The TWSME microgripper shown in Fig. 26 was designed and successfully introduced in a production line of microendoscopes (43). The design consists of crablike tweezers about 1 mm^2 (0.04 inch^2), and the gripper has a moving arm that clamps object against a fixed part. Additional fixtures have been added for visual measurements during calibration and a force limitation in case of mechanical shock to the fixed part.

The material is a Ni-Ti-Cu sheet 0.15 mm (0.006 in) thick. The training process is explained in Fig. 27:

1. The bulk material is annealed at 515°C for 15 min. The microgripper shape is then laser-cut from the sheet. At this point, the gripper does not have any reversible motion: the austenitic shape is the cut shape.
2. The microgripper is fixed on a heat source (thick resistive film, SMD-resistor, thermoelectric devices).

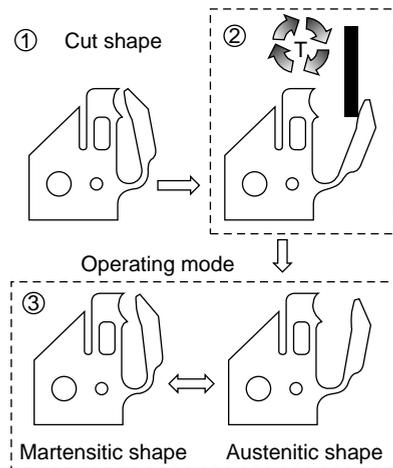


Figure 27. The training process applied to the microgripper (thermal cycles under constraint).

Then, the training process is applied to the device. Different training methods have been reported in the literature (1,2). In this special case, the best method for achieving gripper specifications is to apply thermal cycling under constraint. To perform this task, a shaft is used to deform the gripper up to 4 to 5%. The maximum motion range is reached after 50 cycles. After training cycles, the austenitic shape is not exactly the same as the cut shape. Residual plastic deformation resulting from the training process is observed and has to be considered while designing the gripper.

3. During cooling, the gripper opens. When heating, the gripper recovers its parent shape (high-temperature shape) and closes its jaw. For a hinge thickness of $70 \mu\text{m}$ ($2.7 \times 10^{-3} \text{ in}$), a range of motion of $150 \mu\text{m}$ ($6 \times 10^{-3} \text{ inch}$) was obtained, and a grasping force of 16 mN was measured. If this force is compared with the weight of the gripper itself, the gripper's force/weight ratio is 1000, and if only the active part of the gripper is considered in the gripper weight estimate, this ratio increases to 30,000! However, the force when cooling is about four times lower (approximately 4 mN). This is a limitation of the TWSME: the force induced when cooling is lower than the recovery force generated upon heating. Nevertheless, a gripper needs forces when closing to grasp an object but does not need force when opening. Fatigue experiments consisting of grasping and releasing cycles showed that motion loss saturates after 100,000 cycles. After 200,000 cycles, the motion loss is 10%, and the loss after 1,000,000 cycles is estimated at the same level.

Local Heating of the Material ("Martensite" Spring-Biased Design)

To create a reversible effect, active and passive parts are needed simultaneously within the material. One solution consists of controlling the heating process through the

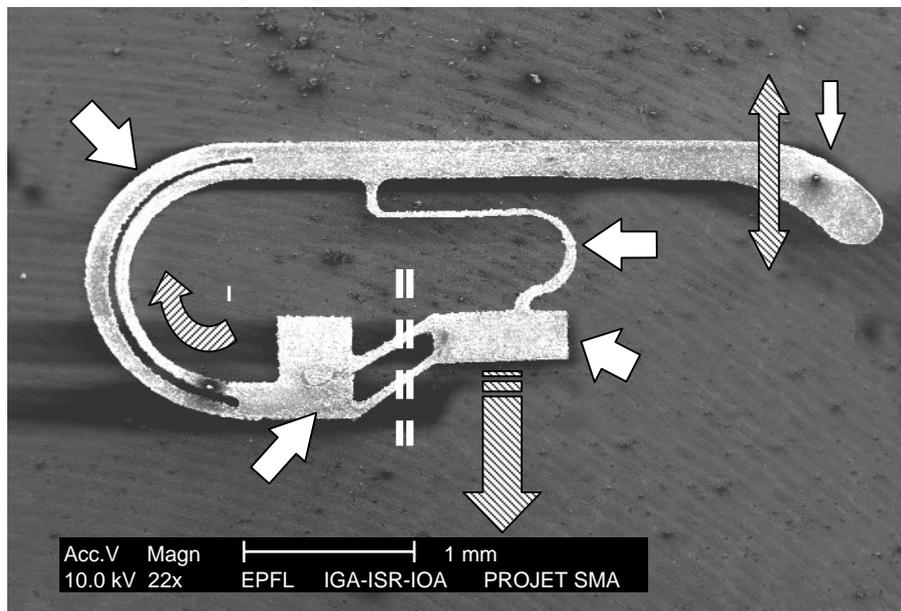


Figure 28. Optical switch using the local heating design principle. (1) actuating part, (2) bias spring, (3) shutter, and (4,5) fixture and connection pads (scale 1 mm = 0.04 in).

structure. By carefully designing the electrical path, for example, part of the material can be in the austenitic state, and the remainder is in the martensitic state. The elasticity of the martensitic region can be used as a bias spring. The optical switch shown in Fig. 28 illustrates this idea.

The device was machined from a $20\ \mu\text{m}$ thick (7.9×10^{-4} inch) Ni–Ti–Cu foil. The motion can be compared to that of a semaphore. During assembly, the structure is prestressed by pulling the fixed part (#5, see Fig. 28). Then, parts #4 and #5 are fixed and separated from each other (dashed line). When a current passes between the two pads (#4 and #5), the part that has the thinnest section (#1) is heated up to the transformation temperature, recovers its original shape, and pulls down the shutter (#3). When cooling down, the bias spring—the part that has the largest section (#2)—pulls back on the structure causing reverse motion. In this design, the shutter range of motion is $190\ \mu\text{m}$ (7.5×10^{-3} in). The time response when heating is less than 50 ms, and the power consumption is typically 5 mW.

Antagonistic or “Push-Pull” Design

The working principle is the same as described before, except that the two SMA actuators are part of the same piece of material. By designing an appropriate electrical path, one zone of the material is heated at a time. Depending on which zone is heated, the device will move in one direction or the opposite. The main difference from the previous case is that the reverse motion is due to alternative heating of actuating zones.

An illustration of this principle is shown in Fig. 29. This device is a linear actuator that has one degree of freedom. During assembly, the two leaf springs (#2, #2*) are prestressed and fixed on the contact pads (#3, #3*). When a current passes through one of the contacts (#3, #3*) and the main body (#5), the corresponding spring is heated up

to the transformation temperature and pulls the mobile part (#4) in its direction. When the springs are mounted opposite each other, two-way motion is obtained. To obtain precise motion in only one direction, a guiding system (#1) has been added to compensate for machining and assembly errors. The schematic of the mechanical equivalent system is shown in Fig. 29. It consists of two serial four-link structures. One four-link structure guides the motion in one direction, but this guiding is not purely rectilinear, rather it is motion that has a high radius of curvature. Adding a second four-link structure in series compensates for undesired motion in the perpendicular direction. This linear actuator illustrates very well how smart a monolithic design can be: this design is a complete linear stage that includes actuators and guiding elements within one single piece of material.

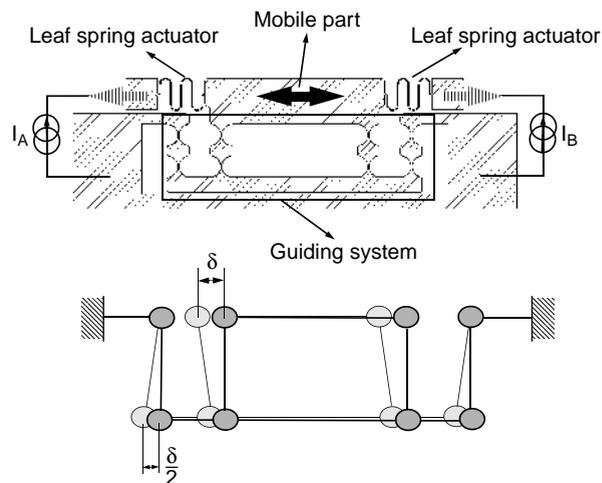


Figure 29. A one-degree-of-freedom linear stage. The mechanical equivalent of guidance is explained.

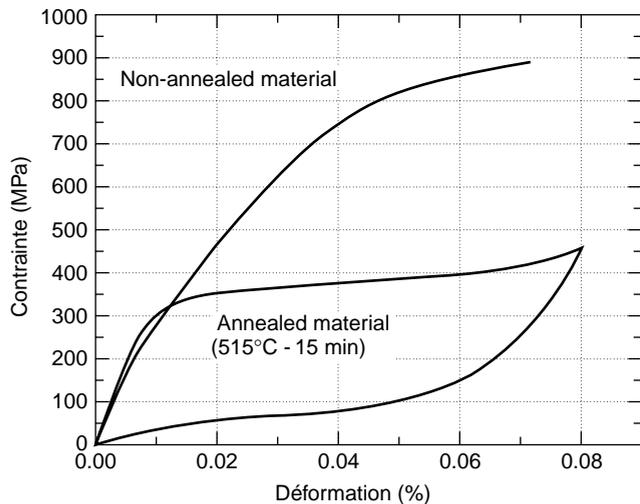


Figure 30. Mechanical behavior before and after annealing a cold-worked sheet having a superelastic behavior at room temperature.

Local Annealing

Basic Principle. During the fabrication of SMA devices, annealing is required to obtain a martensitic transformation within the material. In a sputter-deposited thin film, the material is amorphous after deposition. Annealing is required to crystallize the material. In a cold-rolled sheet, the forming process introduces a large number of dislocations which prevent martensitic transformation. In that case, annealing induces recovery and recrystallization of the material. Typical annealing temperatures range from 400 to 900°C.

The basic idea of local annealing is to limit annealing to selected zones of the material. Martensitic transformation occurs only in annealed zones. In other words, some parts are active, and others are passive.

Figure 30 shows a tensile test on a cold-worked sheet of Ni-Ti before and after annealing (the material is in the austenitic state). The superelastic characteristic can be seen only after annealing. In the framework of a monolithic design, the annealed zone can be used for actuating purposes, and the nonannealed zone can be used as a biasing element for a reversible effect. Two methods can be used for annealing small devices locally (51). The first is direct Joule heating of the material, and the second uses a laser to heat the structure.

Local Annealing by Joule Heating. Direct Joule heating actuates most SMA actuators and devices. It can also be used to heat up the material to the annealing temperature. Kuribayashi et al. (52) used annealing by Joule heating a microactuator consisting of a cantilever working in the bending mode.

An example of a structure locally “Joule annealed” is shown in Fig. 31. The device is a linear actuator that has four leaf springs (51). These are the fabrication steps:

1. The structure is cut out by laser from an “as-received” unannealed Ni-Ti-Cu sheet.

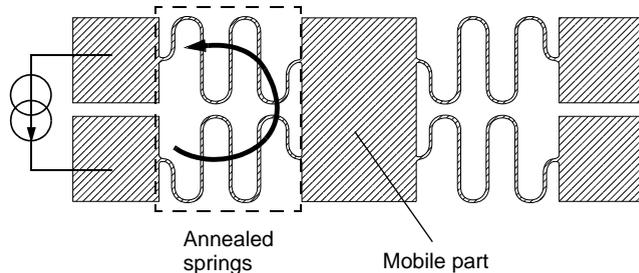


Figure 31. A one-degree-of-freedom linear stage, locally annealed by using an electrical current.

2. An electrical current is passed through the left springs causing the temperature to rise to the annealing temperature (about 500°C). The right springs remain at room temperature. An infrared camera measures the temperature during annealing.
3. Once annealing is over, the springs are prestressed along the axis and are fixed.

Because of this local annealing, one part of the material is passive, and the other is active. The passive part is used as a spring, and the active part is the actuator. The mechanical behavior of this structure is exactly the same as the spring-biased multipart mechanism presented in the previous section.

“Joule annealing” can be a low-cost and efficient method of producing smart SMA microdevices. However, a few limitations exist:

- This method can be applied only to devices that contain loops for an electrical path.
- The cross section along the electrical path has to be carefully designed with respect to temperature distribution during heating. The thinnest section will be the hottest and will be annealed first.
- Because the power dissipated in the structure is directly proportional to the resistance and thus to the cross section, the local temperature and the annealing conditions will depend strongly on machining tolerances.

Laser Annealing of SMA (the LASMA Process). A laser beam is used for local heating. The laser is focused on the point where annealing is desired. In contrast to Joule annealing, laser annealing of SMA (LASMA) can be applied to all kinds of designs (51,55). This method has been successfully used on various types and shapes of SMA microdevices. It can easily be applied to thin films, cold-rolled sheets, cold-drawn wires, or other materials that have been work-hardened.

Figure 32 shows some microstructures in a laser-annealed thin film. The material’s microstructure was observed at several distances from where the laser was focused. Near a distance of about 800 μm (3.15×10^{-2} in), the transition from crystallized region to the amorphous zone can be observed. The circularly shaped bubbles are grains growing in the amorphous matrix.

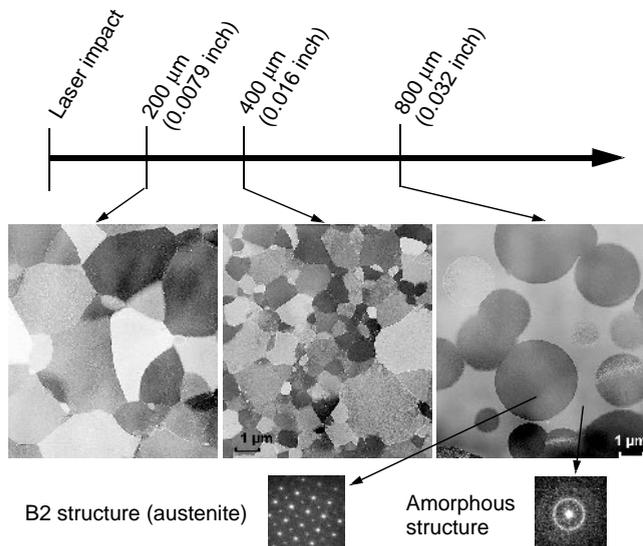


Figure 32. Microstructures along the irradiated zone—observation by transmission electron microscopy.

Two Examples of Laser-Annealed Devices (51,55). The first example is a microcantilever moving up and down, which is shown in Fig. 33.

The structure is machined from a sputtered-deposited Ni–Ti thin film. The LASMA process is used to crystallize the material locally. We have observed that after annealing, the locally affected zone of the material is expanded. This means that a prestrain can be introduced during the process. Therefore, an active structure can be created without

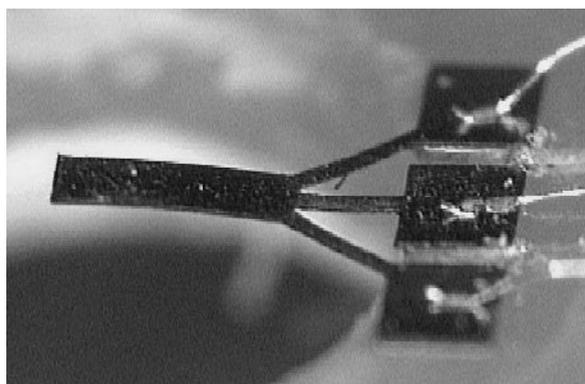


Figure 33. Locally annealed microcantilever: realization and design.

any mechanical interaction. The annealing process is monitored by the resistivity change.

When a current passes through the structure, the annealed cantilever transforms into austenite and pulls the structure downward. Reversible motion is obtained by using the elasticity of the unaffected zone. This example demonstrates that successful reversible motion can be introduced in the device without any mechanical interaction. Moreover, this effect is not limited to bending motion.

Another example of an application is shown in Fig. 34. This device is a microgripper for lens manipulation. The structure is cut out from a cold-worked Ni–Ti sheet. The laser is used to anneal the outer part of the gripper's joint. The inside part is not annealed and is the function of the bias spring. The prestrain is realized by deforming the gripper's unannealed zone well above its plastic limit. When heating the whole structure, only the annealed region will try to recover its original shape. Upon cooling, the inner element, the unannealed portion, will push back the structure and open the gripper. Using the LASMA process, an active element and a bias spring have been introduced in one single piece of cold-worked material. Due to these properties, the LASMA process is a key technology for developing smaller highly integrated microdevices.

Flexible Structures

Flexural designs are a well-established technology in precision engineering (53,54). Their monolithic structure offers many advantages such as freedom from wear, no assembly, and smooth displacements that are contamination-free. However, one of their disadvantages is that flexures are restricted to small displacements for a given size and stiffness in the drive direction. Moreover, like every spring mechanism, the force increases linearly with the displacement. Considering these two limitations, SMA flexures can improve these aspects significantly. The large deformation available enhances the output stroke for a reasonable volume. Moreover, due to their “plateau,” the force will saturate and if the structure is prestrained, a low stiffness linear guiding system can be imagined (55).

Smart Monolithic SMA Microdevices: An Outlook for Design Methods

Table 6 is a summary of some methods for smart design (50). The local hardening mentioned in this table is the exact reverse effect of the LASMA process. Instead of local annealing, the basic idea is to harden or amorphize the material locally. In the design context, one could say that the LASMA process introduces an active region within a passive structure, whereas local hardening introduces a passive element within an active structure.

FUTURE TRENDS

As previously mentioned, scaling down things leads to new design approaches. This rule applies to shape-memory alloy microdevices. Successful breakthroughs have been

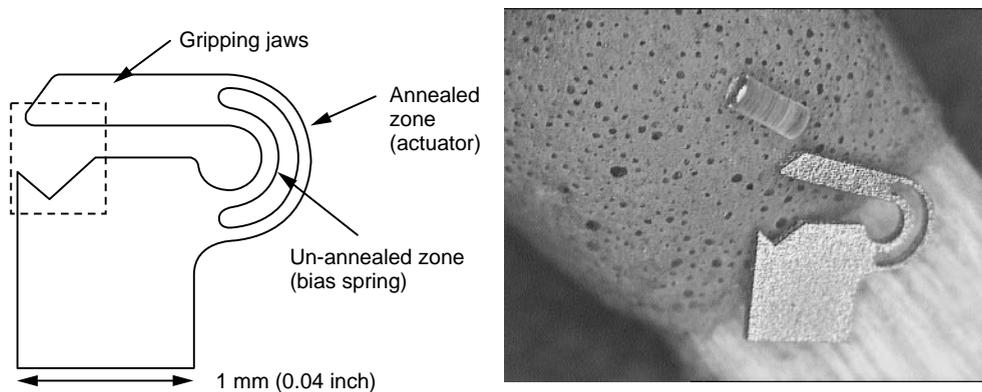


Figure 34. Locally annealed microgripper. (a) Design principle; (b) the microgripper and a cylindrical lens on a match head.

made in this field of research, but there is still a further need to understand the behavior of small devices on the microscale. It is well known that the mechanical behavior of a thin film of a few microns is not the same as the behavior of a sheet that is a few hundred microns thick. As W. Nix mentioned (56), thin film materials are, for instance, much stronger than their bulk counterparts. This may be due to the fine grain sizes commonly found in thin films, but single crystal thin films are also much stronger than bulk materials. Many issues remain unresolved or little addressed for SMA thin films:

- How thin can a functional SMA thin film be? In other words, where is the size limit of the shape-memory effect?
- If a micron-size structure is built, how smooth will the motion be? Will we observe a discontinuous effect?
- How does the scaling law apply to fatigue properties?

Magnetostriction is the spontaneous deformation of a material caused by a change in its state of magnetization. Some materials have both a martensitic transformation and a Curie temperature (ferromagnetic martensites). Depending on which phase has larger saturation

magnetization, a field applied either above or below the martensitic transformation temperature will induce a transformation between austenite and martensite. If the temperature is below the martensitic transformation, the martensitic variants can be redistributed by a magnetic field (57). These materials might be a solution to bypass the main drawback of nonmagnetic SMAs, their response time. Considerable effort is still needed to introduce this material in applications.

The design of a controller for a SMA actuator was not discussed in this article. Nevertheless, a lot of work has been done within the robotics community. Most of them have proposed PID controllers or close cousins (40). Many control schemes use electrical resistance as the sensing variable in the feedback loop, as proposed by Hirose et al. (18). Other methods have been proposed by Gorbet (58) who studied control using a Preisach representation and by Grant and Hayward (44) who designed a variable structure control that consists of switching between a set of output signals in response to an input signal. However, all of the proposed methods have been applied only to simple SMA actuator design consisting of a wire, strip, or spring working against a bias spring. For instance, there still is further room for controlling very complex designs, such as antagonistic designs.

Table 6. A Summary of Design Methods for Smart Monolithic Designs^a

Design Methods	Material State	Prestressed	Advantages	Limitations
TWSME	Annealed	No	Actuation methods, low volume required	Low force in one direction, training process
Local heating	Annealed	Yes	No local treatment, simple method.	Low stiffness of martensite, special heating path
“Push-pull”	Annealed	Yes	High range of motion	Volume required
Local hardening ^b	Annealed	No	Different mechanical hardening	Miniaturization
LASMA	Amorph. Work-Hardened	No Yes	Highly integrated design, different annealing conditions	Managing the annealing process

^aRef. 50.

^bLocal hardening is the exact reverse effect of LASMA applied to a cold-rolled sheet.

CONCLUSION

Robotics gathers skills from mechanical, control, and computer engineering. In addition to these fields, microrobotics needs skills from materials sciences and physics. As you come closer to the atomic scale, interaction between the parts of a mechanism requires a more detailed analysis of the physical phenomena involved in the scaling effect. Microrobotics will never be a simple clone of robotic on a smaller scale. Scaling down things leads to a new field of research where our perception of physical interaction has to be modified. New actuators and sensors have to be invented. In this research trend, shape-memory alloys are one of the most interesting candidates for miniaturization. Due to their unique properties, SMA offers a powerful and a flexible solution in many fields of microrobotics. Smart monolithic SMA devices offer a unique approach to further miniaturization and will contribute to exploring the "bottom" scale using powerful robots and tools.

BIBLIOGRAPHY

1. T. Duerig, K.N. Pelton, D. Stöckel, and C.M. Wayman, *Engineering Aspects of Shape Memory Alloys*. eds., Butterworth Heineman, 1990.
2. K. Otsuka and C.M. Wayman, eds., *Shape Memory Materials*. Cambridge University Press, 1998.
3. K. Funakubo, *Shape Memory Alloy*. Gordon and Breach, 1984.
4. R. Gotthardt and J.-E. Bidaux, *Int. Conf. Displacive Phase Transform. App. Mater. Eng.* TMS, Urban, Ill, May 8–9 1996.
5. P. Scherrer, J.-E. Bidaux, A. Kim, J.-A. E. Manson, and R. Gotthardt, *J. Phys. IV* **9**: Pr9–393 (1999).
6. D.Y. Li and R. Liu, *Wear* **225–229**: 777–783 (1999).
7. SMA Inc., Company data sheet.
8. J.M. Gere and S. Timoshenko, eds., *Mechanics of Materials*. PWS,
9. J.F. Lindberg, *Mater. Res. Soc. Symp., Proc.* **459**: 509–519 (1997).
10. P. Krulevitch, A.P. Lee, P.B. Ramsey, J.C. Trevino, J. Hamilton, and M.A. Northup, *J. MEMS* **5**: 270–282 (1996).
11. R.H. Baughman, *Synth. Met.* **78**: 339–353 (1996).
12. K. Tanaka, *Res. Mech.* **18**: 251–263 (1986).
13. C. Liang and C.A. Rogers, *J. Intelligent Mater. Syst. Struct.* **8**: 303–313 (1997).
14. L.C. Brinson, *J. Intelligent Mater. Syst. Struct.* **4**(2): 229–242 (1993).
15. M. Their, A. Mick, D. Drescher, and C. Bourauel, *J. Mater. Sci.* **26**: 6473–6478 (1991).
16. Y. Liu, Z. Xie, J. Van Humbeeck, and L. Delaey, *Acta Materialia* **46**(12): 4325–4338 (1998).
17. H. Tobushi and K. Tanaka, *JSME Int. J. Ser. I.* **34**(1): 83–88 (1991).
18. S. Hirose, K. Ikuta, M. Tsukamoto, K. Sato, and Y. Umetani, *Proc. Int. Conf. Martensitic Transform.* The Japan Institute of Metals, Nara, Japan, 1986, pp. 1047–1052.
19. K. Escher and E. Hornbogen, in *Progress in Shape Memory Alloys*, S. Eucken, ed., DGM Informationsgesellschaft, 1992, pp. 301–316.
20. R. Kainuma, M. Matsumoto, and T. Honma, *Proc. Int. Conf. Martensitic Transform.* The Japan Institute of Metals, Nara, Japan, 1986, pp. 717–722.
21. D.S. Grummon and R. Gotthardt, *Acta Materialia*, **48**: 635–646 (2000).
22. S. Eucken, in *Progress in Shape Memory Alloys*, S. Eucken, ed., DGM Informationsgesellschaft, 1992, pp. 239–277.
23. T. Lehnert, S. Tixier, P. Böni, and R. Gotthardt, *J. Mater. Sci. Eng.* **A273–275**: 713–716 (1999).
24. V. Brailovski, F. Trochu, and A. Leboeuf, *Proc. 2nd Int. Conf. Shape Memory Superelastic Technol.*, Pacific Grove, CA 1997, pp. 227–232.
25. N.B. Morgan, Ph. D. Thesis, Cranfield University, UK, 1999.
26. D. Stöckel, *Proc. Actuator-92*, Bremen, Germany, 1992, pp. 79–84.
27. R. Feynman, *Annu. Meet. Am. Phys. Soc.*, Dec. 29, 1959.
28. W. Lang, *Mater. Sci. Eng.* **R17**(1): 1–55 (1998).
29. S. Miyazaki and A. Ishida, *Mater. Sci. Eng.* **A273–275**: 106–133 (1999).
30. E. Makino, M. Uenoyama, and T. Shibata, *Sensors and Actuators* **A71**: 187–192 (1998).
31. H.D. Gu, L. You, K.M. Leung, C.Y. Chung, K.S. Chan, and J.K.L. Lai, *Appl. Surf. Sci.* **127–129**: 579–583 (1998).
32. E. Makino, T. Mitsuya, and T. Shibata, *Sensors and Actuators* **79**: 251–259 (2000).
33. J.A. Walker, K.J. Gabriel, and M. Mehregany, *Sensors and Actuators* **A21**: 243–246 (1990).
34. L. Buchailot, S. Nakamura, Y. Nakamura, M. Ataka, and H. Fujita, *Proc. 2nd Int. Conf. Shape Memory Superelastic Technol.*, Pacific Grove, CA, 1997, pp. 183–188.
35. R.H. Wolf and A.H. Heuer, *J. Microelectromechanical Sys.* **4**(4): 206–212 (1995).
36. H. Langen, J.-M. Breguet, H. Bleuler, P. Renaud, and T. Masuzawa, *Int. J. Elect. Mach.* **3**: 65–69 (1998).
37. D. Reynaerts, J. Peirs, and H. Van Brussel, *Sensors and Actuators* **77**: 157–166 (1999).
38. T. Fukuda, S. Guo, K. Kosuge, F. Arai, M. Negoro, and K. Nakabayashi, *IEEE Proc. Int. Conf. Robotics and Autom.*, San Diego, CA, 1994, Vol. 3, pp. 2290–2295.
39. S. Kaneko, S. Aramaki, K. Arai, Y. Takahashi, H. Adachi, and K. Yanagisawa, *J. Intelligent Mater. Syst. Struct.* **7**: 331–335 (1996).
40. K. Ikuta, *Proc. IEEE Int. Conf. Robotics Autom.*, Cincinnati, 1990, pp. 215–216.
41. J. Hesselbach, R. Pittschellis, E. Hornbogen, and M. Mertmann, *Proc. Shape Memory and Superelastic Technol.*, Pacific Grove, CA, 1997, pp. 251–256.
42. A. Lee, D. Ciarlo, P. Krulevich, S. Lehew, J. Trevino, and A. Northrup, *Sensors and Actuators* **A54**: 755–759 (1996).
43. Y. Bellouard, R. Clavel, J.-E. Bidaux, R. Gotthardt, T. Sidler, *Proc. Shape Memory Superelastic Technol.* Pacific Grove, CA, 1997, pp. 245–250.
44. D. Grant and V. Hayward, *IEEE Syst. Control Mag.* **17**(3): 80–88 (1997).
45. K. Kuribayashi, S. Shimizu, T. Nishinohara, and T. Taniguchi, *Proc. Int. Conf. Intelligent Robots Syst.*, Yokohama, Japan, 1993, pp. 1697–1702.
46. K. Kuribayashi, and T. Fujii, *Proc. Int. Symp. Micromechatronics Hum. Sci.* Piscataway, NJ, 1998, pp. 165–170.
47. Shape-Memory Alloys Micro-Actuator, US Pat. 5,061–914. June 27, 1989, J.D. Busch and A.D. Johnson.

48. A.D. Johnson, and V.V. Martynov, *Proc. 2nd Int. Conf. Shape Memory Superelastic Technol.*, Pacific Grove, CA, 1997, pp. 149–154.
49. M. Kohl, K.D. Strobanek, and S. Miyasaki, *Sensors and Actuators A72*: 243–250 (1999).
50. Y. Bellouard, T. Lehnert, T. Sidler, R. Gotthardt, and R. Clavel, *Mater. Res. Soc. Mater. Smart Syst. III 604*: 177–182 (2000).
51. Y. Bellouard, T. Lehnert, J.-E. Bidaux, T. Sidler, R. Clavel, and R. Gotthardt, *Mater. Sci. Eng. A273–275*: 733–737 (1999).
52. K. Kuribayashi, S. Shimizu, M. Yoshitake, and S. Ogawa, *Proc. 6th Int. Symp. Micro Mach. Hum. Sci.* Piscataway, NJ, 1995, pp. 103–110.
53. S.T. Smith and D.G. Chetwynd, Gordon and Breach, 1994.
54. J.M. Paros and L. Weisborg, *Mach. Design 27*: 151–156 (1965).
55. Y. Bellouard, Ph.D. Thesis, Lausanne, EPFL, n°2308 (2000).
56. W. Nix, *Scripta Materialia*, **39**(4/5): 545–554 (1998).
57. R.D. James, *Int. J. Solids Struct.* **37**: 239–250 (2000).
58. R. Gorbet, Ph. D. Thesis, University of Waterloo, 1997.

MICROTUBES

WESLEY P. HOFFMAN

Air Force Research Laboratory
AFRL/PRSM
Edwards AFB, CA

PHILLIP G. WAPNER

ERC Inc.
Edwards AFB, CA

INTRODUCTION

Background

Microtubes are very small diameter tubes (in the nanometer and micron range) that have very high aspect ratios and can be made from practically any material in any combination of cross-sectional and axial shape desired. In smart structures, these microscopic tubes can function as sensors and actuators, as well as components of fluidic logic systems. In many technological fields, including smart structures, microtube technology enables fabricating components and devices that have, to date, been impossible to produce, offers a lower cost route for fabricating some current products, and provides the opportunity to miniaturize numerous components and devices that are currently in existence.

In recent years, there has been tremendous interest in miniaturization due to the high payoff involved. The most graphic example that can be cited occurred in the electronics industry, which only 50 years ago relied exclusively on the vacuum tube for numerous functions. The advent of the transistor in 1947 and its gradual replacement of the vacuum tube started a revolution in miniaturization that was inconceivable at the time of its invention and is not fully recognized even many years later.

Miniaturization resulted in the possibility for billions of transistors to occupy the volume of a vacuum tube or the first transistor, and it was not the only consequence. The

subsequent spin-off developments in allied areas, such as integrated circuits and the microprocessor, have spawned entirely new fields of technology. It is quite likely that other areas are now poised for revolutionary developments that parallel those that have occurred in the electronics industry since the advent of the first transistor.

These areas include microelectromechanical systems (MEMS) and closely related fields, such as microfluidics and micro-optical systems. Currently, these technologies involve micromachining on a silicon chip to produce numerous types of devices, such as sensors, detectors, gears, engines, actuators, valves, pumps, motors, and mirrors on a micron scale. The first commercial product to arise from MEMS was the accelerometer that was manufactured as a sensor for air-bag actuation. On the market today are also microfluidic devices, mechanical resonators, biosensors for glucose, and disposable blood pressure sensors that are inserted into the body.

The vast majority of microsystems are made almost exclusively on planar surfaces using technology developed to fabricate electronic integrated circuits. The fabrication of these devices takes place on a silicon wafer, and the device is formed layer-by-layer using standard clean-room techniques that include electron beams or photolithography, thin-film deposition, and wet or dry etching (both isotropic and anisotropic). Three variations of this conventional electronic chip technology can be used, for example, to make three-dimensional structures that have high aspect ratios and suspended beams. These include the LIGA (lithographie, galvanofornung, abfornung) process (1,2), the Hexsil process (3), and the SCREAM (single-crystal reactive etching and metallization) process (4). The technique most employed, the LIGA process, which was developed specifically for MEMS-type applications, can construct and metallize high-aspect-ratio microfeatures. This is done by applying and exposing a very thick X-ray sensitive photoresist layer to synchrotron radiation. Features up to 600 microns high that have aspect ratios of 300 to 1 can be fabricated by this technique to make truly three-dimensional objects. The Hexsil process uses a mold that has a sacrificial layer of silicon dioxide to form polysilicon structures that are released by removing the silicon dioxide film. A third approach is the SCREAM bulk micromachining process that can fabricate high-aspect-ratio single-crystal silicon suspended microstructures from a silicon wafer using anisotropic reactive ion etching. Note, however, that like the conventional technique used to make electronic circuits, all of these variations use a layered approach that starts on a flat surface.

In addition, there are some disadvantages of the conventional electronic chip fabrication technique and its modifications, even though there have been numerous and very innovative successes using these silicon wafer-based technologies. This is due to the fact that these technologies require building up many layers of different materials as well as surface and bulk micromachining which leads to some very difficult material science problems that have to be solved. These include differential etching and laying down one material without damaging any previous layer. In addition, there are the problems of interconnecting layers in a chip that have different functions. An example of this is a microfluidic device in which there are both fluidic

48. A.D. Johnson, and V.V. Martynov, *Proc. 2nd Int. Conf. Shape Memory Superelastic Technol.*, Pacific Grove, CA, 1997, pp. 149–154.
49. M. Kohl, K.D. Strobanek, and S. Miyasaki, *Sensors and Actuators A72*: 243–250 (1999).
50. Y. Bellouard, T. Lehnert, T. Sidler, R. Gotthardt, and R. Clavel, *Mater. Res. Soc. Mater. Smart Syst. III 604*: 177–182 (2000).
51. Y. Bellouard, T. Lehnert, J.-E. Bidaux, T. Sidler, R. Clavel, and R. Gotthardt, *Mater. Sci. Eng. A273–275*: 733–737 (1999).
52. K. Kuribayashi, S. Shimizu, M. Yoshitake, and S. Ogawa, *Proc. 6th Int. Symp. Micro Mach. Hum. Sci.* Piscataway, NJ, 1995, pp. 103–110.
53. S.T. Smith and D.G. Chetwynd, Gordon and Breach, 1994.
54. J.M. Paros and L. Weisborg, *Mach. Design 27*: 151–156 (1965).
55. Y. Bellouard, Ph.D. Thesis, Lausanne, EPFL, n°2308 (2000).
56. W. Nix, *Scripta Materialia*, **39**(4/5): 545–554 (1998).
57. R.D. James, *Int. J. Solids Struct.* **37**: 239–250 (2000).
58. R. Gorbet, Ph. D. Thesis, University of Waterloo, 1997.

MICROTUBES

WESLEY P. HOFFMAN

Air Force Research Laboratory
AFRL/PRSM
Edwards AFB, CA

PHILLIP G. WAPNER

ERC Inc.
Edwards AFB, CA

INTRODUCTION

Background

Microtubes are very small diameter tubes (in the nanometer and micron range) that have very high aspect ratios and can be made from practically any material in any combination of cross-sectional and axial shape desired. In smart structures, these microscopic tubes can function as sensors and actuators, as well as components of fluidic logic systems. In many technological fields, including smart structures, microtube technology enables fabricating components and devices that have, to date, been impossible to produce, offers a lower cost route for fabricating some current products, and provides the opportunity to miniaturize numerous components and devices that are currently in existence.

In recent years, there has been tremendous interest in miniaturization due to the high payoff involved. The most graphic example that can be cited occurred in the electronics industry, which only 50 years ago relied exclusively on the vacuum tube for numerous functions. The advent of the transistor in 1947 and its gradual replacement of the vacuum tube started a revolution in miniaturization that was inconceivable at the time of its invention and is not fully recognized even many years later.

Miniaturization resulted in the possibility for billions of transistors to occupy the volume of a vacuum tube or the first transistor, and it was not the only consequence. The

subsequent spin-off developments in allied areas, such as integrated circuits and the microprocessor, have spawned entirely new fields of technology. It is quite likely that other areas are now poised for revolutionary developments that parallel those that have occurred in the electronics industry since the advent of the first transistor.

These areas include microelectromechanical systems (MEMS) and closely related fields, such as microfluidics and micro-optical systems. Currently, these technologies involve micromachining on a silicon chip to produce numerous types of devices, such as sensors, detectors, gears, engines, actuators, valves, pumps, motors, and mirrors on a micron scale. The first commercial product to arise from MEMS was the accelerometer that was manufactured as a sensor for air-bag actuation. On the market today are also microfluidic devices, mechanical resonators, biosensors for glucose, and disposable blood pressure sensors that are inserted into the body.

The vast majority of microsystems are made almost exclusively on planar surfaces using technology developed to fabricate electronic integrated circuits. The fabrication of these devices takes place on a silicon wafer, and the device is formed layer-by-layer using standard clean-room techniques that include electron beams or photolithography, thin-film deposition, and wet or dry etching (both isotropic and anisotropic). Three variations of this conventional electronic chip technology can be used, for example, to make three-dimensional structures that have high aspect ratios and suspended beams. These include the LIGA (lithographie, galvanofornung, abfornung) process (1,2), the Hexsil process (3), and the SCREAM (single-crystal reactive etching and metallization) process (4). The technique most employed, the LIGA process, which was developed specifically for MEMS-type applications, can construct and metallize high-aspect-ratio microfeatures. This is done by applying and exposing a very thick X-ray sensitive photoresist layer to synchrotron radiation. Features up to 600 microns high that have aspect ratios of 300 to 1 can be fabricated by this technique to make truly three-dimensional objects. The Hexsil process uses a mold that has a sacrificial layer of silicon dioxide to form polysilicon structures that are released by removing the silicon dioxide film. A third approach is the SCREAM bulk micromachining process that can fabricate high-aspect-ratio single-crystal silicon suspended microstructures from a silicon wafer using anisotropic reactive ion etching. Note, however, that like the conventional technique used to make electronic circuits, all of these variations use a layered approach that starts on a flat surface.

In addition, there are some disadvantages of the conventional electronic chip fabrication technique and its modifications, even though there have been numerous and very innovative successes using these silicon wafer-based technologies. This is due to the fact that these technologies require building up many layers of different materials as well as surface and bulk micromachining which leads to some very difficult material science problems that have to be solved. These include differential etching and laying down one material without damaging any previous layer. In addition, there are the problems of interconnecting layers in a chip that have different functions. An example of this is a microfluidic device in which there are both fluidic

and electronic functions. Clearly, there are numerous materials issues central to this technology.

Other technologies are available that, like conventional lithography, can construct or replicate microscopic features on a flat surface. These approaches include imprint lithography that involves compression molding (5), lasers (6–8), ion beams (9) and electron beam (10) micro-machining, soft lithography (11), writing features into the surface using an atomic force microscope (12,13), and very limited application of deposition using a scanning tunneling microscope (14,15). The majority of these technologies are not discussed in detail because there is not a close link to microtube technology.

In addition to the processing problems mentioned before, there are other limitations inherent in conventional lithographic techniques that are based on planar silicon. For example, in some applications such as those that involve surface tension in fluidics, it is important to have a circular cross section. However, it is impossible to make a perfectly round tube or channel on a chip by conventional technology. Instead, channels on the wafer surface are made by etching a trench and then covering the trench by using a plate (16,17). This process can produce only angled channels such as those that have a square, rectangular, or triangular cross section. Because of the limitations already mentioned, we heartily agree with Wise and Najafi in their review of microfabrication technology (18) when they stated, “The planar nature of silicon technology is a major limitation for many future systems, including microvalves and pumps.”

In the literature, there are at least two technologies in addition to microtubes that remove microfabrication from the flatland of the wafer. One uses “soft lithography,” and the other uses laser-assisted chemical vapor deposition (LCVD). “Soft lithography,” conceived and developed by Whitesides’ outstanding group at Harvard, encompasses a series of very novel related technologies that include microcontact printing, micromolding, and micromolding in capillaries (11). These technologies can fabricate structures from several different materials on flat and curved surfaces. By example, structures can be fabricated using microcontact printing by first making a stamp that contains the desired features. This stamp, which is usually made from poly(dimethylsiloxane) (PDMS) has raised features placed on the surface by photolithographic techniques. The raised features are “inked” with an alkanethiol and then brought into contact with a gold-coated surface, for example, by rolling the curved surface over the stamp. The gold is then etched where there is no self-assembled monolayer of alkanethiolate. Features as small as 200 nm can be formed by this technique. However, the microstructures produced by this technique are the same as those produced by standard techniques, except that the starting surface need not be flat. By using these techniques, submicron features can be fabricated on flat or curved substrates made of materials, such as metals (19), polymers (20), and carbon (21). In addition, these technologies can be used to make truly three-dimensional free-standing objects (22,23).

Another step away from the standard planar silicon technology is the LCVD process (24,25) which can “write in space” to produce three-dimensional microsystems. In

this process, two intersecting laser beams are focused in a very small volume in a low-pressure chamber. The surface of the substrate on which deposition is to occur is brought to the focal point of the lasers. The power to the lasers is adjusted so that deposition from the gas phase occurs only at the intersection of the beams. As deposition occurs on the substrate surface, it is pulled away from the focal point. Under computer control, the substrate can be manipulated so that complex, free-standing, three-dimensional microstructures can be fabricated.

In addition to LCVD and soft lithography, only microtube technology offers the possibility of truly three-dimensional nonplanar microsystems. However, in contrast to these two technologies, microtube technology also offers the ability to make microdevices from practically any material because the technology is not limited by electrodeposition or the availability of CVD precursor materials. In addition, in contrast to these other technologies, microtube technology provides the opportunity to make tubing and also to make it in a variety of cross-sectional and axial shapes that can be used to miniaturize systems, connect components, and fabricate components or systems that are not currently possible to produce.

Microscopic and Nanoscopic Tubes and Tubules

Commercially, tubing is extruded, drawn, pultruded, or rolled and welded which limits the types of materials that can be used for ultrasmall tubes as well as their ultimate internal diameters. In addition, it is not currently possible to control the wall thickness, internal diameter, or the surface roughness of the inner wall of these tubes to a fraction of a micron by these techniques. Using conventional techniques, ceramic tubes are currently available only as small as 1 mm i.d. Copper tubing can be obtained as small as 0.05 mm i.d., polyimide tubing is fabricated as small as 50 μm i.d., and quartz tubing is drawn down as small as 2 μm i.d. This means that quartz is the only tubing commercially available that is less than 10 μm i.d. This quartz tubing is used principally for chromatography.

There are, however, other sources of small tubing that are presently at various stages of research and development. For some time, several groups have been using lipids as templates (26–28) to fabricate submicron diameter tubing. These tubes are made by using electroless deposition to metallize a tubular lipid structure formed from a Langmuir–Blodgett film. Lipid templated tubes are very uniform in diameter, which is fixed at $\sim 0.5 \mu\text{m}$ by the lipid structure. Lengths to 100 μm have been obtained by this technique which is extremely expensive due to the cost of the raw materials.

Other groups are making submicron diameter tubules using a membrane-based synthetic approach. This method involves depositing the desired tubule material within the cylindrical pores of a nanoporous membrane. Commercial “track-etch” polymeric membranes and anodic aluminum oxide films have been used as the porous substrate. Aluminum oxide, which is electrochemically etched, has been the preferred substrate because pores of uniform diameter can be made from 5–1000 nm. Martin (29–31) polymerized electrically conductive polymers from the liquid phase and electrochemically deposited metal in the

pore structure of the membrane. Kyotani et al. (32,33) deposited pyrolytic carbon inside the pores of the same type of alumina substrate. In each case, after the inside walls of the porous membrane are covered to the desired thickness, the porous membrane is dissolved leaving the tubules. A variation of this technique, used by Hoyer (34,35) to form semiconductor (CdS, TiO₂, and WO₃) nanotubes, includes an additional step. Instead of coating the pore wall directly to form the tubule, he fills the pore with a sacrificial material, solvates the membrane, and then coats the sacrificial material with the material for the nanotube wall. The sacrificial material is finally removed to form the nanotube. As in the lipid process, all of the tubules formed by this process in a single membrane are uniform in diameter, length, and thickness. But in contrast to the lipid process, the diameter of the tubules can be varied by the extent of oxidation of the aluminum substrate. Although diameters can be varied in this process, it should be clear that these tubules are limited in length to the thickness of the porous membrane. In addition, the wall thickness is also limited in that the sum of the inside tubule diameter and two times the wall thickness is equal to the starting pore diameter.

Using a sol-gel method, tubules can be made in about the same diameter range as in the membrane approach. By hydrolyzing tetraethylorthosilicate at room temperature in a mixture of ethanol, ammonia, water and tartaric acid, Nakamura and Matsui (36) made silica tubes that had both square and round interiors. The tubules produced by this technique were up to 300 μm long, and the i.d. of the tubes ranged from 0.02 to 0.8 μm . By introducing minute bubbles into the sol, hollow TiO₂ fibers that have internal diameters up to 100 μm have also been made (37) by using the sol-gel approach.

On an even smaller scale, nanotubes are fabricated using a number of very different techniques. The most well-known tube in this category is the carbon "buckytube" that is a cousin of the C₆₀ buckyball (38–42). Since carbon nanotubes were first observed as a by-product in C₆₀ production, the method of C₆₀ formation using an arc-discharge plasma was modified to enhance nanotube production. The process produces tubules whose i.d. is in the range of 1–30 nm. These tubules are also limited in length to about 20 microns. Similar nanotubes of BN (43), B₃C, and BC₂N (44) have been made by a very similar arc-discharge process. In addition, nanotubes of other compositions (45,46) have been prepared using carbon nanotubes as a substrate for conversion or deposition.

An alternative technique for manufacturing carbon tubes that have nanometer diameters has been known to the carbon community for decades from the work of Bacon, Baker, and others (47–50). The process produces a hollow catalytic carbon fiber by pyrolyzing a hydrocarbon gas over a catalyst particle. The fibers, which vary in diameter from 1 nm to 0.1 μm have lengths up to centimeters, can be grown either hollow or has an amorphous center that can be removed by catalytic oxidation after a fiber is formed.

Other nanoscale tubules whose diameters are slightly larger and smaller than buckytubes have been made from bacteria and components of cytoskeletons and by direct chemical syntheses. Chow and others (51) isolated and

purified nanoscale protein tubules called rhabidosomes from the bacterium *Aquaspirillum itersonii*. After the rhabidosomes are metallized by electroless deposition and the bacteria are removed, metal tubules approximately 17 nm in diameter and 400 nm long are produced. Using a similar metallization technique, metal tubes have been fabricated (52) whose inner diameters are 25 nm by using biological microtubules as templates. These microtubules, which are protein filaments of 25 nm o.d. and whose lengths are measured in microns, are components of the cytoskeletons of eukaryotic cells. In contrast to tubules produced from biological templates, the tubules produced by direct chemical synthesis involve using the technique of molecular self-assembly. Some of the nanotubules that fall into this category are made from cyclic peptides (53), cyclodextrins (54), and bolaamphiphiles (55). Cyclic peptide nanotubules have an 0.8 nm i.d. and can be made several microns in length. Other self-assembled nanotubules that range from 0.45 to 0.85 nm i.d. have been synthesized from cyclodextrins (54,56) in lengths in the tens of nanometers.

Although it is clear that individual nanotubules are currently useful for certain applications, such as encapsulation, reinforcement, or as scanning probe microscope tips (57), it is not obvious how individual nanotubules can be observed and economically manipulated for use in devices other than by using a scanning probe microscope (58). Until this problem is solved, the future of individual nanotubes in devices is uncertain. However, this problem can be circumvented if the nanotubules are part of a larger body such as in an array.

If oriented groups or arrays of submicron to micron diameter tubes or channels perpendicular to the surface of the wafer or device are desired, there are at least four means available to make them. Using the technique described before for making anodic porous alumina, a two step replication process (59) can be used to fabricate a highly ordered honeycomb nanohole array from gold or platinum. The metal hole array is from 1–3 micron thick and has holes 70 nanometers in diameter. For smaller tubes or channels, a technique (60) has recently been developed to draw down bundles of quartz tubes to form an array. This process produces a hexagonal array of glass tubes each as small as 33 nm in diameter. This translates to a density of 3×10^{10} channels per square centimeter. Even smaller regular arrays of channels can be synthesized by a liquid crystal template mechanism (61,62). In this process, aluminum silicate gels are calcined in the presence of surfactants to produce channels 2–10 nm in diameters. Finally, channels of ~ 4 nm in cross section can be produced (63) perpendicularly to the surface of an amorphous silica film by forming hematite crystals in a Fe–Si–O film and then etching away the hematite crystals.

Finally, several technologies exist to make channels or layers of channels of desired orientation in solid objects. These technologies are another spin-off of the photolithographic process used for integrated circuits. On a two-dimensional plane, channels that range in size from tens to hundreds of microns in width and depth have been fabricated (16,17) on the surface of silicon wafers by standard microphotolithographic techniques. Forming of microscopic channels and holes in other materials originated

in the rocket propulsion community in 1964. Work at Aerojet Inc. (64) produced metallic injectors and cooling channels in metallic parts using a process that included photolithographic etching of thin metallic platelets and stacking the platelets followed by diffusion bonding of the platelets to form a solid metallic object that has micron-sized channels. The group at Aerojet has recently modified its technique to use silicon nitride. Variations on this technique include electrochemical micromachining and sheet architecture technology.

Electrochemical micromachining (65,66) avoids generating toxic waste from acid etching by making the thin metal part covered with exposed photoresist the anode in an electrochemical cell where a nontoxic salt solution is the electrolyte. Sheet architecture technology (67) developed at Pacific Northwest National laboratory is used to fabricate numerous microscopic chemical and thermal systems, such as reactors, heat pumps, heat exchangers, and heat absorbers. These devices may consist of a single photolithographically etched or laser-machined laminate that has a cover bonded to seal the channels, as described before, or may consist of multiple layers of plastic or metal laminates bonded together.

It is quite apparent from this brief and incomplete review, that a number of very novel and innovative approaches have been used to make microsystems as well as tubes and channels whose diameters are in the range of nanometers to microns. In the next section, the basics of microtube technology which complements these other technologies are discussed.

AFRL MICROTUBE TECHNOLOGY

Properties and Production of Microtubes

Except for self-assembled tubules, the microtube technology developed at the Propulsion Directorate of the Air Force Research Laboratory (AFRL) can produce tubes in the size range of those made by all of the other techniques cited. In contrast to tubing currently on the market and the submicron laboratory scale tubing mentioned before, microtubes can be made from practically any material (including smart materials) and will have precisely controlled composition, diameter, and wall thickness in a great range of lengths. In addition, this technology can produce tubes in a great diversity of axial and cross-sectional geometries. For most materials, there is no upper diameter limit, and for practically any material, internal diameters greater than 5 μm are possible. In addition, for materials that can survive temperatures higher than 400°C, tubes can be made as small as 5 nanometers by using the same process.

To date, tubes have been made from metals (copper, nickel, aluminum, gold, platinum, and silver), ceramics (silicon carbide, carbon, silicon nitride, alumina, zirconia, and sapphire), glasses (silica), polymers (Teflon), alloys (stainless steel), and layered combinations (carbon/nickel and silver/sapphire) in sizes from 0.5–410 μm . Like many of the techniques described before, microtube technology employs a fugitive process that uses a sacrificial mandrel, which in this case is a fiber. High-quality coating

techniques very faithfully replicate the surface of the fiber on the inner wall of the coating after the fiber is removed. By a proper choice of fiber, coating, deposition method, and mandrel removal method, tubes of practically any composition can be fabricated. Obviously, a great deal of material science is involved in making precision tubes of high quality. Some scanning electron microscope (SEM) micrographs of a group of tubes are shown in Fig. 1.

Cross-sectional shapes and wall thickness can be very accurately controlled to a fraction of a micron, which is not possible by using any of the approaches cited before. Numerous cross-sectional shapes have already been made, and some of them are shown in Fig. 2. These micrographs should be sufficient to demonstrate that practically any cross-sectional shape imagined can be fabricated. As seen in Fig. 2, the wall thickness of the tubes can be held very uniform around the tube. It is also possible to control the wall thickness accurately along the length of the individual tubes and among the tubes in a batch or a continuous process. It can be seen in Fig. 2 that the walls can be made nonporous. It will be shown later that the microstructure of the walls and extent of porosity that the walls contain can also be controlled. In addition to the possibility of cross-sectional tube shapes, using a fugitive process also allows fabricating tubes that have practically any axial geometry, as is shown later.

The maximum length in which these tubes can be made has yet to be determined because it depends on many variables, such as the type of tube material, the composition of the sacrificial tube-forming material, and the degree of porosity in the wall. It is possible that there is no limitation in length for a tube that has a porous wall. For nonporous wall tubing, the maximum length would probably be in the meter range because there is a direct relationship between the tube i.d. and the maximum possible length. However, for most applications conceived to date, the length need only be of the order of a few centimeters. Based on a quick calculation, it is apparent that even "short" tubes have a tremendous aspect ratio. For instance, a 10- μm i.d. tube 25 cm long has an aspect ratio of 2500.

Using microtube technology, there is no upper limitation in wall thickness for most materials. To date, free-standing tubes have been made whose wall thickness range from 0.01–800 μm (Fig. 3a). Most of the microtubes tested to date have demonstrated surprising mechanical strength. In fact, preliminary studies of both copper and silver tubes whose wall thickness is in the micron range have shown that microtubes can have up to two times the tensile strength of an annealed wire of the same material of the same cross-sectional area. Besides precise control of the tube wall thickness and composition, the interior surface of these tube walls can have practically any desired texture or degree of roughness. In addition, the walls can range from nonporous to extremely porous, as seen in Fig. 4, and the interior or exterior surfaces of these tubes can be coated by one or more layers of other materials (Fig. 5).

In addition to free-standing microtubes, solid monolithic structures that have microchannels can be fabricated by making the tube walls so thick that the spaces between the tubes are filled (Fig. 6). The microchannels can be randomly oriented, or they can have a predetermined orientation.

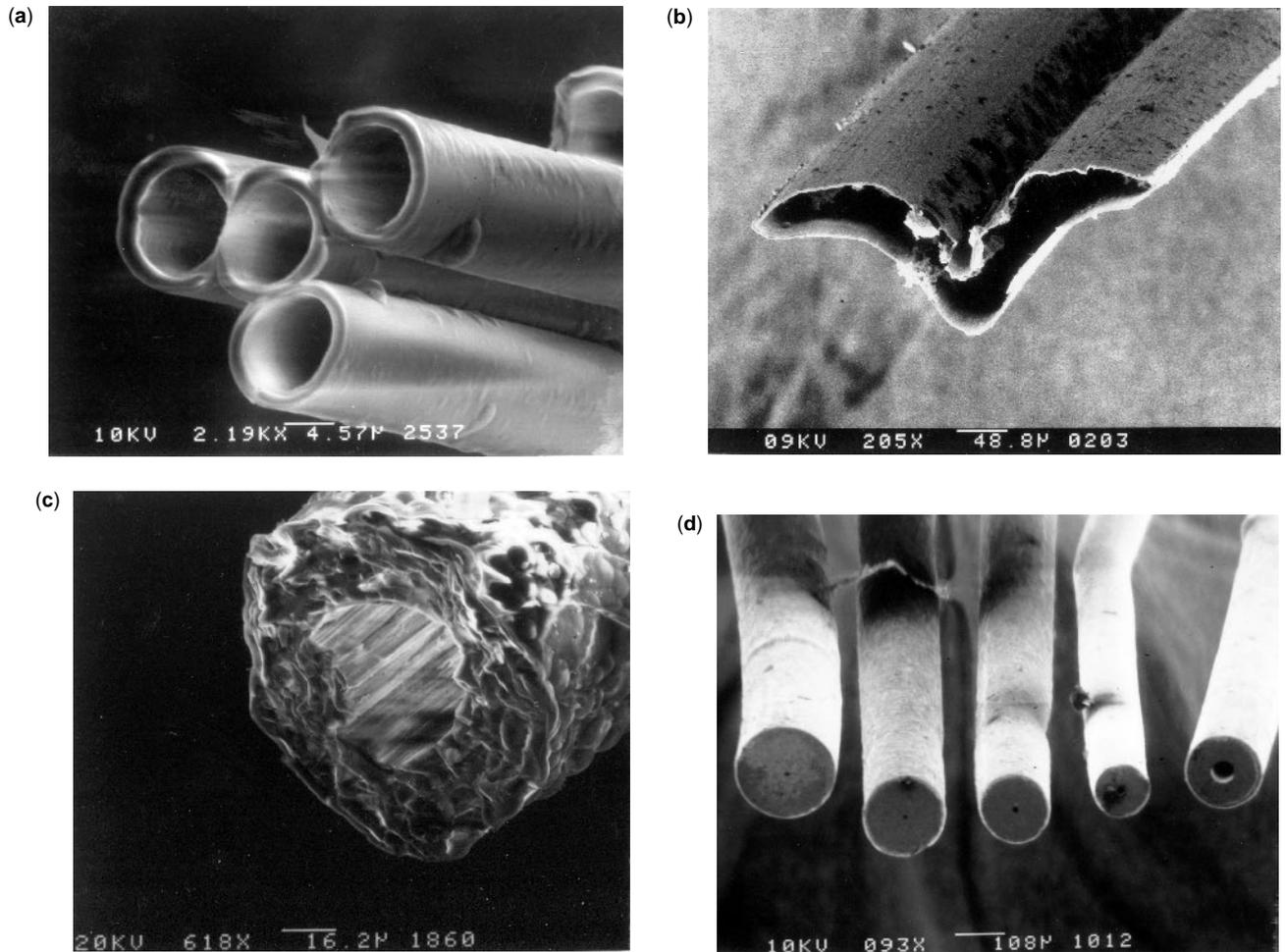


Figure 1. Examples of microtubes: (a) 10- μm silicon carbide tubes; (b) 410- μm nickel tubes; (c) 26- μm silicon nitride tube; and (d) 0.6- μm quartz tube.

Any desired orientation or configuration of microtubes can be obtained by a fixturing process. Alternatively, composite materials can be made by using a material different from the tube wall as a “matrix” that fills in the space among the tubes. The microtubes imbedded in these monolithic structures form oriented microchannels that, like free-standing tubes, can contain solids, liquids, and gases, and as act as waveguides for all types of electromagnetic energy.

Microtube Applications

Discrete thinner walled microtubes are useful in areas as diverse as spill cleanup, encapsulation of medicine or explosives, insulation that is usable across a very wide range of temperature, and as lightweight structural reinforcement similar to that found in bone or wood. The cross-sectional shape of these reinforcing tubes can be tailored to optimize mechanical or other properties. In addition, thinner walled tubes are useful as bending or extension actuators when fabricated from smart materials. Thicker walled tubes (Fig. 3b: nickel and SS) that are just as easily fabricated are needed in other applications, such

as calibrated leaks and applications that involve internal or external pressure on the tube wall.

The ability to coat the interior or exterior surface of these tubes with a layer or numerous layers of other materials enlarges the uses of the microtubes and also allows fabricating certain devices. For example, applying oxidation or corrosion protection layers on a structural or specialty tube material will greatly enlarge its uses. A catalyst can be coated on the inner and/or outer tube surface to enhance chemical reactions. The catalytic activity of the tube can also be enhanced by increasing the porosity in the wall, as shown before in Fig. 4. Multiple alternating conductive and insulating layers on a tube can provide a multiple-path microcoaxial conductor or a high-density microcapacitor.

As stated before, the interior surface of these tube walls can have practically any desired texture or degree of roughness. This control is highly advantageous and allows using microtubes in many diverse applications. For example, optical waveguides require very smooth walls, whereas catalytic reactors would benefit from rough walls. (Because of the fabrication technique, the roughness of the tube wall interior can be quantified to a fraction of a micron by using scanning probe microscopy techniques on the mandrel.)

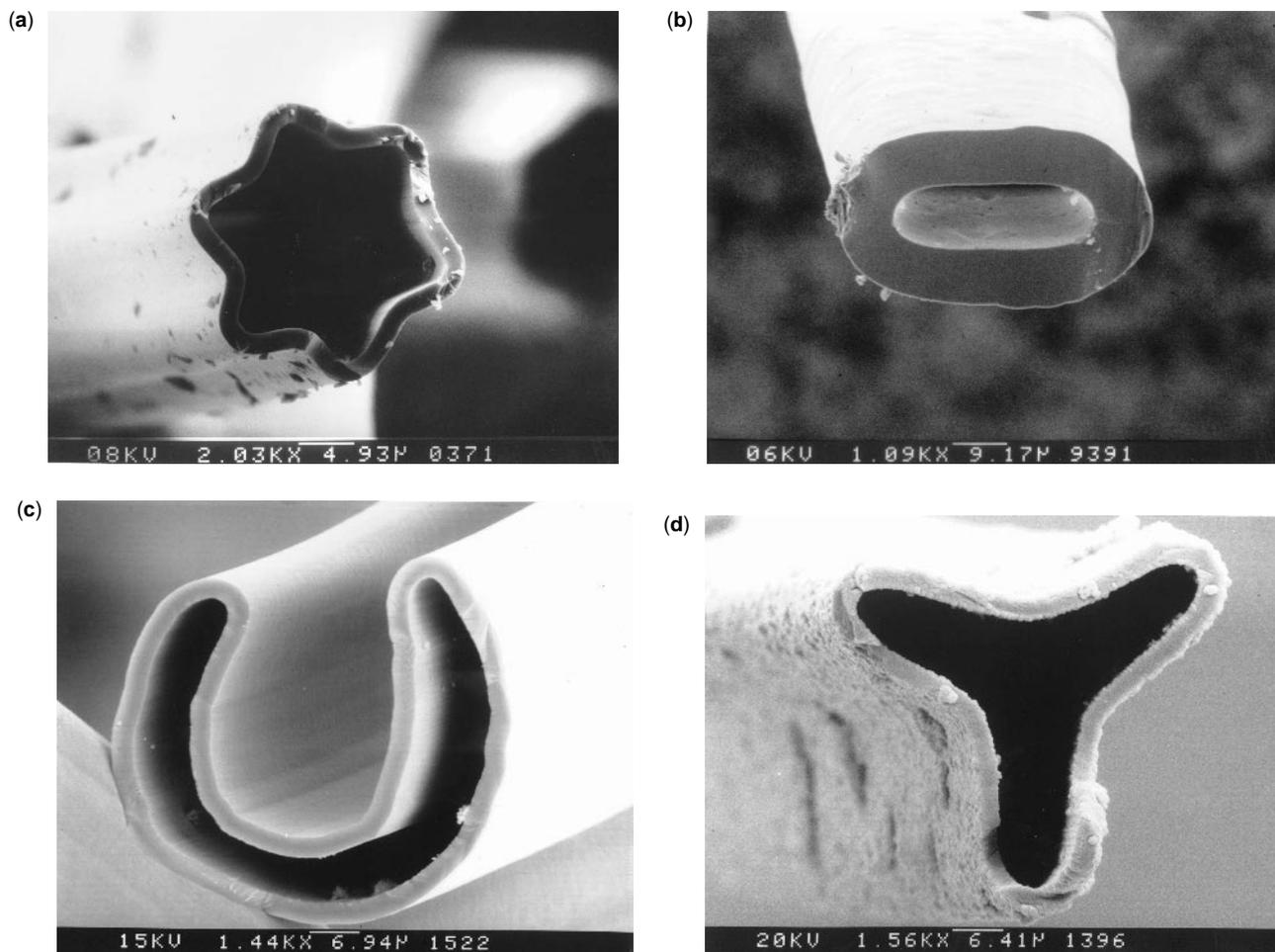


Figure 2. Tubes larger than $1\ \mu\text{m}$ i.d. can be made in any cross-sectional shape such as (a) $17\text{-}\mu\text{m}$ star, (b) $9 \times 34\text{-}\mu\text{m}$ oval, (c) $59\text{-}\mu\text{m}$ smile, and (d) a $45\text{-}\mu\text{m}$ trilobal shape.

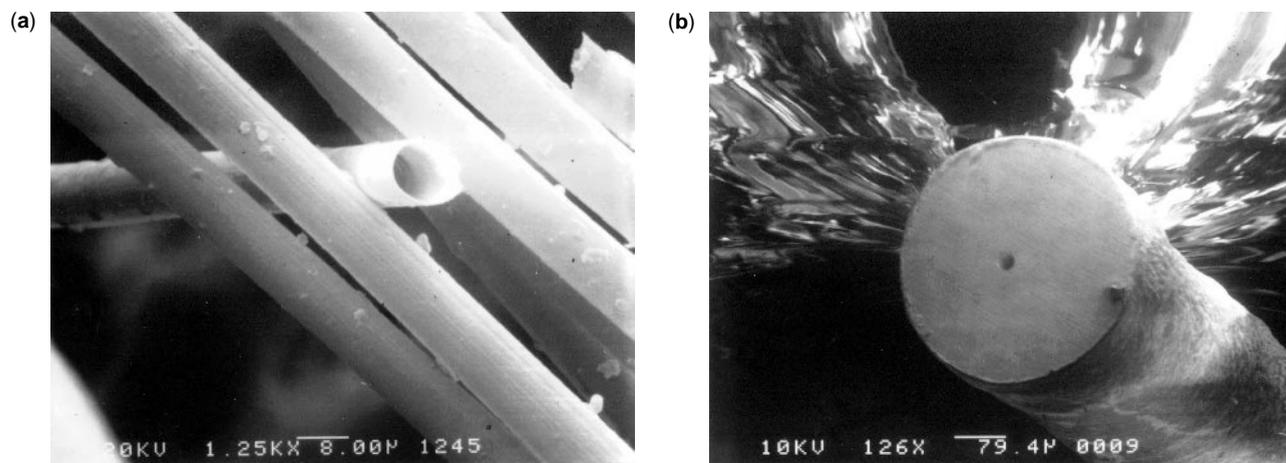


Figure 3. Tubes can be structurally sound and have (a) very thin walls or (b) thick walls.

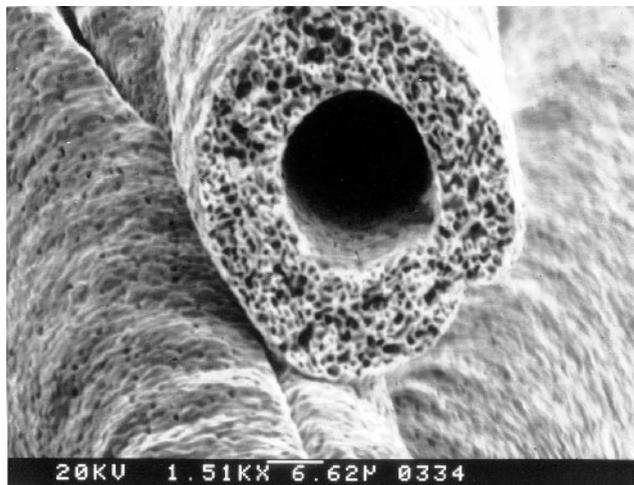


Figure 4. Microtube that has a porous tube wall.

Microtubes can be made straight or curved (Fig. 7), or they can be coiled (Fig. 8). Coiled tubes whose coils are as small as $20\ \mu\text{m}$ can be used, for example, as flexible connectors or solenoid coils. For the latter application, the coils could be of metal or of a high temperature superconductor where liquid nitrogen flows through the

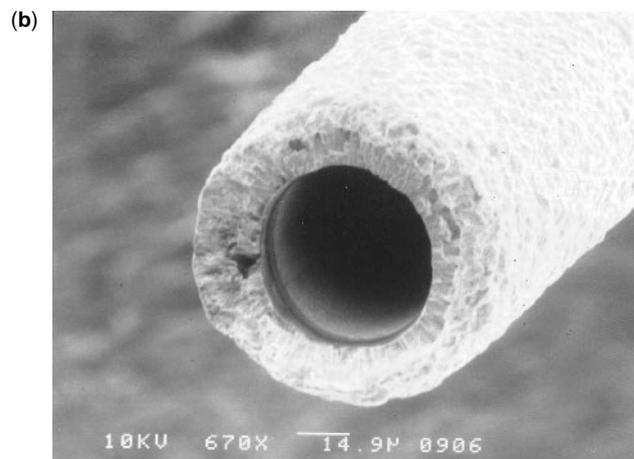
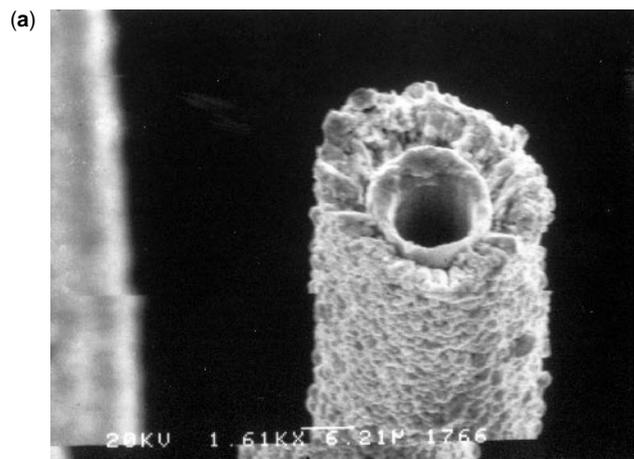


Figure 5. (a) Sapphire tube that has a silver liner. (b) Nickel tube that has a silver liner.

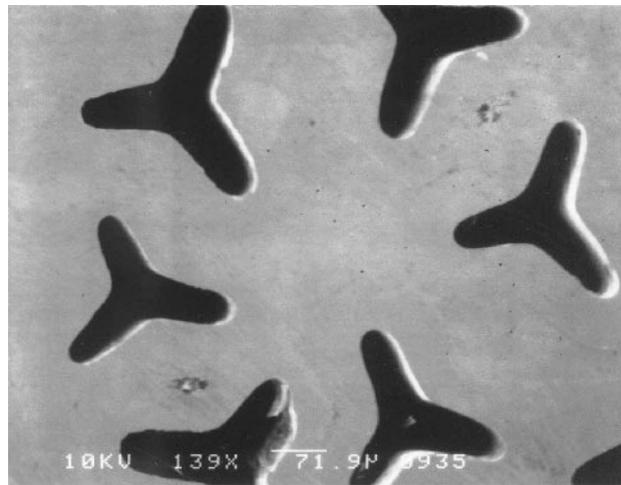


Figure 6. Solid nickel structure that has oriented microchannels.

tube. Another application for coils is for force or pressure measurement. No longer are we limited to quartz microsprings. Using microtube technology, the diameter and wall thickness of the tube, the diameter of the coil, the tube material, and the coil spacing can be very precisely

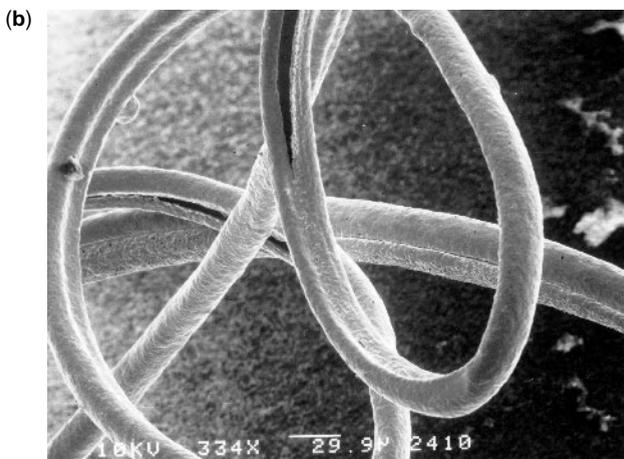


Figure 7. Examples of curved silver tubes: (a) single tube; (b) multiple tubes.

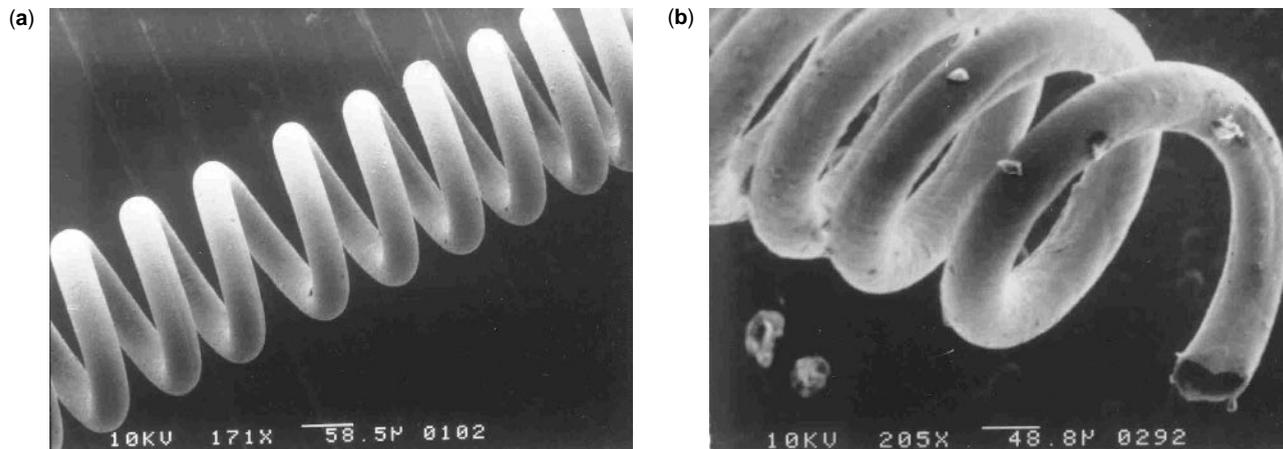


Figure 8. (a) Section of "large" coiled tube. (b) Open end of coiled tube.

controlled to give whatever spring constant is needed for the specific application. In addition, these microcoils can be made from a variety of smart materials and used as actuators or sensors. For example, the length of a spring made from Nitinol[®] can easily be changed by applying heat. It is also possible to wrap one or more coiled spring tubes around a core tube (Fig. 9). Applications for this kind of device range from a counterflow heat exchanger to a screwdrive for micromachines. (For the screw application, the wrapped coil cross section could be made rectangular.)

Like coiled spring tubes, bellows can be used as microinterconnects, sensors, and actuators and can be made in practically any shape imaginable. Figure 10a shows a bellows that has a circular cross section, and the bellows in Fig. 10b has a square cross section and aligned bellows segments. The bellows in Fig. 10c is square and has a twist. A slightly more complex bellows shown in Fig. 10d is a tapered-square camera bellows that has a sunshade to demonstrate the unique capability of this technology. It demonstrates the ability to control cross section and axial shape and to decrease and increase the cross-sectional

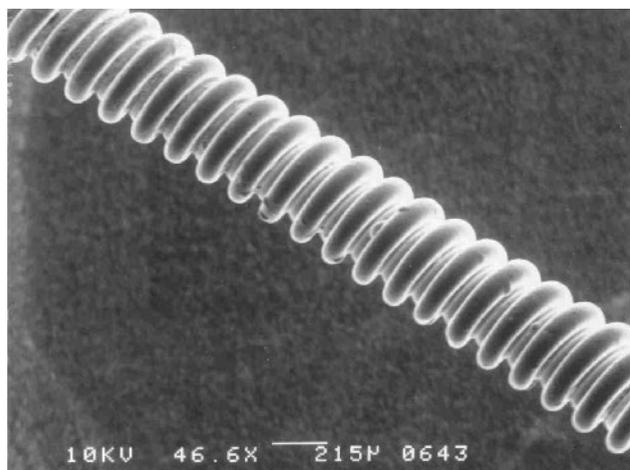


Figure 9. A coiled tube wrapped around a tube or fiber that can be used as a heat exchanger or as a microscopic screwdrive.

dimension in the same device. Bellows fabricated by microtube technology can have a variety of shaped ends for connections to systems for use, for example, as finned heat exchangers, hydraulic couplings for gas and liquid, or static mixers for multiple fluids. The bellows in Fig. 10e has a thicker transitional region and a dovetail on the end for connection to a device machined on a silicon wafer. The female dovetail to mate with this bellows is a commercially available trench design (68) on a silicon wafer that provides a way to attach the bellows to the wafer, which can be pressurized by using proper sealing. (No other technology available can join a fluidic coupling to a wafer for pressurization to relatively high pressures.)

If one end of the bellows is sealed, an entirely new group of applications becomes possible. For example, if a bellows end is sealed, the bellows can be extended hydraulically or pneumatically. In this configuration, a bellows could be used as a positive displacement pump, a valve actuator, or for micromanipulation. As a manipulator, a single bellows could be used for linear motion, three bellows could be orthogonally placed for 3-D motion, or three bellows could be attached at several places externally along their axes (Fig. 11) and differentially pressurized to produce a bending motion. This bending motion would produce a microfinger, and several of these fingers would make up a hand. The large forces and displacements possible by using this technique far surpass those currently possible by electrostatic or piezoelectric means and fulfill the need expressed by Wise and Najafi (18) when they stated that "In the area of micro-actuators, we badly need drive mechanisms capable of producing high force and high displacement simultaneously."

For most applications, it is necessary to interface microtubes and the macroworld. This is possible in a number of ways. For example, a tapering process can be used in which the diameter is gradually decreased to micron dimensions. Alternatively, the tubes and the macroworld can be interfaced by telescoping or numerous types of manifolding schemes (Fig. 12). An example of a thin-walled 5- μm i.d. tube telescoped to a 250- μm o.d. tube is shown in Fig. 13. A tube of this type could be used as a micropitot tube and, of course, could be made more robust by thickening the walls.

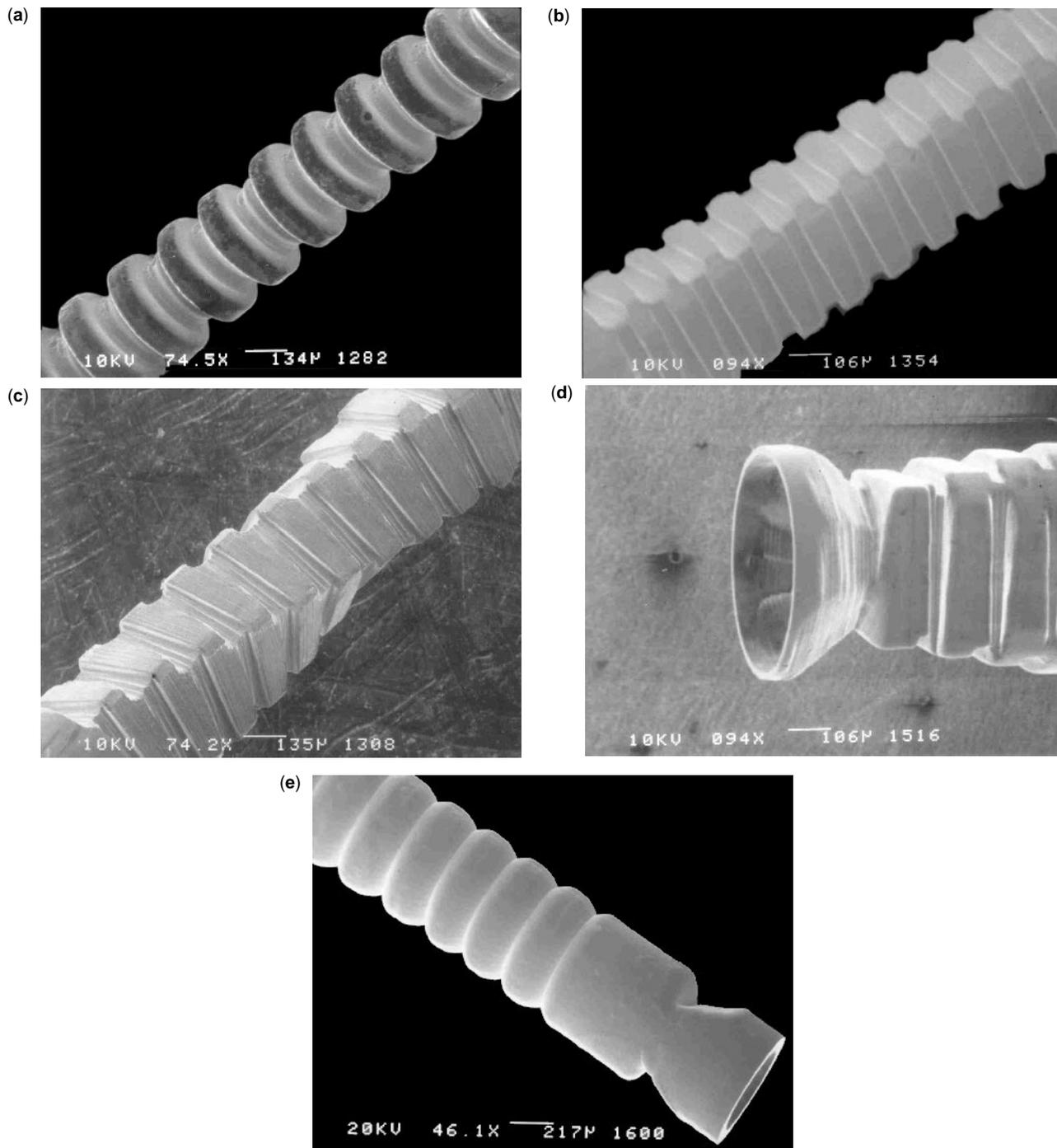


Figure 10. (a) A conventional round bellows. (b) A straight bellows that has a square cross section (c) A square bellows that has a twist. (d) A tapered square camera bellows that has a sun shade to demonstrate the versatility of the technique. (e) A round bellows that has a dovetail connector.

Although microtube technology has unique capabilities, it should be obvious that no single technology can fill all of the requirements imposed by diverse applications. Thus, microtube technology cannot easily compete with other technologies in certain applications. One of these involves gas and liquid separation such as in chromatography. For example, quartz tubing that can be extruded and drawn

in very long lengths is inexpensive and available in micron dimensions. However, note that even in areas such as separation, there are niches for microtubes that involve the composition of the tube material, the cross-sectional shape, or the inner wall coating. For example, Fig. 14 shows microtubes manifolded to a tubular frame for a specific gas separation that requires microtubes

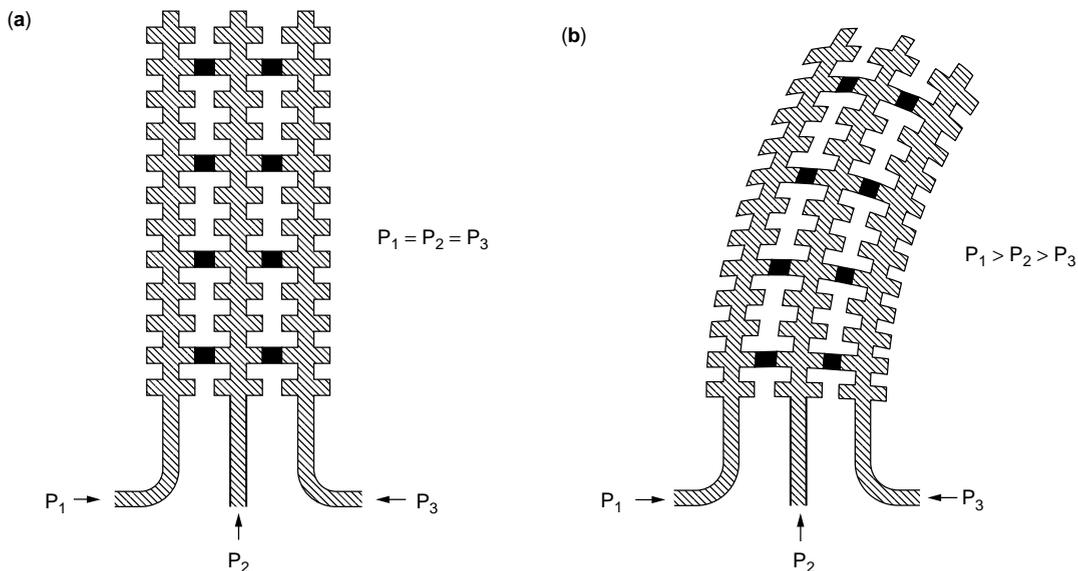


Figure 11. Microtube bellows finger: (a) unpressurized; (b) pressurized.

of a specific composition, precise diameter, and wall thickness.

Currently, these tubes have been made by a batch process in the laboratory, but the technique is equally suited to a continuous process which would be more efficient and also much easier in some cases. Obviously, a continuous process would reduce costs. For most materials, costs are already rather low because, unlike some other processes, expensive tooling is not required. For many materials such as quartz, aluminum, and copper, the anticipated cost is

~\$0.01/cm for thin-walled tubes. For precious metals such as gold or platinum, the cost would be significantly higher due to the cost of raw materials.

Microtubes have almost universal application in areas as diverse as optics, electronics, medical technology, and microelectromechanical devices. Specific applications for microtubes are as diverse as chromatography, encapsulation, cross- and counterflow heat exchange, injectors, micropipettes, dies, composite reinforcement, detectors, micropore filters, hollow insulation, displays, sensors,

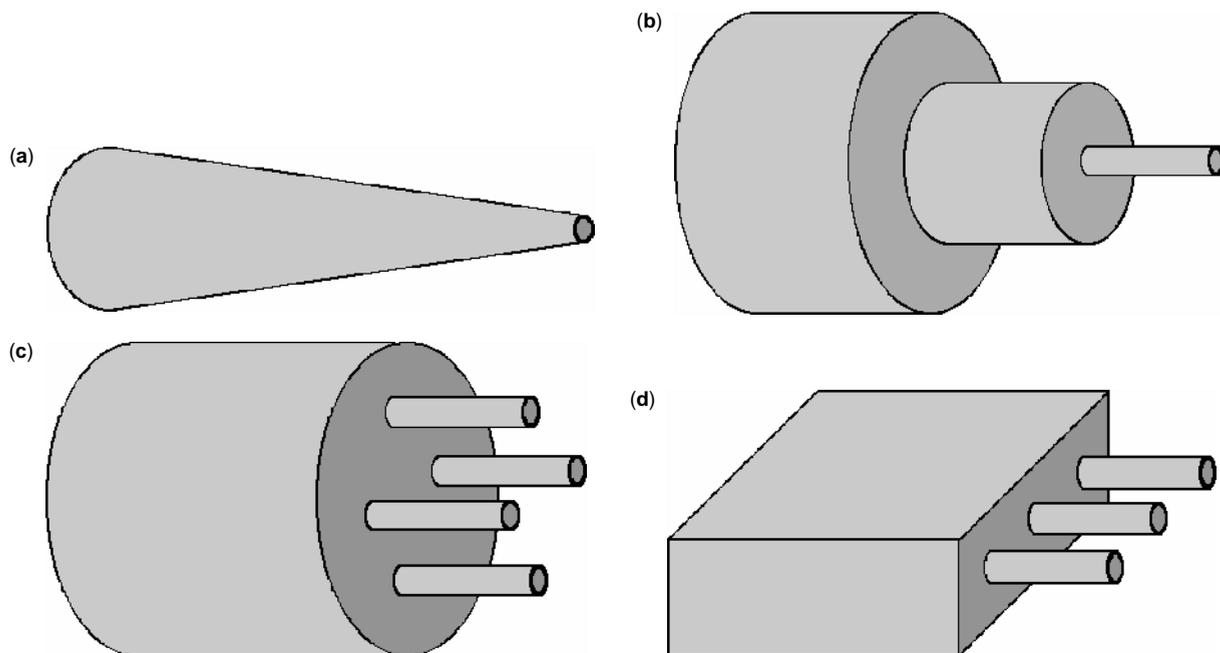


Figure 12. Different ways of transitioning microtubes to the real world: (a) taper, (b) telescope, (c) bundle, and (d) manifold.

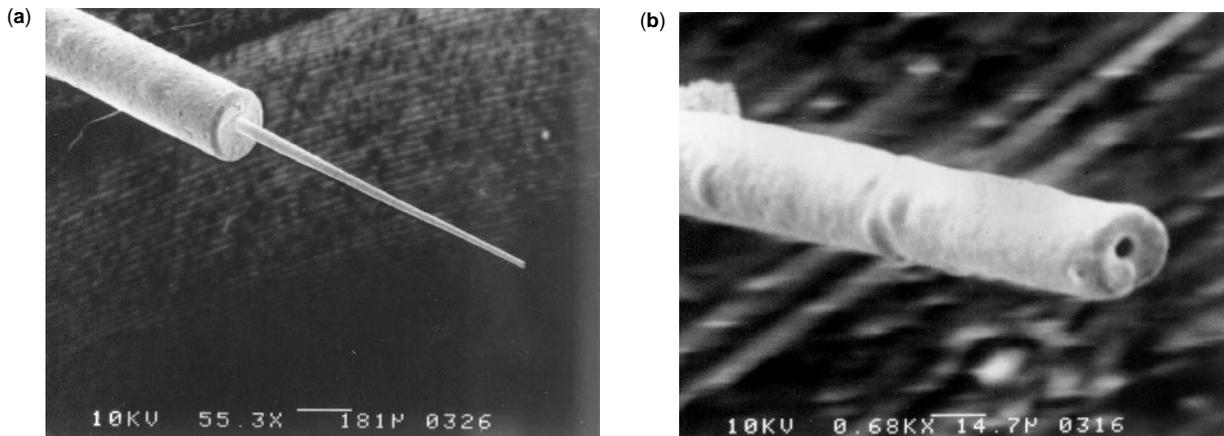


Figure 13. (a) A thin-walled 5- μm i.d. tube telescoped to a 250- μm o.d. tube. (b) View of the small open end of the telescope.

optical waveguides, flow control, pinpoint lubrication, microsponges, heat pipes, microprobes, and plumbing for micromotors and refrigerators. The technology works equally well for high- and low-temperature materials and appears feasible for all applications that have been conceived to date. As can be seen, there are numerous types of devices that have become possible as a result of microtube technology. One category of devices that is highlighted is that based on surface tension and wettability.

MICROTUBE DEVICES BASED ON SURFACE TENSION AND WETTABILITY

Now, there is great interest in developing microfluidic systems to decrease the size of current devices, increase their speed and efficiency, and decrease their cost because microfluidic systems have the potential, for example, for drastically decreasing the cost of certain health tests, allowing implantable drug delivery systems, and very significantly reducing the time needed to complete the Humane Genome Project. Microtube technology based on surface tension and wettability is unique in its capabilities and is truly an enabling technology in the microfluidic field.

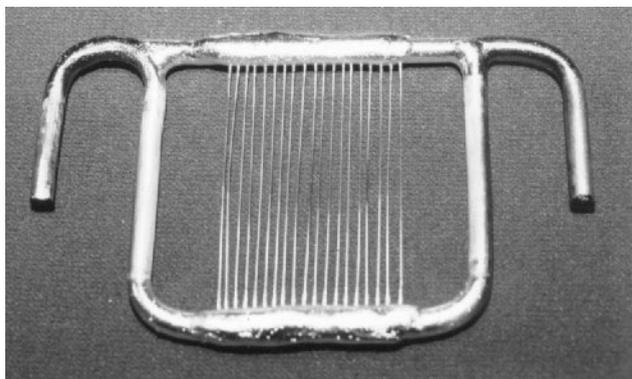


Figure 14. Microtubes are manifolded to a tubular frame for gas separation.

As miniaturization of mechanical, electrical, and fluidic systems occurs, the role of physical and chemical effects and parameters has to be reappraised. Some effects, such as those due to gravity or ambient atmospheric pressure, are relegated to minor roles or can even be disregarded entirely as miniaturization progresses. Meanwhile, other effects become elevated in importance or, in some cases, actually become the dominating variables. This “downsizing reappraisal” is vital to successful miniaturization. In a very real manner of speaking, new worlds are entered into in which design considerations and forces that are normally negligible in real-world applications become essential to successful use and application of miniaturized technology.

Surface tension and wettability are closely related phenomena that are greatly elevated in importance as miniaturization proceeds. Surface tension involves only the strength of attraction of droplet molecules for one another (cohesive forces), but wettability also includes the strength of attraction of droplet molecules to molecules of the wall material (adhesive forces). It is important to realize that surface tension and wettability are usually not comparable in effect to normal physical forces at macroscopic levels. For example, surface tension is usually ignored when determining fluid flow through a pump or tube. Its effect is many orders of magnitude smaller than pressure drop caused by viscosity because the difference in pressure ΔP between the inside of a droplet and the outside is given by the Young and Laplace equation of capillary pressure (69,70):

$$\Delta P = 2\gamma/r. \quad (1)$$

In this equal-radii form of the capillary pressure law used for a spherical droplet, γ is surface tension and r is droplet radius. The pressure inside the droplet can be thought of as caused by a surface “skin,” similar to a balloon that holds air in. Instead of a thin membrane of rubber as in the case of balloons, however, confining forces in surface tension are caused by the affinity of molecules of droplet material for one another. Because molecules are missing a

binding partner looking outward on the surface of a drop, they pull on their nearest neighbors.

Normally, droplet dimensions in most macroscopic applications are measured in thousands of microns. Therefore, pressure differences due to surface tension are inconsequential and typically measure far less than atmospheric pressure. For comparison, pressure drops resulting from viscous flow are typically of the order of magnitude of tens of atmospheres. When r is of the order of microns, however, pressure differences due to surface tension become enormous and frequently surpass tens of atmospheres. This is precisely the reason that fine aerosol droplets are so difficult to form. However, the formation of tiny droplets is not specifically the focus of discussion here, but rather their behavior in miniature voids, such as cavities, capillaries, and channels that are shaped so that they partially confine the droplet. The position of droplets within such microvoids is governed by the surface tension of the droplet fluid, the wettability of the fluid with respect to microvoid walls contacted during displacement, the geometric configuration of the walls that confine the fluid droplets, and any pressure external to the droplet. Microdevices fabricated from these microvoids can be made to operate when wettabilities are greater than or less than 90° , but not exactly 90° . They can operate using either nonwetting or wetting fluids. The difference between wetting and nonwetting fluids in capillaries can be explained by using Fig. 15.

In Fig. 15a, a nonwetting fluid droplet is forced into a single microtube. An insertion pressure has to push the nonwetting droplet inside the microtube because of the repulsion between the droplet and the walls. Once it is inside, however, no further pressure is necessary. In fact, any pressure simply moves the nonwetting droplet along the microtube at a velocity determined by the applied pressure and the frictional forces between the droplet and the microtube wall. Note that the nonwetting droplet becomes elongated when it is constrained in the capillary and has a convex-shaped interface along the axis of the capillary. In addition, it can be seen that the radius of the nonwetting droplet is now greater than the radius of the microdevice tube and that the contact angle θ with the capillary surface is between 90° and 180° , which is the contact angle for a totally nonwetting droplet. In contrast, the situation is very different if the fluid totally wets the microtube surface, as seen

in Fig. 15b. In this case, the fluid is sucked into the microtube, and fluid flow is governed only by frictional forces. This is the situation in normal macroscopic applications. For wetting fluids, the ends of the droplets are concave because the walls of the microdevice are wet by the droplet and attract the droplet molecules. The contact angle for wetting fluids is between 0 and 90° , 0° indicates a totally wetting fluid. In this article, the term nonwetting refers to a contact angle greater than 90° , and the term wetting means a contact angle less than 90° .

The behavior of a microtube device that employs nonwetting droplets is easily understood if one compares it to the mercury intrusion method (71–73) of measuring the pore-size distribution within porous solids. This technique is based on the understanding that the pressure needed to force a nonwetting fluid into a capillary or a pore in a solid is given by the relationship proposed by Washburn (71):

$$P = 2\gamma \cos \theta / r \quad (2)$$

where θ is the contact angle of the fluid with the material under test, P is the external pressure applied to the nonwetting fluid, and r is the radius of the capillary or pore which, act is assumed for simplicity, is spherical and has a constant diameter. This equation is valid for any fluid in contact with a capillary or porous solid whose contact angle is greater than 90° . Once the external applied pressure exceeds that needed to insert the nonwetting fluid into a constant-diameter capillary or pore, the nonwetting fluid flows into that particular diameter capillary or pore until it fills it. Then, the volume of the intruded fluid is a direct measure of that particular capillary's or pore's void volume. If a smaller capillary or pore branches off the larger diameter void, it remains unfilled until the insertion pressure is raised sufficiently high that Eq. (2) is again satisfied, and the process repeats itself.

In contrast to the mercury intrusion method of determining pore volume, instead of determining the pore volume, the emphasis in devices based on microtube technology is placed on the movement and the position of the droplet in the confining voids. These droplets can be wetting or nonwetting. As will be apparent later, a myriad of smart microdevices are based on surface tension and wettability.

Because these microdevices have no moving mechanical parts, they are very reliable, can be used in both static and dynamic applications, and are very rugged. They can experience pressures or forces far beyond their normal operating range and still return to their original accuracy and precision. In addition, unlike technology built up on a silicon wafer, these microdevices can be made from practically any material. Thus, high-temperature microdevices can be fabricated by properly choosing the device and droplet material.

Devices That Use the Interaction of Nonwetting Droplets and Gases and Wetting Fluids

This group of devices uses the surface properties of materials, primarily surface tension and wettability, as the principal means of actuating and controlling motion by and

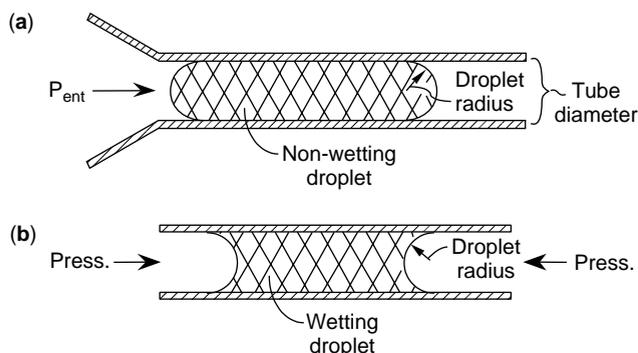


Figure 15. Behavior of fluid droplets in capillaries: (a) nonwetting droplet; (b) wetting droplet.

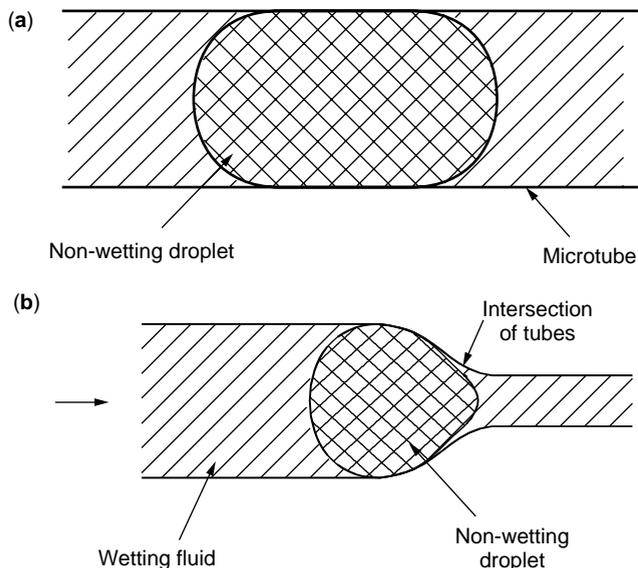


Figure 16. Nonwetting droplet inserted into a microtube under pressure. (a) Constant diameter tube. (b) Tube that has a transition to a smaller diameter.

within microtube devices. These devices, which have no moving mechanical parts, can perform mechanical tasks whose scale of motion is measured in microns.

These devices, similar to other microdevices based on surface tension and wettability, are composed of various sizes of nonwetting droplets inserted into microscopic voids of various shapes and sizes. These voids can be in the form of cavities, capillaries, and channels that are shaped so that they partially confine the droplet. Gas or wetting fluid is placed in the microcavities along with the nonwetting fluid. During operation, the nonwetting droplets move in response to fluid or gas pressure or vice versa. Specifically, these nonwetting droplets may translate within a void of the microtube device that is filled with the gas or wetting fluid, translate from one void space to another, or rotate in a fixed position. Microtube devices of this type can stop fluid flow or act as a check valve, a flow restricter, a flow regulator, or a gate, for example. The minimum dimension of the voids in these devices typically ranges from about 20 nm to about 1000 μm .

In Fig. 16a, a non-wetting fluid droplet is forced through a single microtube. An initial insertion pressure has to push the nonwetting droplet inside the microtube. If the diameter of the microtube in Fig. 16a decreases at a certain point to form a telescoping microtube (Fig. 16b), a considerably higher pressure must be applied by a gas or wetting fluid to squeeze the nonwetting drop into the smaller section of the microtube. In contrast, if a wetting fluid is employed instead of a gas, as before, it is also sucked into the smaller diameter section, completely filling all the available space in the microcavity. By inserting an appropriately sized nonwetting droplet into a tapered microtube or a microtube that has a transition to a smaller dimension that is filled by a second fluid that wets the tube walls, all flow of the wetting fluid can be stopped by applying a pressure that forces the nonwetting droplet to block the

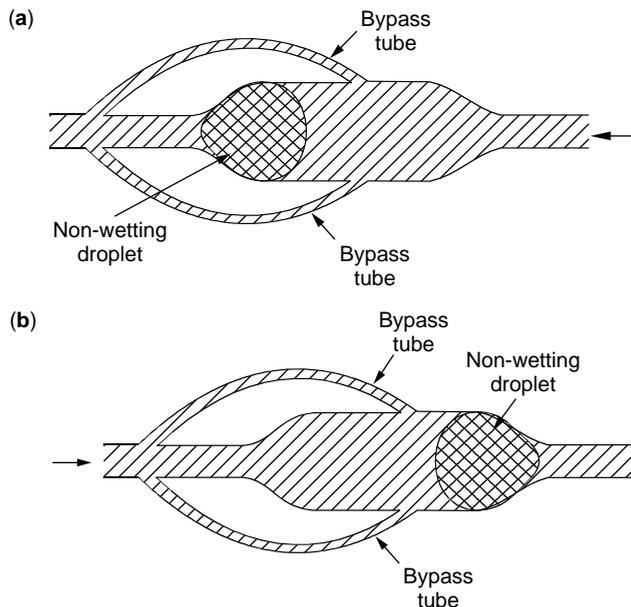


Figure 17. Microtube check valve: (a) flow possible through bypass tubes; (b) flow is blocked.

entrance to the smaller section of the cavity. This is the situation in Fig. 16b where the nonwetting droplet has been forced to the intersection of the larger and smaller microtube sections by the flowing gas or wetting fluid. Figure 17a,b illustrates an extension of this concept. By adding additional small-diameter bypass-flow paths to one end of a doubly constricted tube, flow is possible only in the direction of the end that has the added flow paths attached to the cavity. Of course, these bypass tubes must be properly sized to prevent nonwetting droplets from squeezing into them. This microtube device in Fig. 17 acts as a check valve and has no solid moving parts. This cannot be achieved at the macroscopic level because forces that arise from surface tensions of fluids are too small due to the much larger geometries employed.

Figures 18 and 19 are further extensions of this same concept. In Fig. 18, bypass tubes are left off the microtube check valve and convert it to either a microtube flow limiter (Fig. 18) or a microtube flow restricter (Fig. 19). In Fig. 18, because the nonwetting droplet and the larger tube wall form a seal, the only wetting fluid flow that can occur in either direction when the nonwetting droplet travels back and forth is equal to the volume of the larger tube section minus the volume of the nonwetting droplet. In Fig. 19, the diameter of the nonwetting droplet is now smaller than the diameter of the larger microtube section but larger than

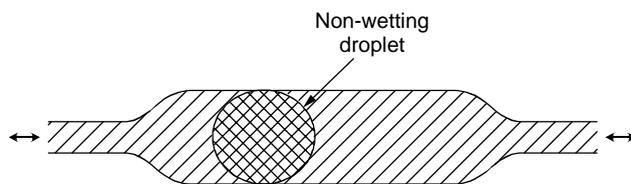


Figure 18. Microtube flow limiter.

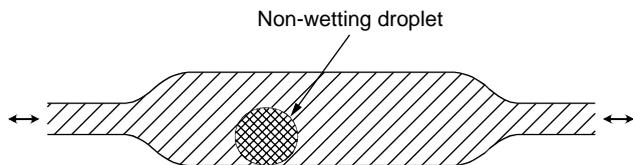


Figure 19. Microtube flow restricter.

the diameter of the smaller microtube. Thus, flow can take place around the nonwetting droplet. However, fluid flow is not merely restricted, but is entirely stopped if there is enough flow to push the drop to one end that blocks the smaller tube.

Figure 20 illustrates a microtube pressure/flow regulator. In this device, bypass tubes have openings or are open along their entire lengths to a conically shaped transitional region placed between the larger and smaller diameter tubes. Furthermore, the lengths of the joined bypass tubes (now better described as bypass channels) up to the conical transitional region can be varied. Increased pressure or flow forces the nonwetting droplet farther into the conical transitional region and exposes more flow channel openings to wetting fluid. The result is increased flow of the gas or wetting fluid as a function of pressure. By suitably sizing the nonwetting droplet, properly orienting the device, correctly shaping the transitional cone, and precisely positioning bypass channels, this device can also function as a microtube pressure-relief valve; no flow occurs until some predetermined pressure is exceeded. Then, flow takes place as long as pressure is maintained. Note that only two bypass flow channels are shown in Fig. 20. This was done to simplify the drawing. Any convenient number, one or more, of channels can be employed. Finally, by making bypass-flow channels vary in cross-sectional area, uniformly increasing or decreasing flow can be produced as a function of pressure.

In addition to a check valve, it is possible to use nonwetting fluids to make a positive closure valve that has zero dead space to control a gas or wetting fluid. Figure 21a,b

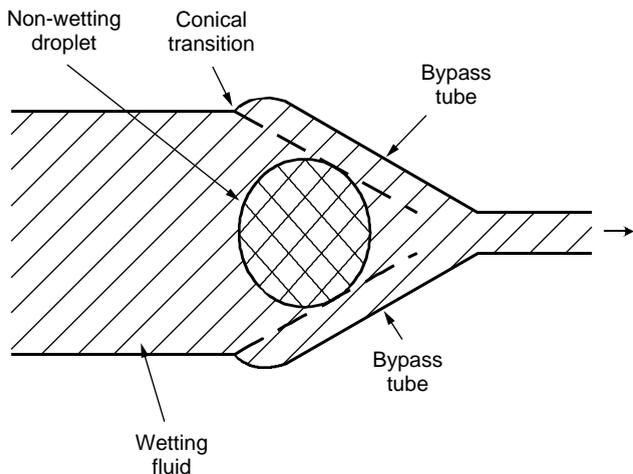


Figure 20. Microtube flow or pressure regulator.

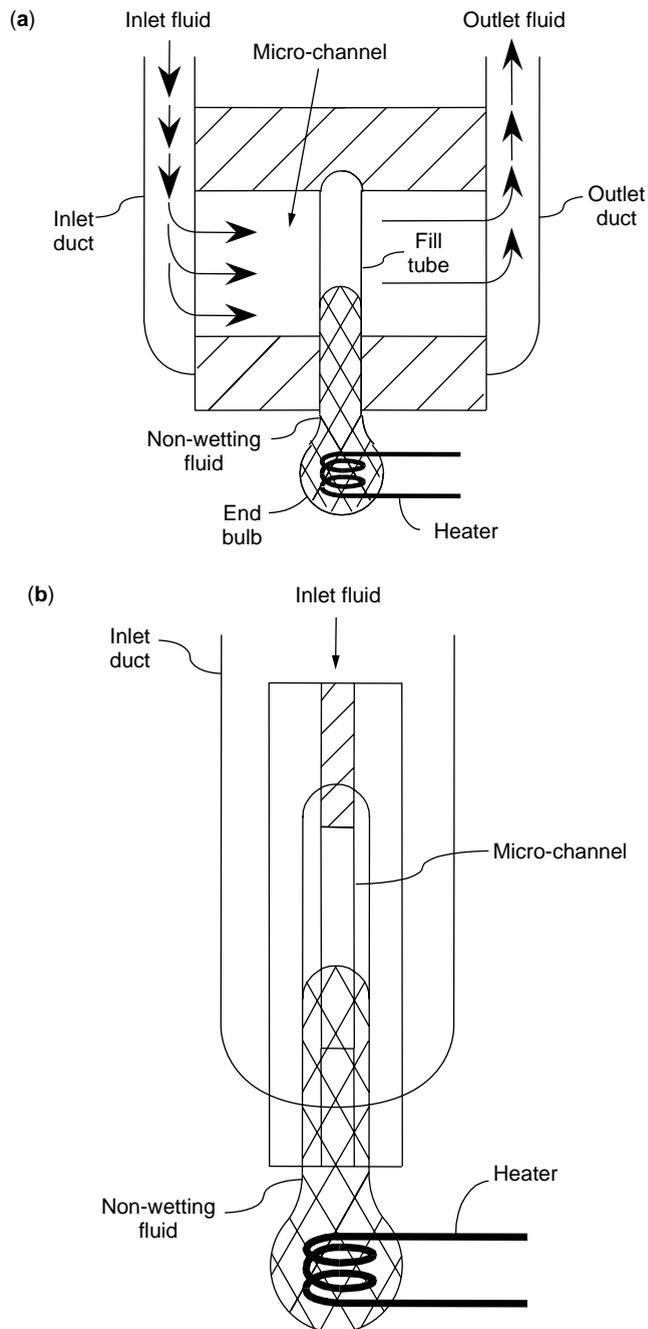


Figure 21. Positive closure microtube valve that has zero dead space: (a) top view; (b) side view.

illustrates a microvalve composed of a fill tube joined to an end bulb, where two microchannels are attached to the fill tube. In this example, the nonwetting droplet controls the flow of a wetting fluid or gas through a microchannel whose thickness is less than that of the fill tube. In this microvalve, an inlet fluid flows through an inlet duct and then into one of the microchannels. If the fill tube is not blocked by the nonwetting droplet, the inlet fluid traverses the unblocked fill tube at the point where both microchannels attach to it. Then, the fluid exits the microvalve through an outlet duct as outlet fluid. In Figure 21a,b, the nonwetting

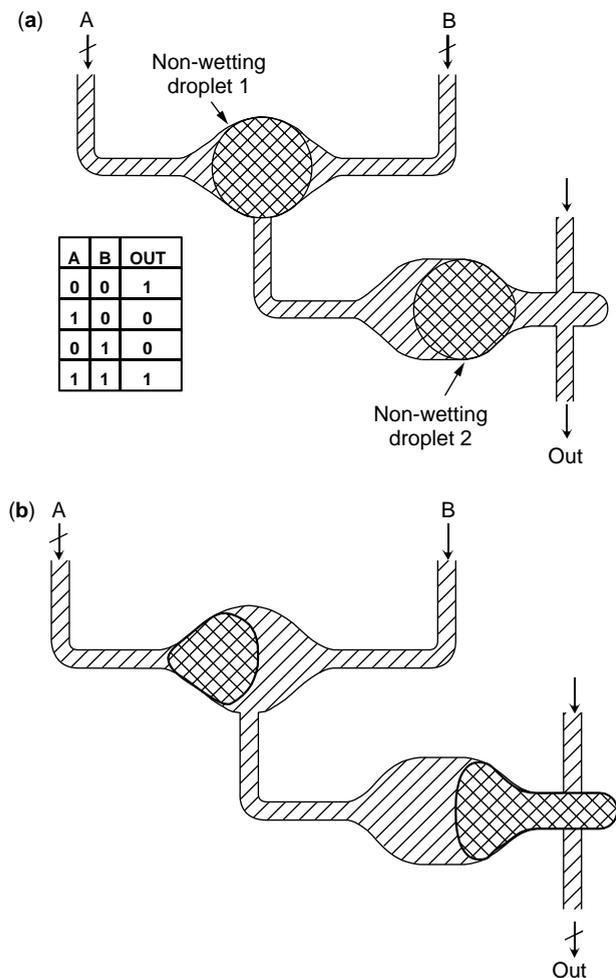


Figure 22. Microfluidic logic circuit that acts as comparator: (a) output; (b) no output.

droplet is activated by a heater and only partially fills the fill tube. Obviously, many other forms of activation are possible, and it is possible to assemble these microvalves in parallel to control large flows of liquids or gases.

A final category of microfluidic devices in which one fluid controls another can be understood by the more complex examples given in Fig. 22. These figures present microtube devices that use surface tension and wettability in fluidic logic circuits that are fully digital, not analog. The device in Fig. 22 functions as a comparator; there is an output only if the inputs are equal. Thus, if pressure is applied to either branch A or B, the gate (nonwetting droplet 2) closes, as in Fig. 22b, and no flow occurs (and no pressure is transmitted) between branches C and D. If equal pressure is applied to A and B or no pressure is applied to A and B, the gate remains open as in Fig. 22a, and flow occurs (and pressure is transmitted) between C and D. Nonwetting droplet 1 is returned to the center position whenever pressure is removed because surface tension always minimizes droplet surface area and a sphere has the lowest surface area per unit volume of any object. Only at the center position can it be a sphere, and unless placed under unbalanced force by pressure from A or B, it remains at

the center. Numerous other types of logic circuits, such as OR, NOR, AND, and NAND gates, can also be fabricated in this manner. By combining a number of these logic components in a suitable arrangement, digital operations can be performed identically to those of electrical devices. Instead of electricity being on or off in a circuit, pressure is applied or not applied, and fluid flow does or does not occur.

Microdevices Based on the Positions and Shapes of Nonwetting Droplets

In this group of microdevices, the basic principle of operation is the movement or shape change of nonwetting droplets in tubes, channels, or voids that have at least one microscopic dimension. This movement or shape change results from external or internal stimuli. The change in droplet shape depends on the cavity shape and always minimizes the surface free energy of the droplet. The cavity that constrains the droplet in these devices can be sealed or can have one or more openings. The shape of this cavity determines the reaction of the droplet to a stimulus, as well as the use of the microdevice, and the output that can be obtained from it. Uses for these microdevices are as diverse as sensors, detectors, shutters, and valves.

As just stated, microtube sensors based on surface tension and wettability are one type of device in this group. Some of these sensors respond to one or more external stimuli such as pressure, temperature, and gravity or acceleration by changes in the displacement or shape of liquid interfaces contained within microtubes and/or microchannels that have either fixed or variable axial geometries and circular or noncircular cross-sectional profiles. Other sensors respond to internal stimuli, such as a change in surface tension of the liquid droplet or a change in the wettability of the microdevice's internal walls. Some of these sensors can quantify the displacement or change in shape of the constrained droplet that results from external or internal forces acting on it.

An example of one of the simplest microdevices in this group of devices is a microtube pressure sensor (Fig. 23) that uses a nonwetting fluid in the form of a droplet. Figure 23a illustrates the position of the droplet when the entrance pressure P_{ent} is equal to the device pressure P_{dev} . Figure 23b illustrates the position of the droplet when the entrance pressure is greater than the device pressure P_{dev} , and Fig. 23c illustrates the position of the droplet when there is a much higher entrance pressure. This sensor demonstrates the reaction of such a device to an outside stimulus which in this case is an increase in externally applied pressure P_{ent} . As can easily be seen, the shape of the nonwetting droplet changes in reaction to increases in the applied external pressure P_{ent} . More precisely, increasing the external pressure P_{ent} , that acts through an entrance microtube squeezes the droplet into ever smaller diameter locations within a microcavity, which results in displacing the nonwetting interface toward the smaller diameter end of the device. For this type of sensor, this microcavity may be tear shaped, circular, or have practically any shape, as long as there is a change in at least one dimension and this dimension is from 0.003–1000 μm . For simplicity, the pressure on the smaller side of the microdevice P_{dev} , which

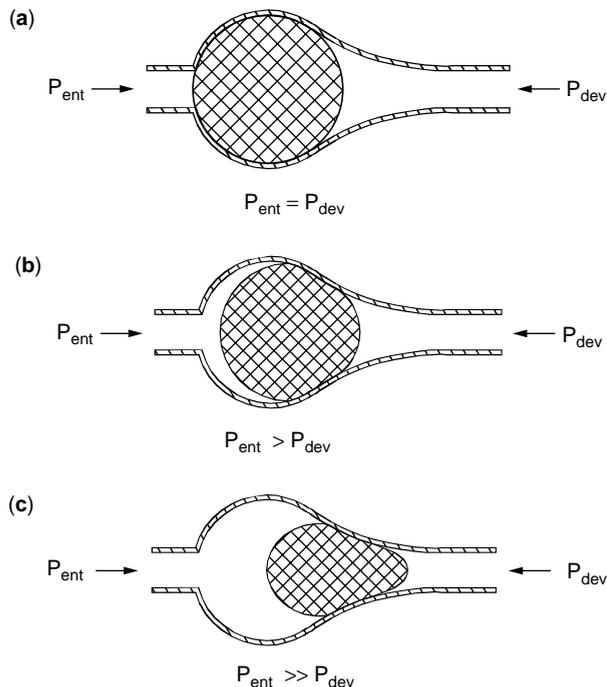


Figure 23. Microtube pressure sensor based on surface tension and wettability.

opposes the external pressure P_{ent} , is set at zero in this figure. P_{dev} is most easily thought of as a residual gas pressure left over inside the device from the actual fabrication process. It does not have to be zero, as shown later. The only requirement for this type of sensor is that P_{dev} be less than P_{ent} and that both pressures be smaller than the bursting strength of the walls of the microtube pressure sensor. It should be apparent from Fig. 23 that an overpressure of the device will push the droplet further into the device tube than designed for, but when the pressure is released, the droplet will return to its equilibrium position. As long as the walls of the device have not been damaged and maintain their original shape, the sensitivity and accuracy of this device and the others that are described following will be unaffected by overpressure.

The reaction of the droplet to external pressure is easily calculable from surface tension theory (the change in radius of the smaller end of the droplet is inversely proportional to applied the external pressure), but the actual decrease in radius and resulting displacement of the nonwetting interface can be understood only intuitively or observed visually by microscope, as presented in Fig. 23. Figure 24 illustrates a modification of this microtube pressure sensor based on surface tension/wettability which enables nonvisual determination of the displaced interface. This nonvisual response to the reaction (movement or shape change) of the droplet interface caused by a stimulus can take many forms. One of these is a change in electrical resistance. A center contact, which has a measurable electrical resistance, is inserted through the microtube device and establishes electrical contact with the nonwetting droplet. This center contact can be a wire, tube,

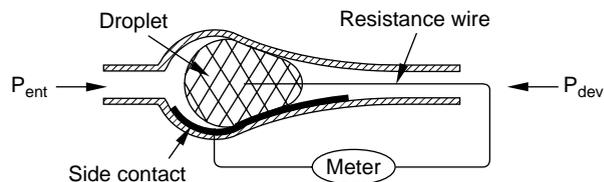


Figure 24. Microtube pressure sensor that has a resistance wire continuous readout.

tape, or any other elongated geometry desired. A second or side contact is likewise placed in a position to make electrical contact with the conductive droplet, but not make direct contact with the center contact. These two contacts are then connected to an apparatus that measures resistance. If the nonwetting droplet material composition has been selected so that it is electrically conductive as well as nonwetting, any displacement of the nonwetting interface that reduces the length of the center contact not touching the droplet thereby results in reducing the center contact's resistance measured by the resistance measuring apparatus. To maximize this effect, the center contact should have a very high resistance per unit length compared to the side contact, the actual droplet itself, and compared to the remainder of the circuit that connects both contacts to the resistance measuring apparatus.

As stated previously, the opposing pressure P_{dev} need not be zero. It has been set at zero thus far for simplicity. For this type of sensor, it merely needs to be less than the externally applied pressure P_{ent} ; otherwise, the nonwetting droplet could be expelled from the microtube pressure sensor.

Note here that the devices shown schematically can measure a variety of external or internal stimuli. The pressure sensor in Fig. 24, for example, could also measure acceleration and oscillation along the device axis as well as rotation and temperature, which affect both the thermal expansion and the surface tension of the droplet. If another center contact is also placed in the device on the end opposite the present center contact, the device can measure acceleration in two directions. In addition, it should be apparent that to measure parameters such as temperature, rotation, acceleration, or oscillation, it is not even necessary for the entrance tube and the device tube to be open to the atmosphere. Thus, to measure these external stimuli or some internal stimulus, a totally sealed cavity would function as well as the open pressure sensor in Fig. 24.

For simplicity, only pressure sensors are shown schematically, and it should be understood that the devices work equally well in reaction to many other stimuli. A partial list that includes vibration, acceleration, rotation, temperature, electromagnetic fields, and ionizing radiation demonstrates the broad scope of this sensor technology.

When a wetting droplet is employed in place of a nonwetting droplet, instead of needing P_{ent} to reach some value given by Eq. (2) to force the droplet into the microtube, it goes in automatically. This behavior is often referred to as "wicking." In contrast to the nonwetting droplet, no pressure is needed to get the drop into the tube. The fact that wetting droplets behave similarly to nonwetting droplets in

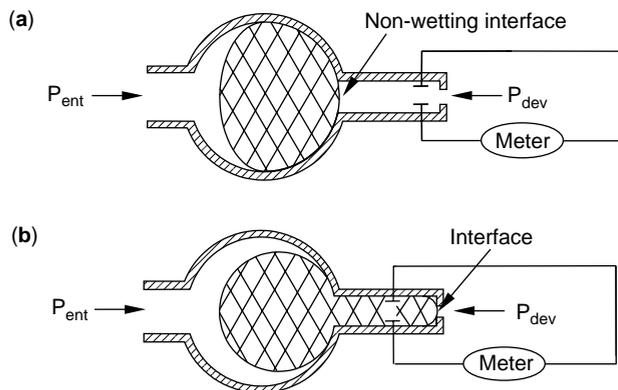


Figure 25. Microtube pressure sensor that has a “Yes-No” straight port.

some respects means that microdevice sensors can employ either wetting or nonwetting fluids as the droplet material and still serve the same function. For microtube pressure sensors, this would require switching the side of the droplet that actually “feels” the applied pressure. Microdevices that sense other stimuli would also need similar kinds of modification. These would be specific for the actual sensing application.

Figure 25a,b illustrates the relationship between the radius of curvature of the small end of the drop and the various pressures involved for a microtube pressure sensor that has a digital type of response; this will hereafter be referred to as a “Yes-No” response. In Fig. 25a, the applied pressure P_{ent} is not sufficient, compared to P_{dev} , to force the nonwetting droplet into the straight port. Therefore, the continuity apparatus registers an open circuit, or “No” response. In Fig. 25b, the applied pressure P_{ent} is sufficient to force entry of the nonwetting droplet into the straight port. Once the droplet has entered the straight port, it completely fills it, and the continuity apparatus registers a closed circuit, or “Yes” response. As mentioned previously, this same kind of “Yes-No” response can be duplicated by using wetting fluids. However, because a wetting fluid would spontaneously wick into the smaller diameter tube, the only difference would be that now an applied pressure of sufficient magnitude would need to be directed to P_{dev} to expel the wetting droplet from that same straight tube. Therefore, for a wetting fluid, the continuity apparatus would work in reverse to the “Yes-No” response for a nonwetting droplet. In this case, P_{dev} must be greater than P_{ent} , and therefore, an open circuit signifies “Yes,” and a closed circuit signifies “No.” However, because a wetting droplet adheres to the microdevice walls, including the straight port, fluid remaining on these surfaces might compromise the accuracy of the continuity apparatus. Therefore, it is preferable to use nonwetting droplets in this kind of sensor. This same logic applies to most microdevices based on surface tension and wettability, and so in the discussion that follows, only nonwetting behavior is illustrated.

There are obviously many other means for measuring displacement of a nonwetting droplet. Other basic electrical parameters that can be employed are capacitance

and inductance. Note here that for all of the aforementioned techniques for measuring displacement of a nonwetting droplet by using electrical means, the electrical properties of the nonwetting droplet must, of course, be suitable for the measurement technique employed. For some applications, the resistance of the nonwetting droplet must be sufficiently low to permit measuring the resistivity of the center contact accurately enough for the application, at hand. For other applications, the conductivity must be high enough to enable measuring capacitance accurately. For certain applications, permeability must be sufficiently different between the nonwetting droplet and its surrounding medium in the microdevice to allow measuring inductance accurately enough to satisfy the demands of the desired application. These electrical property requirements are most likely to be different, depending on the measuring technique employed and the particular application being developed. Note that it is also possible to combine two or more readout techniques in a single device.

For either multirange or redundancy-driven applications, a great deal of variation is possible. These variations are in the form of identical or different devices, cavity, and/or channel or tube configurations, as well as identical or different types of readouts. Many different types of device channel or tube configurations are possible that will give either linear or nonlinear responses, as well as analog or digital responses to the stimuli being sensed. For example, a gradual taper would produce a linear response, whereas a very rapid taper would give a nonlinear response. In another example, a device such as that shown in Fig. 25 could be modified with a tapered section to follow the constant dimension tube. This would result in a digital response followed by an analog response. In addition to these differences in individual sensors, multiple sensors could all be used together simultaneously or switched on or off as needed.

As mentioned previously, the presence or absence of nonwetting material in a straight tube (Fig. 25) enables the pressure sensor or other microdevice that derives its capabilities from surface properties of materials to function in a digital or a “Yes-No” mode of response. Figure 26 illustrates another very simple kind of “Yes-No” readout response for a pressure sensor that does not have a straight tube. Now, the center contact in Fig. 24 has been truncated, so that it does not make contact with the nonwetting droplet for low values of the pressure difference between P_{ent} and P_{dev} . This lack of contact, or gap, is shown in Fig. 26. The truncated center contact makes contact only with the nonwetting droplet once a predetermined

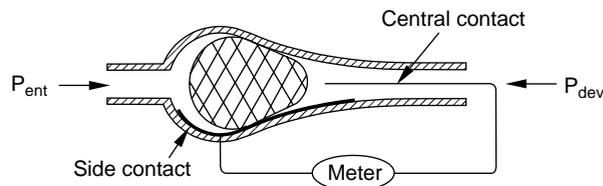


Figure 26. Microtube pressure sensor that has a “Yes-No” central contact.

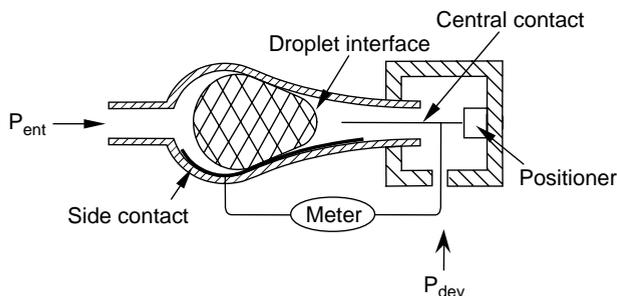


Figure 27. Microtube pressure sensor that has a variable “Yes-No” central contact.

pressure difference exists (P_{ent} greater than P_{dev}). Once this occurs, the continuity meter signals that contact has been made, and the desired “Yes-No” readout response is provided. Once sufficient pressure difference has been established, the truncated center contact can then function as a center resistance contact, center contact, some other kind of readout implement, or some combination thereof. The truncated center contact should not move relative to the microdevice walls in this simple form of “Yes-No” readout microdevice. It should be apparent that devices of this type that have a truncated center contact can also serve as an electrical switch, based on surface tension and wettability, that can be made to operate independently of gravity, can be impervious to radiation, and can be activated by numerous stimuli.

A more sophisticated “Yes-No” readout microdevice pressure sensor is illustrated in Fig. 27. Now the truncated center contact is attached to a positioner, which, in turn, is attached to a positioner holder, which is itself held firmly in place relative to the actual microdevice walls. In Fig. 27, the positioner holder is shown attached to the microdevice walls. The positioner is any type of device that can move the truncated center contact relative to the microdevice walls in a predetermined fashion. Examples of such positioning devices are numerous. They can be the type where the operator sets the gap and thereby controls the device’s sensitivity, such as a pressurized microbellows and a piezoelectric crystal. Alternatively, the positioning device can be the type that is influenced by its environment, such as those made from photostrictive, chemostriptive, electrostrictive, or magnetostrictive materials, which change length due to light, a chemical environment, or an electric or magnetic field. In these types of “smart” materials, the positioner can be controlled in real time by its environment, and thus the device can respond to two stimuli simultaneously. Moreover, using any such positioning devices, the gap can be changed by a feedback circuit. By altering the size of the gap either before or during actual operation of the microdevice, the amount of pressure difference needed between P_{ent} and P_{dev} to establish contact and thereby evoke the “Yes-No” readout response or continuity, as measured by the continuity apparatus, can be changed. Thus, the sensitivity of this device can be changed by an operator, by its environment, or by a feedback circuit. As before, once continuity has been established for the simple “Yes-No” readout response microdevice of Fig. 27, the truncated center contact

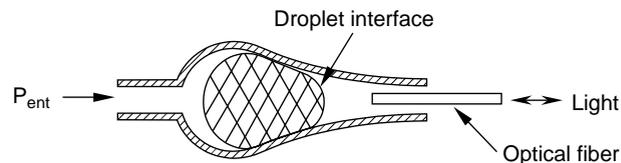


Figure 28. Microtube pressure sensor that has an interferometer readout.

can be used for other kinds of readout purposes. For example, the continuity apparatus can be modified to function as the resistance measuring apparatus shown in Fig. 24. If this is done, both digital and analog readout responses can be garnered from the same sensor. More than one center contact of different lengths and/or more than one side contact can also be employed in Fig. 27, thereby providing multiple digital responses from one device.

Note that the sensing techniques mentioned thus far have all been relatively simple and have employed principles of physics that are intuitively easy to understand: changes in resistance, capacitance, or inductance. Another simple technique for detecting the position of a droplet interface is using an electromagnetic beam impinging on a detector that is blocked by the advancing surface of the droplet. This type of arrangement can basically give only a “Yes”-“No” response. Two other techniques that can also be employed to monitor displacement of the nonwetting droplet interface are optical interference and electron tunneling. These techniques are capable of much higher levels of resolution of the nonwetting droplets’ displacement, which results in greater levels of sensitivity.

Figure 28 illustrates the readout technique that employs optical interference. The only additional requirement that must be imposed to use this technique is that the nonwetting droplet must reflect at least some of the electromagnetic radiation input through the fiber-optics input/output cable back through the same cable. If these conditions are met, an interference pattern can then be generated between the incoming and outgoing rays of radiation that can be detected by a suitable apparatus located at the opposite end of the fiber-optics input/output cable. This interference pattern will be highly dependent on the position of the internal interface of the nonwetting droplet, as well as on the wavelength of radiation employed. Therefore, it is an extremely accurate technique for monitoring any displacement of that interface.

Figure 29 illustrates the readout technique for electron tunneling. There are two primary differences between this readout technique and the previous readout technique that employs a truncated center contact, as illustrated in Fig. 27. In this apparatus, the truncated center contact is replaced by a very sharp needle-shaped electrode. In addition, the continuity apparatus is replaced by a much more sensitive electron tunneling current detector that can measure the tiny electrical currents generated when the needle-shaped electrode moves very close to the internal interface and creates gaps of the order of atomic dimensions. As in optical interference, tunneling current measurements are many times more sensitive to

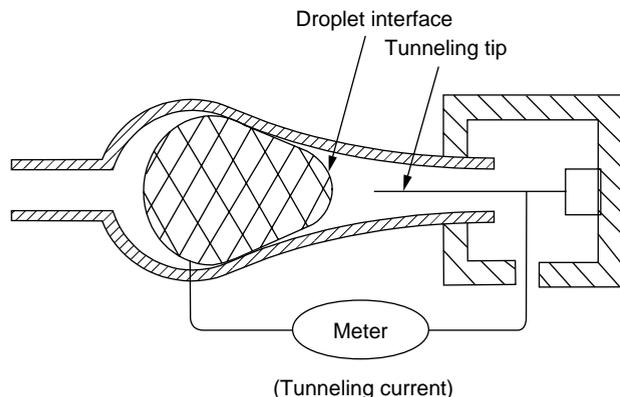


Figure 29. Microtube pressure sensor that has a tunneling current readout.

displacements of the internal interface than simpler readout techniques discussed initially. Obviously, any other detecting technique used in scanning probe microscopy, such as atomic force, magnetic force, or capacitance, can be used in place of the needle-shaped electrode and the current-detection circuit. In addition, the sensitivity of these devices, as in the device in Fig. 27, can be changed by varying the gap between the tip and the droplet.

At this point, it is well worth mentioning again that the movement of a droplet in a microscopic tube, channel, or void and the process of remote measurement of displacements of the internal interface itself is of critical significance to the devices shown, not the actual kind of remote measurement technique employed. Whatever technique is employed affects only the accuracy of the remote readout. Thus, regardless of the measurement technique, displacements of the internal interface in reaction to external stimulus will be the same and will depend only on the surface tension and wettability of the sensor components and on sensor geometry.

Until now, it has been tacitly assumed that motion of the internal interface during any remote readout of its displacement is negligible. This is not necessarily so. Any measurement of the position of the internal interface will take some finite amount of time. If there is motion of the internal interface during this finite measurement time, the position of the internal interface will be some sort of average readout. If this is acceptable to the designer of the microdevice, all is well. If it is not, either the method of remote readout or the level of precision of the analytical instruments employed must be changed to increase the speed of readout to the degree required. Once this has been done, microdevices based on surface tension and wettability that have remote readout capabilities can function either as static or dynamic analytical detectors or sensors.

Until now, it has also been assumed that the shaped or tapered microtubes or microchannels within which droplets move or flow under the influence of surface tension and wettability and some external forcing agent such as pressure or acceleration, had circular cross sections. This does not have to be so. Figure 30a, b illustrates flow of an elongated nonwetting mercury droplet constrained on

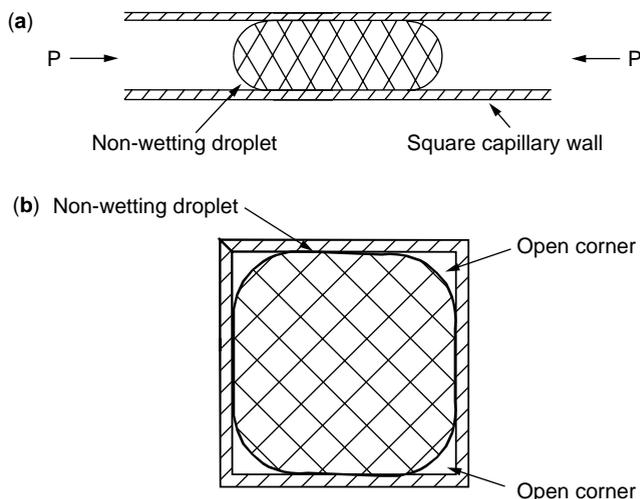


Figure 30. Nonwetting droplet in square channel: (a) side view; (b) end view.

four sides by walls that form a square cross section. For mercury and other high contact angle liquids whose contact angles are greater than 135° , there will always be open corners in a channel that has right-angle corners. These corners remain unfilled because infinite internal pressure would be required in these high contact angle liquids, the result of setting r equal to zero in the relationship given in Eq. (2), to fill in all corners completely. This can never be true for two reasons. First, there is no such entity as infinitely high pressure. Second, in Fig. 30a,b, bypass flow of externally applied pressure P_{ent} will occur through all open corners, thereby reducing the actual pressure applied to the nonwetting droplet. An analogous situation occurs when a child shoots an irregularly shaped pea through a circular straw. Even though gaps equivalent to the open corners in Fig. 30a,b exist around the pea, the child can still expel the pea from the straw simply by blowing hard enough, thereby producing a sufficiently effective pressure on the pea to accomplish the purpose. This is exactly the situation that exists for flow of nonwetting droplets in noncircular microtubes or microchannels. Therefore, all previous arguments for remote sensing of droplet interfaces in circular cross-sectioned microtubes or microchannels apply equally well to remote sensing of droplet interfaces in microtubes, microchannels, or voids that have any type of noncircular profile. Moreover, noncircular microtubes or microchannels can certainly be used in conjunction with circular microtubes or microchannels in the same microdevice. In fact, there is very good reason to do so. Noncircular microtubes or microchannels can be fabricated relatively easily by using techniques such as photolithography and LIGA on a surface. This is currently done on silicon wafers by a sequence of deposition and/or etching techniques in a number of different ways, two of which will be given. A noncircular channel can be formed, for example, by etching the channel in the surface and then covering the channel by sealing a glass plate over it. Alternatively, for example, the noncircular channel can be formed by etching a channel in the surface and then filling it with

a sacrificial material. Another material is deposited over the filled channel and then the sacrificial material is removed to leave a microchannel. However, no matter how the noncircular channel is formed in the surface, a bypass flow of gases occurs through its open corners, as illustrated in Fig. 30b, if the liquid has a high contact angle. As just mentioned, this makes it difficult to apply pressure accurately and reproducibly to some nonwetting droplets contained within such microtubes or microchannels. This is not true for circular microtubes or microchannels. Thus, the presence of small circular microtubes or microchannels at appropriate positions in any device fabricated from noncircular microtubes or microchannels will allow either gases or wetting fluids to apply hydrostatic pressure to microdevices that contain nonwetting droplets at 100% efficiency. The reverse is also true. A circular cross section in the device will also allow the nonwetting droplet to apply force to the gas or wetting fluid at 100% efficiency. This also means that it is possible to have wetting and nonwetting fluids in the same microdevice. Finally, regardless of the cross-sectional shape of the microtubes, microchannels, or voids, all wetting fluids will have 100% efficiency.

It is extremely important to realize that the previous discussion also illustrates that an elongated non-wetting droplet confined within a microtube or microchannel that has, for the sake of illustration, square walls can serve purposes other than remote sensing. For example, it can be used to act as a shutter in optical applications, where the presence or absence of the droplet controls whether or not light or other electromagnetic radiation is allowed to pass through the square microchannel walls. In this instance, the nonwetting droplets function in much the same fashion as a window blind by controlling whether or not light is let through a window depending on whether or not the blind is up or down. It could also control particle beams in a similar manner.

Figure 31 illustrates a much more familiar looking shutter mechanism that could very easily function identically to traditional mechanical shutters. An end bulb is connected to a fill tube, and both are filled by a nonwetting liquid, which is called the working fluid and is opaque for the particular application. A rectangular void is also connected to the fill tube, but its thickness is less than the diameter of the fill tube. (The thickness of the void in this figure is exaggerated for clarity.) The shutter that has constant void thickness is illustrated in the open configuration in Fig. 31a,b, where the incident radiation or particle beam passes through the shutter, and is closed in Fig. 31c,d where the incident beam or radiation is blocked by the shutter. The void width and void length can be much greater than the void thickness and only one void dimension has to be macroscopic to carry out a shutter's function. This illustrates an extremely important point: although all of the dimensions of a device can be microscopic, only one dimension of a device must be in the range where surface tension and wettability become dominant factors in the device's reaction to internal or external stimuli to consider the device a microdevice. The rectangular shutter of constant void thickness illustrated in Figure 31 must be considered a microdevice because of the microscopic dimensions of its thickness, even though

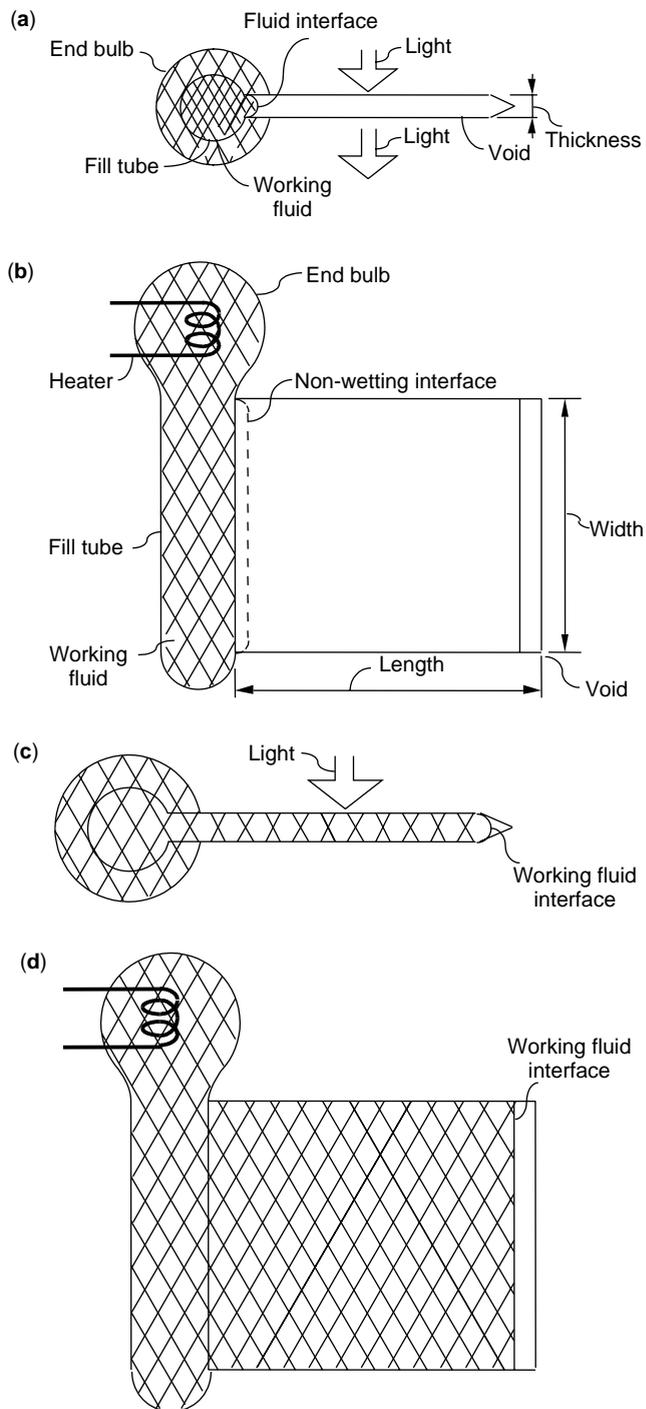


Figure 31. Macroscopic microtube rectangular shutter: (a) side view of open shutter; (b) top view of open shutter; (c) side view of closed shutter; (d) top view of closed shutter.

it can have very macroscopic dimensions for one or more of its other features. This is true for all microdevices based on surface tension and wettability. In this example, the method of actuation of the rectangular shutter shown in Fig. 31 is derived by an externally generated electrical current input through a heater contained within the working fluid. As the working fluid expands due to this heat input,

any gas bubbles or other gas-filled voids contained within the working fluid become compressed, thereby raising the internal pressure P_{int} within the working fluid. This thrusts the internal interface farther and farther into the rectangular void. At some point, the radius will decrease sufficiently, so that the internal interface will shoot across the void length, providing that the volume of compressed gas-filled voids is much larger than the volume of the shutter. This will close the shutter in the same fashion as the “Yes-No” devices described earlier in Fig. 25. If gas bubbles or other gas-filled voids are not present in the working fluid, then the heat input will not cause a “Yes-No” type of reaction, but rather will enable the shutter to close more gradually. In this way, a partially closed shutter can be maintained by controlling the heat input appropriately. The gradual closing capability can be obtained for a pressure-activated shutter by employing a void thickness that has a decreasing taper. A different tapered end is shown at the end of the rectangular void where the working fluid stops in the closed position. This is to minimize the water hammer effect and does not have to be present for the shutter to work. Of course, external pressure or some other external stimuli as well as an internal stimuli could also be used in place of the heater, and the shutter would still function. If pressure were employed, it would then be a pressure sensor that has some macroscopic dimensions that would be very easy to observe. Filling of the void would signify that a certain pressure had been reached. Obviously, there are numerous other applications of this technology but only two others will be mentioned. One involves using a reflective nonwetting fluid, so that a mirror results when the void space is filled, and the second application encompasses a much larger microscopic void area. If the void space is made as large in area as a window pane, solar energy acting on the reservoir could be used to force liquid into a void of microscopic dimensions in the window pane and thus block sunlight from going through the window if an opaque nonwetting liquid is employed.

The void shown in Fig. 31 has constant thickness and is rectangular. Neither parameter is necessary. Figure 32 illustrates a circular shutter, which has a straight top face and a curved bottom face that make up the void. Now, depending on the amount of expansion of the working fluid caused by electric power supplied to the heater, the shutter can be completely open when all of the working fluid is contained within the outside bulb, completely shut and have no circular gap in the center at all, or anywhere in between, as Fig. 32 illustrates. Certainly, both the top face and bottom face can be curved or straight, and virtually any shutter geometry can be employed. Likewise, actuating techniques other than heat input to the working fluid by an internal heater can be used. External heat input by radiation or conduction or changes in the internal pressure of the working fluid by any other means can be used to achieve the same resulting shutter behavior.

As mentioned earlier, surface tension and wettability govern the position of droplets within microdevices. Thus, in addition to the external stimuli already mentioned, any external stimuli that changes either the surface tension of the droplet or the wettability of the surface can be detected by a suitably designed microdevice sensor. Some, but

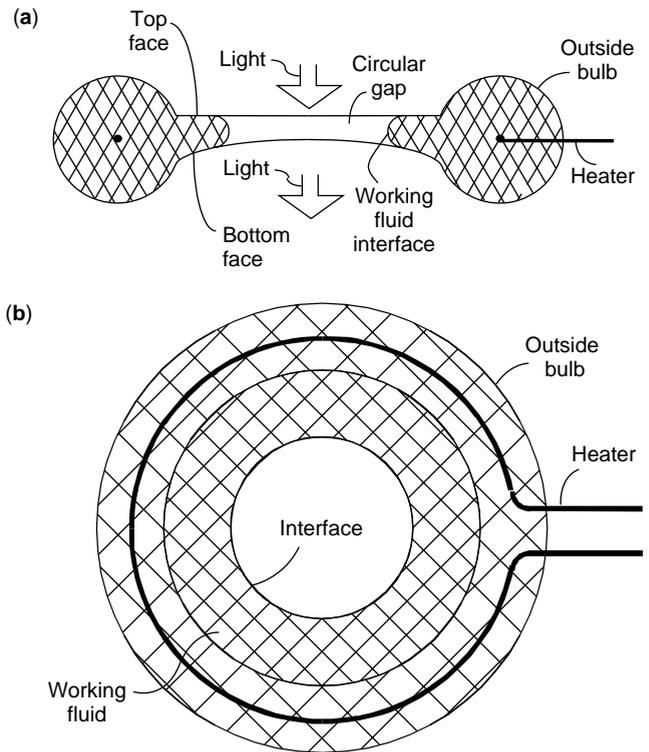


Figure 32. Circular shutter or iris: (a) side view; (b) top view.

not all, such stimuli include the following: temperature, magnetic field, electrical field, rotation, radiation, and beams of particles.

Until now, all of the microdevice sensors illustrated have been designed to respond to external stimuli. This is not the only mechanism for displacing microdevice droplets. Any compositional change that occurs within droplets themselves or on the walls of microdevices can also change surface tension or wettability. These changes can be either reversible or irreversible and can be caused by a gas or wetting fluid in the device along with the nonwetting droplet. In addition, the surface tension of the droplet increases as both the temperature and rotation increase. If these or any other internally induced change in a microdevice's surface tension or wettability occurs, it can be detected and monitored remotely using any of the techniques described previously. Obviously, these internally induced changes in a microdevice's surface tension or wettability can also be used to move the droplet(s) to perform work.

In addition, no actual dimensions of either microtubes or microchannels have yet been discussed. Assuming a nonwetting fluid such as mercury, which has a surface tension at room temperature of approximately 470 dynes/cm and a contact angle on glass microdevice walls of roughly 140° , one can calculate the following droplet radii for the indicated internal pressures using Laplace's equation modified to include the effect of wettability (Table 1):

In this section, we have shown that the flow of droplets within microtubes and microchannels that is controlled by surface tension and wettability can be used to sense, qualitatively and quantitatively, any environmental factor that acts on a droplet or affects either its surface tension or

Table 1. Calculation of Droplet Radii

r_{dev} (μm)	P_{int} (lb/in^2)	P_{int} (kPa)
950	0.1	0.69
95	1.0	6.89
9.5	10	68.9
0.95	100	689
0.095	1,000	6890
0.003	100,000	689,000

wettability, or both. This sensing can be performed remotely by a variety of techniques. It has also been demonstrated that wetting droplets can sometimes be used with only minor device modifications. Static as well as dynamic remote sensing can be performed, and microtubes or microchannels that have circular or noncircular cross sections and variable axial geometries can be employed either individually or together. The reaction of these devices to any stimuli can be tailored by the device geometry and the method of sensing and result in linear and nonlinear analog output, as well as digital output. The use of this technique in nonsensing applications that perform mechanical functions was also demonstrated. Finally, whenever at least one microdevice dimension lies between $1000 \mu\text{m}$ and $0.003 \mu\text{m}$, it has been shown that it can perform all of the various tasks that have been discussed.

Nonwetting Droplets That Perform Work

The preceding discussion has indicated that the movement of nonwetting droplets within microtubes or microchannels can be used for sensing and controlling fluids. In addition, microscopic nonwetting droplets can be used for mechanical control and manipulation within microdevices, including tasks such as position control, moving objects, deforming objects, pumping fluids, circulating fluids, and controlling their flow. Obviously, complex machines and engines can be produced by properly joining actuator and pumping elements.

An application of a microscopic nonwetting droplet for low friction position control is a microtube liquid bearing shown in Fig. 33. Referring to Fig. 33a, for example, the bearing assembly is a microtube that has one or more circular channels on its circumference that actually join the microtube's interior void space in a narrow ring-shaped opening. A center rod only slightly smaller in diameter than the bearing assembly is supported by nonwetting fluid that fills the circular channels. This fluid cannot leak out around the center rod if the gap is small enough because too much pressure (Table 1) is required to form the droplet of smaller radius that would be able to leak. Therefore, the center rod is free to either rotate or translate axially within the bearing assembly. It is called an external bearing because of this outside configuration. The only restraining forces involved are frictional forces between the center rod and the non-wetting fluid.

Figure 33b illustrates a reciprocal situation called a microtube internal bearing. A straight walled microtube is used. A central rod has at least one groove about the circumference, and the nonwetting fluid fills this groove, which allows both rotational and translational motion.

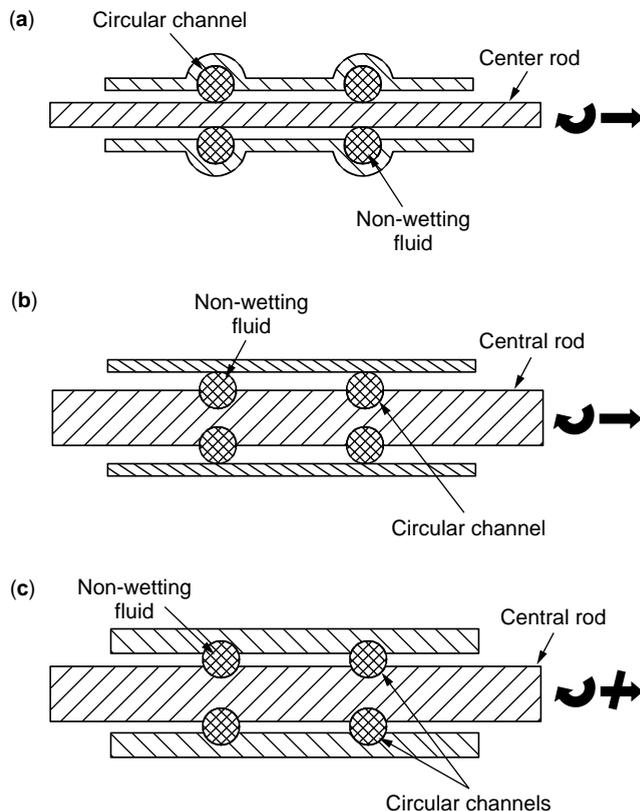


Figure 33. Shaft supported by nonwetting fluidic bearings: (a) rotating and translating outer bearing; (b) rotating and translating inner bearing; (c) inner/outer bearing that rotates but does not translate.

Figure 33c is a mixed combination of internal and external microtube liquid-bearing locations. In this configuration, however, only rotational motion is easily achieved. For translation to occur, shearing of a wetting droplet must take place. Although this is not as difficult as forming a small-radius annular droplet, it still involves generating new droplet surface area and therefore requires more force to produce translation than for either the purely internal or purely external bearings.

Figure 34 illustrates a microtube liquid bearing that will not allow significant translational motion. It is a

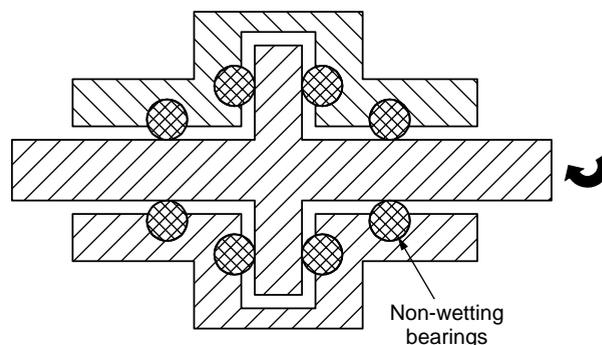


Figure 34. Thrust bearing that incorporates four nonwetting fluid bearings.

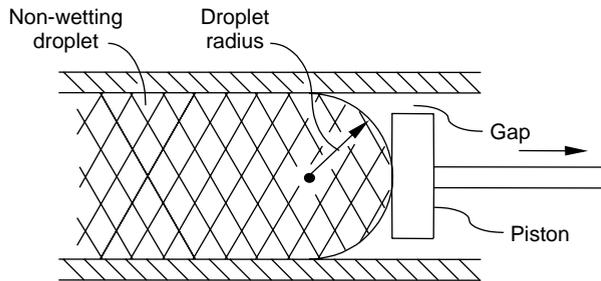


Figure 35. Piston actuated by nonwetting fluid.

thrust bearing that uses four separate microtube liquid bearings in an external configuration. As before, an internal or mixed configuration is also possible, and additional microtube liquid bearings that use surface tension/wettability effects can be employed.

Figure 35 illustrates how a nonwetting droplet can be employed as an actuator; its motion down a microchannel or microtube is used to move a loosely fitting piston. It is important to note that as long as the droplet does not wet the piston and walls, none of the droplet will squeeze by the piston if the clearance is less than the radius of the nonwetting droplet. The nonwetting droplet is deliberately shown only in part in Fig. 35 to demonstrate that the actual mechanism, external stimuli, or internal stimuli, or wetting fluid, that causes it to push on the piston is unimportant. Its ability to function to transmit force mechanically and thereby perform work is all that matters.

CONCLUSIONS

Microtubes appear to have almost universal application in areas as diverse as optics, electronics, medical technology, and microelectromechanical devices. Specific applications for microtubes are as wide ranging as chromatography, encapsulation, cross- and counterflow heat exchange, injectors, micropipettes, dies, composite reinforcement, detectors, micropore filters, hollow insulation, displays, sensors, optical waveguides, flow control, pinpoint lubrication, microsponges, heat pipes, microprobes, and plumbing for micromotors and refrigerators. The technology works equally well for high- and low-temperature materials and appears feasible for all applications that have been conceived to date.

The advantage of microtube technology is that tubes can be fabricated inexpensively from practically any material in a variety of cross-sectional and axial shapes in very precise diameters, compositions, and wall thicknesses of orders of magnitude smaller than is now possible. In contrast to the other micro- and nanotube technologies currently being developed, microtubes can be made from a greater range of materials in a greater range of lengths and diameters and far greater control over the cross-sectional shape. These tubes will provide the opportunity to miniaturize (even to nanoscale dimensions) numerous products and devices that currently exist, as well as allowing the fabrication of innovative new products that have to date been impossible to produce.

Space only allowed presenting one application of microtube technology to new innovative products in greater detail. This one application comprised those devices that are based on surface tension and wettability. The few devices that were shown as examples demonstrated the breadth of this technology only in one field. The application of microtube technology to other fields is considered equally rich and limited only by a designer's imagination.

ACKNOWLEDGMENTS

The invaluable help provided by Hong Phan in fabrication, Marietta Fernandez in microscopy, and Tom Duffey in artwork is greatly appreciated. Financial support from Dr. Alex Pechenik of the Chemistry and Materials Science Directorate, Air Force Office of Scientific Research was responsible for much of this work.

BIBLIOGRAPHY

1. G. Stix, *Sci. Am.* **267**: 106–117 (1992).
2. W. Menz, W. Bacher, M. Harmening, and A. Michel, *1990 Int. Workshop on Micro Electro Mechanical Syst. (MEMS 90)*, 1990, pp. 69–75.
3. C.G. Keller and R.T. Howe, *Transducers 95*, Stockholm, Sweden, 1995, pp. 376–381.
4. W. Hofmann, C.S. Lee, and N.C. MacDonald, *Sensors and Materials* **10**: 337–350 (1998).
5. S.Y. Chou, P.R. Krauss, and P.J. Renstrom, *Science* **272**: 85–87 (1996).
6. S. Weiss, *Photonics Spectra* **108** (Oct. 1995).
7. M. Mullenborn, H. Dirac, and J.W. Peterson, *Appl. Phys. Lett.* **66**: 3001–3003 (1995).
8. D.Y. Kim, S.K. Tripathy, L. Li, and J. Kumar, *Appl. Phys. Lett.* **66**: 1166–1168 (1995).
9. K.H. Schlereth and H. Bottner, *J. Vac. Soc. Technol.*, **B10**: 114–117 (1992)I.
10. B.D. Terris, H.J. Mamin, M.E. Best, J.A. Logan, D. Rugar, and S.A. Rishton, *Appl. Phys. Lett.* **69**: 4262–4264 (1996).
11. Y. Xia and G.M. Whiteside, *Angew. Chem. Int. Ed.* **37**: 550–575 (1998).
12. M. Ishibashi, S. Heike, H. Kajayama, Y. Wada, and T. Hashizume, *J. Surf. Anal.* **4**: 324–327 (1998).
13. H.J. Mamin and D. Rugar, *Appl. Phys. Lett.* **61**: 1003–1005 (1992).
14. W. Li, A. Virtanen, and R.M. Penner, *J. Phys. Chem.* **96**: 6529–6532 (1992).
15. E.E. Ehrichs and A.L. de Lozanne, *Nanostructured Physics and Fabrication*. Academic Press, NY, 1989, pp. 441–445.
16. J.C. Harley, M.S. Thesis, University of Pennsylvania, 1991.
17. D.J. Harrison, K. Fluri, K. Seiler, Z. Fan, C.S. Effenhauser, and A. Manz, *Science* **261**: 895–897 (1993).
18. K.D. Wise and K. Najafi, *Science* **254**: 1335–1342 (1991).
19. P.C. Hidber, W. Helbig, E. Kim, and G.M. Whitesides, *Langmuir* **12**: 1375–1380 (1996).
20. E. Kim, Y. Xia, and G.M. Whitesides, *Nature* **375**: 581–584 (1995).
21. O.J.A. Schueller, S.T. Brittain, and G.M. Whitesides, *Sensors and Actuators A* **72**: 125–139 (1999).

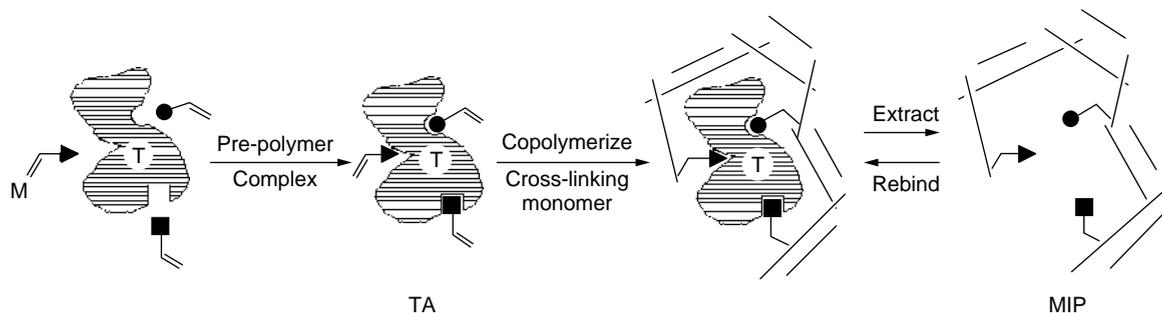
22. R.J. Jackman, S.T. Brittain, A. Adams, H. Wu, M.G. Prentiss, S. Whitesides, and G.M. Whitesides, *Langmuir* **15**: 826–836 (1999).
23. B. Xu, F. Arias, and G.M. Whitesides, *Adv. Mater.* **11**: 492–495 (1999).
24. O. Lehmann and M. Stuke, *Science* **270**: 1644–1646 (1995).
25. F.T. Wallenberger, *Science* **267**: 1274–1275 (1995).
26. J. Schnur, *Science* **262**: 1669–1676 (1993).
27. P. Yager and P. Schoen, *Mol. Cryst. Liq. Cryst.* **106**: 371–375 (1984).
28. G.M. Chow, M.A. Markowitz, and A. Singh, *JOM* **45**: 62–65 (1993).
29. C.R. Martin, L.S. Van Dyke, Z. Cai, and W. Liang, *J. Am. Chem. Soc.* **112**: 8976–8977 (1990).
30. C.R. Martin, *Science* **266**: 1961–1966 (1994).
31. J.C. Hulteen and C.R. Martin, *J. Mater. Chem.* **7**: 1075–1087 (1997).
32. T.K. Kyotani, L. Tsai, and A. Tomita, *Chem. Mater.* **7**: 1427–1428 (1995).
33. T.K. Kyotani, L. Tsai, and A. Tomita, *Chem. Mater.* **8**: 2109–2113 (1996).
34. P. Hoyer, *Adv. Mater.* **8**: 857–859 (1996).
35. P. Hoyer, N. Baba, and H. Masuda, *Appl. Phys. Lett.* **66**: 2700–2702 (1995).
36. H. Nakamura and Y. Matsui, *J. Am. Chem. Soc.* **117**: 2651–2652 (1995).
37. M. Aizawa, Y. Nakagawa, Y. Nosaka, N. Fujii, and H. Miyama, *J. Non-Crystal Solids* **124**: 112–115 (1990).
38. Y. Ando and S. Iijima, *Jpn. J. Appl. Phys.* **32**: L107–L110 (1993).
39. S. Iijima, *Nature*, **354**: 56–58 (1991).
40. T.W. Ebbesen and P.M. Ajayan, *Nature* **358**: 220–223 (1992).
41. S. Iijima, T. Ichihashi, and Y. Ando, *Nature* **356**: 776–779 (1992).
42. T.W. Ebbesen, *Phys. Today* **49**(6): 26–32 (1996).
43. N.C. Chopra, R.J. Luyken, K. Cherrey, V.H. Crespi, M.L. Cohen, S.G. Louie, and A. Zettl, *Science* **269**: 966–967 (1995).
44. Z. Weng-Sieh, K. Cherrey, N.G. Chopra, X. Blase, Y. Miyamoto, A. Rubio, M.L. Cohen, S.G. Louie, A. Zettl, and R. Gronsky, *Phys. Rev. B* **51**: 11229 (1995).
45. C.N.R. Rao, B.C. Satishkumar, and A. Govindaraj, *Chem. Commun.* 1551–1552 (1997).
46. P.M. Ajayan, O. Stephan, P. Redlich, and C. Colliex, *Nature* **375**: 564–567 (1995).
47. R. Bacon, *J. Appl. Phys.* **31**: 283–290 (1959).
48. R.T.K. Baker, M.A. Barber, P.S. Harris, F.S. Feates, and R.J. Waite, *J. Catal.* **26**: 51–62 (1971).
49. R.T.K. Baker, *Carbon* **27**: 315–323 (1989).
50. R.T.K. Baker, P.S. Harris, and S. Terry, *Nature* **253**: 37–39 (1975).
51. G.M. Chow, M. Pazirandeh, S. Baral, and J.R. Campbell, *Nanostructured Mater.* **2**: 495–503 (1993).
52. R. Kirsch, M. Mertig, W. Pompe, R. Wahl, G. Sadowski, K.J. Boehm, and E. Unger, *Thin Solid Films* **305**: 248–253 (1997).
53. M.R. Ghadiri, J.R. Granja, R.A. Milligan, D.E. McRee, and N. Khazanovich, *Nature* **366**: 324–327 (1993).
54. A. Harada, J. Li, and M. Kamachi, *Nature* **364**: 516–518 (1993).
55. J.-H. Fuhrhop, D. Spiroski, and C. Boettcher, *J. Am. Chem. Soc.* **115**: 1600–1601 (1993).
56. G. Li and L.B. McGown, *Science* **264**: 249–251 (1994).
57. H. Dai, J.H. Hafner, A.G. Rinzler, D.T. Colbert, and R.E. Smalley, *Nature* **384**: 147–150 (1996).
58. L. Langer, L. Stockman, J.P. Heremans, V. Bayot, C.H. Olk, C. Van Haesendonck, Y. Bruynseraede, and J.-P. Issi, *J. Mater. Res.* **9**: 927–932 (1994).
59. H. Masuda and K. Fukuda, *Science* **268**: 1466–1468 (1995).
60. R.J. Tonucci, B.L. Justus, A.J. Campillo, and C.E. Ford, *Science* **258**: 783–785 (1992).
61. J.S. Beck, J.C. Vartuli, W.J. Roth, M.E. Leonowicz, C.T. Kresge, K.D. Schmitt, C.T. Chu, D.H. Olsen, E.W. Sheppard, S.B. McCullen, J.B. Higgins, and J.L. Schlenker, *J. Am. Chem. Soc.* **114**: 10834 (1992).
62. C.T. Kresge, M.E. Leonowicz, W.J. Roth, J.C. Vartuli, and J.S. Beck, *Nature* **359**: 710–712 (1992).
63. S. Kondoh, Y. Iwamoto, K. Kikuta, and S.-I. Hirano, *J. Am. Ceram. Soc.* **82**: 209–212, (1999).
64. H.H. Mueggenburg, J.W. Hidahl, E.L. Kessler, and D.C. Rousar, *AIAA/SAE/ASME/ASEE Joint Propulsion Conf. Exhibit Proc.* Nashville, TN 1992, #AIAA 92-3127.
65. M. Datta, *J. Electrochem. Soc.* **142**: 3801–3805 (1995).
66. M Datta, *Electrochem. Soc. Interface*, 32–35 (Summer 1995).
67. Microcomponent Chemical Process Sheet Architecture, U.S. Pat. 5,811,062, Sept. 22, 1998, R.S. Weneng, M.K. Drost, C.J. Call, J.G. Birmingham, C.E. McDonald, D.K. Kurath, and M. Friedrich.
68. C. Gonzalez, S.D. Collins, *Transducers 97*, Chicago IL, 1997, pp. 273–274.
69. T. Young, *Trans. R. Soc. London* **95**: 65 (1805).
70. P. S. de Laplace, *Mechanique Celeste*. Supplement to Book 10, 1806.
71. E.W. Washburn, *Proc. Nat. Acad. Sci. USA* **7**: 115 (1921).
72. H.L. Ritter and L.C. Drake, *Ind. Eng Chem. Anal. Ed.* **17**: 782 (1945).
73. A.W. Adamson, *Physical Chemistry of Solid Surfaces*. Wiley, NY, 1976, pp. 529–532.

MOLECULARLY IMPRINTED POLYMERS

BRADLEY R. HART
 KENNETH J. SHEA
 University of California
 Irvine, CA

INTRODUCTION

Molecular imprinting is a process for synthesizing materials that contain highly specific recognition sites for small molecules (scheme 1). The imprinting process consists of a template molecule (*T*) that preorganizes functional, polymerizable monomers (*M*). This template assembly (*TA*) is copolymerized with a large excess of cross-linking monomers to produce an insoluble network material. Extraction of the template molecule leaves behind a region that is complementary in size, shape, and functional group orientation to the template molecule. The resulting polymeric materials are referred to as molecularly imprinted polymers (MIPs):



Scheme 1

Intermolecular forces between the template, functional monomers and polymer matrix determines MIP structure. Pre-organization of the functional monomers has been achieved with covalent (1–4), electrostatic (5,6), and hydrogen bonding interactions (7–9). The most common method of polymer formation employs free radical initiation. These materials are tough, easy to handle, and have been shown to retain their recognition properties for over five years without loss of selectivity or capacity (10,11). As a result, they have advantages over more fragile biological systems for molecular recognition and have been referred to as “plastic antibodies” (12). Under the proper conditions, MIPs have shown the ability to function as selective binding materials (13–15), microreactors (3,16,17), facilitated transport membranes (18,19), and catalysts (20–23). Their utility, coupled with their longevity and robustness, make MIPs candidates for applications in separation science, as industrial catalysis, and for medical diagnostics (24,25).

Molecular imprinting began as an attempt to develop “synthetic antibodies.” Initially referred to as “specific adsorption,” a technique developed by Dickey utilized silica polymers to create synthetic receptors for the dye molecules methyl orange and ethyl orange (26,27). Dickey found that a material prepared using methyl orange was able to re-adsorb that molecule 1.4 times better than ethyl orange, and that the selectivity could be reversed by using ethyl orange during the polymerization. These results were followed by the first imprinted chiral stationary phase (CSP), which was prepared in 1952 by Curti and Columbo. Dickey’s method was used to imprint silicate polymers with *D*-camphorsulfonic acid and enantiomers of mandelic acid (28). Curti and Columbo were able to achieve chromatographic separation of camphorsulfonic acid enantiomers and mandelic acid enantiomers using this process.

The first report of an imprinted organic polymer was made by Wulff in 1972 (29). A *D*-glyceric acid template was covalently bonded to *p*-amino styrene and 2,3-*O*-*p*-vinylphenylboronic ester and was then incorporated into a divinylbenzene polymer (29,30). This important development increased the specificity of interaction between template and functional groups within the polymer and allowed for more controlled design of functionality within the imprinted polymer. Another major contribution to the field of molecular imprinting was made by Mosbach and co-workers in 1984 when they introduced a new

method of forming the pre-polymerization complex based on noncovalent interactions (31). Instead of covalently imprinting target molecules, a template assembly was formed utilizing electrostatic and hydrogen bonding interactions to specifically position functional monomers. Since this time, there has been a tremendous amount of work in the area of molecular imprinting, much of which has been reviewed in the literature (11,32–34,35–38).

POLYMER CHEMISTRY

Organic Materials

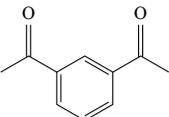
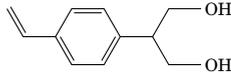
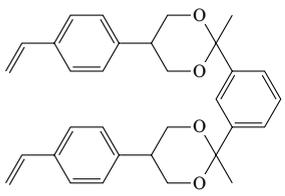
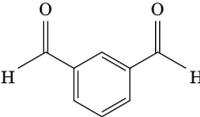
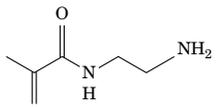
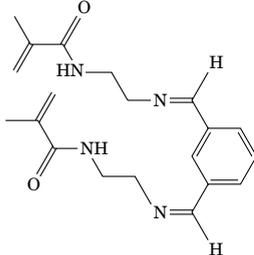
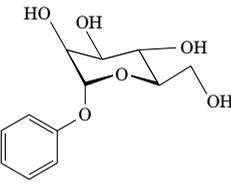
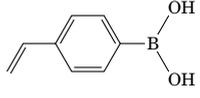
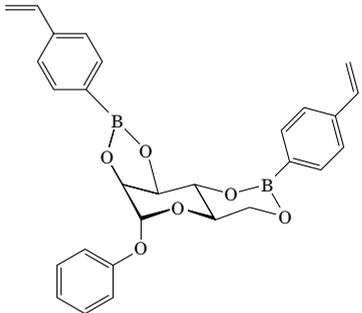
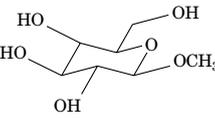
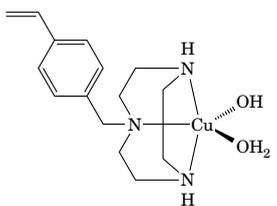
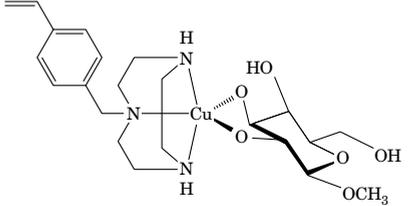
Molecularly imprinted polymers are almost exclusively composed of highly cross-linked organic polymers. These polymers may be prepared using a variety of monomers and polymerization methods. By far, the most common method has been the free radical polymerization of acrylate functionalized monomers. A detailed description of the chemistry of free radical polymerization is available in the literature (39).

Typical formulations consist of equal parts of monomer/template assembly and an inert deluent, referred to as a porogen. The choice of porogen is important since it must solubilize the monomers and, in the event of noncovalent imprinting, must not suppress interactions between the imprinting molecule and the functional monomers. The porogen also serves as an important determinant in the formation of the final polymer morphology.

The resulting network polymers are macroporous in nature. Their macroporosity arises from the phase separation of solid polymer from solvent and unreacted monomer. The timing of the phase separation is dependent on the extent of polymerization and the polymer solubility in the porogen. The phase separated polymer particles aggregate and create a high internal surface area (40,41). This is particularly important since the highly cross-linked networks present a barrier to facile transport of solvent and reagents to the functional sites. The high internal surface area (200–400 m²/g) exposes a large fraction of polymer mass to the external solvent and reagents.

The high percentage of cross-linking monomer (80 to >90 mol %) that is usually employed results in a polymer with a rigid three-dimensional structure. In addition to providing a rigid structure, the cross-linking provides a scaffold for positioning functional groups within voids

Table 1. Representative Examples of Templates, Monomers, and Their Associated Complexes Used for Covalent Imprinting

Type of Linkage	Template	Monomer	Complex	Reference
Ketal				(13,46)
Schiff base				(2)
Boronic ester				(43–45)
Metal complex				(48)

enabling recognition of the imprinted molecules. The interaction between the template molecule and the functional monomers can be covalent or noncovalent in nature. The former can consist of any number of interactions including electrostatic, hydrogen bonding, hydrophobic, and π - π interactions.

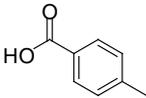
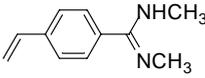
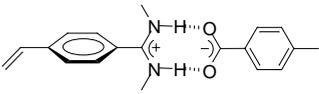
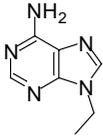
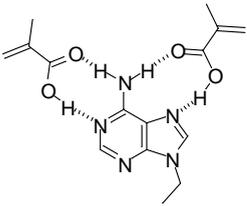
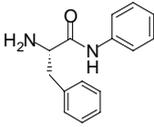
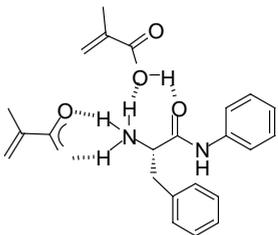
Covalent Imprinting

The covalent method of imprinting requires the synthesis of a polymerizable derivative of the template molecule. Covalent bonds are chosen so as to permit their cleavage following the polymerization. After the cleavage step, the monomers should present functionality that can interact with the imprint molecule either by reforming the covalent

linkage or by noncovalent interactions. Some common covalent bonds used for this purpose are carboxylic (42) and boronic esters (43–45), ketals (13,46), and Schiff bases (2). Also included in this area are metal complexes that bind and orient the template (47–49). Representative examples of covalent template assemblies are given in Table 1.

The covalent method of pre-organization has several advantages over noncovalent methods. Perhaps the most important is that there is no requirement for excess functional monomer. This minimizes the amount of nonspecific bonding that can occur between the analyte and the polymer matrix. A nonspecific interaction in this case is any bonding that occurs other than between the analyte and the defined binding site. In addition, covalent imprinting may produce more “well-defined” binding sites, thus reducing the

Table 2. Representative Examples of Templates, Monomers, and Their Associated Complexes Used for Noncovalent Imprinting

Type of Linkage	Template	Monomer	Complex	Reference
Electrostatic				(51)
Hydrogen bond				(8,15)
				(31,52)

distribution of binding sites with varying affinities for the template molecule. Potential disadvantages of this technique include the additional steps required to synthesize the template-monomer complex and for the chemical cleavage of the template from the polymer. If rebinding is to occur by reformation of covalent bonds, the rate of rebinding may be slow and not suitable for certain applications such as separation media for chromatography.

Noncovalent Imprinting

Non-covalent imprinting employs weaker interactions between the template, functional monomers and cross-linking monomers to position the functional groups and define the binding cavity (31,50). The interactions that have been most commonly utilized are hydrogen bonds, electrostatic, π - π , and hydrophobic interactions. Some representative examples of noncovalent template assemblies are given in Table 2.

The noncovalent method has the benefit of being relatively simple to carry out. It requires fewer steps than the covalent approach and is more compatible with automated or combinatorial methods. In addition, the types and number of potential templates is greater than what can be achieved using the covalent approach. This method is limited, however, to templates that can interact somewhat strongly with functional monomers. In addition, an excess of functional monomers is usually required in the polymerization mixture to ensure optimal results. In part because of this requirement, polymers prepared in this manner contain large numbers of functional groups, which

can contribute to nonspecific bonding and in general may provide lower affinity sites than can be achieved with covalently imprinted polymers.

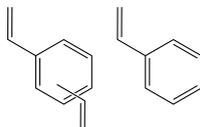
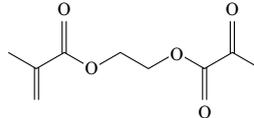
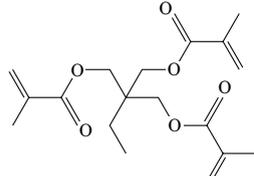
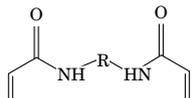
Polymer Networks

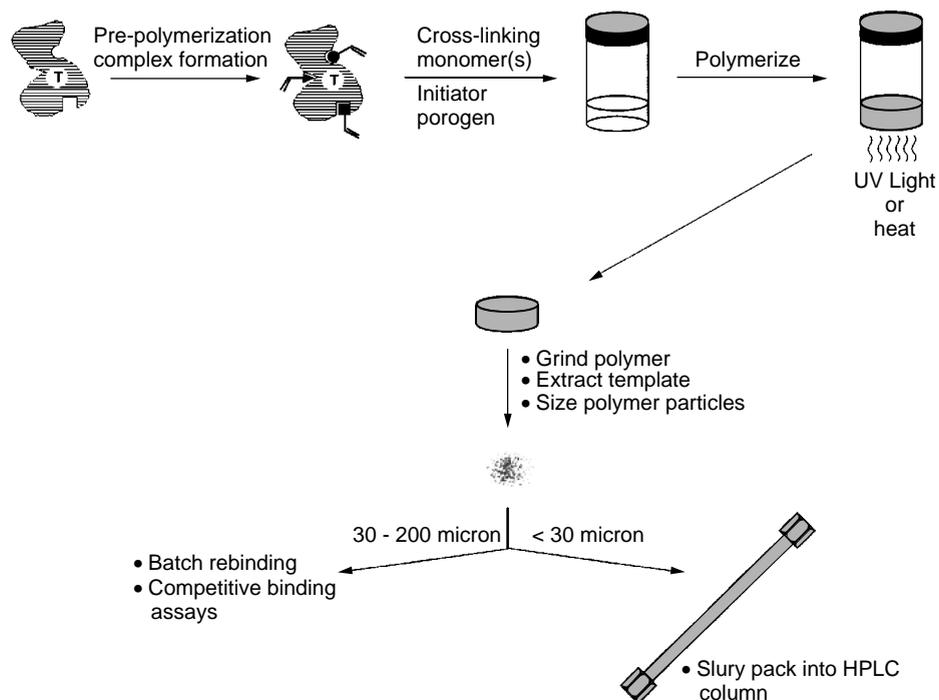
In most imprinting systems the bulk of the polymer is made up of difunctional cross-linking monomers. A wide variety of these have been developed, with the majority being acrylate or styrene based (Table 3). Cross-linking monomers play an important role in determining the bulk properties of the materials. In addition, compatibility between the cross-linking monomers and all other monomers in the formulation must be established.

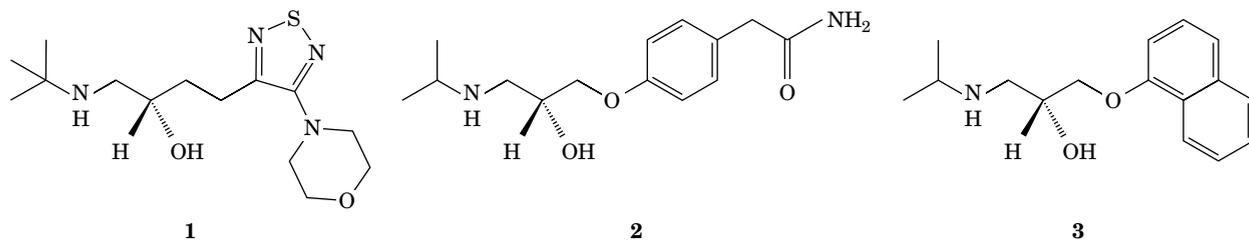
Preparation and Processing. The initial step in the imprinting protocol is formation of a prepolymerization complex between the molecule to be imprinted and the functional monomers which will provide recognition (Fig. 1). This complex can be formed through either covalent or noncovalent interactions. In both cases, the formation of the complex must be reversible to allow extraction/rebinding of the template molecule. This complex is combined with the cross-linking monomer(s), initiator, and the porogen. Oxygen, which can interfere with free-radical polymerizations, is removed from the formulation either by freeze-pump-thaw methods or by displacing it from the solution by sparging with an inert gas.

Free-radical polymerization of the monomer/porogen mixture is initiated either photochemically or thermally.

Table 3. Representative Polymer Networks Used in Molecular Imprinting

Monomer System	Description	Structure	Reference
Styrene/divinylbenzene	Nonpolar/ hydrophobic		(29,53)
Ethylene glycol Dimethacrylate (EGDMA)	Polar aprotic		(9,54)
2,2-bis(hydroxymethyl) butanol trimethacrylate (TRIM)	Polar aprotic trifunctionalized		(55)
<i>N,N'</i> -1,2-Ethanediybis (2-methyl-2-propenamide),	Proteinaceous		(2,56)
Tetramethoxy silane	Silica (sol-gel)	$\text{Si}(\text{OMe})_4$	(26,57,58)

**Figure 1.** Schematic representation of the molecular imprinting process.



Scheme 2

Due to the large number of initiators available, the range of conditions in terms of radiation intensity, temperature, and time that is required is large. The choice of conditions and method of initiation is dictated by the sensitivity of the template to temperature and/or photochemical radiation.

The polymer that is produced is often a monolithic solid. For noncovalent imprinting, releasing the template molecule from the polymer is often achieved by extracting the polymer using a solvent such as methanol. In some cases, it is beneficial to use a small quantity of acid to aid in the removal of the template. For covalent imprinting, the conditions for removal of the template are dictated by the requirements for chemical cleavage of the template–monomer bond.

The next step in the processing sequence involves crushing the polymer followed by sizing of the resulting particles. Many of the chromatographic applications of imprinted polymers require the polymer to be ground into small particles (25–150 μm) and sized to a uniform range.

APPLICATIONS

To date, the applications of MIPs have fallen into four areas: (1) separation technology, (2) catalysis (enzyme mimics), (3) recognition elements for sensors, and (4) anti-body and receptor site mimics.

Separation Technology

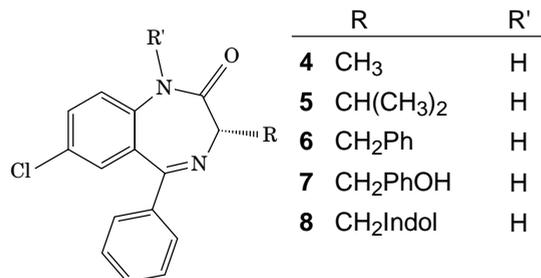
The use of MIPs as stationary phases for chromatography is the most studied application of this technology. The list of molecules that have been examined includes biomolecules such as amino acids, small peptides, proteins, carbohydrates, nucleotides, nucleotide bases, steroids, and a number of small drug molecules. The majority of separation studies have focused on the preparation of stationary phases for high performance liquid chromatography (HPLC) (59). MIPs have also been used in thin layer chromatography (TLC) (60), capillary electrophoresis (61,62), membranes (18,19) and solid-phase extraction (SPE) (63–65).

Mosbach and coworkers demonstrated direct enantioselective separation of β -blocker (-)-S-timolol (template 1; scheme 2) using a chiral stationary phase (CSP) prepared by molecular imprinting (66).

Macroporous copolymers of EGDMA and MAA, prepared with template 1 as the imprinting molecule, were employed as a chiral stationary phase. Following extraction of the template with acetic acid, the MIP stationary phase was used for separation of racemic (template 1) by HPLC. The imprinted polymer stationary phases

gave baseline separation of racemic timolol with a separation factor (α) of 2.9. The separation factor (α) is a measure of the separation of two peaks in a chromatogram and is given by $\alpha = k_2/k_1$, where k' is equal to the capacity factor of each peak. The capacity factor is a measure of binding strength and is defined as $k' = (t - t_0)/t_0$, where t is the retention time of the solute, and t_0 is the void time. The structurally related racemic derivatives atenolol (template 2) and propranolol (template 3) were *not* separated into their enantiomers, nor were they retained on the column to any substantial degree.

Shea et al. explored the ability of MIPs to separate enantiomers of a series of benzodiazepine molecules (9). Enantiomerically pure benzodiazepine derivatives (molecules 4–8, Scheme 3) were used to imprint EGDMA polymers using MAA as the functional monomer. The resulting materials were ground into particles and analyzed by HPLC.



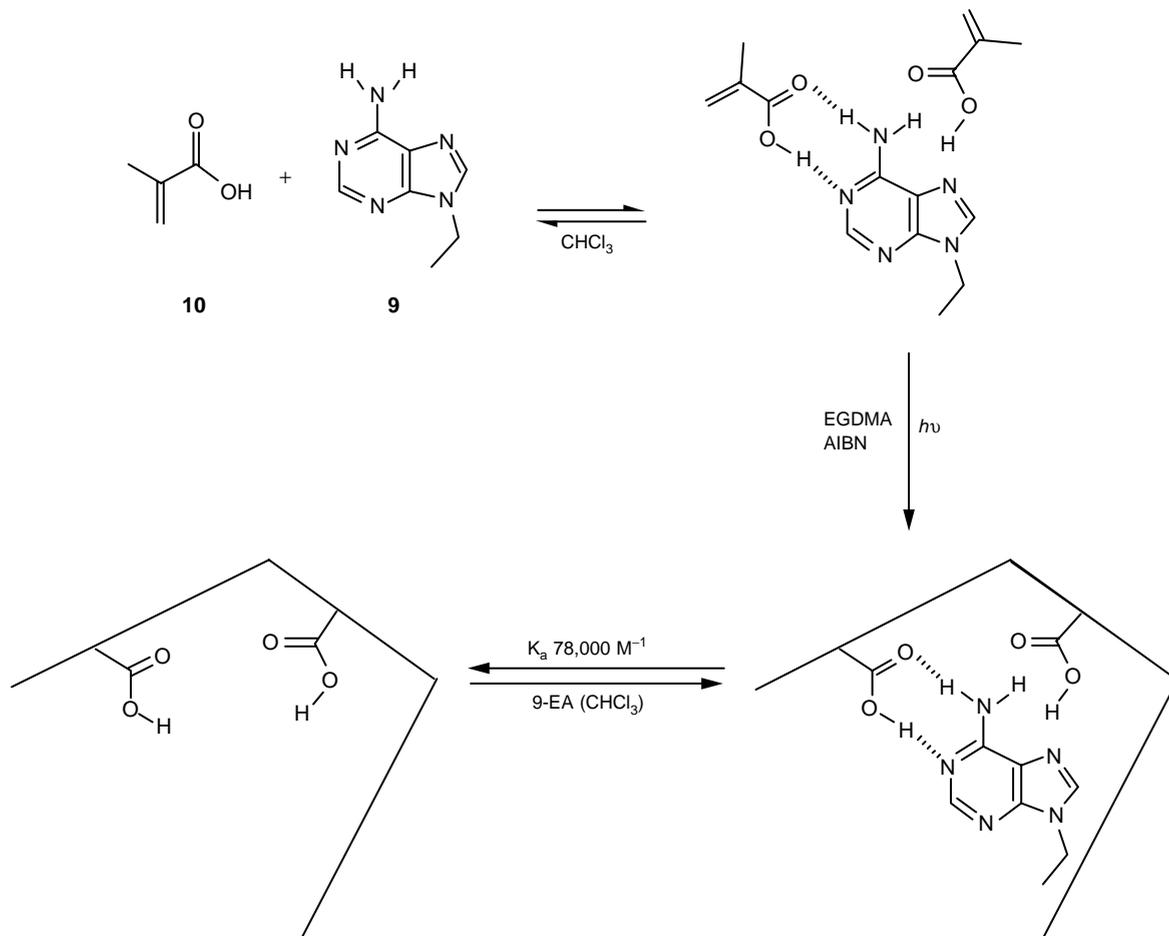
Scheme 3

Regardless of the benzodiazepine used for the imprinting, enantioselectivity could be achieved for the entire family of benzodiazepines. In each case, the imprinted enantiomer was retained over the nonimprinted isomer ($\alpha > 1$; Table 4). Additionally, the greatest separation for each polymer corresponded to the benzodiazepine which had been imprinted (bold entries). The ability of benzodiazepine MIPs to produce separation for a family of molecules in favor of the absolute configuration of the template is an important result of this study.

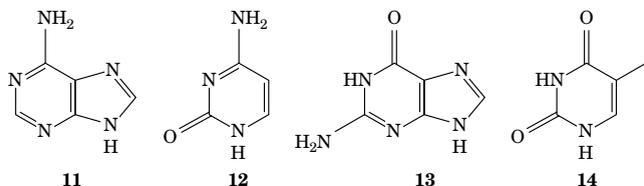
Synthetic polymer complements to biologically important nucleosides have been prepared using 9-ethyladenine (9-EA; polymer 9, scheme 4) as the imprinting molecule (15). The molecular basis for the imprinting arises from the fact that adenine and its derivatives are known to form complexes with carboxylic acids. 9-EA was employed as the template; methacrylic acid (MAA; polymer 10) was employed as the functional monomer (scheme 4). The formulation included EGDMA as the cross-linking monomer in chloroform solution. Polymerization was initiated photochemically at 12°C in the presence of AIBN.

Table 4. Benzodiazepine α Values for Retention on MIP Stationary Phases

MIP Template	Benzodiazepine Derivative Injected				
	4-Hydroxybenzyl (7)	Benzyl (6)	Indolyl (8)	Isopropyl (5)	Methyl (4)
(S)-7 4-Hydroxybenzyl	3.8	2.5	1.3	1.6	1.4
(S)-6 Benzyl	2.3	3.0	1.7	1.6	1.7
(S)-8 Indolyl	1.5	1.8	2.3	1.3	1.4
(S)-5 Isopropyl	1.2	1.3	1.3	2.2	1.6
(S)-4 Methyl	1.0	1.1	1.1	1.1	1.3

**Scheme 4**

When used as the stationary phase for HPLC, the 9-EA imprinted column packing was found to have a strong affinity for adenine and its derivatives. The remaining purine and pyrimidine bases elute close to the void volume.

**Scheme 5**

An illustration of the selectivity is given in Fig. 2. An HPLC trace of a sample mixture of adenine (A; template 11), cytosine (C; 12), guanine (G; 13), and thymine (T; 14) reveals C, G, and T elute close to the void volume (~ 2 min at 1 ml/min flow) while adenine elutes in 23 minutes. In contrast, polymers made by replacing 9-EA with benzylamine show no affinity for the nucleoside bases: all elute between 1.6 and 2.0 minutes.

Most studies with imprinted polymer have been carried out using polymer particles either in the batch or chromatographic mode. There have been several reports however of the preparation of imprinted polymers fabricated

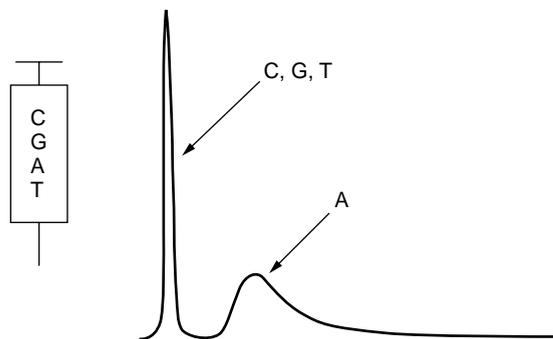


Figure 2. HPLC race of a 4.6 mm \times 10 cm column packed with polymer templated with 9-ethyladenine (9-EA). A mixture of adenine (A), thymine (T), guanine (G), and cytosine (C) was injected. Adenine elutes at 23 minutes; all other nucleosides elute close to the void volume (2 min). The mobile phase is MeCN:AcOH:H₂O in the ratio 92.5:5:2.5.

into membranes for the selective transport and separation of molecules. One example involves the preparation of MIPs with 9-ethyladenine (template 9) as the template and methacrylic acid (10) as the polymerizable functional monomer (scheme 4) (18).

The free-standing films of imprinted polymers were obtained by the polymerization of films of a DMF solution of ethylene glycol dimethacrylate and methacrylic acid with 9-ethyladenine (template 9) as the template.

Transport studies were carried out with an H-shaped two-compartment cell. The concentration of substrate in the receiving phase as a function of time was quantified by HPLC. Under steady-state conditions, a linear correlation between amount of substrate transported and time was observed. The transport rates of adenine, thymine,

and cytosine were measured; in all cases it was found that adenine was transported at a higher rate than other nucleic acid bases through membranes imprinted with template 9. Furthermore, the imprinted membranes transported adenosine (15) at a higher rate than guanosine (16) (Fig. 3). In contrast, reference polymer membranes, prepared in the absence of template (9), transported adenine at a comparable rate to those of cytosine and thymine. A reference polymer membrane of the same monomer composition prepared without template 9, on the other hand, transports adenine at rates comparable to those of thymine and cytosine.

Sensors

MIPs have been used as the recognition element in chemical sensing devices. Biosensors utilize a recognition element, such as an antibody or enzyme, in conjunction with a transducer. A chemical signal results from the binding of the analyte to the receptor that is then transformed into an electrical or optical signal that can be monitored. In many cases, the development of recognition elements lags far behind the transduction methods available, such as optical, resistive, surface acoustic wave (SAW) or capacitance measurements. There is considerable potential for new chemical sensing devices if suitable recognition elements were available.

There are several potential advantages that may be gained by utilizing MIPs as the recognition element in place of biological receptors. Since MIPs are synthetic receptors, they have a virtually unlimited pool of analytes. In addition, MIPs are stable to harsh conditions that may not be compatible with biologically based sensors. Furthermore, the ability to incorporate a signaling element, such as a fluorescent probe, in the vicinity of the binding

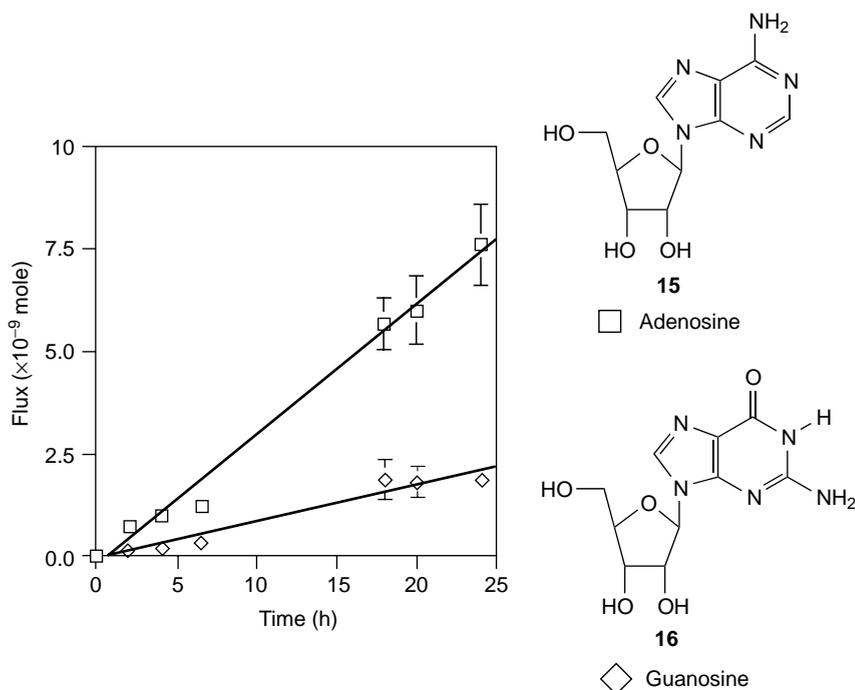


Figure 3. Plot of the facilitated transport of adenosine from an equimolar methanol/chloroform (6:94 v/v) solution (adenosine = guanosine = 26 μ M). The membrane was imprinted with 9-ethyladenine. The selectivity factor is 3.4.

Table 5. Summary of Sensor Applications Using Molecularly Imprinted Polymers

Class	Typical Analyte	Functional Monomer	Detection Range	Reference
Fluorimetry	Dansyl-L-phenylalanine	MAA, 2VPy	0–30 $\mu\text{g/mL}$	(68)
	cAMP	DMASVBP + HEMA	0.1–100 μM	(69)
Conductometry	Atrazine	DEAEM	0.01–0.5 mg/L	(70)
Amperometry	Morphine	MAA	0.1–10 $\mu\text{g/mL}$	(71)
Luminescence	PMP	Eu(III) + DVMB	0.125–150000 $\mu\text{g/L}$	(49,72)
pH	Glucose	STACNCu	0–25 mM	(48)

site could be utilized in sensor applications. It may also be possible to prepare sensors with an array of polymers imprinted for any number of analyte molecules for fabrication of a single sensor capable of detecting many substances. Some examples of sensor applications using MIPs and their detection ranges are shown in Table 5.

The use of luminescence spectroscopy combined with fiber optics can provide systems for sensor applications. Utilizing molecularly imprinted polymers in combination with these systems can add chemical selectivity to these types of sensors. This technology has been applied to the detection of many types of compounds including nerve agents (49), herbicides (72,73), drugs molecules (70), and biomolecules (68,74).

Mosbach et al. demonstrated the use of a MIP based fiber-optic sensing device for a fluorescence-labeled amino acid (dansyl-L-phenylalanine; template 17, scheme 6) (67). These materials were prepared using MAA and 2-vinyl pyridine as the functional monomers and EGDMA as the cross-linking monomer. A system was devised in which polymer particles were held against the fiber tip using a nylon net (Fig. 4).

In both systems the polymers showed an increased affinity for the imprinted enantiomer. This was demonstrated as an increase in the measured fluorescence in the fiber-optic system. In addition, the ability to follow increases in concentration by monitoring the increases in fluorescence was demonstrated. Binding of the analyte to the polymer

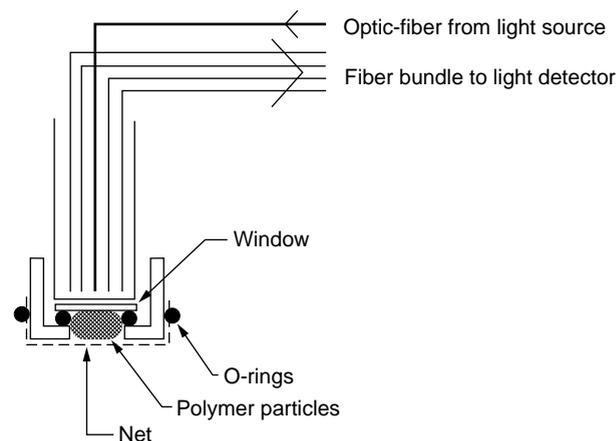
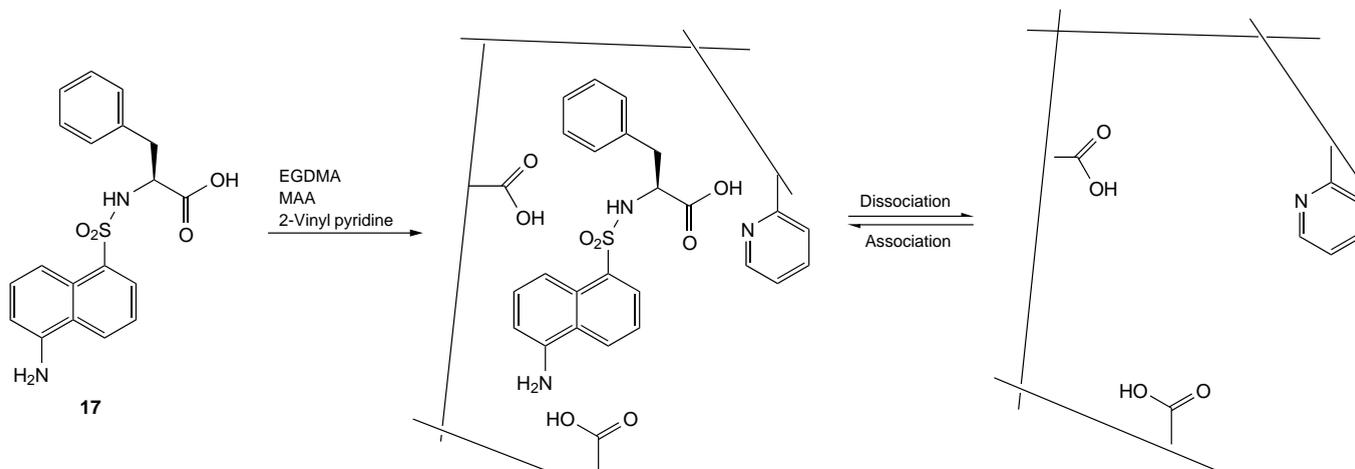


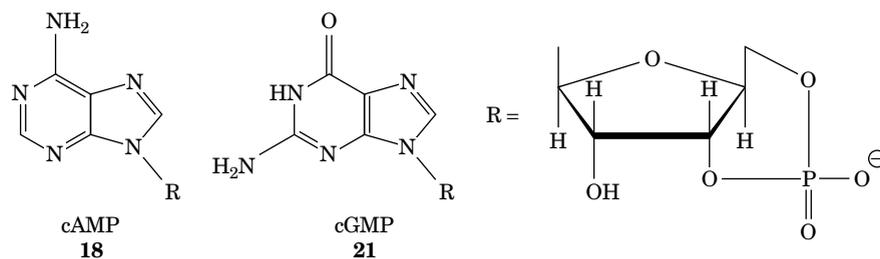
Figure 4. Fiber-optic sensing device configuration. A layer consisting of 2 mg of polymer particles was held below the quartz glass window. The layer was kept in place by a nylon net (20 μm pores) and an O-ring. An optical fiber was connected through the quartz window onto the polymer surface. Part of the total light emitted from the polymer was collected and guided through the fiber bundle to the detector (visible light-sensitive photodiode).

could be followed in real time as the time required to reach a steady-state response was 4 hours. This rebinding time correlates with the equilibration time of this imprinted system.

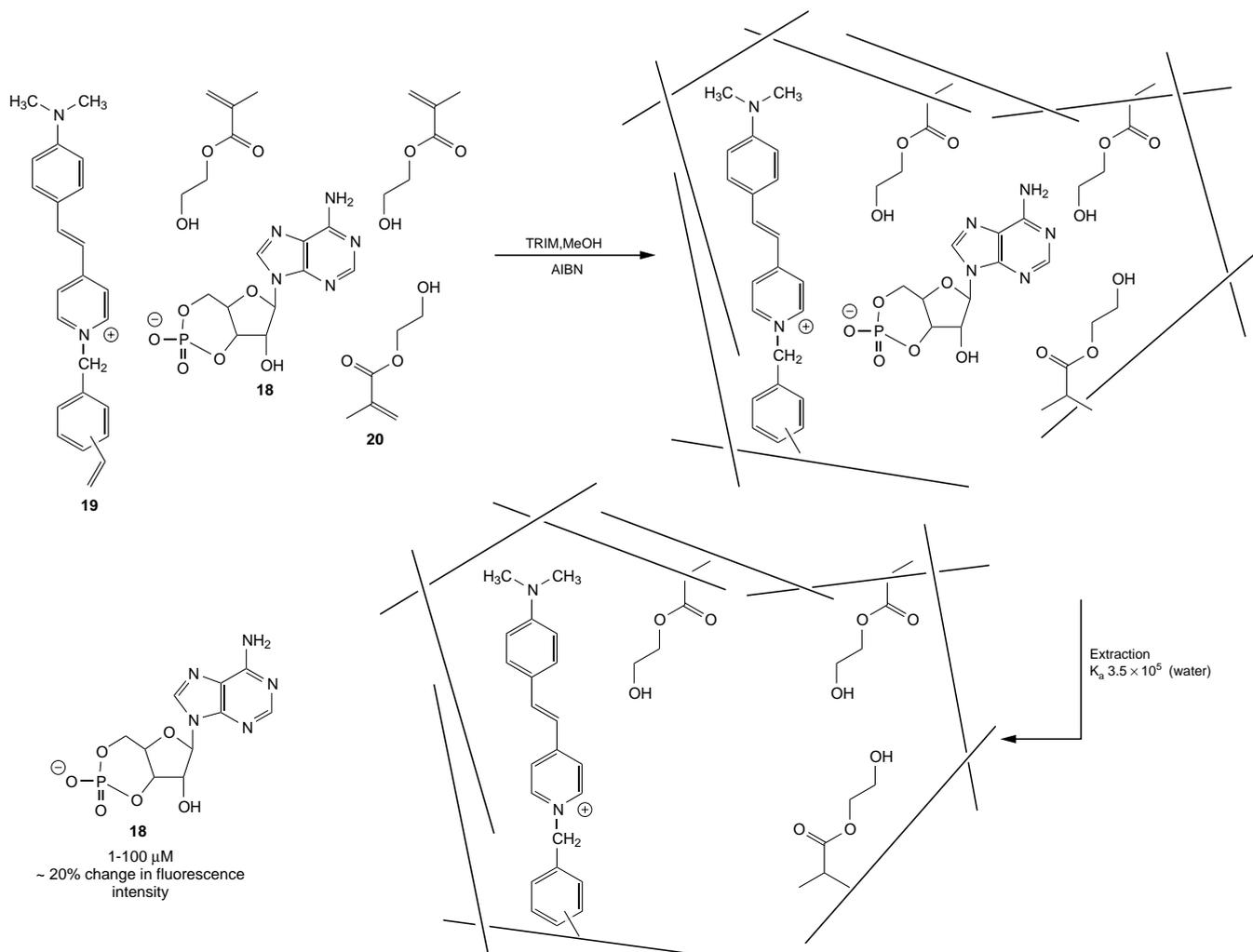


Scheme 6

Powel et al. prepared imprinted polymers for the detection of cAMP (polymer 18) (68). These polymers are unique in that they contain a fluorescent dye, *trans*-4-[*p*-(*N,N*-dimethylamino)styryl]-*N*-vinylbenzylpyridinium chloride (19, scheme 8), as an integral part of the binding site. The polymer serves as both the recognition element and the transducer for the detection of cAMP. The polymer was prepared by polymerization of trimethylol trimethacrylate (TRIM), 2-hydroxyethyl methacrylate (HEMA; 20), and the fluorescent functional monomer, template 19, in the presence of cAMP. The polymerization was initiated photochemically at room temperature.



Scheme 7



Scheme 8

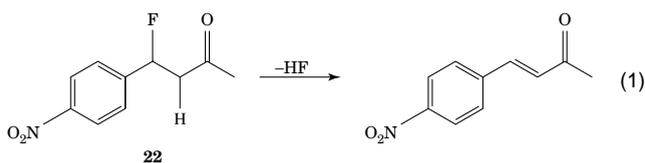
Possible interactions between the template and the polymer include aryl stacking and electrostatic interactions between 19 and the template. HEMA is able to provide additional sites in the polymer for hydrogen bonding with the template.

The fluorescence of the imprinted polymer was diminished when in the presence of aqueous cAMP. The quenching was found to reach a maximal level between 10 and 100 μ M cAMP. In addition, no quenching was observed in a control polymer prepared without template. Addition of cGMP (21) had little effect on the fluorescence of both the imprinted and control polymer.

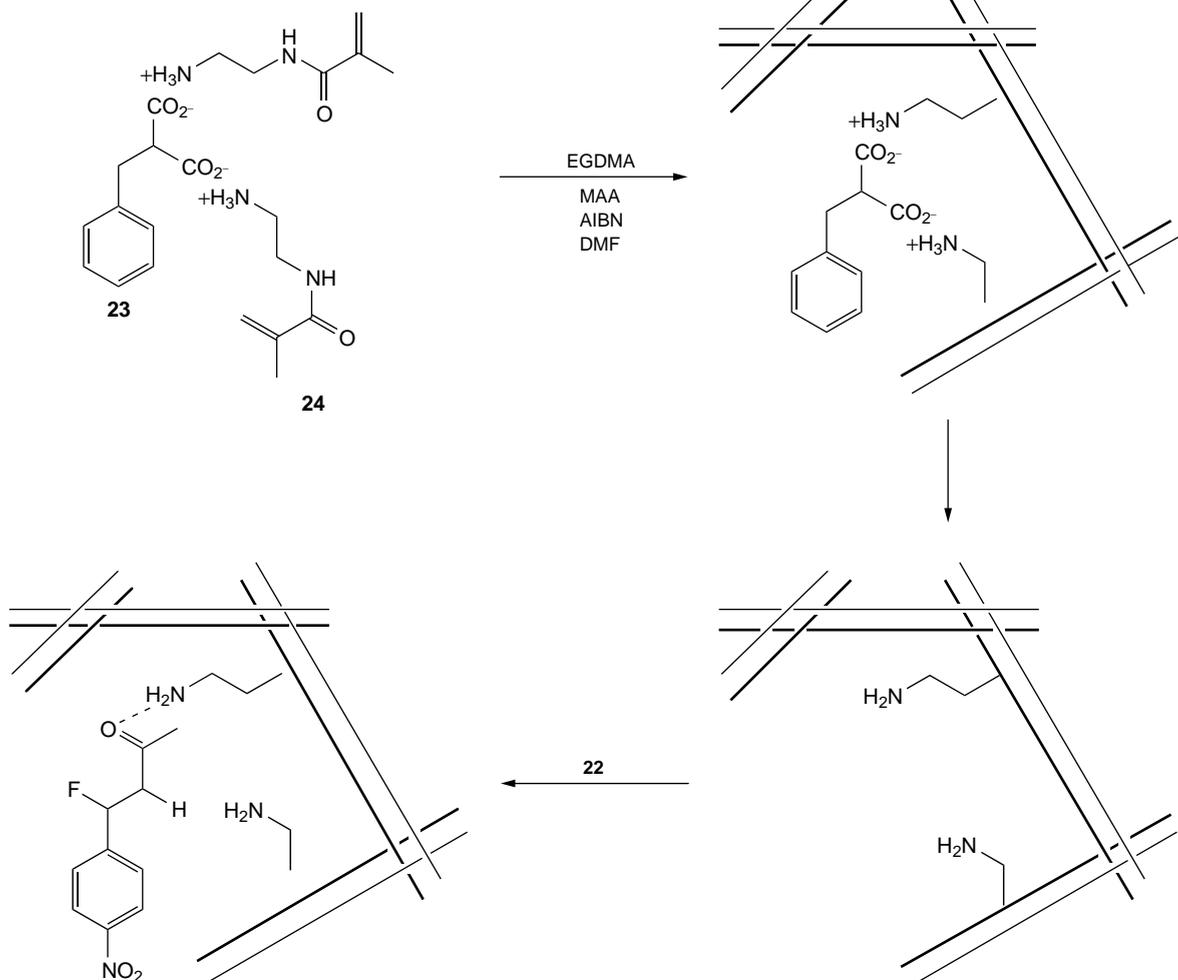
Catalysis

One of the most challenging applications of molecularly imprinted polymers is the fabrication of designed catalysts and the preparation of artificial enzymes. Because of their ability to prearrange functional groups within the polymer matrix, MIPs lend themselves to this application. Methods for preparing these MIPs include (1) utilizing a template that pre-arranges the catalytic groups within the binding site, this method has been referred to as the "bait-and-switch" method, (2) the use of a transition state analog as the template, (3) the use of coordination compounds to effect the catalytic reaction (32,75).

The use of templated polymers as catalysts for chemical transformations has been demonstrated using the dehydrofluorination of 4-fluoro-4-(*p*-nitrophenyl)butan-2-one (22, scheme 9) (21).



Scheme 9



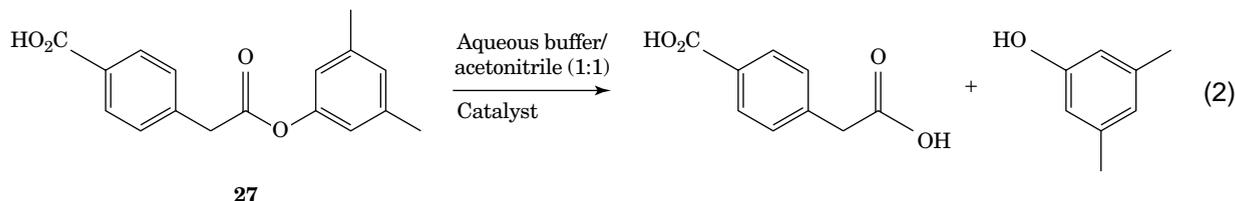
Scheme 10

The strategy for preparing the MIP catalyst is illustrated in scheme 10 below. The template molecule, benzylmalonic acid (template 23), was used to orient *N*-(2-aminoethyl)-methacrylamide (24) monomers prior to polymerization. The bulk of the polymerization solution consisted of a mixture of EGDMA, methyl methacrylate, monomer 24, and DMF as the porogen.

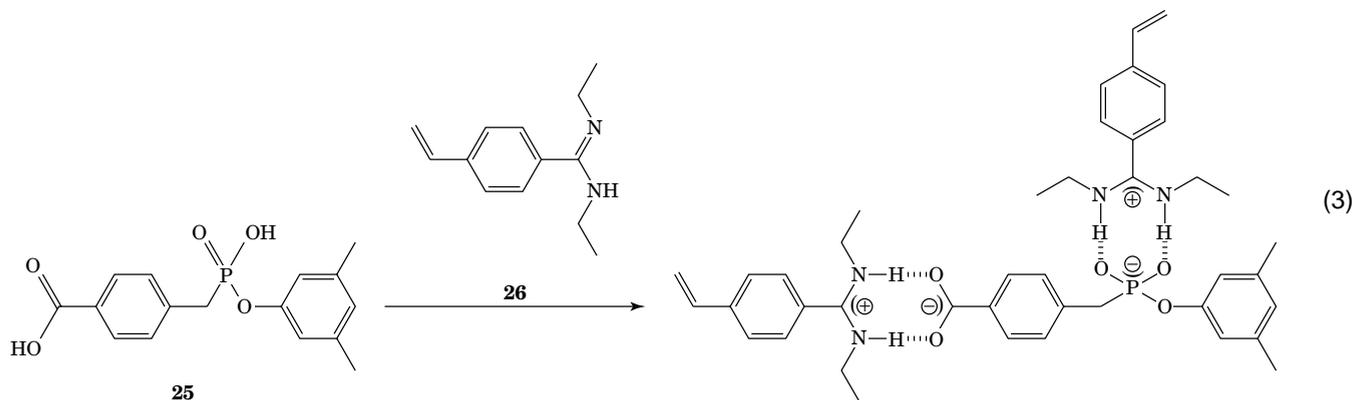
A survey of the catalytic performance of the templated polymer and that of a control polymer with randomly oriented amino groups indicated that both polymers were found to catalyze the reaction shown in scheme 9; however, the rate of dehydrofluorination by the templated polymer (k_{template}) under pseudo-first-order conditions was 3.2 time greater than k_{control} (Fig. 5).

The imprinted polymer exhibits Michaelis-Menten kinetics in benzene with a K_m of 27 mM and a k_{cat} of $1.1 \times 10^{-2} \text{ min}^{-1}$. In comparison, an antibody designed to catalyze the dehydrofluorination of 22 in water shows a K_m of 0.182 mM and a k_{cat} of 0.192 min^{-1} .

Wulff et al. have reported the preparation of molecularly imprinted polymers which catalyze the hydrolysis shown in scheme 11 (23). The polymers were prepared using a phosphonic acid monoesters 25 as a transition state analog template (scheme 12).



Scheme 11



Scheme 12

Coordination of monoester **25** with two equivalents of amidine **26** provides a polymerizable complex, Eq. (3). The bis(amidinium) complex was polymerized with a large excess of EGDMA in THF.

Removal of the phosphoric acid monoester transition state analogue left a cavity bearing two *N,N*-diethylamidinium groups (Fig. 6). Hydrolysis of ester **27** by the polymer imprinted with phosphoric acid monoester **25** was accelerated by nearly 100-fold over the reaction of ester **27** with amidine **26** in homogeneous solution. In addition, the catalyzed hydrolysis showed typical Michaelis-Menten kinetics while the homogeneous solution has a linear relationship between reaction rate and concentration. The catalysis could also be inhibited by addition of phosphoric acid monoester **25** which has a higher affinity ($K_i = 0.025$ mM) for the polymer than acid **27**. The activity of these catalysts is only one to two orders of magnitude below that of catalytic antibodies.

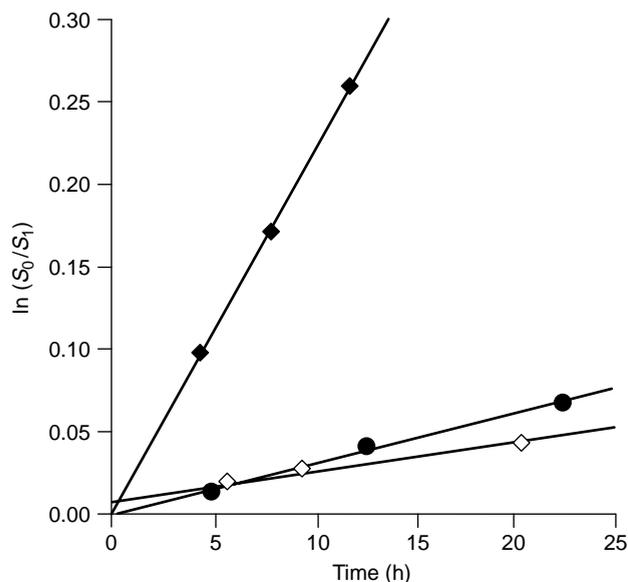


Figure 5. Pseudo-first-order plot of the dehydrofluorination of polymer **22** (scheme 9) under conditions of excess substrate to catalyst. The log of the ratio of the initial concentration of substrate **22** (S_0) to that at time t (S_t) is plotted against time. In the presence of pyridine as a scavenger, the catalyst demonstrates seven turnovers after 23 hours. Imprinted polymer catalyst (◆); control polymer (●); homogeneous reaction (2-aminoethyl acetamide) (◇).

Antibody and Receptor Site Mimics

Binding assays are used for the determination of small quantities of compounds in blood serum and other systems and for screening compounds for affinity to a specific receptor. Assays of this type require a receptor, often an antibody that specifically binds the molecule being studied. Antibodies often show extremely good recognition properties for their associated molecule (antigen); however, their use is hampered by their sensitivity to harsh conditions and what can be an involved and costly production process.

Several studies have been performed that utilize imprinted polymers as artificial antibodies for use in binding assays. These studies have been performed using competitive radioligand binding protocols and simple batch binding methods (24,76,77). The advantages of using these

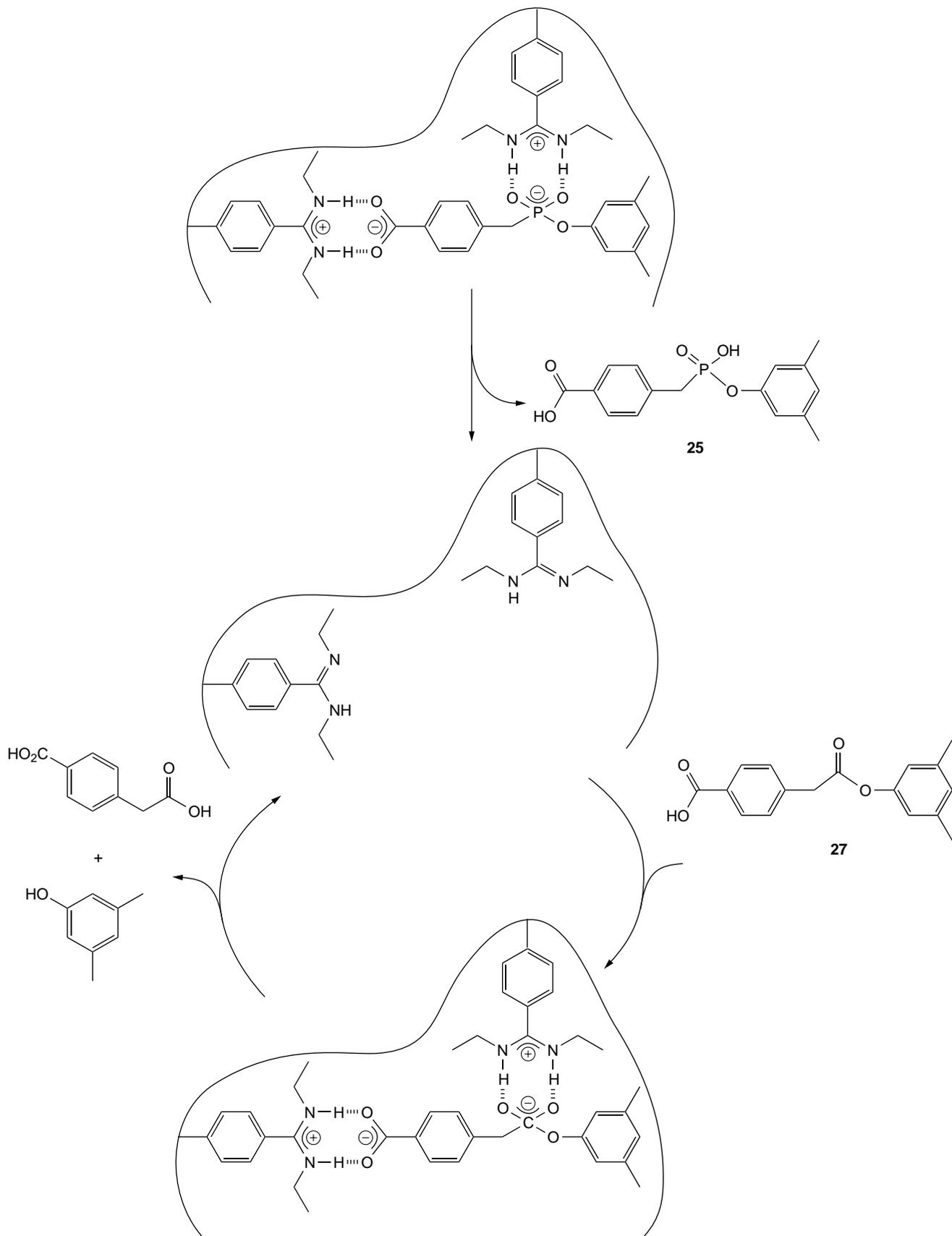


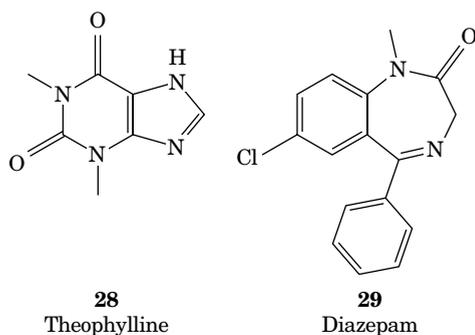
Figure 6. Schematic representation of catalytic cycle for the hydrolysis of ester 27 (scheme 11). A cavity bearing two *N,N*-diethylamidinium groups remains after removal of the template 25 (middle) scheme 10. The alkaline hydrolysis of ester 27 is catalyzed by the polymer that stabilizes the transition state of the reaction (bottom).

Table 6. Cross-reactivity of Xanthine and Uric Acid Derivatives for Binding of ³H-Theophylline and Benzodiazepines for Binding of ³H-Diazepam to Imprinted Polymers

Theophylline Antibodies	Cross-reaction	Diazepam Antibodies	Cross-reaction
<i>Competitive Ligand</i>	(%)	<i>Competitive Ligand</i>	(%)
Theophylline (1,3-dimethylxanthine)	100	Diazepam	100
3-Methylxanthine	7	Alprazolam	40
Xanthine	<1	Desmethyldiazepam	27
Hypoxanthine	<1	Clonazepam	9
7-(β-hydroxyethyl-1,3-dimethylxanthine)	<1	Lorazepam	4
Caffeine (1,3,7-trimethylxanthine)	<1	Chlordiazepoxide	2
Theobromine (3,7-dimethylxanthine)	<1		
Uric acid	<1		
1-Methyluric acid	<1		
1,3-Dimethyluric acid	<1		

materials as the recognition elements in assays include their ease of preparation, the stability of the polymers under harsh conditions such as high temperature, extreme pH, and organic solvents. In addition, the ability to prepare MIPs when a biological receptor is unavailable or difficult to isolate provides further impetus for their use.

The first use of imprinted polymers as the recognition element in a radiolabeled ligand binding assay was demonstrated with assays for Diazepam (28) and Theophylline (29), in human serum (24) (scheme 13). The results of this assay are comparable to established antibody assay protocols for these drugs. The method employed for this study uses inhibition of radiolabeled ligand binding by the serum analyte. The amount of radioligand bound to the polymer is inversely proportional to the analyte concentration in the serum.

**Scheme 13**

Both assays showed linear responses over ranges satisfactory for the therapeutic monitoring of both drugs (14–224 μM for Theophylline and 0.44–28 μM for Diazepam). The cross-reactivity of major metabolites and compounds of similar structure to Theophylline and Diazepam was determined to evaluate the specificity of the synthetic receptors. Theophylline imprinted materials proved to be highly specific with only 3-methylxanthine showing cross-reactivity (Table 6). The Diazepam assay indicated that cross reactivity of the synthetic receptor was

very similar to antibody assays. In this case, several structurally similar benzodiazepines showed some cross-reactivity (Table 6). In fact, both assays displayed very good agreement with the corresponding antibody assays.

CONCLUSIONS

Molecular imprinting is a method with which materials with high affinities for a wide range of organic compounds can be made. Since its inception, the number and types of monomers and templates that have been used successfully has increased dramatically. In addition, our understanding of the mechanisms through which the materials operate is increasing. While there is still enormous room for development, some areas of application have become well developed and commercial applications are on the horizon.

The applications discussed in this article are only a few examples of what has been achieved. Other uses for MIPs are being explored, as well as new ways to develop and optimize their capabilities. The true utility of molecularly imprinted materials is limited only by the ingenuity of those developing them.

BIBLIOGRAPHY

- G. Wulff and G. Kirstein. *Angew. Chem. Int. Ed. Eng.* **29**: 684–686 (1990).
- K.J. Shea, G.J. Stoddard, D.M. Shavelle, F. Wakui, and R.M. Choate. *Macromol.* **23**: 4497–4507 (1990).
- S.E. Bystrom, A. Borje, and B. Akermark. *J. Am. Chem. Soc.* **115**: 2081–2083 (1993).
- M.J. Whitcombe, M.E. Rodriguez, P. Villar, and E.N. Vulfson. *J. Am. Chem. Soc.* **117**: 7105–7111 (1995).
- B. Sellergren and K.J. Shea. *J. Chromatogr.* **654**: 17–28 (1993).
- J. Matsui, T. Kato, T. Takeuchi, M. Suzuki, K. Yokoyama, E. Tamiya, and I. Karube. *Anal. Chem.* **65**: 2223–2224 (1993).
- M. Kempe and K. Mosbach. *J. Chromatogr.* **691**: 317–323 (1995).
- D. Spivak, M.A. Gilmore, and K.J. Shea. *J. Am. Chem. Soc.* **119**: 4388–4393 (1997).

9. B.R. Hart, D.J. Rush, and K.J. Shea. *J. Am. Chem. Soc.* **122**: 460–465 (2000).
10. D.A. Spivak. Thesis. University of California at Irvine, 1995.
11. G. Wulff. *Angew. Chem. Int. Ed. Eng.* **34**: 1812–1832 (1995).
12. K. Haupt and K. Mosbach. *Trends Biotechnol.* **16**: 468–475 (1998).
13. K.J. Shea and D.Y. Sasaki. *J. Am. Chem. Soc.* **113**: 4109–4120 (1991).
14. K. Mosbach. *Trends Biochem. Sci.* **19**: 9–14 (1994).
15. K.J. Shea, D.A. Spivak, and B. Sellergren. *J. Am. Chem. Soc.* **115**: 3368–3369 (1993).
16. K.J. Shea, E.A. Thompson, S.D. Pandey, and P. Beauchamp. *J. Am. Chem. Soc.* **102**: 3149 (1980).
17. G. Wulff and J. Vietmeier. *Makromol. Chem.* **190**: 1727–1735 (1989).
18. J. Mathew-Krotz and K.J. Shea. *J. Am. Chem. Soc.* **118**: 8154–8155 (1996).
19. S. Marx-Tibbon and I. Willner. *J. Chem. Soc. Chem. Commun.* 1261–1262 (1994).
20. D.K. Robinson and K. Mosbach. *J. Chem. Soc. Chem. Commun.* 969–970 (1989).
21. J.V. Beach and K.J. Shea. *J. Am. Chem. Soc.* **116**: 379–380 (1994).
22. B. Sellergren and K.J. Shea. *Tetrahedron: Asymmetry* **5**: 1403–1406 (1994).
23. G. Wulff, T. Gross, and R. Schonfeld. *Angew. Chem. Int. Ed. Eng.* **36**: 1962–1964 (1997).
24. G. Vlatakis, L.I. Andersson, R. Muller, and K. Mosbach. *Nature* **361**: 645–647 (1993).
25. P.K. Owens, L. Karlsson, E.S.M. Lutz, and L.I. Andersson. *Trends Anal. Chem.* **18**: 146–154 (1999).
26. F.H. Dickey. *Proc. Natl. Acad. Sci.* **35**: 227 (1949).
27. F.H. Dickey. *J. Phys. Chem.* **59**: 695 (1955).
28. R. Curti and U. Colombo. *J. Am. Chem. Soc.* **74**: 3961 (1952).
29. G. Wulff and A. Sarhan. *Angew. Int. Ed.* **11**: 341 (1972).
30. G. Wulff, A. Sarhan, and K. Zabrocki. *Tetrahedron Lett.* **44**: 4329 (1973).
31. L. Anderson, B. Sellergren, and K. Mosbach. *Tetrahedron Lett.* **25**: 5211 (1984).
32. K.J. Shea. *Trends Polym. Sci.* **2**: 166–173 (1994).
33. O. Ramstrom and R.J. Ansell. *Chirality* **10**: 195–209 (1998).
34. K. Mosbach and O. Ramstrom. *Bio-Technology* **14**: 163–170 (1996).
35. S. Al-Kindy, R. Badia, J.L. Suarez-Rodriguez, and M.E. Diaz-Garcia. *Critical Reviews in Analytical Chemistry* **30**: 291–309 (2000).
36. H. Asanuma, T. Hishiya, and M. Komiyama. *Advanced Materials* **12**: 1019–1030 (2000).
37. K. Haupt and K. Mosbach. *Chemical Reviews* **100**: 2495–2504 (2000).
38. T. Takeuchi and J. Haginaka. *Journal of Chromatography B* **728**: 1–20 (1999).
39. G. Odian. *Principles of Polymerization*. Wiley, New York, 1991, ch. 3.
40. A. Guyot. In *Synthesis and Separations Using Functional Polymers*. P. Hode, eds., and D.C. Sherrington, Wiley-Interscience, New York, 1988, ch. 1.
41. J.R. Millar, D.G. Smith, W.E. Marr, and T.R.E. Kressman. *J. Chem. Soc.* 218 (1963).
42. B. Sellergren and L. Andersson. *J. Org. Chem.* **55**: 3381–3383 (1990).
43. G. Wulff and M. Minarik. *J. Liq. Chromatogr.* **13**: 2987–3000 (1990).
44. G. Wulff, J. Vietmeier, and H.G. Poll. *Makromol. Chem.* **188**: 731 (1987).
45. G. Wulff, W. Vesper, R. Grobe-Einsler, and A. Sarhan. *Makromol. Chem.* **178**: 2799 (1977).
46. K.J. Shea, G.J. Stoddard, and D.Y. Sasaki. *Macromol.* **22**: 4303–4308 (1989).
47. P.K. Dhal and F.H. Arnold. *Macromol.* **25**: 7051–7059 (1992).
48. G.H. Chen, Z.B. Guan, C.T. Chen, L.T. Fu, V. Sundaresan, and F.H. Arnold. *Nat. Biotechnol.* **15**: 354–357 (1997).
49. A.L. Jenkins, O.M. Uy, and G.M. Murray. *Anal. Chem.* **71**: 373–378 (1999).
50. B. Sellergren. In R.A. Bartsch and M. Maeda, eds., *ACS Symp. Series 703*. American Chemical Society, Washington, DC, 1998.
51. G. Wulff and R. Schonfeld. *Adv. Matls.* **10**: 957+ (1998).
52. L. Andersson, B. Ekberg, and K. Mosbach. *Tetrahedron Lett.* **25**: 5211–5214 (1985).
53. K.J. Shea and E.A. Thompson. *J. Org. Chem.* **43**: 4253 (1978).
54. B. Sellergren and K.J. Shea. *J. Chromatogr.* **635**: 31–49 (1993).
55. M. Kempe. *Anal. Chem.* **68**: 1948–1953 (1996).
56. O. Norrlov, M. Glad, and K. Mosbach. *J. Chromatogr.* **299**: 29–41 (1984).
57. A. Katz and M.E. Davis. *Nature* **403**: 286–289 (2000).
58. N.K. Raman, M.T. Anderson, and C.J. Brinker. *Chem. Mat.* **8**: 1682–1701 (1996).
59. V.T. Remcho and Z.J. Tan. *Anal. Chem.* **71**: 248A–255A (1999).
60. D. Kriz, C.B. Kriz, L.I. Andersson, and K. Mosbach. *Anal. Chem.* **66**: 2636–2639 (1994).
61. J.M. Lin, T. Nakagama, K. Uchiyama, and T. Hobo. *J. Pharm. Biomed. Anal.* **15**: 1351–1358 (1997).
62. J.M. Lin, T. Nakagama, X.Z. Wu, K. Uchiyama, and T. Hobo. *Fresenius J. Anal. Chem.* **357**: 130–132 (1997).
63. J. Olsen, P. Martin, I.D. Wilson, and G.R. Jones. *Analyst* **124**: 467–471 (1999).
64. B. Sellergren. *Trends Anal. Chem.* **18**: 164–174 (1999).
65. D. Stevenson. *Trends Anal. Chem.* **18**: 154–158 (1999).
66. L. Fischer, R. Muller, B. Ekberg, and K. Mosbach. *J. Am. Chem. Soc.* **113**: 9358–9360 (1991).
67. D. Kriz, L. Ramstrom, A. Svensson, and K. Mosbach. *Anal. Chem.* **67**: 2142–2144 (1995).
68. P. Turkewitsch, B. Wandelt, G.D. Darling, and W.S. Powell. *Anal. Chem.* **70**: 2025–2030 (1998).
69. S.A. Piletsky, E.V. Piletskaya, A.V. Elgersma, K. Yano, I. Karube, Y.P. Parhometz, and A.V. Elskaya. *Biosens. Bioelectron.* **10**: 959–964 (1995).
70. D. Kriz and K. Mosbach. *Anal. Chim. Acta.* **300**: 71–75 (1995).
71. A.L. Jenkins, O.M. Uy, and G.M. Murray. *Anal. Commun.* **34**: 221–224 (1997).
72. B. Bjarnason, L. Chimuka, and O. Ramstrom. *Anal. Chem.* **71**: 2152–2156 (1999).
73. S. Kroger, A.P.F. Turner, K. Mosbach, and K. Haupt. *Anal. Chem.* **71**: 3698–3702 (1999).
74. K. Hirayama and K. Kameoka. *Bunseki Kagaku* **49**: 29–33 (2000).
75. K. Mosbach. *Protein Eng.* **8**: 54–54 (1995).
76. K. Haupt. *React. Funct. Polym.* **41**: 125–131 (1999).
77. L. Ye, P.A.G. Cormack, and K. Mosbach. *Anal. Commun.* **36**: 35–38 (1999).

NEURAL NETWORKS

JERRY A. DARSEY

University of Arkansas at Little Rock
Little Rock, AR

A SHORT TUTORIAL ON ARTIFICIAL NEURAL NETWORKS

The first part of this article introduces the reader to the fundamentals of artificial neural network (ANN) computing, but it is, in no way, comprehensive. Many books and articles give more detailed knowledge of this subject, and this author strongly recommends that the reader refer to them. Some of the references at the end of this article will be of great value in this regard. In the two succeeding sections, we present examples of the way an artificial network may be used for predictive purposes. It is felt that the most efficient way to familiarize a reader with the power and beauty of artificial neural networks is through presentation of examples.

The next section concerns the use of an ANN for extrapolating heat capacity data of polymers to a temperature range where measurements are difficult. It is an example of using ANN to predict parameters outside the range of data on which it was originally trained. This section shows that predictions can be made accurately, even outside the training range of the data.

The final section illustrates the “power” of ANN in correlating vast amounts of input parameters with output parameters oriented to physical properties. The basic principles used in this section can generally be applied to a wide range of problems in material sciences for a large number of chemical and physical systems.

INTRODUCTION

The volume of information that is available for any given physical or chemical system is becoming increasingly unmanageable. We are literally in danger of being “swamped” by this profusion of available data. It is now very common that a system we are interested in studying may have a terabyte of information associated with it, that is, one or more trillion bytes of information that need to be processed. Artificial intelligence is one of the only techniques by which we have any chance of sorting through this voluminous amount of data and extracting meaningful information about the system under study, and an artificial neural network is one of the more powerful techniques of artificial intelligence.

So what exactly is an artificial neural network (ANN)? It can be defined as “a massively parallel distributed processor that has a natural propensity for storing experiential knowledge and making it available for use” (1). Figure 1 shows a typical fully connected, three-layered ANN. These networks were first developed as computer

models derived primarily from an attempt to model the human brain (2). However, it was not until the 1980s that interest in them was renewed (3). Of the various types of networks researched, probably the most widely used is called the backpropagation network. A backpropagation neural network is constructed from a simple totally connected set of neurons. Neurons are the basic element of neural networks. Each neuron’s purpose is to accumulate the sums of weighted inputs and then transfer its result of summation to the next level of neurons through an activation function. Figure 2 shows an example of a typical neuron. The most common activation function applied is commonly referred to as the sigmoid function, defined as

$$f(x) = \frac{1}{1 + \exp(-ax)}, \quad (1)$$

where a is the slope of the sigmoid function. Other activation functions sometimes used are the hyperbolic tangent and a scaled arctangent function. In theory, any bounded, differentiable nonpolynomial function can be used as a transfer function. The important feature of these functions, however, is their nonlinearity. It is actually this nonlinearity of these functions that gives rise to the “learning” capabilities of artificial neural networks. For information on other transfer functions which have been suggested, see (4) and references within.

Each neuron of each layer is connected in the forward direction, that is, from the input layer to the output layer. However, when the results of propagating the input data of a network are compared to the data provided in the output nodes, an error is calculated between the output data and the numbers calculated for each cycle. When this error is higher than a given limit, the weights between the input layer and the hidden layer and between the hidden layer and the output layer are modified, and the cycle is repeated. This feature of propagating the error backward through the network is responsible for calling the network a backpropagation neural network.

Neural networks are not programmed in the conventional sense. These networks use a large amount of data as examples and work by trial and error. This trial-and-error phase is called training. Before starting the training process, all of the weights between the various nodes are set to a small random number. Next, each training pattern is presented to the input nodes. The input from each input node is then multiplied by the appropriate weight, and the resulting number is distributed to the appropriate hidden node in the hidden layer. One number from each input node will be distributed to each hidden node, that is, for a six-input-node network such as shown in Fig. 1, each hidden node receives six weighted numbers, all of which are summed. The summed value of each node in the hidden layer is then transformed by an activation function and passed on for processing by the output layer of nodes. The nodes in the output layer process numbers that it receives

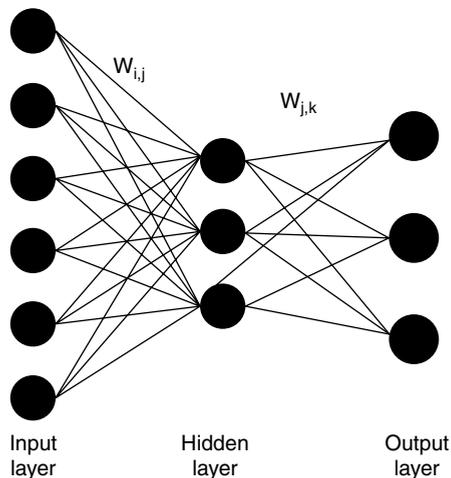


Figure 1. Typical architecture of an artificial neural network.

similarly to the numbers received by the middle or hidden layer of nodes. The processed numbers in the output layer are compared to the numbers submitted to the network for training (the answers), and an error is calculated. If the error calculated is greater than some convergence limit, the weights are changed in a prescribed manner, and the entire procedure is repeated until the error is below the set limit.

In a feed-forward network, the processing elements (PEs) are interconnected through unidirectional information channels or connections. A typical processing element is shown in Fig. 2. The output y is given by the nonlinear transformation,

$$y = f \left[\sum_{ij} w_i x_i - \theta \right], \quad (2)$$

where θ is a node bias and w_i are the connection weights. Various types of threshold functions are used in networks; the most common is the sigmoidal function shown in Eq. (1). This satisfies the nonlinear differential equation 3,

$$\frac{df}{dz} = \beta f(1 - f). \quad (3)$$

Thus, the relationship between two sets of information can be approximated, in general, by any degree of nonlinearity. The key to developing a neural network to model this

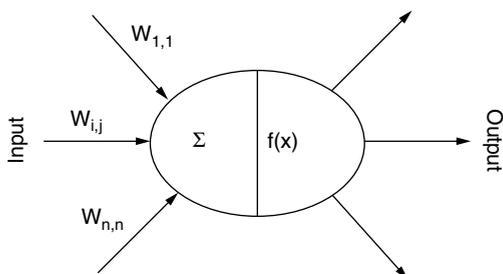


Figure 2. A typical processing element of a neural network.

relationship is to train the network by using as many examples as possible. An artificial neural network should not simply memorize patterns but should generate an implicit functional relationship between “input-output” data. This unique property of neural networks has been exploited in this article.

Most neural networks use the training of a three-layer, feed-forward, and fully connected network incorporating the backpropagation algorithm first developed by Rumelhart and McClelland (7). In experiments using three- and four-layer perceptrons, there were indications that a three-layer perceptron was more than adequate for interpolative problems. This also indicated that the mapping interfaces were fairly regular. The backpropagation network algorithm was an iterative gradient technique that minimizes the global error between the actual output and the network output. The algorithm required a continuously differentiable nonlinear function, such as the sigmoidal function [see Eq. (1)]. The “error function” δ_j^p in the computation of connection weights between hidden layers is given by Eq. (4):

$$\delta_j^p = x_j^p(1 - x_j^p)(t_j - x_j^p) \quad (4)$$

Note that δ_j^p at lower level p is a function of δ_j^{p+1} corresponding to the error in layer $p + 1$. Because of this property, the algorithm is called the backpropagation network. t_j is the target value of the j th output node. The output of node j in layer p is given by

$$x_j^p = \left[1 + \exp \left(- \sum_i w_{ij} x_i^{p-1} + \theta_j^p \right) \right]^{-1}. \quad (5)$$

The iterative procedure continues until the global error is less than a set tolerance. The evolution of the correction weights w_{ij} during training or their pattern for a trained network is not studied in this article. This is a very important aspect of quantifying network performance. For more detail, see (8–10) and references therein.

SELECTING INPUT DATA

The most effective input data for training an ANN are those data that span the space from which predictions are to be made. These are not the only data, however, that can be used. An ANN also work for prediction outside the data space. However, it usually requires much more data for this purpose. For example, if one wishes to predict the molecular weight of a group of molecules based on one or more of the physical properties of these molecules such as boiling point or density, then the values used as inputs should be only for molecules whose molecular weights are between the highest and lowest molecular weights to be predicted. Let us say that you have molecules whose molecular weights are between 50 and 250 U. You should train only on the physical properties corresponding to molecules whose masses are between 50 to 250 U. If you train on properties of molecules corresponding to this range of molecular weights and then try to predict the molecular weight of a

molecule of a mass of, say, 350 or 400 U from its corresponding physical properties, your prediction will probably not be very accurate. To make this prediction, you need to have some examples of data of physical properties from molecules whose masses are between 350 or 400 U. However, this is not the whole story. It is possible to predict outside the range of values to some extent. It all depends on how much data is available and how much of a "connection" exists between the input data and output values. If the data are sufficient for the neural network to make a strong connection, then it can predict outside the training range. If the connection is weak, then the prediction will not be good.

PREPARATION OF DATA

Now that you have selected your training set of data, you need to process your data into a format compatible with your neural network program. Many programs available handle this function automatically, but there will be times when you may need or want to process the data yourself. As previously discussed, there are several transformative functions that are applied in ANNs. The most common transfer function is the sigmoid function shown in Eq. (1). This function requires that all of the input data be in the range of zero to one. For practical reasons, most data is scaled between 0.1 and 0.9. An example of an equation for scaling data is given by

$$x'_i = [a(x_i - x_{\min}) / (x_{\max} - x_{\min})] + b \quad (6)$$

where x'_i is the scaled input data, x_i is the original input data, x_{\min} is the minimum value of the input data, x_{\max} is the maximum value of the input data, and a and b are constants allowing fixed limits of the scaled data. For example, if $a = 0.90$ and $b = 0.05$, the scaled values of the input data will range from 0.05 to 0.95.

NUMBER OF NEURONS IN HIDDEN LAYER(S)

There are probably as many ways of selecting the number of nodes in a hidden layer as there are people who use ANNs. No hard-and-fast rules can be used to choose either the number of hidden layers or the number of nodes in the hidden layer or layers. For almost all purposes, one hidden layer should be sufficient. Jack Aurada however has pointed out in his book, *Introduction to Artificial Neural Systems*, that there sometimes can be advantages of a two hidden layer architecture from the viewpoint of efficiency (11). The convergence in the error "with two hidden layers yielded superior results" compared to a similar network that had only one hidden layer and the same total number of hidden nodes (11). Several authors have proposed a rule of thumb for selecting the number of hidden nodes in a three-layer system (12,13). However, probably, the procedure that is most commonly used is either a bottom-up or a top-down trial-and-error procedure (14). The top-down approach starts with too many neurons and removes neurons until the network can no longer converge in a reasonable time. The bottom-up approach starts with too few neurons

and add neurons as necessary until the network converges in a reasonable time. Most of the time, it is a compromise between the time it takes to converge and the number of neurons in the hidden layer. The one condition you must avoid is not to have too many neurons in the hidden layer. This could lead the network to "memorize" the results of the learning process and not generalize. This is sometimes called overfitting.

How do you know how many nodes to start with in the hidden layer? Some authors propose some type of empirical rule. For example, Eberhart and Dobbins found that a reasonable number of nodes in the hidden layer at starting in a three-layer network is the square root of the number of input plus output nodes, plus a few (15). Leigh uses a procedure that 'dictates that the number of hidden nodes in one hidden layer network should not exceed 30% of the size of the input layer (16). Another empirical rule proposed by Weiss and Kulikowski uses the procedure that "a network should average at least ten samples per weight in the network"(17).

SUPERVISED VERSUS UNSUPERVISED NEURAL NETWORKS

So far, we have discussed only supervised ANNs. To summarize, these networks "learn" to make associations between input data and output parameters (e.g., physical properties, desired characteristic) through training. This training is performed by presenting numerous examples to the network, and the ANN makes numerous adjustments of its weight space until a given set of input data can elicit a correct output within some predetermined error limit.

There is another major type of ANN called an unsupervised ANN. An unsupervised ANN, sometimes called a self-organized ANN, perform its function by doing a multidimensional clustering of the input data. What typically happens is that the input data is presented to the network and the network adjusts its internal parameters so that similar inputs produce similar outputs. In essence, different clusterings of data inputs are mapped into different "regions" of a multidimensional space that allow quick and easy analysis of very large and seemingly unrelated amounts of data. Some of the more notable supervised networks are based on adaptive resonance theory. Examples of programs based on this theory are ART1, ART2, AR2, Fuzzy ART, etc. Other programs based on the unsupervised model are Hopfield nets, bidirectional associative memory, Kohonen self-organizing maps, fuzzy associative memory, and many others. For additional details on unsupervised ANNs, see (18) and references therein. For a list of some additional programs that use supervised or unsupervised algorithm, see Table 1 of (19).

ARTIFICIAL NEURAL NETWORK EXTRAPOLATIONS OF HEAT CAPACITIES OF POLYMERIC MATERIALS TO VERY LOW TEMPERATURES

As a first example of applying an ANN to materials, we present the use of neural network computing to predict the temperature dependence of heat capacities C_p , in the

Table 1. Polymers Used in Training Heat Capacities to Very Low Temperatures

Nylon 6,6	Poly(ethylene terephthalate)
Poly(oxymethylene)	Poly(<i>p</i> -methacryloyloxybenzoic acid)
Poly(propylene)	Poly(<i>p</i> -hydroxybenzoic acid)
Poly(ϵ -caprolactone)	Poly(vinyl alcohol)
Poly(chlorotrifluoroethylene)	Poly(<i>p</i> -phenylene)
Poly(vinyl benzoate)	Poly(oxy-1,4-phenylene)
Poly(oxyoctamethylene)	Poly(tetrafluoroethylene)
Poly(methyl methacrylate)	

temperature range 10–100 K of a series of 15 polymeric materials and relate these values to the temperature dependence of C_p in the temperature range 100–360 K (20). In this way, newly measured values of the heat capacity C_p in the higher temperature range (100–360 K) can be extrapolated to the low-temperature domain where measurements are difficult. Note that this is a different use of an artificial neural network because we are predicting values of a physical property outside the range of values in which the training occurred. Remember, this is an example of an application of artificial neural networks. For more details on the results of the work being presented, the reader should refer to (21).

Introduction

The field of thermal analysis of polymers has greatly enlarged in about the last 20 years since the introduction of simple and comparatively inexpensive instrumentation. The measurement of the heat capacities of polymers is fundamental to understanding the theory of glass transition temperatures, for example, because the measure of the heat absorbed provides a direct measure of the increase in the molecular motion of molecules. This knowledge can be used to help polymer scientists understand a wide range of properties of these systems in response to temperature and temperature changes. These could include, but are not limited to, polymerization, degradation, and other chemical changes associated with temperature.

Application of Neural Network Computing to the Extrapolation of Heat Capacities

Figure 3 shows the three-layered backpropagation network used in this specific project. The input layer consist of 25 neurons, a middle or “hidden” layer of 15 neurons, and an output layer with 10 neurons; the network has been trained using the backpropagation learning algorithm (22).

As we stated in the previous section, a set of correct input/output pairs, which we call examples, was presented to the ANN considered in this work and the internal weights were adjusted to represent these examples best through the backpropagation learning algorithm. The name backpropagation is due to the procedure for updating the weights during training in which the network corrects its internal parameters by starting with the output values and propagating backward through the hidden layer to the input layer.

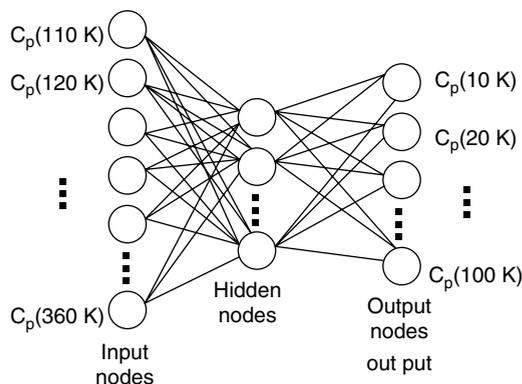


Figure 3. Neural net work used in predicting heat capacities of polymers.

Each neuron of an ANN constitutes a processing element (PE), and it is connected through various weights to other PEs. The processing element sums the product of each input, and the connection weights form the previous layer of PEs and then filter it by a nonlinear function, usually called the activation function. The form of this function is given by Eq. (1).

The weights w_{ij}^p (weighing the signal between the i th PE of layer $p - 1$ and the j th PE of layer p are initialized randomly and then adjusted during training. Training the network means establishing the values of these weights between pairs of processing elements. For a multilayer network, the output layer of the j th PE in layer p is given, in general, by

$$x_j^p = f \left[\sum_i w_{ij}^{p-1} x_i^{p-1} + \theta_j^p \right], \quad (7)$$

where the nonlinear function f has the sigmoidal form of Eq. (1) and θ_j^p is the bias of the j th PE layer of layer p . The connection weights, w_{ij}^p and the biases θ_j^p are initially uniformly distributed in the range $\{-1, +1\}$. They are updated each time an example is presented to the network during training, according to the backpropagation algorithm described in detail in (21,23). Several variations and enhancements of the backpropagation algorithm are aimed at accelerating the learning process and improving its accuracy. In the variant employed at the University of Tennessee and used for this study, the parameter β in Eq. (3) is adjusted during training. Different values for β in the range 0.1 to 0.9 are used; the one that demonstrates the fastest convergence is selected.

During training, the input pattern and a target output pattern are associated. To make the outputs as close as possible to the desired value, we minimize the sum of the squared errors

$$E(w) = 1/2 \sum (t_j - a_j)^2 \quad (8)$$

with respect to the set weights of w , where t_j is the target (i.e., the output part of an example) and a_j is the

actual value of the output layer. The local error at each mode is computed, and the connection weights are adjusted through a gradient descent algorithm to minimize the global error (20). The computational burden in applying multilayered ANNs lies primarily in network training. The weights of the network are fixed after training is completed. These values are then used during "recall" sessions, when an unknown input is presented to the network and it computes the appropriate network.

RESULTS

In our calculations, we trained the network by using input-output pairs for values of C_p versus temperatures from 10–360 K in increments of 10 K for polymers taken from the ATHAS data bank of heat capacities of more than 100 macromolecules (24). The input layer of our network consisted of 25 neurons, a middle layer of 15 neurons, and an output layer of 10 neurons. These heat capacities are based on experimental data, fitted to approximate vibrational frequency spectra. The computed heat capacities are smooth functions of temperature and contain no statistical errors. The input consisted of C_p in the temperature range of 110–360 K; using backpropagation, the output was C_p for $T = 10, 20, \dots, 100$ K. The 15 polymers used in training are given in Table 1. It can be seen that the input data covered a wide variety of polymers.

All calculations were implemented on a VAX 2000 workstation. After the network was trained, five additional polymers were chosen to test the ability of the network to predict the low-temperature C_p from the high-temperature C_p . The polymers chosen for our test cases were PEN [poly(ethylene 2,6-naphthalenedicarboxylate)], PE (polyethylene), PVF [poly(vinyl fluoride)], PS (polystyrene) and nylon 6 [poly(ϵ -caprolactam)]. From Table 2, it can be seen that typical errors ranged from 0.80% to 0.01%, and the average error was 0.34%. Errors were calculated as the difference between the "expected value" and the value calculated by the neural network.

Errors of this order of magnitude are far less than experimental errors that are typically 1–5%; the larger errors occur at lower temperatures where heat capacities are of low magnitude. Knowing this, the heat capacity in the easily accessible temperature range from 100–360 K permits reasonably precise extrapolation to low temperatures. The low-temperature data are needed to evaluate the total heat content of the polymeric materials.

CONCLUSIONS

This example shows the feasibility of using neural networks for extrapolating heat capacities of polymers in the low-temperature region. The exciting possibility of this method is that by "teaching" a neural network a sufficient quantity of known experimental data, an excellent extrapolation to unknown data may be possible. This procedure can also be applied to a wide range of properties of many materials, including but not limited to polymeric materials, as demonstrated in this section. For additional examples

Table 2. Examples of Measured and Predicted Heat Capacities

Polymers	Temperature (K)	Predicted Value C_p	Measured Value C_p	Error (%)
Run 1: [poly(ethylene 2,6-naphthalenedicarboxylate)]				
	10	5.1826	5.1987	-0.31
	20	19.7206	19.8736	-0.77
	30	32.5057	32.7414	-0.72
	40	43.1375	43.3324	-0.45
	50	52.8020	52.8549	-0.10
	60	62.1784	62.0543	+0.20
	70	71.5633	71.2073	+0.50
	80	80.9630	80.4163	+0.68
	90	90.3907	89.7000	-0.78
	100	99.7745	99.0121	+0.77
Run 2: [polyethylene]				
	10	0.0995	0.1002	-0.67
	20	0.7411	0.7440	-0.39
	30	1.9143	1.9230	-0.45
	40	3.2316	3.2475	-0.49
	50	4.5225	4.5434	-0.46
	60	5.7424	5.7695	-0.47
	70	6.8786	6.9083	-0.43
	80	7.9231	7.9597	-0.46
	90	8.8866	8.9190	-0.35
	100	9.7150	9.7835	-0.70

of applications of ANN to polymeric systems, see (21, 23) and references therein.

ARTIFICIAL NEURAL NETWORK MODELING OF MONTE CARLO SIMULATED PROPERTIES OF POLYMERS

Our second example looks at a neural network used to learn features of the rotational potential energy surface of a number of polymers and relate those features to conformational properties of the polymers. Specifically, we modeled Monte Carlo simulations of 20 polymers (see Table 3) in which we calculated two statistical properties, the characteristic ratio and the temperature coefficient of the characteristic ratio of each polymer. This particular example is illustrative of any modeling study where one can relate a theoretically calculated parameter or series of parameters and a property of the materials being studied. The reader can find additional examples by this author and others in the references at the end of this article.

Table 3. Polymers Used in Training and Testing in the Monte Carlo Study

Polydimethylsiloxane (II)	Poly(vinyl alcohol) (I)
Polyethylene	Poly(vinyl alcohol) (II)
1,4- <i>cis</i> -Polyisoprene	Poly(oxyethylene)
Polyacrylonitrile	Poly(propylene)
Polydimethylsiloxane (I)	Polyurethane
Polyalanine	Polysiloxane
Polyglycine	Poly(vinyl chloride)
Polyisobutylene	1,4- <i>trans</i> -Polybutadiene
Polyleucine	1,4- <i>trans</i> -Polychloroprene a
Poly(oxyethylene)	1,4- <i>trans</i> -Polyisoprene

INTRODUCTION

Monte Carlo techniques have been used to study chemical systems for many years. It has been a particularly effective technique for studying the statistical properties of polymers. Flory used the Monte Carlo technique in effectively predicting many properties of polymers (25). Coupling neural networks (26–28) with Monte Carlo techniques provides a powerful new tool for predicting the properties of molecules. Accurate knowledge of the rotational potential energy surface is fundamental to the overall accuracy of the predictions from a Monte Carlo study. In this light, we first created 20 rotational potential energy surfaces corresponding to the 20 polymers in this study by using the semiempirical technique MNDO (modified neglect of differential overlap) and then used these energy surfaces corresponding to the 20 polymers in this study to produce 10,000 chains of each polymer, each chain 100 bonds long. From these results, we calculated the mean-square end-to-end distance, the characteristic ratio, and its corresponding temperature coefficient. A neural network was then used to model the results of these Monte Carlo calculations. We found that artificial neural network simulations were highly accurate in predicting the outcome of the Monte Carlo calculations for polymers for which it was not trained. The overall average error for predicting the temperature coefficient was 0.89%.

Previous discussions of using a neural network has focused on predicting heat capacity relationships (28,29) between the potential surface and quantum mechanical eigenvalues (30), molecular dynamic results (31,32), and frequencies computed by normal mode analysis (33).

In this project, we first created 20 rotational potential energy surfaces for 20 different polymers using the semi-quantum mechanical program MNDO (34). Next, we used these surfaces in our Monte Carlo program to calculate various average statistical properties of our 20 polymers. Last, we submitted our 144 potential energy points of each polymer to the 144 input nodes of our network and trained the network to predict the statistical properties that we obtained from our Monte Carlo calculations. Our calculation section provides more detail.

CALCULATIONS

We began our calculations by first optimizing the geometry of each of the 20 polymers using the Gaussian-92 (35) ab initio SCF-MO (self-consistent field molecular orbital) program that has a 4–31G basis set of functions. Of course, we optimized on only a fragment of each polymer, that was small enough so that we could run it on our PC computer but large enough so that end effects were small. We also made sure that the two ends of each fragment of each polymer were identical. The fragments were usually between 8 and 12 main chain bonds long. We also optimized the geometry of each molecule when the main chain was in the plane's zigzag (all main chain dihedrals were 180°) conformation.

Next, we took the optimized geometry of each polymer and calculated 144 rotational potential energy states using

the semiquantum mechanical program MNDO. This program was run on our (primitive) Zenith 386 PC that has a math coprocessor available from C-graph software, Inc. of Austin, Texas (36). Each of the 144 rotational energy states was produced by rotating most interior dihedral angles by 30 through 360°. It should be emphasized that we were not attempting to produce the most accurate rotational potential energy states. To do so would have required a much more accurate quantum mechanical program such as Gaussian 92 (12) or higher, and would have required several hundred additional CPU hours of computational time, probably using a supercomputer.

Next, we used our rotational potential energy surfaces in Monte Carlo simulations (37) to calculate several average statistical properties for each polymer. The procedure adopted was to produce a Markov chain in which each atom k along the backbone of the chain was added in a configuration dictated by its conditional probability calculated from its corresponding Boltzman factor $\exp[-E_k/(RT)]$, where E_k is the k th rotational potential energy, R is the gas constant, and T is the temperature in degrees K. The properties calculated were the mean-square end-to-end distance from which we calculated the characteristic ratio and the temperature coefficient of the characteristic ratio. We took the last two numbers and used them as the output parameters to train two neural networks. The first network used the 144 rotational potential energy points of each polymer as input and the mean-square end-to-end distance as output (see Fig. 4). The second network was used to train the temperature coefficient of the characteristic ratio. This network used eight input nodes corresponding to four temperatures and four mean-square end-to-end distances, and one output node corresponding to the temperature coefficient of the characteristic ratio (see Fig. 5). For more detail on the way the neural network program worked and how neural networks work in general, see (25–27, 38) and the references therein.

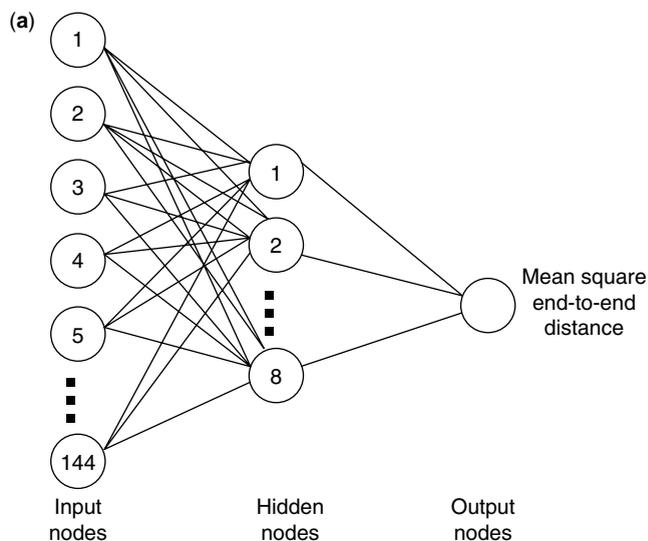


Figure 4. Architecture of neural network for calculating the mean-square end-to-end distance of polymers.

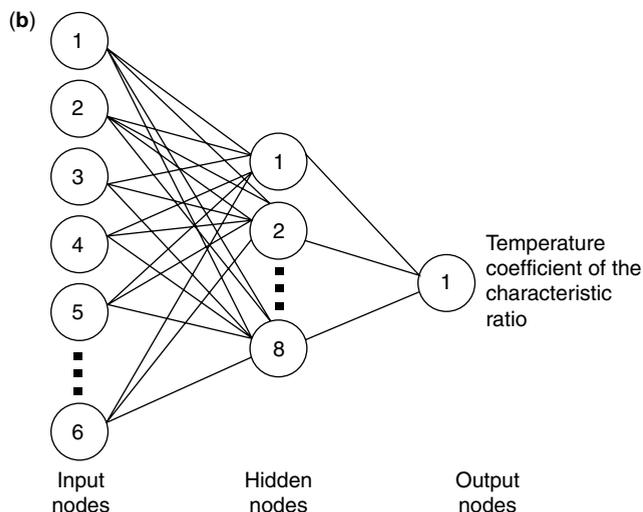


Figure 5. Architecture of neural network for calculating the temperature coefficient of the characteristic ratio.

RESULTS

Table 3 lists the polymers used in this study for training and testing. Note that two of the polymers have two different isomers, I and II. These correspond to rotations about two different consecutive bonds in these two polymers. We calculated a total of 20 polymers, trained on 16, and tested on four. Table 4 presents the results for seven runs in which we randomly chose four polymers from the list of 20 in Table 3 to test our training. Each run represents a different set of four polymers. As can be seen, the relative error ranged from a high of 15.4% to a low of 0.33%, and the overall average error in the prediction of the characteristic ratio (C.R.) was 4.82%. We feel that this error was very acceptable considering that most experimental error for measuring the characteristic ratio is about 5% or higher. It should also be mentioned that no attempt

Table 5. Calculated Characteristic Ratios (C.R.) for 20 Polymers at Several Temperatures

	Polymer	T(K)	C.R.
Run 1	Polyleucine	225	8.8772
		300	7.2709
		375	6.1158
		450	5.2585
Run 2	Polyacrylonitrile	225	12.0099
		300	8.2179
		375	6.3012
		450	5.3106
Run 3	Polyglycine	225	10.4083
		300	6.7474
		375	5.2177
		450	4.4124
Run 4	Polyisobutylene	225	6.3543
		300	4.0702
		375	3.3472
		450	3.0631

was made to reproduce the experimental characteristic ratio. The purpose of this work was to determine whether a characteristic ratio could be accurately predicted from a given rotational potential energy surface, whether or not that surface could predict the experimental characteristic ratio and/or temperature coefficient. However, if the rotational potential energy surface was an accurate representation of the molecule, then this procedure should accurately predict the various macromolecular properties. We feel that the results of this study verify this conjecture.

Table 5 lists the characteristic ratios of various polymers calculated at several temperatures. The training and testing of the characteristic ratio (C.R.) for those numbers were generated at 300 K. The C.R. was also calculated at 255, 325, and 450 K. From this data, we calculated the temperature coefficient of the characteristic ratio in Table 6. This number was found by determining $d[\ln(r_{0,n}^2)]/dT$

Table 4. The Predicted Characteristic Ratio (C.R.) for Selected Polymers at 300 K

Polymer	Calculated C.R.	Predicted C.R.	Relative % Error
Run 1:			
Polyalanine	3.6713	3.7252	1.47
Poly(vinyl chloride)	4.2046	4.1587	1.09
Poly(vinyl alcohol) (I)	7.3291	7.6339	4.16
1,4- <i>cis</i> -Polyisoprene	6.6070	7.2346	9.50
Run 2:			
Polyalanine	3.6713	3.6489	0.61
Polyglycine	6.7474	5.7083	15.4
Polyleucine	7.2709	7.1342	1.88
1,4- <i>trans</i> -Polybutadiene	4.2144	4.3332	2.82
Run 3:			
Poly(oxyethylene)	3.7047	3.1664	14.53
Polydimethylsiloxane (I)	3.9599	4.1737	5.40
Polyleucine	7.2709	7.3683	1.34
Poly(oxyethylene)	7.5822	7.6929	1.46

Table 6. Predicted Temperature Coefficients for Selected Polymers at 300 K

Polymer Coeff. $\times 10^3$ (K^{-1})	Calc. Temp. Coeff. $\times 10^3$ (K^{-1})	Predicted Temp., % Error	Relative Error
Run 1:			
Poly(oxyethylene)	-1.60	-1.6022	0.14
Polyleucine	-2.30	-2.3013	0.06
Poly(oxyethylene)	-2.61	-2.6026	0.28
Polydimethylsiloxane	-1.90	-1.8962	0.20
Run 2:			
Poly(propylene)	-4.24	-4.1577	1.94
Poly(vinyl alcohol) (II)	-0.91	-0.9149	0.54
Polydimethylsiloxane (II)	1.13	1.0930	3.27
Polyethylene	-2.36	-2.3531	0.29
Run 3:			
Polyacrylonitrile	-3.62	-3.6442	0.67
Polyurethane	-1.46	-1.4602	0.02
Polyisobutylene	-3.17	-3.1113	1.85
Polyglycine	-3.79	-3.7687	0.56

where n , the number of main chain bonds, was 101 (39). As for the characteristic ratio, there was no attempt to obtain an accurate reproduction of the experimental temperature coefficient by training it on a series of characteristic ratios. As can be seen from Table 6, the predictions were exceptionally good. The errors ranged from 0.02% to 3.27%, and the average error was only 0.89%. Noted that the network was trained on the characteristic ratio, not $\langle r^2 \rangle$ or $\ln \langle r^2 \rangle$. The network was able very quickly to "learn" the correct functional association.

CONCLUSIONS

In this study, we demonstrated the feasibility of using artificial neural network procedures to model macromolecular properties of polymeric systems. Although the results were not perfect, this author feels they are very satisfactory. The maximum error for predicting the characteristic ratio from its corresponding rotational potential energy surface was about 15%, and the overall average error, however, was only 5%. Using more polymers in the training, this error would be reduced. The temperature coefficient, however, was substantially better. Its maximum error was only about 3%, and the average error was less than 1%. It should be emphasized that we were not attempting to produce the most accurate experimental characteristic ratios or temperature coefficients. To do so would have required producing much more accurate rotational potential energy surfaces for the 20 polymers, and this would have required a much more powerful quantum mechanical program than MNDO and an immense amount of computer time. The major emphasis was to see whether, given a potential energy surface, one can predict the corresponding characteristic ratio and temperature coefficient by using a trained artificial neural network.

The advantage of using neural networks rather than other more traditional techniques is that once you properly train your network, it requires no more than one second or less of computer time to obtain results for other polymers for which the network has not been trained. If the original

choice of polymers was sufficiently diverse and the data set for a trained neural network was sufficiently accurate, predictions of any properties for which you trained should be very good. Therefore, this could reduce or eliminate the costly and time-consuming Monte Carlo simulations (or other simulations such as molecular dynamics) and, therefore, greatly reduce the time and cost of computations.

SUMMARY

We have seen how artificial neural networks provide a very powerful tool to material scientists. The training phase of these networks may be very time-intensive, but the predictive powers are immense. This article only scratches the surface of ANN. There are numerous journal articles as well as books that may be taped to teach the reader more about this technique as well as many other techniques involved in the general field called artificial intelligence. There are also numerous Internet sites where additional information can be obtained. Some of these useful sites where the reader can find additional information on neural networks, as well as on other artificial intelligence procedures, in addition to some free software, are listed here:

<http://www.neurodynamx.com>
<http://www.zsolutions.com/index.htm>
<http://www.stats.ox.ac.uk>
<http://www.ncrg.aston.ac.uk/Welcome.html>
<http://nastol.astro.lu.se/~henrik/neuralnet1.html>
<http://cns-web.bu.edu/pub/snorrason/bookmarks/neural.html>
<http://www.neci.nj.nec.com>
<http://chemeng.iisc.ernet.in/gopa/neural/neural.html>
<http://www.spss.com/software/neuro/>
<http://www.tropheus.demon.co.uk/>
<http://www.calsci.com/index.html>
<http://www.wardsystems.com/>
 and many, many more . . .

ACKNOWLEDGEMENT

The author gratefully acknowledges Dr. Don W. Noid, of Oak Ridge National Laboratory, who was instrumental in involving the author in this research area. Some of the material in the tutorial was also discussed in a paper previously published in (8), coauthored by B. R. Upadhyaya. In addition, Dr. Bobby Sumpter was greatly involved in many of the research projects involving artificial neural networks. Some of the research results presented, in part, in this work were sponsored by the Division of Materials Sciences, Office of Basic Energy Sciences, U.S. Department of Energy, under contract DE-ACO5-84OR21400 with Martin Marietta Energy Systems, Inc., and the materials Research Division of the NSF, Grant #DMR-8818412.

BIBLIOGRAPHY

1. S. Haykin, *Neural Networks: A Comprehensive Foundation*. Macmillan, Englewood Cliffs, NJ (1994).
2. T. Kohonen, *Neural Networks 1*: 3 (1988).
3. J.J. Hopfield, *Proc. Natl. Acad. Sci. USA* **79**: 2554 (1982).
4. S.K. Kenue, *Proc. Intell. Robots Comput. Vision X: Neural Biol. 3-D Methods* SPIE, Bellingham, WA, (1991).
5. R.P. Lippmann, *IEEE ASSP Mag.* **42**: 4 (1987).
6. W.S. McCulloch and W. Pitts, *Bull. Math. Biophys.* **5**: 115 (1943).
7. D. Rumelhart and J. McClelland, *Parallel Distributed Processing*. Bradford Books MIT Press, Cambridge, MA, (1986), Vol. 1.
8. J.A. Darsey, D.W. Noid, and B.R. Upadhyaya, *Chem. Phys. Lett.* **177**: 189 (1991).
9. M. Caudill, *AI Expert* p. 46 (December, 1987).
10. M. Caudill, *AI Expert* p. 55 (February, 1988).
11. J. Aurada, *Introduction to Artificial Neural Systems*. West, St. Paul, MN, 1992.
12. B. Amirikian and H. Nishimura, *Neural Network 7*: 321 (1994).
13. T.M. Nabhan and A.Y. Zomaya, *Neural Network 7*: 89 (1994).
14. J. Devillers, Ed., *Neural Networks in QSAR and Drug Design*. Academic Press, NY, 1996.
15. R.C. Eberhart and R.W. Dobbins, *Neural Network PC Tools. A Practical Guide*. Academic Press, San Diego, 1990, p. 414.
16. D. Leigh, in *Intelligent Systems for Finance and Business*, S. Goonatilake and P. Treleaven, eds., J Wiley, NY, 1995, pp. 61–69.
17. S.M. Weiss and C.A. Kulikowski, *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. Morgan Kaufmann, San Francisco, 1991, p. 104.
18. H. Ritter, T. Martinetz, and S. Klaus, *Neural Computation and Self-Organizing Maps*. Addison-Wesley, NY, 1992.
19. B. Sumpter, C. Getino, and D. Noid, *Annu. Rev. Phys. Chem.* **45**: (1994).
20. For a description, see B. Wunderlich, *Thermal Analysis*. Academic Press, NY, 1990.
21. J.A. Darsey, D.W. Noid, B. Wunderlich, and L. Tsoukalas. *Makromol. Chem. Rapid Commun.* **12**: 325 (1991).
22. D. Rumelhart and J. McClelland, *Parallel Distribution Processing*. Bradford Books/MIT Press, Cambridge, MA 1986, Vol. 1.
23. D.W. Noid and J.A. Darsey, *Comput. Polym. Sci.* **1**: 157 (1991).
24. S.F. Lau and B. Wunderlich, *J. Thermal Anal.* **59**: 28 (1983).
25. P.J. Flory, *Statistical Mechanics of Chain Molecules*. Wiley-Interscience, NY (1969).
26. P.D. Wasserman, *Neural Computing: Theory and Practice*. Van Nostrand Reinhold, NY, 1989.
27. J.M. Zurada, *Introduction to Artificial Neural Systems*. Weit, NY, 1992.
28. J.A. Darsey, A.G. Soman, and D.W. Noid, *Makromol. Chem., Theory Simulation 2*: 711 (1993).
29. D.W. Noid, M. Varma-Noir, B. Wunderlich, and J.A. Darsey, *J. Thermal Anal.* **37**: 2295 (1991).
30. J.A. Darsey, D.W. Noid and B.R. Upadhyaya, *Chem. Phys. Lett.* **177**: 189 (1991).
31. D.W. Noid and J.A. Darsey, *Comput. Polym. Sci.* **1**: 157 (1991).
32. B.G. Sumpter, C. Getino, and D.W. Noid, *J. Phys. Lett.* **96**: 2761 (1992).
33. B.G. Sumpter and D.W. Noid, *Chem. Phys. Lett.* **192**: 455 (1992).
34. M.J.S. Dewar and W. Thiel, *J. Am. Chem. Soc.* **99**: 4907 (1977).
35. *Gaussian 92*, Revision, B.M.J. Frisch. G.W. Trucks, M. Head-Gordon, P.M.W. Gill, M.W. Wong, J.B. Foresman, B.G. Johnson, H.B. Schlegel, M.A. Robb, E.S. Replogle, R. Gomperts, J.L. Andres, K. Raghavachari, J.S. Binkley, C. Gonzalez, R.L. Martin, D.J. Fox, D.J. Defrees, J. Baker, J.J. P. Stewart, and J.A. Pople. Gaussian, Inc., Pittsburgh, PA, 1992.
36. For a copy of software, contact C-Graph Software, Inc., P.O. Box 5641, Austin, TX 78763.
37. A. Baumgartner, in *Applications of the Monte Carlo Method: Topics in Current Physics*, K. Binder, ed., Springer-Verlag, NY, 1987.
38. NNets Program. P.T. Baffes. Johnson Space Center Report. No. 23366, Sept. 1989.
39. J.E. Mark, A. Einsberg, W.W. Graessley, L. Mandelkern, and J.L. Koenig. *Properties of Polymers*. American Chemical Society. Washington DC, 1984, pp. 18–21.

NONDESTRUCTIVE EVALUATION

MAHESH C. BHARDWAJ
SecondWave Systems
Boalsburg, PA

INTRODUCTION

Ultrasound is widely used in health care for noninvasive diagnostics and in industry for nondestructive testing. In the human body, it generates visual images from inside the test medium: the fetus, malignant tissue, stones, etc. In industrial applications, besides defect detection, ultrasound is also useful for determining significant material characteristics such as density, thickness, mechanical properties, and level sensing. Knowledge of ultrasonically analyzed information is important for human health as well as for cost-effective production of quality industrial materials.

Ultrasound operates on the same principle as other characterization methods also based on wave-material interactive phenomena, such as optical, X ray, infrared,

Table 1. Categories of Ultrasonic Measurements and Their Applications

Measurement Category	Measured Parameters	Applications
Time domain	Times-of-flight and velocities of longitudinal, shear, and surface waves	Density, thickness, defect detection, elastic and mechanical properties, interface analysis, anisotropy, proximity and dimensional analysis, robotics, remote sensing, etc.
Attenuation domain	Fluctuations in reflected and transmitted signals at a given frequency and beam size	Defect characterization, surface and internal microstructure, interface analysis, etc.
Frequency domain	Frequency dependence of ultrasound attenuation, or ultrasonic spectroscopy	Microstructure, grain size, grain boundary relationships, porosity, surface characterization, phase analysis, etc.
Image domain	Time-of-flight, velocity, and attenuation mapping as functions of discrete point analysis by raster C-scanning or by synthetic aperture techniques	Surface and internal imaging of defects, microstructure, density, velocity, mechanical properties, true 2-D and 3-D imaging.

Raman spectroscopy; nuclear magnetic resonance; and neutron, γ -ray, and mass spectrometry. By propagating a wave in a given medium, useful information about the medium can be generated by analyzing the transmitted or reflected signals. Ultrasound differs from other wave-based methods because it does not require sample preparation, is nonhazardous, and provides the means to determine mechanical properties, microstructure, imaging, and microscopy. Ultrasonic equipment is also portable and cost-effective. Most significantly, ultrasound is applicable to all states of matter, except plasma and vacuum. Furthermore, propagation of ultrasound in a material is not affected by its transparency or opacity. Table 1 provides a comprehensive introduction to ultrasound measurements and to the information revealed either directly or through correlation with measurements.

Since about 1980, both ultrasound and its applications have grown substantially. Uses in industry have gone beyond overt defect detection in metals to include characterization of elastic and mechanical properties; delaminations in multilayered, particulate and fibrous materials; proximity and dimensional analysis; measurements of anisotropy and heterogeneity; surface profiling, chemical corrosion, crystallization and polymerization; liquid and gas flow metering; imaging of surface and internal features of materials; viscosity of liquids; texture and microstructure of granular and cellular materials; applied and residual stresses; high temperature, pressure, and radiation environment applications; and robotics, artificial intelligence. These highly desirable applications have attracted the attention of a wide range of industries: structural and electronic materials and components manufacturers, aircraft and aerospace, chemical and petroleum, plastics and composites, lumber and construction, highways and aircraft landing strips, bridges and railroads, rubber and tire, food, and pharmaceutical products.

In medical diagnostics where the sophistication of ultrasound is more advanced than in industry, it can replace harmful X rays in many critical instances. Ultrasound is useful for visualizing a fetus, measuring the cornea, tissue

characterization, imaging of plaque in arteries and gum disease, brain wave measurements, monitoring of the heart beat, skin and breast cancer detection, blood flow metering, etc.

In 1980, we were content if ultrasound could detect a 1-mm defect and 0.5-mm resolution in a given test material. Today, after much R&D, we have developed short-pulse and high-frequency transducers by using advanced electronics and signal processing to the point that we can measure resolution and detectability in the micrometer range. Obviously, ultrasound has come a long way since the discovery of piezoelectricity by Pierre and Jacques Curie in 1876 (1) and its first application by Richardson in 1913 for sonar (2).

Ultrasound and its applications have grown phenomenally in recent years, but the mode by which it is propagated in a given test medium is severely limited. Due to extremely high attenuation of ultrasound by air, its transmission in a test medium is done by physically contacting (coupling) the transducer to the test medium. Therefore, all conventional ultrasound has the severe limitation of physical contact between the transducer and the test medium by a liquid gel (3). If this contact could be eliminated, then we could diagnose burnt or malignant skin damage without discomfort to the patient. Similarly, a number of industrial materials sensitive to liquid contact could be tested to measure thickness, density, mechanical properties, defect detection, etc. This is significant in ensuring materials quality and process control and for cost-effective production. The development of a noncontact ultrasound (NCU) mode would allow many more useful applications of ultrasound. For example, using NCU, characteristics of materials that are porous and hygroscopic could be determined. Similarly, materials in the early stages of formation (uncured plastics, green ceramics, and powder metals) and those that are continuously rolled on a production line (plastics, rubber, paper, construction, and lumber) could also be tested under manufacturing conditions. NCU could also be applied to medical problems where contact with a patient can be harmful, as in the evaluating wounds and diagnosis of the eye.

However, for NCU to become a reality, we first need transducers and electronic systems sensitive enough to transmit and detect ultrasound without contact with the test medium. And herein lies a big problem. Conventional wisdom stipulates that ultrasound (from ~ 200 kHz to >5 MHz) cannot be propagated through solids or liquids without physical contact between the transducer and the test medium. Therefore, NCU has been generally considered an impossible dream due to the phenomenal mismatch of acoustic impedance between the coupling air and the test media. This mismatch can run as high as six orders of magnitude when we consider propagation of ultrasound from air to materials such as steels, superhard alloys and dense ceramics, cermets, diamond, and diamond-like materials. To realize the NCU mode, this barrier of acoustic impedance must be broken. And for this to happen, it is imperative that ultrasonic transducers be characterized by phenomenally high sensitivity. Achieving NCU is analogous to “*throwing a helium-filled rubber balloon so that it can pierce a stainless steel wall.*”

REALITY THAT DEFIES NONCONTACT ULTRASOUND

The exorbitant mismatch between the acoustic impedance of the coupling medium, air, and that of the test material generates enormous resistance to ultrasound propagation in materials. This, in conjunction with the extremely high attenuation of ultrasound (in MHz region) by air, further compounds the problem of the NCU mode. In simple terms, when ultrasound travels from a medium of low acoustic impedance to one of high acoustic impedance, only a fraction of the energy is transmitted in the latter. The fraction of ultrasound transmission and energy transferred at the air-material interface is given by

$$T \text{ (Transmission coefficient in the propagation medium)} \\ = 4Z_1Z_2/(Z_1 + Z_2)^2 \quad (1)$$

where Z_1 is the acoustic impedance of the ultrasound carrier medium (for example, air for the NCU mode) and Z_2 is the acoustic impedance of the test medium. The transmission coefficient is derived as the ratio of transmitted acoustical energy V (measured in volts) and the input

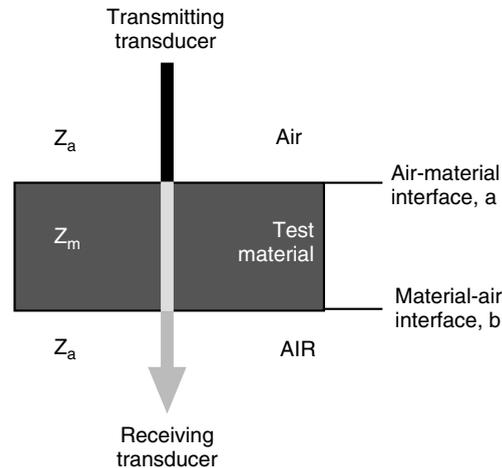


Figure 1. Interfaces to be crossed by ultrasound (shown by arrow) in the noncontact transmission mode to propagate ultrasound through a test material. Interface “a” corresponds to air–material (from acoustic impedance Z_a to Z_m) transmission. Interface “b” corresponds to material–air (from acoustic impedance Z_m to Z_a) transmission

energy V_0 of a plane wave when refracted at 0° incidence on the interface between the two media:

$$T \propto V^2/V_0^2 \quad (2)$$

This relationship can also be described by a decibel scale:

$$T = 20 \log V/V_0 \text{ (dB)}, \quad (3)$$

$$\text{Energy transferred in the propagative medium} \\ = 20 \log T \text{ (dB)}. \quad (4)$$

For more details and the significance of plane wave transmission and reflection at a number of interfaces in terms of acoustical pressure and intensity, see (4).

Ultrasound in noncontact transmission must propagate from air into the test material and then again into air so that the transmitted wave can be detected by a receiving transducer (Fig. 1). Therefore, the high energy loss at the air–material interface is doubled by further loss at the material–air interface. Table 2 provides the transmission

Table 2. Transmission Coefficients and Energy Transfer in Selected Materials at Various Interfaces in the Noncontact Mode, per Fig. 1. As a Reference, Similar Data for Water Are Also Shown. Z (air) = 440 Rayl; Z (water); 1.5 Mrayl; 1 Rayl = $\text{kg/m}^2 \text{ s}$.

Material Z_m (Mrayl)	Interface (Fig. 1)	Transmission Coefficient [Eq. (1)]		Energy Transfer [Eq. (4)] (dB)		Total Energy Loss at Interfaces $a + b$ (dB)	
		In Air	In Water	In Air	In Water	In Air	In Water
Steel	Air – steel, a	0.000034	0.11	–89	–19	178	38
51.0	Steel – air, b	0.000034	0.11	–89	–19		
Aluminum	Air – aluminum, a	0.0001	0.3	–79	–10	158	20
17.0	Aluminum – air, b	0.0001	0.3	–79	–10		
Acrylic	Air – acrylic, a	0.0005	0.84	–66	–1.5	132	3
3.5	Acrylic – air, b	0.0005	0.84	–66	–1.5		
Silicone	Air – rubber, a	0.018	0.96	–35	–0.35	70	<1
rubber	Rubber – air, b	0.018	0.96	–35	–0.35		
1.0							

coefficients and energy losses in selected test materials in the noncontact ultrasound mode calculated by using Eqs. (1) and (4). As a reference, similar losses for water immersion (contact technique) are also provided. The following conclusions can be drawn from the data:

1. Transmission losses decrease as the acoustic impedance of the test material comes within the vicinity of that of the coupling medium, whether it is air or water.
2. Total energy loss at various interfaces in the noncontact transmission mode can run as high as six orders of magnitude compared to similar losses by using water as the coupling medium.

While conducting this exercise, we did not address the issues of the interrogating frequency of the transducer and the frequency dependence of ultrasound attenuation by air. When these factors are combined with the inherent loss of ultrasound energy at various interfaces in the NCU mode, the problem of noncontact transmission in solids is only exacerbated. Relatively speaking, the attenuation of ultrasound in air is intrinsically high compared to that in solids and liquids. And because attenuation in a medium increases as a function of the fourth power of the frequency, transmission of megahertz frequencies in air becomes almost incomprehensible. To overcome these NCU-defying realities, first we need to create ultrasonic transducers that have high sensitivity (or very low insertion loss). Sensitivity is needed to overcome interfacial transmission losses (Table 2) and also to facilitate transducer excitation by relatively low power voltages. This will help avoid unwanted heating of transducers and their subsequent destruction. Once optimum sensitivity is achieved, we can increase the transducer frequency to make it comparable to that used in conventional contact testing. Accomplishing this task has captivated the imagination of materials and transducer researchers.

PURSUIT OF NONCONTACT ULTRASONIC TRANSDUCERS

A few researchers have tried to develop noncontact material characterization by using wave phenomena, which include optics, thermal, infrared, X ray, and nuclear magnetic resonance. In pursuit of bulk ultrasonic wave propagation in 1963, White (5) reported generating elastic waves in solid materials by momentarily heating a material surface. This technique eventually led to the development of the thermographic method which has been used for surface and subsurface imaging of composites, metals, etc. by sensing minute temperature fluctuations as a function of material texture, microstructure, defects, and other variables. This method has been applicable to those materials that can sustain heat or emanate heat during the testing.

Next came laser-induced ultrasound (6) that was used to characterize Rayleigh waves in metals (7) and for subsurface materials evaluation (8). The laser-based method has been applied to those materials that can withstand

Table 3. Transmission Coefficients and Energy Transferred in Air as a Function of the Final Acoustic Impedance Matching Layer on the Piezoelectric Element. As a Reference, Similar Data Are Also Shown Using Water as the Coupling Medium. Z (air): 440 Rayl; Z (water): 1.5 MRayl.

Final Layer on Piezoelectric Element Z (Mrayl)	Transmission Coefficient, T		Energy Transferred, $20 \log T$ (dB)	
	In Air	In Water	In Air	In Water
Bare piezoelectric, PZT, 31.0	0.00006	0.17	-85	-15
Hard epoxy, 4.0	0.0004	0.79	-68	-2
Silicone rubber, 1.0	0.001	0.92	-58	-0.7
Porous rubber, 0.9	0.002	0.94	-54	-0.5
^a Pressed fiber, 0.1	0.018	-	-35	-

^aWorldwide patents pending and in process.

the impact of a high-power laser beam. Laser-based ultrasound has become acceptable for high melting point metals and ceramics. The nondestructiveness of this laser-based ultrasound method is questionable when analyzing heat and shock-sensitive materials, such as polymers, green ceramics and powder metals, pharmaceutical and food products, and tissue. Ultrasound generated by electromagnetic acoustic transducers has been used in the NCU mode for nondestructive testing (9). This method is applicable only to ferromagnetic materials.

The various noncontact analytical methods outlined do provide useful information about the test materials. However, all of them are limited to specific materials and are partially destructive, complex, or expensive. The difficulty of propagating ultrasound in test media by the noncontact mode, as shown in Table 3, presents limited alternatives for achieving this mode in practical terms. These involve the techniques of ultrasound generation based upon true production of ultrasound, so that its propagation in the test medium is not affected by its (medium) exclusive properties.

Researchers and transducer experts have been designing piezoelectric devices by manipulating the acoustic impedance transitional layers in front of the piezoelectric element. In the materials industry, one of the early applications of noncontact ultrasound was testing Styrofoam blocks by using a 25-kHz frequency (10). A precursor to high-frequency, noncontact transducers was the 1982 development of piezoelectric dry coupling longitudinal and shear wave transducers up to 25-MHz frequency. Since 1983, these transducers have been commercially available for characterizing the thickness, velocity, and elastic and mechanical properties of green, porous, and dense materials (11–13). Dry coupling transducers feature a solid compliant and acoustically transparent transitional layer in front of the piezoelectric materials such as lead metaniobates and lead zirconate–lead titanate. These devices, which eliminate the use of a liquid coupling agent, do require contact with the material.

An important by-product of dry coupling devices was the development of air/gas propagation transducers, which use less than a 1-Mrayl acoustic impedance matching layer of a nonrubber material on the piezoelectric material. These

commercially available transducers have been successfully produced in planar and focused beam configurations for transmitting ultrasound in air up to ~ 5 MHz frequency and receiving up to 20 MHz. Air/gas propagation transducers, between 250 kHz to ~ 5 MHz, quickly found applications in aircraft/aerospace industries for imaging and for defect detection in fibrous, low- and high-density polymers, and composites. For such applications, these transducers have been used with high energy or tone burst excitation and high signal amplification systems. However, for applications such as level sensing and surface profiling, the low energy spike or square wave transducer excitation mechanism has been sufficient.

Similar transducers of 1-MHz and 2-MHz frequency were also produced at Stanford University by using silicone rubber as the front acoustic impedance matching layer (14). By using such a transducer at 1 MHz, the distance in air could be measured from 20 to 400 mm at an accuracy of 0.5 mm. Further improvements in transduction efficiency were shown by planting an acoustic impedance matched layer that is composed of tiny glass spheres in the matrix of silicone rubber on piezoelectric elements (15,16). Researchers at Strathclyde University (17) have reported air-coupling transducers based on piezoelectric composites between 250-kHz to 1.5-MHz frequencies. By using tone burst transducer excitation, they have been successful in producing millivolt level transmitted signals through a composite laminated honeycomb structure at 500 kHz.

More recently, piezoelectric transducers featuring perfect acoustic impedance matched layers for optimum transduction in air have been successfully developed from <100 kHz to 5 MHz (18).¹ The sensitivity of these new transducers in air is merely 30 dB lower than their conventional contact counterparts. As a result, ultrasound in the megahertz region can be easily propagated through practically any medium, including even very high acoustic impedance materials such as steel, cermets, and dense ceramics. This advancement, the major focus of this paper, is discussed in detail along with the various medical and industrial applications in the noncontact ultrasound mode.

Air-coupled transducers based upon capacitance (electrostatic) phenomena have also undergone substantial developments in recent years. Researchers at Kingston and Stanford Universities have successfully produced micro-machined capacitance air transducers; the latter claim a high 11-MHz frequency (19,20). These transducers that are characterized by high bandwidths have been used to evaluate composites and other materials. Ultrasound experts at the University of Bordeaux, have reported generating and detecting Lamb waves in the noncontact mode in anisotropic viscoelastic materials by using capacitive transducers (21).

Though much progress has been made in enhancing the transduction efficiency of transducers based on piezoelectric and capacitive phenomena, from a practical standpoint these advancements have by no means reached a saturation point. In the subsequent sections of this article, we describe the successful development of piezoelectric

transducers that are characterized by extraordinarily high sensitivities in the frequency range from 100 kHz to 5 MHz. The evidence of the high sensitivity of these new transducers can be seen from the fact that even very high frequencies such as 2–5 MHz can be propagated through a number of solids in the NCU mode.

PIEZOELECTRIC TRANSDUCERS FOR UNLIMITED NONCONTACT ULTRASONIC TESTING

The efficiency of an ultrasonic transducer depends on the coupling coefficients and other electromechanical properties of the piezoelectric material. It also depends on the mechanism by which ultrasound is transferred from the piezoelectric material to the medium in which ultrasound needs to be propagated. In the noncontact mode, this medium is air. Because the acoustic impedances of piezoelectric materials are several orders of magnitude higher than that of air, it is usually necessary to implant transitional (acoustic impedance matching) layers of various materials in front of it (the piezoelectric material). Ultimately, the characteristics of the final layer determine the transduction efficiency of a transducer device. The significance of the final acoustic impedance matching layer in the noncontact transducers cannot be overemphasized. Because the properties of a given piezoelectric material can be considered constant for a given device, the ultimate transfer of ultrasonic energy in air is entirely controlled by the acoustic characteristics of the final matching layer on the piezoelectric material (Fig. 2). To understand this, we examine the effect of the final acoustic impedance matching layer on the piezoelectric material with respect to transmitting ultrasound from it (piezoelectric element) to air, as per Eqs. (1) and (4). Table 3 shows the transmission coefficients and the transfer of ultrasonic energy in air as well as in water (as a reference) by assuming a number of final acoustic impedance matching materials on the piezoelectric material. The following conclusions can be drawn from the data:

1. The transmission coefficient in air increases as the acoustic impedance of the front layer on the piezoelectric material is reduced.

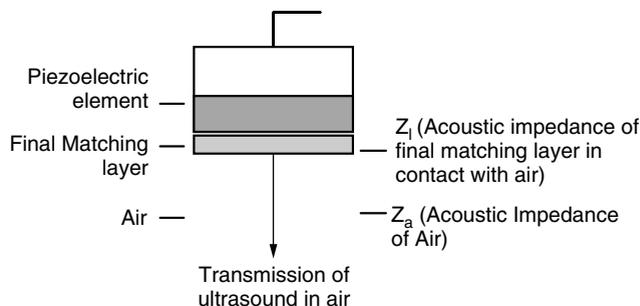


Figure 2. Schematic of an ultrasonic transducer showing the critical final acoustic impedance matching layer relative to the piezoelectric element and the coupling medium, air.

¹U.S. and international patents pending and in process.

- There is a significantly high transfer of ultrasonic energy in water compared to that in air due to better acoustic impedance matching of the final layer on the piezoelectric element with that of water.

In the light of (14–17) and this author's 1983 design (commercially available as air/gas propagation transducers from Ultrason Laboratories), the best final piezoelectric matching layer for maximum ultrasound transmission in air is composed of soft polymers. The polymer layer in these transducer designs can be porous or nonporous or can have embedded hollow spheres (in the polymer layer). For the sake of simplicity, we will identify all polymer acoustic impedance matched layer transducers as air-coupled (AC) transducers. These transducers yield a -58 to -54 dB transfer of ultrasonic energy in air, which is significant for propagating up to ~ 2 -MHz ultrasound in some solids in the noncontact mode. For example, by using a suitable transducer excitation mechanism and high received signal amplification, ultrasound can be transmitted in low acoustic impedance materials (typically materials that are lower than ~ 3 Mrayl) by using AC transducers. However, ultrasound propagation by such transducers in materials > 3 Mrayl is arduous, if not impossible.

The AC transducers based on polymer matching layer transducers do demonstrate the feasibility of noncontact ultrasound, but they are far from the most efficient. To enhance transduction efficiency in air, this author has been developing and applying a number of low acoustic impedance final matching layer materials since 1978. In 1995, we produced and evaluated (18) a transducer that had compressed fiber as the final matching layer. For the sake of simplicity and comparison, we will identify them as noncontact (NC) transducers. This transducer design exhibited unprecedented and phenomenal transduction in air which was found sufficient for NCU transmission in practically all material types. Perfected in 1997, the NC transducers can increase ultrasonic energy transfer from the transducer to air from -54 dB (AC transducers) to -35 dB; see Table 2 (18). An increment of sensitivity by an order of magnitude is extremely significant and warrants special attention. After initial trials at 200 kHz, 500 kHz, 1 MHz, and 2 MHz, NC transducers have been produced up to ~ 8 MHz. However, from a practical standpoint, it has been shown that they propagate up to 5-MHz ultrasound in nearly all material types in the NCU mode in ambient air. Leaving aside transmission in plastics and composites, the sensitivity of NC transducers is also high enough for transmission in materials that have extremely high acoustic impedance such as steel, dense ceramics, and cermets. In the following sections, we provide detailed observations about NC transducers and their sensitivity compared to AC types.

Transducer Characterization Scheme

NC transducers, like their contact or water-coupled counterparts, can also be characterized in the transmission or in the reflection (pulse-echo) modes. Figure 3 shows the setup for characterizing transducers in the transmission mode which is used to analyze NC and AC transducers. Here, the

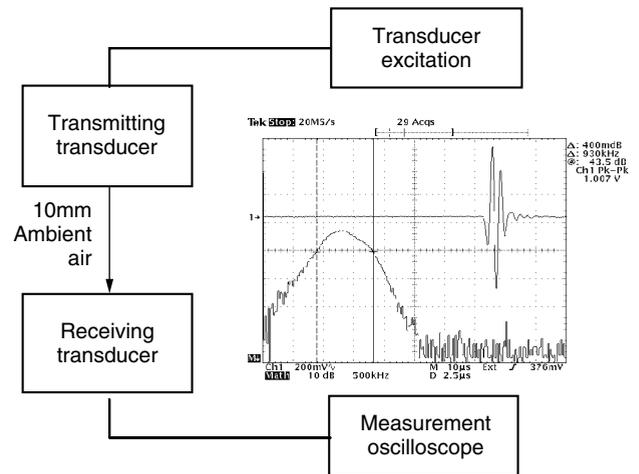


Figure 3. Noncontact ultrasonic transducer characterization scheme in transmission mode. Frequency, bandwidth, pulse width, and signal-to-noise ratio are read directly from the oscilloscope. Sensitivity = $20\log(V_x/V_0)$. V_0 = Excitation in volts. V_x = Received signal amplitude in volts.

transmitting and receiving transducers are aligned and separated by a 10-mm (or more, depending on the transducer frequency) column of ambient air. In this case, the transmitting transducer is excited by a pulse of known voltage, V_0 . The output from the receiving transducer is directly fed into a measurement oscilloscope that has a mechanism to measure the frequency domain characteristics. Frequency and bandwidth are measured directly from the frequency domain envelope, and V_x , the received signal amplitude in volts, and the pulse width are measured from the time domain RF trace. The signal-to-noise-ratio (SNR) is determined by the following relationship when measurements are made without signal averaging. It is understood that while doing so, the instrument and cable noise also factor into the measurement.

$$\text{SNR} = 20 \log V_x/V_n \text{ (dB)}, \quad (5)$$

where V_n is the amplitude of the noise in volts. Sensitivity (insertion loss) is determined by

$$S = 20 \log V_x/V_0 \text{ (dB)}. \quad (6)$$

By using this characterization scheme, several NC transducers were analyzed in ambient air. Figures 4–7 show typical time and frequency domains for 200-kHz, 1.5-MHz, 3.0-MHz, and 5.0-MHz NC transducers and their salient acoustic characteristics.

Sensitivity Comparison of NC and AC Transducers with Conventional Contact Transducers

Because sensitivity is the most critical requirement for NC or AC transducers, it is important to develop some kind of comparison scheme. To this effect we chose conventional contact, water immersion transducers as a reference. A number of NC transducers were characterized for sensitivity in the transmission mode according to Fig. 3. Similar

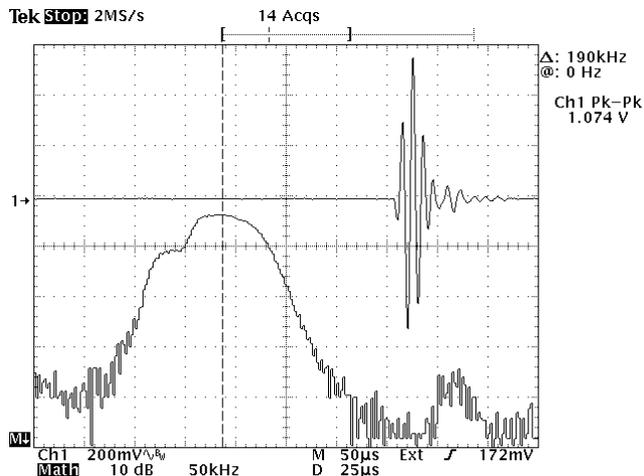


Figure 4. Time and frequency domain of broadband 200-kHz, 50-mm active area transducers in transmission mode. T-R air separation: 100 mm. Bandwidth center frequency: 200 kHz. Bandwidth at -6dB: 100 kHz (50%). Pulse width: <20 μs. Sensitivity: -46 dB. SNR: 50 dB.

transducers, suitable for conventional contact water immersion operation where acoustic impedance matches that of water, were characterized in water for sensitivity measurements. The setup for such transducer characterization is the same as in Fig. 3, except that in this case the 10-mm air column was replaced by a 10-mm column of water. Sensitivities for both transducer types, calculated according to Eq. (6), are shown in Table 4. From this comparison, it is quite clear that the sensitivity of the new noncontact transducers is approximately 30 dB below their contact counterparts. AC transducers were approximately 50 dB below their conventional contact water immersion counterparts.

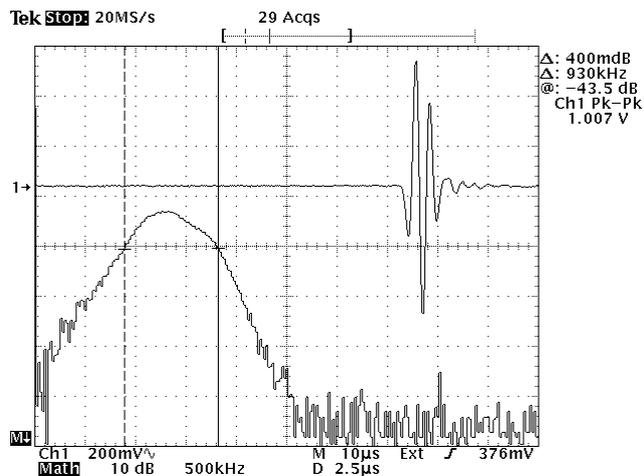


Figure 5. Time and frequency domain of broadband 1.5-MHz, 25-mm active area transducers in transmission mode. T-R air separation: 10 mm. Bandwidth center frequency: 1.4 MHz. Bandwidth at -6 dB: 0.92 MHz (65%). Pulse width: <2 μs. Sensitivity: -58 dB. SNR: 46 dB.

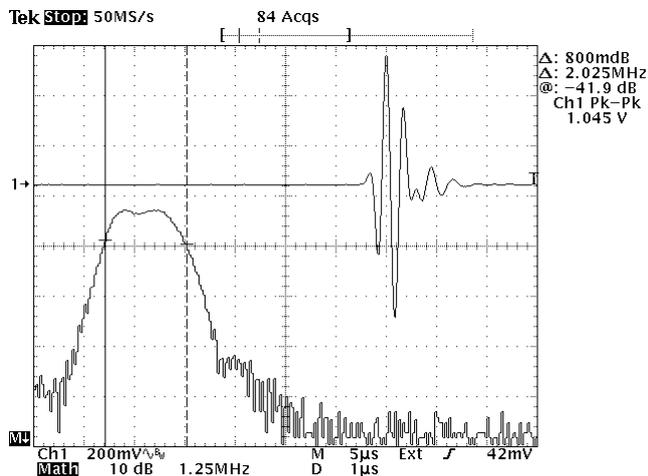


Figure 6. Time and frequency domain of broadband 3.0 MHz, 12 mm active area transducers in transmission mode. T-R air separation: 10 mm. Bandwidth center frequency: 2.6 MHz. Bandwidth at -6 dB: 2.0 MHz (75%). Pulse width: <700 ns. Sensitivity: -62 dB. SNR: 40 dB.

Application Related Experiments and Sensitivity Comparison

In the previous section, we demonstrated the high sensitivity of new NC and AC transducers by analyzing them according to a transducer characterization scheme and by comparing them to similar observations of conventional contact transducers. Although this comparison provides substantial evidence of the superiority of NC transducers over the AC types, it still does not present a convincing argument. Graphical evidence is needed to prove this point. To this effect, we performed several application-related experiments aimed at propagating ultrasound in an NCU

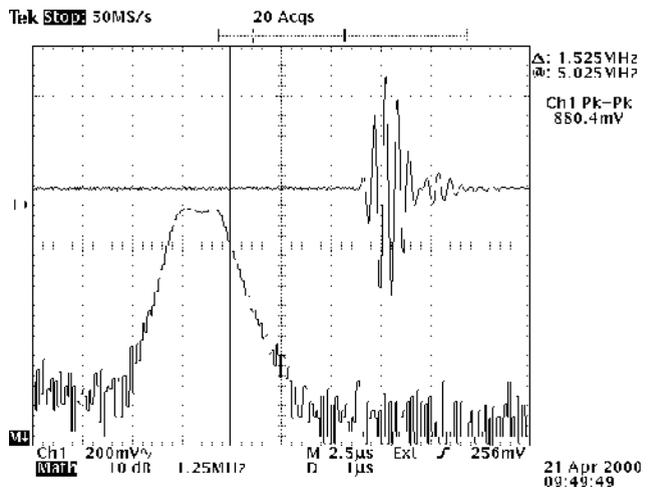


Figure 7. Time and frequency domain of narrowband 5.0 MHz, 12-mm active area transducers in transmission mode. T-R air separation: 5 mm. Bandwidth center frequency: 4.5 MHz. Bandwidth at -6 dB: 1.5 MHz (33%). Pulse width: <800 ns. Sensitivity: -68 dB. SNR: 40 dB.

Table 4. Sensitivity Comparison of Noncontact and Conventional Contact Transducers.^a Mode of Testing: Transmission. Medium of Testing: 10 mm Ambient Air for Noncontact and 10 mm Water for Contact Transducers

Frequency (MHz)	Active Diameter (mm)	Sensitivity in Ambient Air (dB)	Sensitivity w.r.t Water Immersion ^b (dB)
0.25	50	-38	Below 18
	25	-46	Below 26
0.5	50	-44	Below 24
	25	-50	Below 30
1.0	25	-52	Below 32
	19	-54	Below 34
	12.5	-56	Below 36
	3.2	-62	Below 38
1.5	12.5	-58	Below 38
2.0	12.5	-58	Below 38
	1.5	-66	Below 40
3.0	12.5	-62	Below 40
5.0	12.5	-68	Below 44

^aSensitivities reported here were obtained by exciting the transmitting transducer using a broadband and 15-ns pulse. Tone burst excitation sensitivities will be 12 dB higher.

^bFor some contact transducers, a -20-dB sensitivity is assumed.

mode through a solid material by using AC and NC transducers. Figures 8 and 9 present observations in support of this. Both observations correspond to NCU transmission through 20-mm thick aluminum by 1-MHz and 20-mm active area diameter transducers in the direct transmission mode. In both cases, transducers are separated from the test material surfaces by an ~5-mm air column. Furthermore, in Figs. 8 and 9, the transmitting transducer was excited by a high-energy 400-volt (into 4Ω input impedance) pulser, and the signal received from the

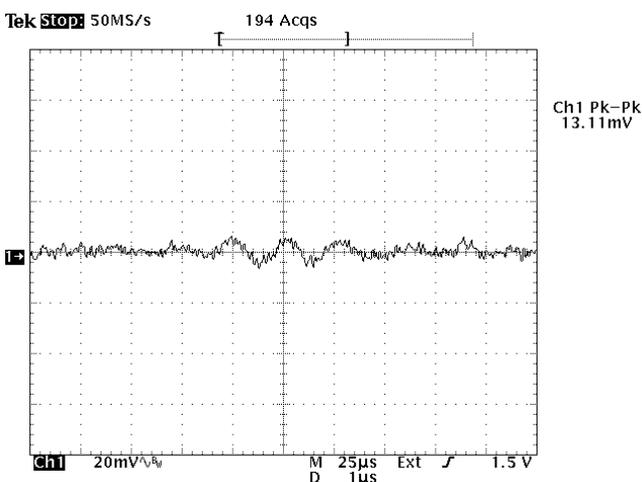


Figure 8. A 1-MHz noncontact transmitted signal through 20-mm aluminum by using transducers based on a soft, porous, polymer matching layer. Excitation of the transmitting transducer: 400 V into 4-Ω input impedance. Receiving transducer amplification: 64 dB. Transmitting and receiving transducers are 5 mm away from the material surfaces. Under these conditions, the transmitted signal amplitude is 13.1 mV. Compare with Fig. 9.

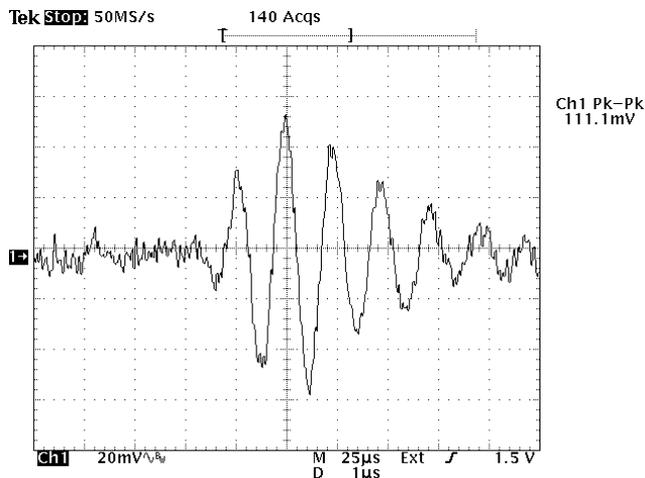


Figure 9. A 1-MHz noncontact transmitted signal through 20-mm aluminum by using new transducers based on a compressed fiber matching layer. Excitation of the transmitting transducer: 400 V into 4-Ω input impedance. Receiving transducer amplification: 64 dB. Transmitting and receiving transducers are 5 mm away from the material surfaces. Under these conditions, the transmitted signal amplitude is 111.1 mV. Compare with Fig. 8.

receiving transducer was amplified by a 64-dB gain. The key difference is that Fig. 8 was obtained by AC transducers and Fig. 9 by NC transducers. Under these conditions, the amplitude of the transmitted signal through 20-mm aluminum by AC transducers is 13.1 mV, whereas it is 111.1 mV for NC transducers. This clearly establishes the superiority of the new noncontact transducer design over the other air-coupled transducers described in (14–17),

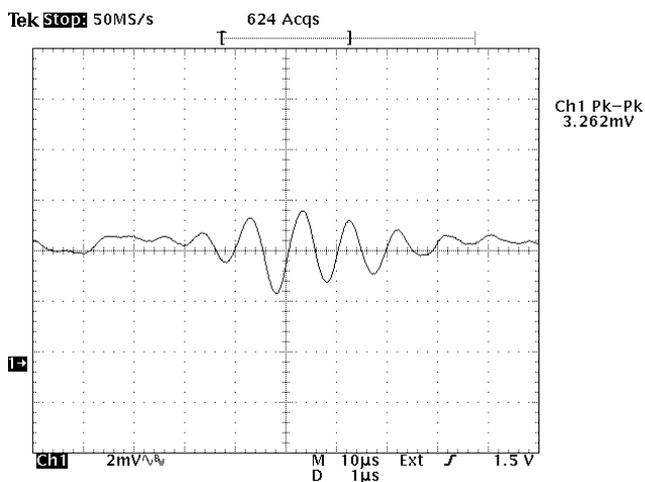


Figure 10. A 1-MHz noncontact transmitted signal through 20-mm aluminum by using new transducers based on compressed fiber matching layer. Excitation of the transmitting transducer: one burst 16-V sine wave. Receiving transducer amplification: 64 dB. Transmitting and receiving transducers are 5 mm away from the material surfaces. Under these conditions, the transmitted signal amplitude is 3.26 mV. No other air-coupled transducer can transmit ultrasound under similar conditions in high acoustic impedance materials by very low level excitation.

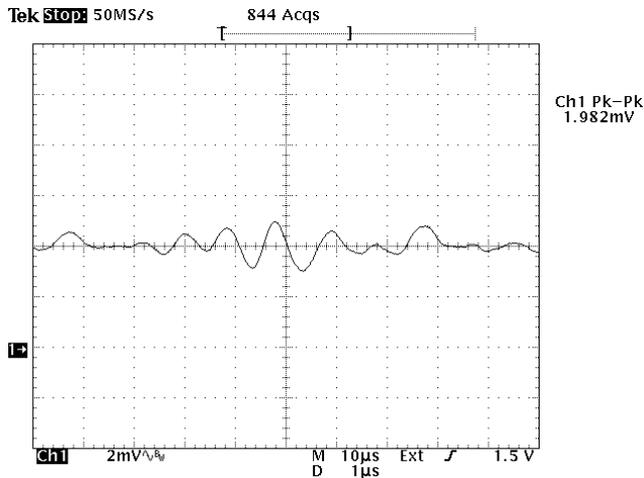


Figure 11. Ultrasound transmission through 25-mm carbon steel (Z , 51Mrayl) by using the new noncontact transducers (1-MHz, 20-mm active area diameter) with a single burst of 16-volt sine wave excitation and 64-dB amplification of the received signal by an equivalent transducer. No other air-coupled transducer can transmit ultrasound under similar conditions in high acoustic impedance materials by very low level excitation.

including this author's air/gas propagation transducers that are commercially available from Ultran Laboratories since 1983.

To demonstrate further the exceptionally high sensitivity of NC transducers, we decided to conduct an experiment that would normally be considered impossible! An experiment analogous to that described before was performed, except that in this case the NC transmitter was excited merely by a single burst of a 16-volt sine wave and 64-dB amplification of the received signal. Figure 10 presents the observation from this, showing a 3.26-mV signal transmitted through 20-mm aluminum in the noncontact mode. AC transducers (based on soft polymer matching layers, that is, porous or nonporous or have hollow spheres) were unsuccessful in generating ultrasound transmission through 20-mm aluminum by 16-volt excitation, despite high signal averaging! It is also interesting to note that by low energy excitation, using NC transducers, we propagated megahertz frequencies even in steel, whose acoustic impedance is 51 Mrayl, six orders of magnitude higher than in air! We show an example of this startling conclusion in Fig. 11. It is important to note that the purpose of observations shown in Figs. 8–11 was to demonstrate the high sensitivity of NC transducers relative to any other similar device. The purpose of these experiments is not to recommend or suggest the usage of low-energy transducer excitation for testing materials.

NONCONTACT ULTRASONIC ANALYZER

As is evident from the preceding sections, NC transducers can be used with any suitable commercially available pulser-receiver to transmit and to detect ultrasound through any material. However, our ultimate goal was to generate an NCU mode that would rival the performance

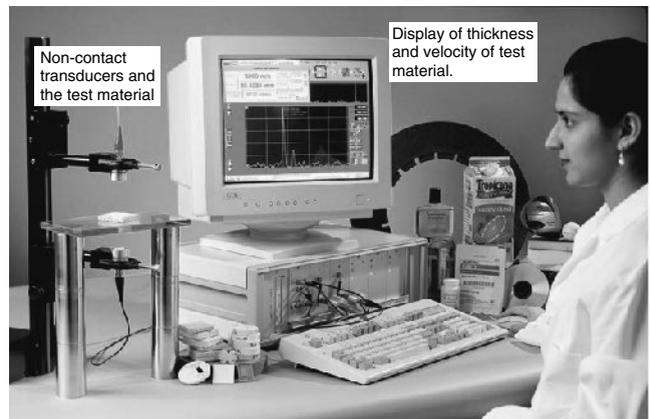


Figure 12. Ultrasonic noncontact analysis system, the NCA 1000, using noncontact transducers and a monitor screen displaying the thickness and velocity through the test material.

of conventional contact or immersion ultrasound. The high transduction of NC transducers is not enough to accomplish this task. For example, we still need to overcome the natural barrier of acoustic impedance mismatch between the coupling air and the test medium. From Table 2, we see that losses due to this mismatch are formidable. To circumvent this and not jeopardize our objective of equating NCU performance with that of conventional contact-based ultrasound, a new mechanism of transducer excitation and signal amplification was needed. Nevertheless, this seemingly impossible task, too, was overcome. In 1997–1998, a novel ultrasonic system was conceived and produced. Identified as the NCA 1000,² this instrument was developed by Leon Vandervalk and Ian Neeson of VN Instruments, Canada. It is based on synthesizing a computer-generated chirp combined with the best attributes of noncontact transducers. Signal processing in the NCA 1000 is done by synthetic aperture imaging techniques. The NCA 1000 is characterized by a dynamic range of >150 dB and a time-of-flight (TOF) measurement accuracy of ± 10 ns in ambient air and better than ± 1 ns in closed conditions. The NCA 1000 (Fig. 12) measures the TOF, thickness, velocity, and integrated response (area underneath transmitted or reflected signals in dB) of materials in the time domain. By using the FFT mechanism of this system, it is also possible to conduct noncontact ultrasonic spectroscopy. Furthermore, by raster scanning the transducers or the test material, we can generate surface or internal images of the test material. Such images can be representative of the material surface roughness, TOF, transmission attenuation, velocity, or thickness.

REFLECTION AND TRANSMISSION IN NONCONTACT MODE

Analogous to conventional contact or immersion ultrasound, ultrasound in the non-contact mode is also reflected and transmitted at various interfaces as well as through a test medium. In this section, we provide examples of

²U.S. patent pending and in process.

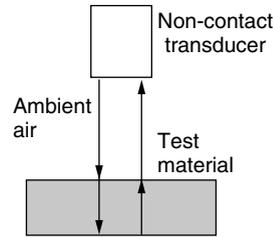
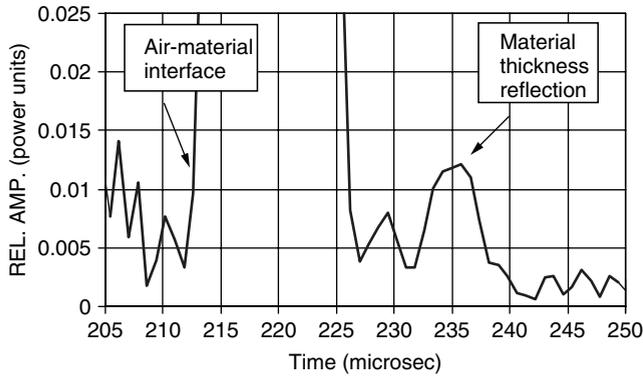


Figure 13. Noncontact ultrasound pulse-echo reflection from 9-mm thick silicone rubber. Transducer: 1-MHz, 12.5-mm active area diameter, and 76-mm point focused. Transducer to material distance in air: 40 mm. It is important to note that the material thickness reflection is ~ 40 dB lower than that of the air-material interface.

various paths of ultrasound reflection and transmission as functions of test material's interfaces and its volume.

Single-Transducer Operation (Pulse-Echo)

By operating one transducer simultaneously as a transmitter and receiver (analogous to the pulse-echo technique), it is very easy to generate reflection from an air-material interface due to the extremely high reflection coefficient at this interface. However, in this mode, it is nearly impossible to produce a far side reflection corresponding to the test material thickness in ambient air. This difficulty stems from several factors, such as the extremely small transmission of ultrasound in the test material, the extremely high beam spread on the surface of the material, and the inherent electrical noise associated with single transducer operation from the initial pulse. To a degree, the adverse effects of these factors can be minimized by a focused transducer, which will reduce the beam spread and focus ultrasound (thus intensify the reflected energy) within the test material. Figure 13 shows pulse-echo reflection signals from a 9 mm thick silicone rubber sample. This observation was generated by using a 1-MHz focused NC transducer. Similar results have also been observed for other plastic materials. However, at this time, we have no concrete proof of generating these observations from high acoustic impedance materials, such as metals and ceramics. Because reflection from an air-material interface is extremely strong in single transducer operation, ultrasonic reflectivity can be used to characterize the surface characteristics of the material. Such applications include surface acoustic impedance, surface roughness, particle size measurement, surface texture and microstructure, distance, proximity, and level sensing, and any other surface conditions that are sensitive to noncontact ultrasound.

Separate Transmitter and Receiver Operation on the Same Side (Pitch-Catch)

By using two noncontact transducers, one a transmitter (T) and the other a receiver (R), on the same side of the test material, (Fig. 14), it is possible to launch and measure the characteristics of longitudinal, shear, and surface waves in practically all types of material. Generation of these waves

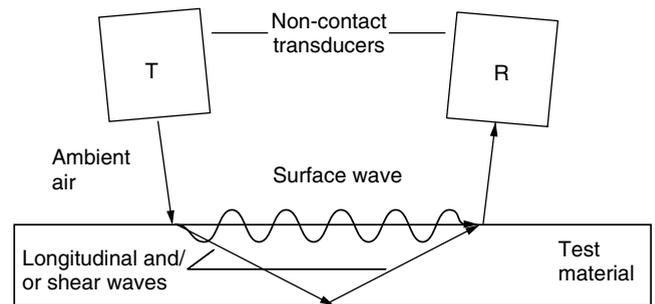


Figure 14. Schematic of a setup for same side operation of separate transmitting (T) and receiving (R) transducers, also showing various types of wave propagation in the test material.

is determined by Snell's law,

$$\sin i / \sin r = V_a / V_m, \quad (7)$$

where i is the incident angle in air, r the refraction angle in the test material, V_a is the ultrasound velocity in air, and V_m the velocity in the test material.

By manipulating the incident angle in air, a variety of wave types can be produced in a test material. Figure 15 shows the far side thickness reflection of a longitudinal

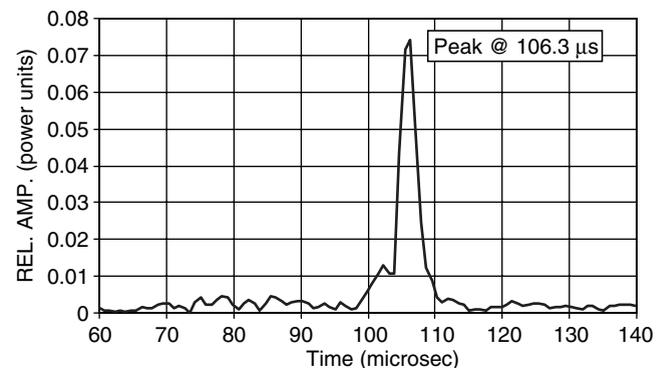


Figure 15. Thickness reflection of longitudinal wave propagation in 12-mm thick aluminum per Fig. 14 setup. Distance between the transducers and the material surface: ~ 15 mm ambient air. Distance between T and R: ~ 50 mm. Incident angle in air: $\sim 3.0^\circ$.

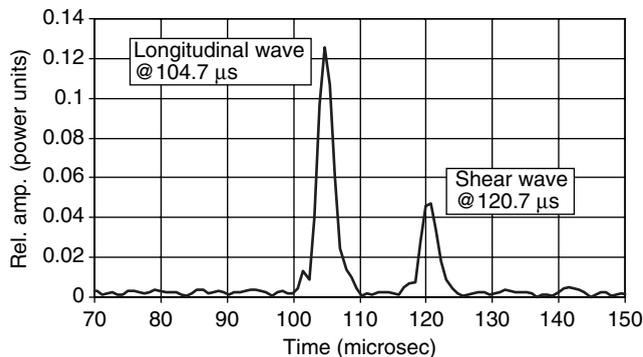


Figure 16. Thickness reflections of longitudinal and shear waves in 32-mm thick transparent polystyrene per Fig. 14 setup. Distance between the transducers and the material surface: ~ 13 mm ambient air. Distance between T and R: ~ 50 mm. Incident angle in air: $\sim 6.0^\circ$.

wave from 12-mm thick aluminum. Figure 16 shows the far side thickness reflection of longitudinal and shear waves from 32 mm thick transparent polystyrene. Figure 17 is a longitudinal wave refracted surface wave in aluminum produced by an incident angle equal to the first critical angle (i.e., total reflection of a longitudinal wave) which is 3.16° for aluminum. Figure 18 shows a shear wave refracted surface wave in aluminum, generated by an incident angle equal to the second critical angle (i.e., total reflection of a shear wave), which is 6.3° for aluminum. It is important to note that while performing these experiments, distances—corresponding to transducers and the test materials and angles of transmitting and receiving transducers—were not measured accurately. The primary function of this exercise is to show the feasibility of various types of bulk and surface wave generation by noncontact ultrasound. Applications of such measurements include NCU evaluation of materials from one side, defect detection, anisotropy measurements, and relationships of ultrasonic velocities to test material elastic and mechanical properties.

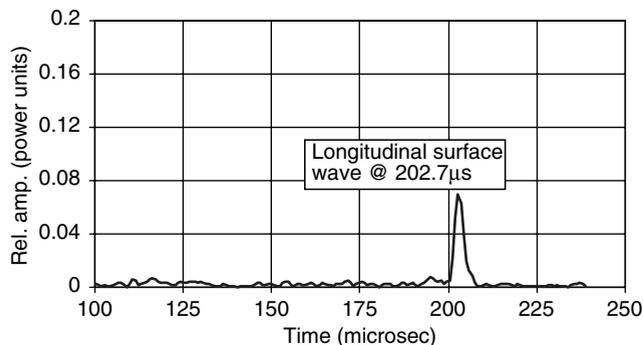


Figure 17. Surface wave in aluminum produced by total reflection of longitudinal wave when the incident angle is equal to the first critical angle. Setup for this is shown in Fig. 14. Distance between the transducers and the material surface: ~ 15 mm ambient air. Distance between T and R: ~ 100 mm. Incident angle in air: $\sim 3.2^\circ$.

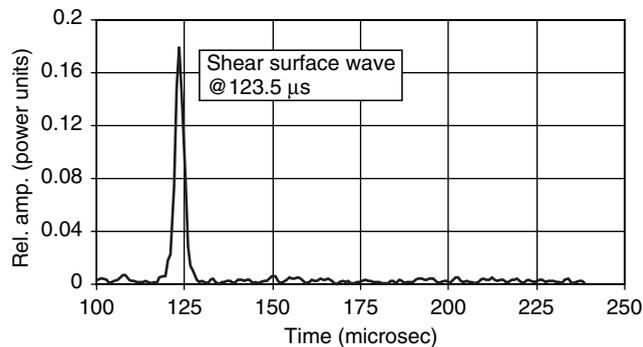


Figure 18. Surface wave in aluminum produced by total reflection of shear wave when the incident angle is equal to the second critical angle. Setup for this is shown in Fig. 14. Distance between the transducers and the material surface: ~ 15 mm ambient air. Distance between T and R: ~ 100 mm. Incident angle in air: $\sim 6.5^\circ$.

Direct Transmission

When a test material is inserted between two noncontact transducers facing opposite each other in air, then ultrasound is transmitted and reflected from all interfaces corresponding to air and the material. Details of this are shown in Figs. 19 and 20. Direct transmission is relatively the easiest technique in noncontact ultrasound. Therefore, it has been quite extensively studied and developed. Applications of this technique are numerous: thickness and velocity measurements, defect detection, textural and

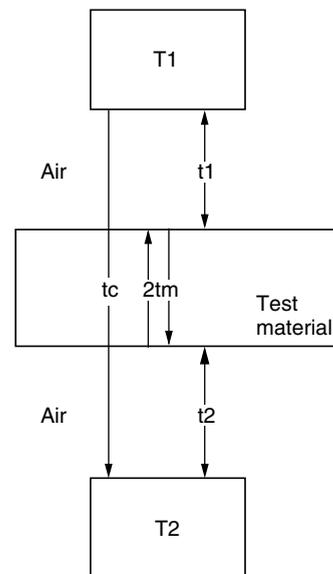


Figure 19. Propagation of ultrasound in the direct transmission noncontact mode. Here, t_c is the complete time of flight (TOF) corresponding to the propagation of ultrasound in air and the test material, $2t_m$ is the round-trip TOF through the test material thickness, t_1 and t_2 are, respectively, TOFs from transducer T_1 to the top surface of the material, and transducer T_2 to the bottom surface of the material. For the further significance of this, see Fig. 20.

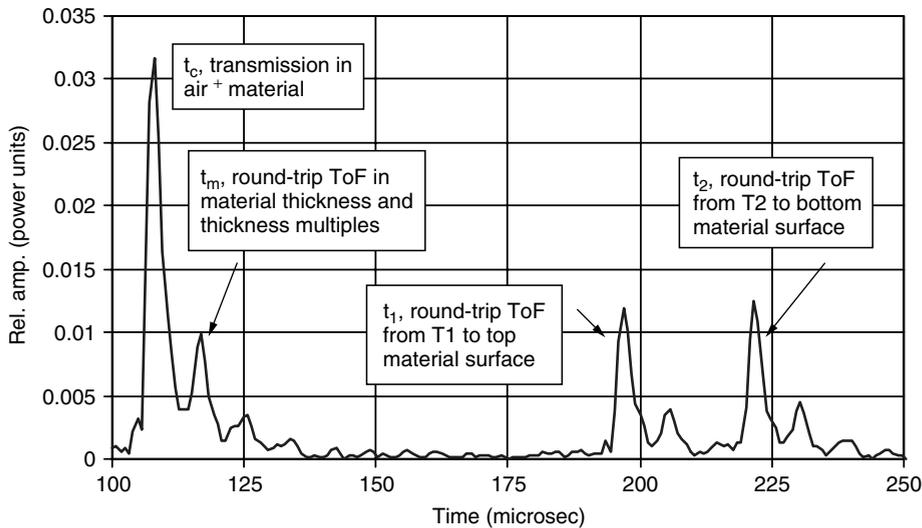


Figure 20. Direct transmission non-contact ultrasound propagation through a test material (7-mm thick cheese) per Fig. 19 setup.

microstructural evaluation, transmission, velocity, thickness, and TOF imaging, detection of the presence or absence of liquids in containers, and many more.

VERY HIGH FREQUENCY NCU PROPAGATION IN MATERIALS

By using NC transducers in ambient air, we have amply demonstrated that frequencies as high as 2 MHz can be easily propagated through a variety of materials, including fibrous and particulate, plastics, ceramics, metals, and composites. However, frequencies even higher than 2 MHz have been successfully investigated for propagation through solids. Figure 21 shows an example of 4-MHz propagation through 4.5-mm thick multilayer graphite fiber-reinforced plastic composite in ambient air. Similar observations have also been made for soft polymers, thin

metals, and fibrous and particulate materials. As expected, the magnitude of transmitted signals through solids increases substantially when examined under high gas pressures. Figure 22 shows the propagation of 4-MHz transmission through 8-mm thick steel under 60 bars nitrogen by using a single transducer in the pulse-echo technique.

NONCONTACT ULTRASONIC MEASUREMENTS

Because ultrasound can be reflected and transmitted through a test material and its surfaces, one can use the respective signals to make significant measurements in the time and frequency domains. Such measurements can be further related to important test material characteristics, such as velocity, thickness, defects, internal and surface texture or microstructure, and other ultrasound-dependent parameters.

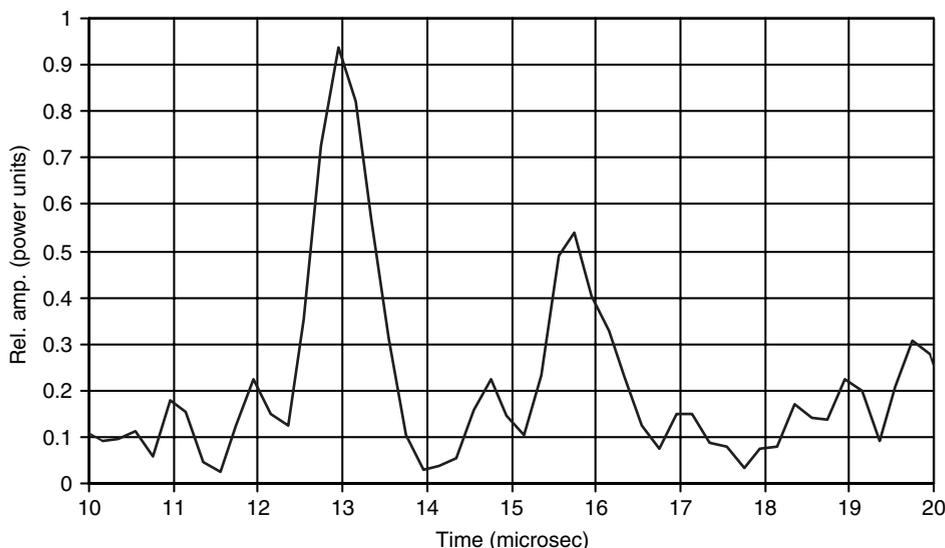
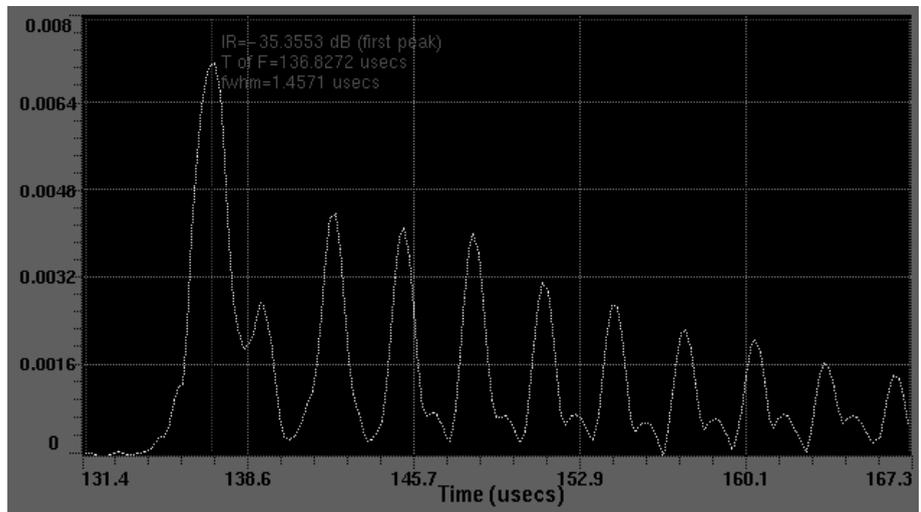


Figure 21. Noncontact ultrasound direct transmission through 4.5-mm thick multilayer graphite fiber-reinforced plastic composite at 4 MHz in ambient air per Fig. 19 setup. Transducers are 4-MHz and 12.5-mm active area diameter. Distances from transducers to test materials surfaces are ~3 mm. First peak: directly transmitted signal through air and the test material. Subsequent peaks: test material thickness reflection and its multiples.

Figure 22. Single transducer (pulse-echo) signal through 8-mm thick section of steel at 4 MHz under 60 bar nitrogen pressure. Transducer used is 4-MHz and 12.5-mm active area diameter. Distance from transducer to the surface of test material is ~25 mm. First peak: gas-materials interface. Subsequent peaks: thickness and thickness multiples corresponding to the test material.



Velocity and Thickness Measurements

There are several ways to determine longitudinal wave velocity in test materials by noncontact ultrasound. For example, if multiple reflections corresponding to the thickness of the test material are observed (Fig. 23), then one can measure the TOF, t_m , between the two successive peaks to determine the velocity of a material of known thickness. The TOF measured this way corresponds to the round-trip TOF in the test material. Therefore,

$$V_m = 2d_m/t_m. \quad (8)$$

For example, t_m measured between any two successive peaks from Fig. 23 is $10.4 \mu\text{s}$ for a 13.5-mm thick material (in this case, isotropic graphite); thus, the velocity is 2595 m/s.

It is important to note that the appearance of multiple thickness reflections in the NCU mode depends on the attenuation and acoustic impedance of the test material and the frequency of transducers. For example, the lesser

the attenuation, the lower the acoustic impedance, and at thicknesses greater than one wavelength in the test material, material thickness reflections are observable. On the other hand, when only the transmission signal is observed (i.e., without thickness multiples for attenuative materials), one can determine the TOF of the known thickness of the test material similarly to using contact delay line transducers. In such a case, the solid delay lines on the transducers are replaced by air columns in front of the transmitting and receiving transducers as functions of air distances between the test material surfaces.

As seen in the aforementioned technique, to determine the test material velocity, its thickness must be known. However, in the NCU mode, the thickness of the test material can also be measured. For simultaneous measurements of material thickness and its velocity in the NCU mode, we must examine all paths of ultrasound transmission and reflection to and from the test material. These paths relate to propagation of ultrasound relative to transmitting and receiving transducers “taking to each other” in the air column, ultrasound transmitted through the test

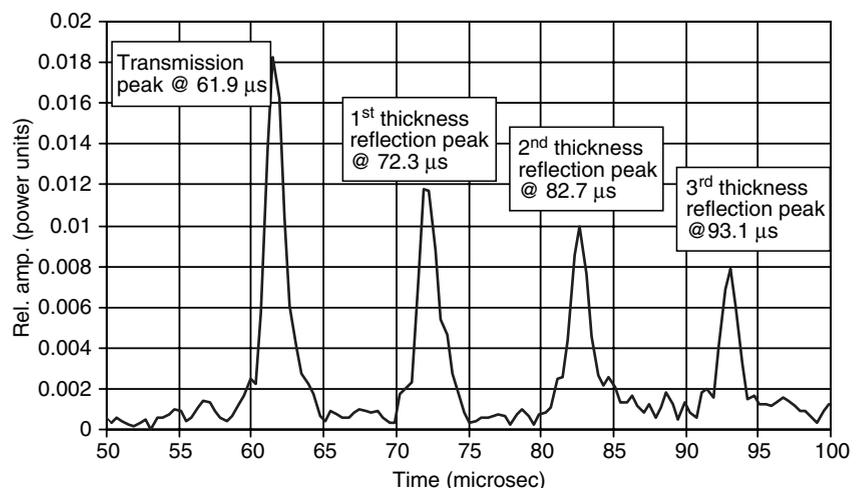


Figure 23. Velocity in material determined by direct transmission NCU propagation in a test material (in this case, 13.5-mm thick isotropic graphite) characterized by multiple thickness reflections. Note that the TOF (round-trip TOF through the material) between any two successive peaks is $10.4 \mu\text{s}$; thus, the velocity in the material is 2595 m/s.

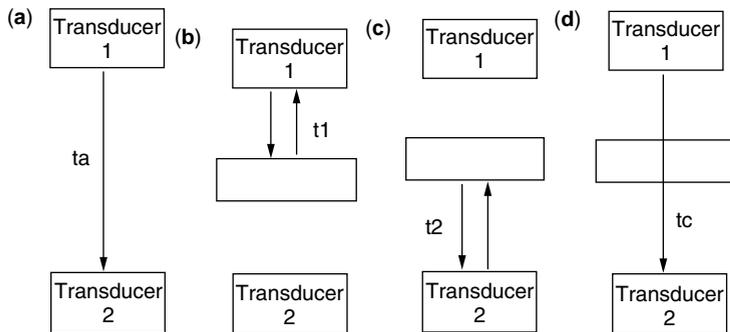


Figure 24. Propagation of ultrasound with respect to a test material in the noncontact transmission mode. The NCA 1000 measures and calculates all times of flight shown in this figure and then displays them on the monitor screen (Fig. 25). See text for details.

material, and ultrasound reflections from the test material surfaces in air. These paths of ultrasound propagation in the NCU transmission mode—needed to determine the test material thickness and velocity—are shown in Fig. 24. The signals generated by these paths of propagation and their significance are as follows:

Path (a) is the transmission from transducer 1 to transducer 2 in air—measures TOF, t_a . If ultrasound is propagated from transducer 2 to transducer 1, the same TOF is measured. Path (b) is the reflection from transducer 1 to the material surface in air—measures TOF, t_1 . Path (c) is the reflection from transducer 2 to the material surface in air—measures TOF, t_2 . Path (d) is the transmission from transducer 1 to transducer 2, and the test material is in between—measures TOF, t_c . If ultrasound is propagated from transducer 2 to transducer 1, the same TOF is measured.

From these times of flight measurements, the test material thickness and its velocity are determined according to the following relationships:

$$V_m = d_m/t_m \quad (9)$$

$$d_m = V_a t_{am} \quad (10)$$

$$t_{am} = t_a - (t_1 + t_2)/2 \quad (11)$$

$$t_m = t_{am} - (t_a - t_c) \quad (12)$$

In these equations, d_m is the test material's thickness, V_m , the velocity of ultrasound in the test material V_a , the velocity of ultrasound in air (determined from a reference material), t_{am} , the time of flight in air corresponding to the test material thickness d_m , and t_m is the time of flight in the test material.

By proper substitutions,

$$d_m = V_a [t_a - (t_1 + t_2)/2] \quad (13)$$

$$V_m = d_m/t_{am} - (t_a - t_c) \quad (14)$$

As an example, Fig. 20 shows actual transmitted and reflected signals when a test material is examined in the noncontact transmission mode. Identification and location of these signals with respect to the test material are shown in Fig. 19. As can be seen from Eqs. (13) and (14), to determine the thickness and velocity according to this scheme, one only needs the measurements of four times of flight (t_a , t_1 , t_2 , t_c) and the velocity of ultrasound in air. These parameters

were measured and calculated by the NCA 1000 computer and are displayed with the velocity and thickness of the test material on the monitor screen (Fig. 25).

Integrated Response, Transmissivity, and Reflectivity Measurements

In the time domain, the NCA 1000 measures and displays the times of flight of the signals and also shows the integrated response (IR) of these signals. IR is a measurement of the area underneath a particular peak in power dB units. Due to the very high, >150-dB, linear dynamic range of the NCA 1000, the IR can be used to measure the amount of ultrasonic energy transmitted (transmissivity) or reflected (reflectivity) from a test material and relate it (IR) to subtle changes in the material. For example,

$$IR_m = IR_c - IR_a, \quad (15)$$

where IR_m is the amount of ultrasonic energy transmitted in the test material, IR_c is the ultrasound transmission through air and the material (between the transmitting and receiving transducers), and IR_a is the amount of ultrasound energy transmitted only through air. IR_c and IR_a are measured directly by the NCA 1000.

If a given test material is homogeneous, then the measurement of IR_m , it has been found, is related to the transmission coefficient [Eq. (1)]. To illustrate this, we evaluated a flat polished specimen of polystyrene. Figure 26 shows the IR_c (−21.72 dB) of the transmitted peak of ultrasound from air into the specimen, and Fig. 27 shows a similar peak, but only through air, corresponding to IR_a (+41.7 dB), thus yielding IR_m −63.42 dB for the specimen. It is important to note that this measurement corresponds very closely to

$$IR_m = 20 \log T, \quad (16)$$

where T , the transmission coefficient, is defined by Eq. (1).

For example, the calculated value for ultrasonic energy transmitted in polystyrene [Eqs. (1) and (4)] is −63.34 dB, which is very close to −63.42 dB determined by measuring integrated response peaks. It should be pointed out that the transmission coefficient is assumed to be independent of ultrasonic attenuation and the thickness of the test medium. In reality, this is not absolutely true. For example, varying thicknesses of polystyrene samples at different frequencies yield different values of T . Though these variations are

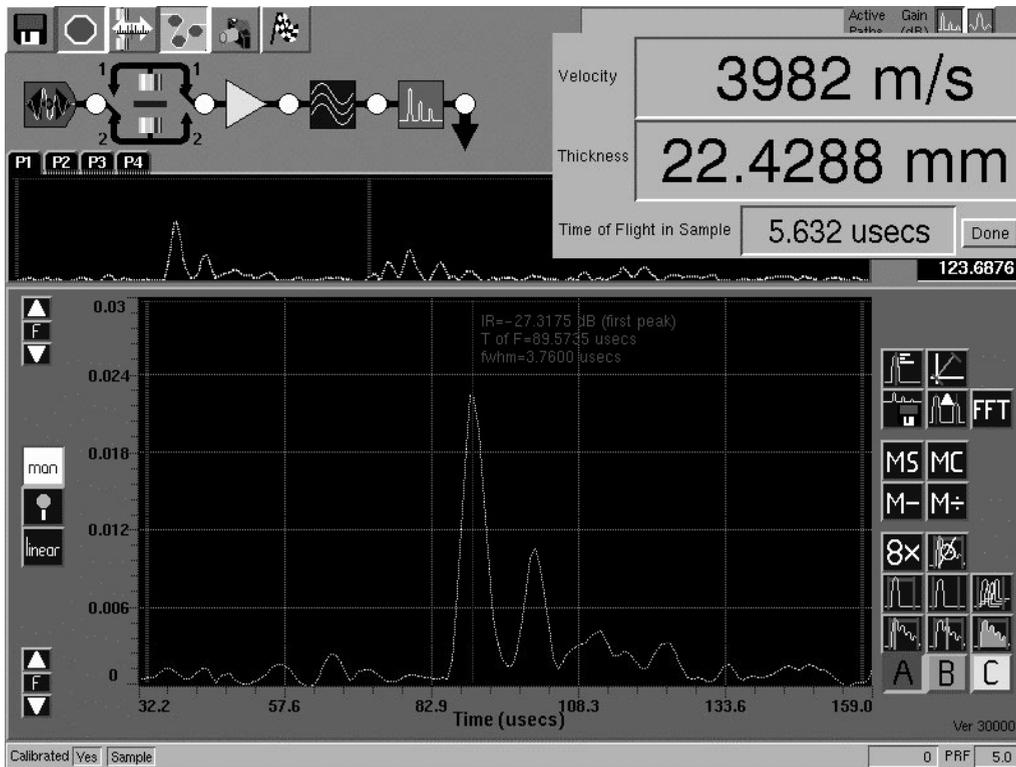


Figure 25. The NCA 1000 screen displaying velocity of ultrasound and thickness of a material. The test material is a 22.5-mm porous sintered ceramic.

very small, yet they are measurable. On the other hand, if the transmission coefficient can be measured with a very high degree of certainty and precision, then it should also be possible to measure the absolute density of the test material by first determining the acoustic impedance Z_m of the test material:

$$Z_m = \frac{Z_1}{T} [(2 - T) + 2(1 - T)^{1/2}], \quad (17)$$

$$\rho_m = Z_m / V_m. \quad (18)$$

Measurement of IR_m and solving Eqs. (17) and (18) provide approximate ideas about the acoustic impedance and

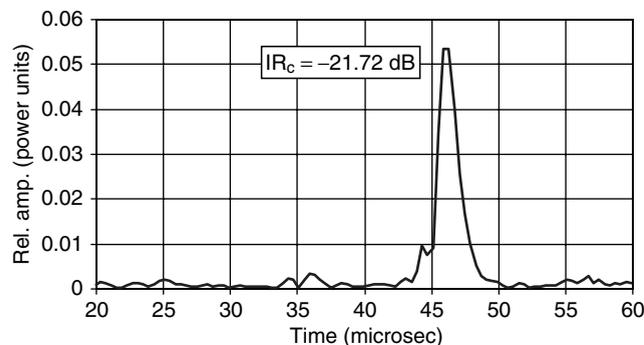


Figure 26. Noncontact transmission through a 20-mm thick flat polished polystyrene sample showing the integrated response IR_c of the transmitted peak.

density of the test material. For accurate determination of these characteristics, factors such as ultrasound attenuation (analogous to absorption coefficient in X-ray absorption) and material thickness must also be considered:

$$T = I_2 / I_0 = Z_1 Z_2 / (Z_1 + Z_2)^2 = \exp(-\mu \rho x), \quad (19)$$

where T is the transmission coefficient, I_2 the ultrasound energy transmitted into the material (acoustic impedance

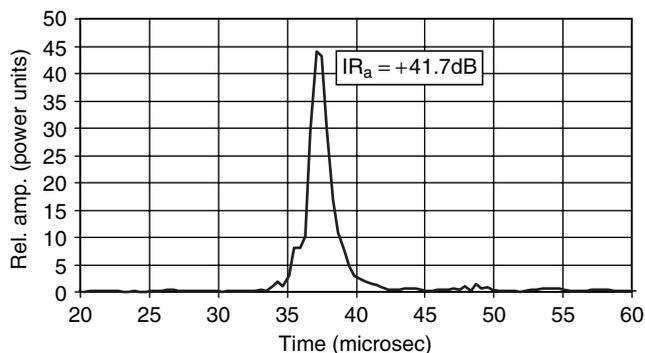


Figure 27. Noncontact transmission through air column showing the integrated response IR_a of the transmitted peak. Note that while making this measurement, the distance between the transmitting and receiving transducers was compensated for the 20 mm thickness of the specimen in Fig. 26.

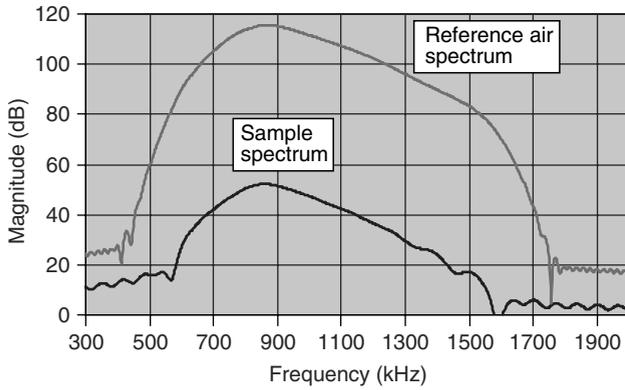


Figure 28. Procedure for noncontact ultrasonic spectroscopy. Top: FFT magnitude spectrum of ultrasound transmission through air as a reference. Bottom: FFT magnitude spectrum of ultrasound transmission where the test material (composite rubber) is between the transmitting and receiving transducers in air.

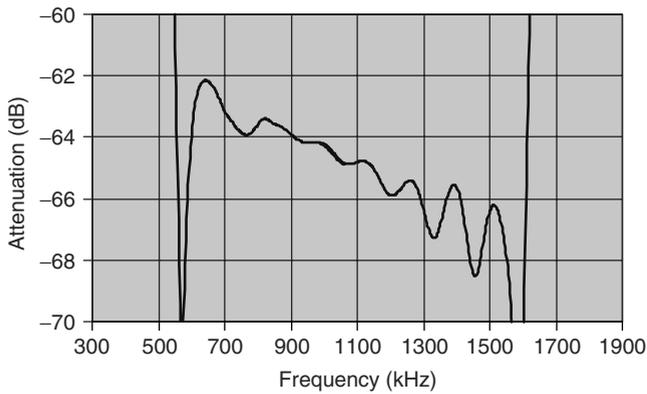


Figure 29. Frequency dependence of ultrasonic attenuation by subtracting the air reference from that of the sample spectrum (Fig. 28).

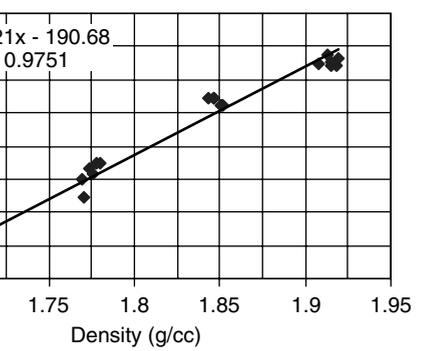
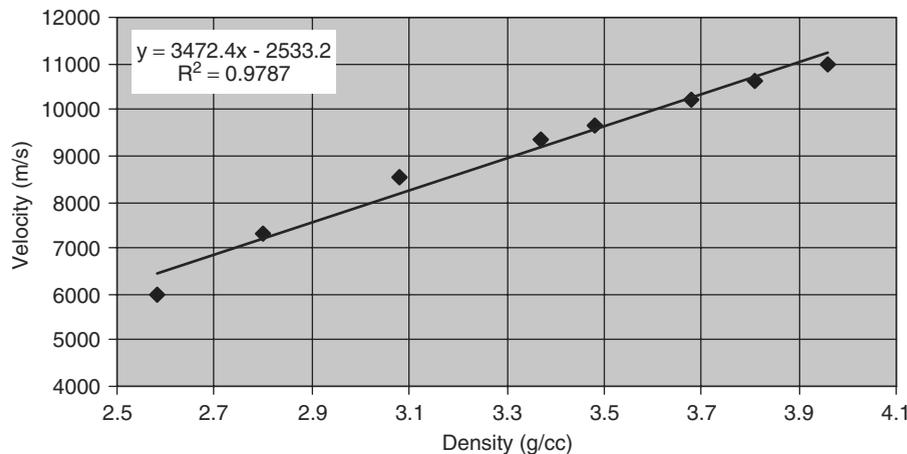


Figure 30. Relationship between density and noncontact ultrasonic velocity in low-density green alumina. Transducer: 1-MHz, 12.5-mm active area diameter.

Z_2) from air (acoustic impedance Z_1), I_0 the input ultrasound energy, μ the material ultrasound attenuation coefficient, ρ the material density, and x the material thickness. At this time, the development of these relationships and techniques for measuring T , Z , μ , and ρ by noncontact ultrasound are in progress (21).

Because the measurement of the transmission coefficient still needs to be validated, it is best to refer to IR_m as transmissivity (when propagating ultrasound through the material in the direct transmission mode) or reflectivity (when ultrasound is reflected from the surface of the material). Such measurements are useful in evaluating the test material's internal and surface characteristics, such as defects, texture, microstructure, and roughness.

Noncontact Ultrasonic Spectroscopy

By performing the fast Fourier transformation (FFT) of transmitted or reflected time domain signals, test materials can also be characterized to investigate the frequency dependence of ultrasonic attenuation. Such examinations are important while testing microstructurally complex materials or those for which time domain velocity measurements are not sensitive enough. The first step for

Figure 31. Relationship between density and noncontact velocity in sintered alumina. Transducers: 1 MHz for samples less dense than 3.5 g/cc and 2 MHz for samples denser than 3.5 g/cc.

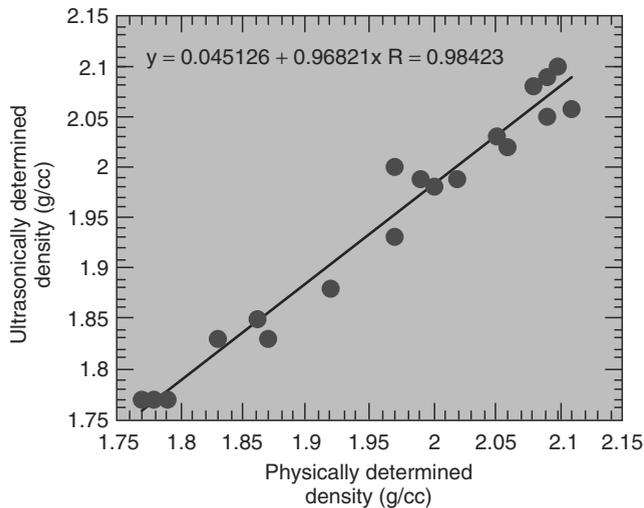


Figure 32. Comparison of ultrasonically and physically determined density of green alumina.

frequency dependence of ultrasonic characterization is the acquisition of a reference frequency magnitude spectrum of a transmitted signal in air without the test material. The next step is to do the same when the test material is inserted between the transducers. As an example, the FFT magnitude spectra for air and the test material are shown in Fig. 28. To generate the frequency dependence of ultrasonic attenuation, the sample spectrum is subtracted from that of the reference air spectrum (Fig. 29). By performing a similar analysis, the surfaces of materials can also be analyzed in the frequency domain. Frequency-dependent attenuation can be related to the test material's internal and surface characteristics, such as defects, texture, microstructure, and roughness.

APPLICATIONS OF NONCONTACT ULTRASOUND

Noncontact transducers have now been successfully produced in the frequency range of 100 kHz to >5 MHz.

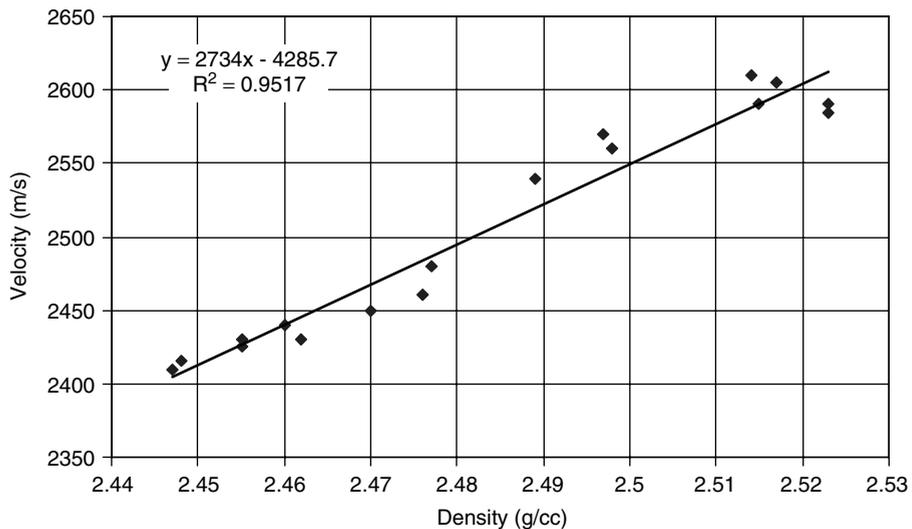


Figure 33. Relationship between velocity and density for isostatically pressed, high-density green alumina. Transducers: 1-MHz, 12.5-mm active area diameter.

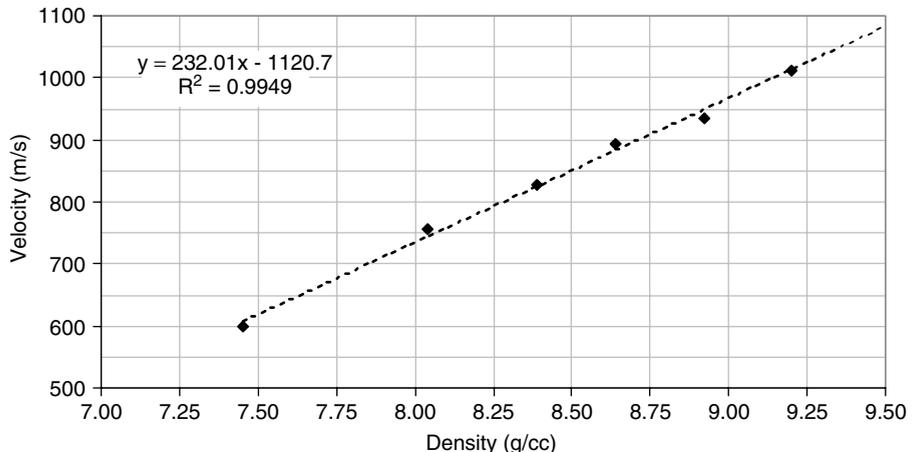


Figure 34. Relationship between velocity and density for pressed green tungsten carbide. Transducers: 500-kHz, 12.5-mm active area diameter.

Applications of noncontact transducers greater than 3MHz in ambient air are limited, but transducers between 200 kHz and 3 MHz have been extensively used for several industrial and medical applications (22–29). In this section, we present selected examples of NCU applications for materials testing and other objectives.

Materials Characterization and Defect Detection

Figures 30 and 31, respectively, show the relationships between densities of green and sintered alumina and the noncontact velocities in these materials. Figure 32 shows a correlation between ultrasonically (from reference samples) and physically determined densities of green alumina. The noncontact ultrasonic technique has been successfully applied to characterize density and defects in a variety of green materials such as ceramics tiles, multilayer electronic packages, powder metals, cements, and concrete. Figures 33 and 34, respectively, show the velocity–density relationships for isostatically pressed high-density green alumina and tungsten carbide. Examples of defect detection in green and sintered ceramics are shown in Figs. 35 and 36, and similar observations for aluminum are shown in Fig. 37. Figures 38 and 39, respectively, show trend plots of direct transmission and same side reflection in graphite fiber-reinforced plastic (GFRP) composites bonded to a honeycomb structure. The same side observations (Fig. 39)

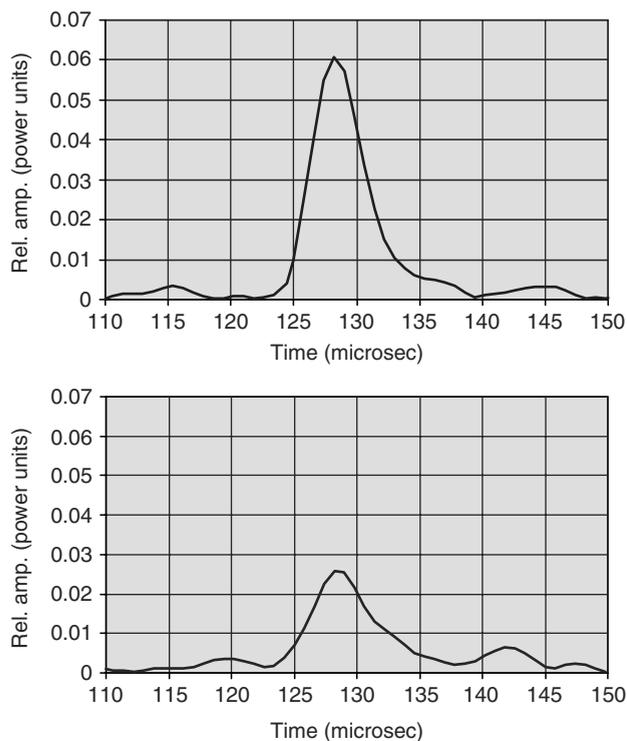


Figure 35. Defect detection in a sample of 14-mm thick green porcelain. Top: ultrasound transmission through a defect-free region. Bottom: ultrasound transmission through a region that has 1.5-mm diameter side-drilled cylindrical hole. Compare the amplitudes of the transmitted ultrasound intensity of the two regions. Transducers: 1-MHz and 12.5-mm active area diameter.

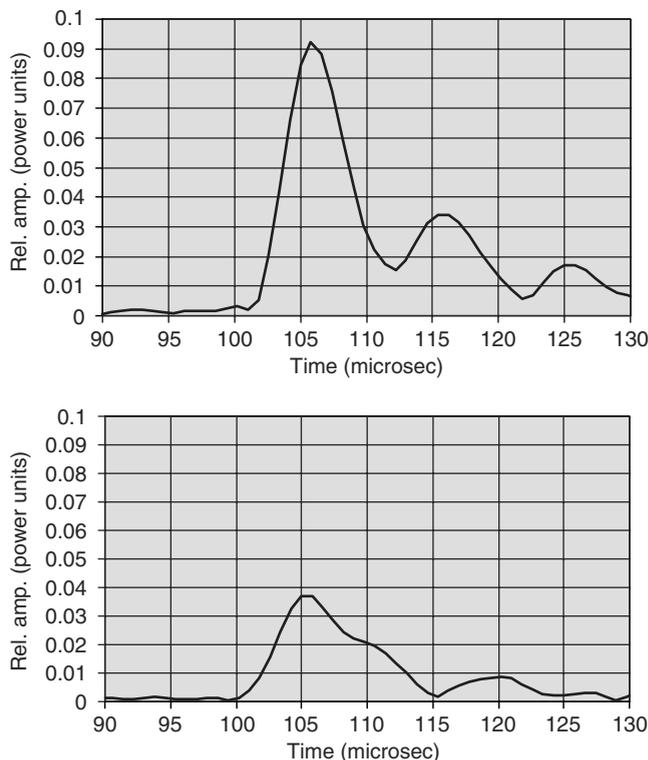


Figure 36. Defect detection in a sample of 20-mm thick, 20% porous, low thermal expansion sintered ceramic. Top: ultrasound transmission through a defect-free region. Bottom: ultrasound transmission through a region that has 1.5-mm diameter side-drilled cylindrical hole. Compare the amplitudes of the transmitted ultrasound intensity of the two regions. Transducers: 1-MHz and 12.5-mm active area diameter.

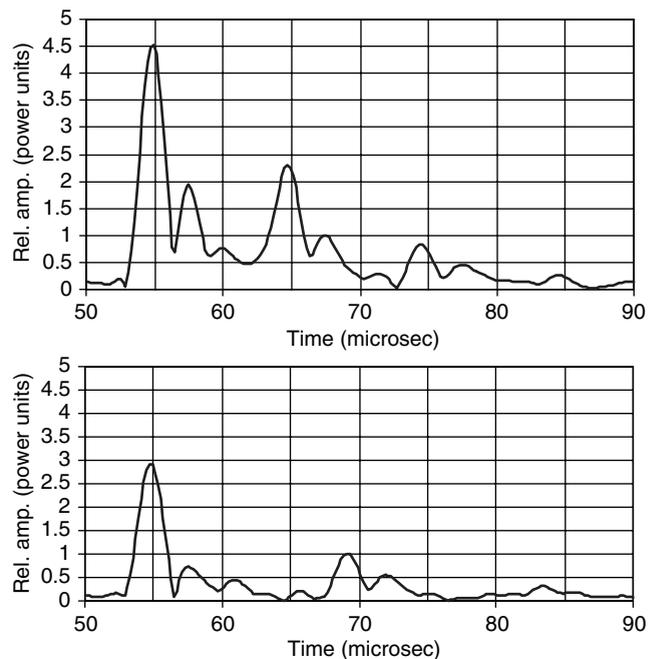


Figure 37. Detection of defects in an 8-mm thick sheet of aluminum by noncontact transmission mode. Top: defect-free region. Bottom: region that has a 1.5-mm cylindrical hole. Transducers: 2-MHz and 12.5-mm active area diameter.

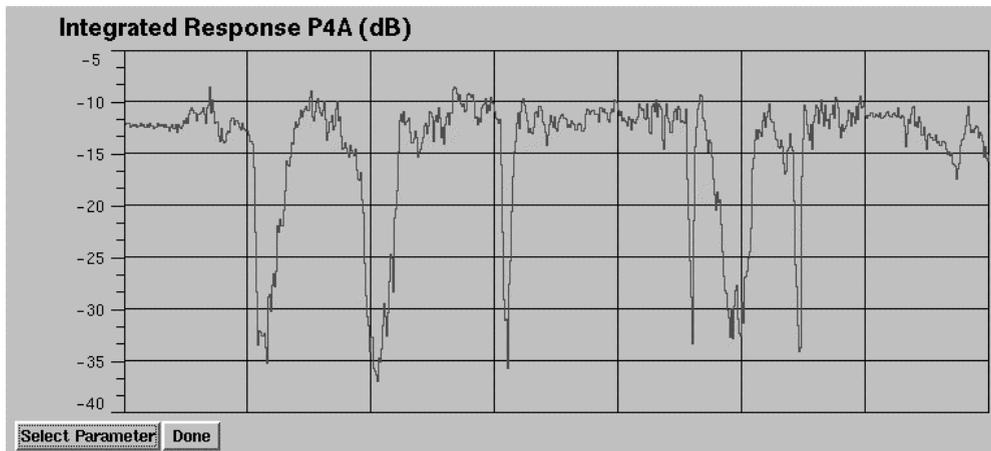


Figure 38. NCA 1000 trend plot showing direct transmission through a GFRP composite bonded to a honeycomb structure at various points. Regions showing the sharp drop in transmission are indicative of defects, such as delaminations. Transducers: 500-kHz and 12.5-mm active area diameter, separated from the material surfaces by ~ 40 mm ambient air.

correspond to the bond between the GFRP and the honeycomb.

Examples of textural and microstructural analysis by noncontact ultrasonic spectroscopy are shown in Figs. 40 and 41. Figure 40 shows the frequency dependence of ultrasound attenuation by three specimens of extremely porous ceramics (in this case, space shuttle tiles), and Fig. 41 shows similar observations from two samples of packaging foam whose cell dimensions vary. Figure 42 shows very high frequency noncontact transmission spectroscopy of two samples of paper towels. Similar observations have

also been made to detect bubbles and pores in liquids and other materials.

To evaluate surface characteristics by noncontact ultrasonic spectroscopy, several grinding discs of SiC varying in particle size were chosen. These discs were placed at a fixed air distance of 10 mm from a 2-MHz, 12.5-mm active area diameter transducer. Reflection from a polished sample of steel was assumed as a reference signal. Reflected signals from the reference and test materials were analyzed by performing FFT, and the frequency dependence of ultrasound was measured by subtracting the sample FFT spectra from

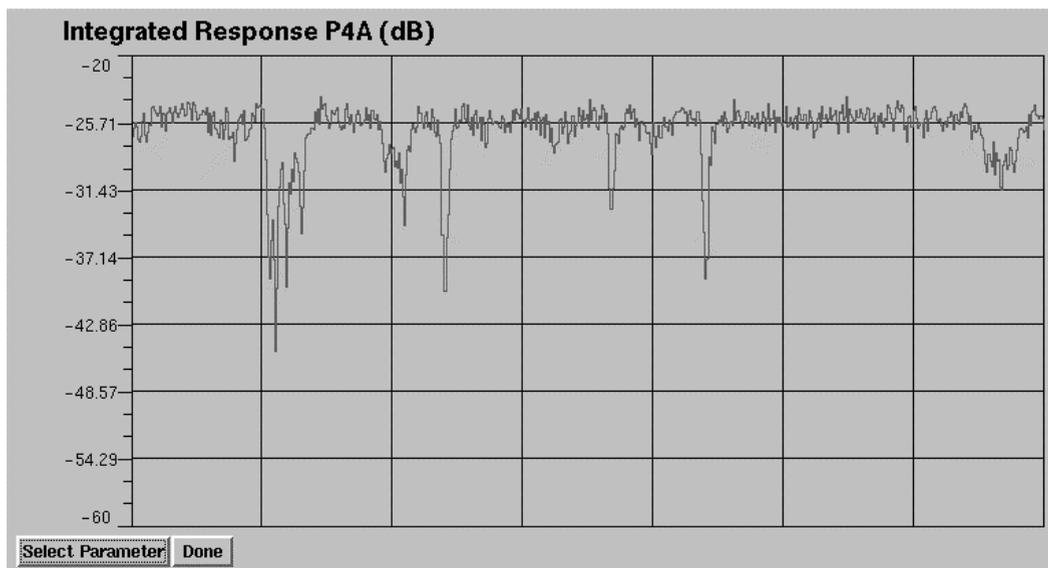


Figure 39. NCA 1000 trend plot showing the T-R reflection from same side in a GFRP composite bonded to a honeycomb structure at various points. Regions showing the sharp drop in reflected ultrasound from composite to honeycomb bond are indicative of defects, such as delaminations. Transducers: 500-kHz and 12.5-mm active area diameter, separated from the test material surface by ~ 10 mm ambient air.

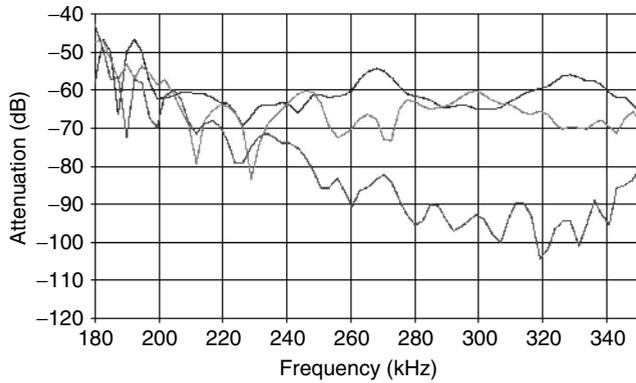


Figure 40. Noncontact transmission ultrasonic spectroscopy of extremely porous ceramics (space shuttle tiles) for microstructural characterization. Top: 0.38 g/cc. Middle: 0.28 g/cc. Bottom: 0.1 g/cc. Transducers: 250-kHz, 25-mm active area diameter.

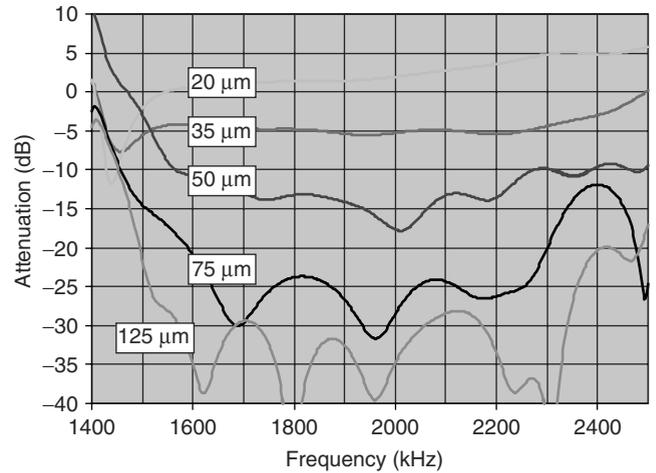


Figure 43. Noncontact reflection ultrasonic spectroscopy for surface characterization of materials. Note that as the surfaces become rough, the attenuation of ultrasound increases. Transducer: 2-MHz and 12.5-mm active area diameter.

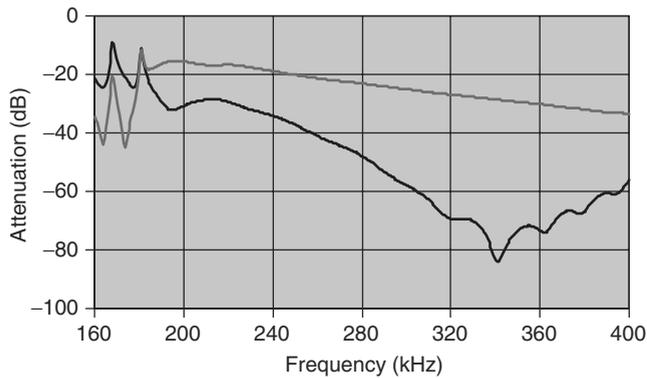


Figure 41. Noncontact transmission ultrasonic spectroscopy of packaging foam. Top: small cell. Bottom: large cell. Transducer: 250-kHz and 25-mm active area diameter.

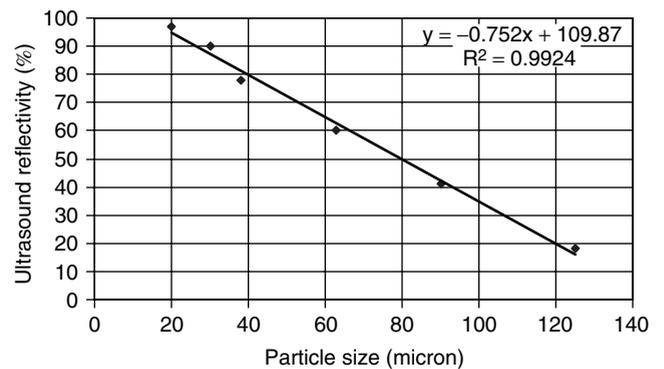


Figure 44. Noncontact reflectivity measurement for surface characterization of materials. Compare the reduction in reflectivity and high ultrasound attenuation (Fig. 36) as functions of increasing particle size. Transducer: 2-MHz and 12.5-mm active area diameter.

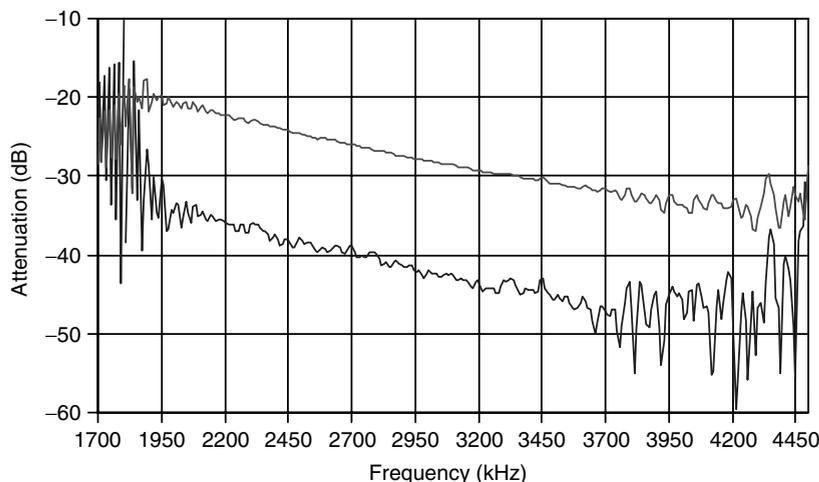


Figure 42. Very high frequency noncontact transmission ultrasonic spectroscopy of two different paper towels. Top: 0.2-mm thick, relatively hard, shallow dimpled texture. Bottom: 0.4-mm thick, relatively soft and deeply dimpled texture.

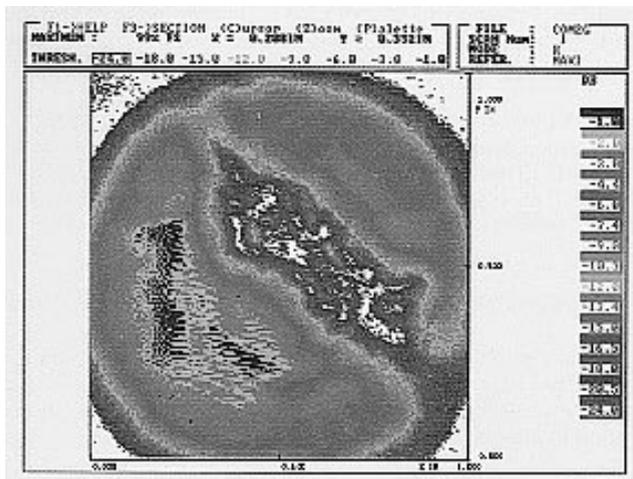


Figure 45. Partial contact ultrasonic image of an impact-damaged eight ply graphite-fiber-reinforced plastic composite 1.5 mm thick. Area scanned: 25 × 25 mm. This image was generated by using a 2-MHz and 25-mm active area diameter transmitter in contact with the material surface, and a 2-MHz, 3-mm diameter receiver was raster scanned in noncontact mode across the other surface of the material. A conventional pulser that sent 400 V into a 4- Ω input impedance square wave and 64-dB receiver gain was used.

that of the steel reference (Fig. 43). It is quite evident that as the particle size increases, the frequency-dependent attenuation also increases. A similar experiment was performed in which the integrated response (IR) of reflected ultrasound from the test and reference materials surfaces was measured. The reflectivity of ultrasound (as a function of particle size) was determined by comparing the sample

IR with that of a steel reference specimen (Fig. 44). Once again, it is quite clear that as the particle size increases, the ultrasound reflectivity decreases (due to the scattering of ultrasound). By performing these simple experiments, it is relatively easy to characterize the material's surface texture or its roughness.

Noncontact Ultrasonic Imaging

Analogous to the conventional water immersion technique, ultrasonic transducers in the noncontact mode can be raster scanned to generate images corresponding to the internal and surface characteristics of test materials. Figure 45 shows a partial contact ultrasonic image of an impact-damaged 1.5-mm thick multilayered graphite-reinforced plastic composite. The test material in this case was placed on a large stationary transmitting transducer, and a small receiver in the noncontact mode was scanned across the other surface of the material. Figure 46 shows noncontact transmission images of a thick, mild-impact-damaged glass-fiber-reinforced plastic composite by monitoring signals corresponding to transmission and thickness reflection through the material. To demonstrate the analytical ability of the NCA 1000 system further a thick sample of an iron powder compact was imaged by monitoring the transmission integrated response and the material velocity. These images are shown in Fig. 47. Figure 48 shows an image of defects in aluminum. Similar images for materials, such as steel welds, fiber webs, cheese, meats, wood, and other materials have been generated by using the NC transducers with the NCA 1000 and other commercial instruments.

Because the NCA 1000 interprets ultrasound reflections of both the transmitting and the receiving transducers from material surfaces, it is now also possible to measure

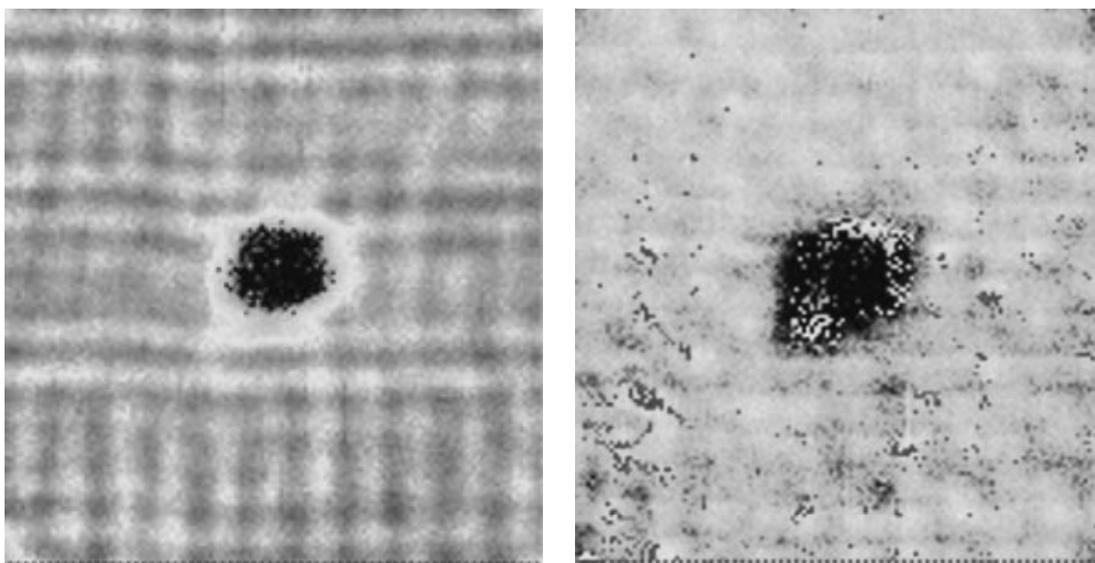


Figure 46. Noncontact ultrasonic imaging of a mild-impact-damaged 6.4-mm thick multilayered glass-fiber-reinforced plastic composite by using NCA 1000 and 1-MHz, 12.5-mm active area diameter transducers with a 1-mm aperture. Left: Direct transmission image. Right: First thickness reflection image.

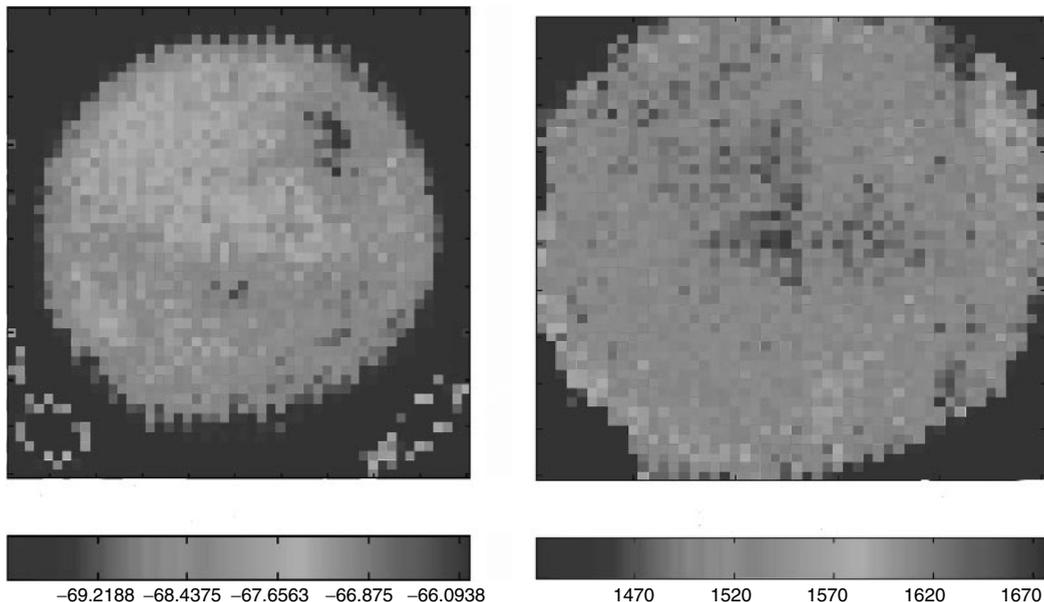


Figure 47. Noncontact imaging of a green 14-mm thick iron powder compact by using the NCA 1000 and 500-kHz, 12.5-mm active area diameter transducers with a 3-mm aperture in direct transmission mode. Left: Relative attenuation of integrated response (dB). Right: Velocity (m/s). Area scanned: 50 × 50 mm. Note that the outer high velocity region is also characterized by high attenuation (low IR) and the inner region of low velocity by lower attenuation.

the thicknesses of materials that are continuously rolled on a production line and are too wide for micrometers.

Food, Beverage, and Pharmaceutical

Figure 49 shows transmitted noncontact ultrasonic signals from regions with and without almonds in milk chocolate. Figure 50 is an example of fat content measurement

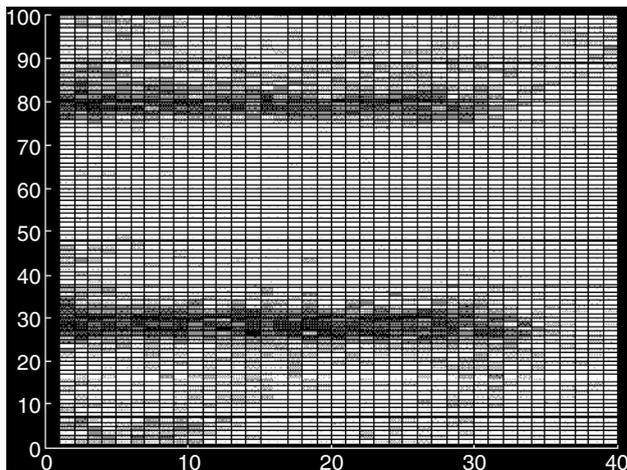


Figure 48. Noncontact ultrasonic image of a 8-mm thick aluminum sheet in transmission mode showing 1.5-mm (top) and 2.0-mm (bottom) side-drilled cylindrical holes. Also compare with Fig. 37. Transducers: 2-MHz and 12.5-mm active area diameter. Image provided by E. Blomme, Katholieke Hogeschool, Kortrijk, Belgium.

in milk and milk products. Similarly, Fig. 51 shows the measurement of sugar content in water. We have also applied noncontact ultrasound to detect beverages and other liquids in plastic, metal, and cardboard containers. The quality of heat and vacuum seals in pharmaceutical

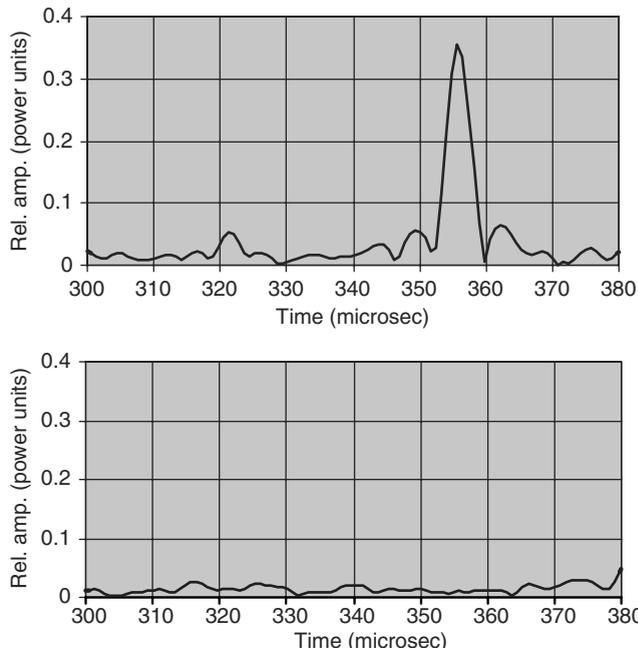


Figure 49. Detection of absence or presence of almonds in milk chocolate. Top: region without almonds. Bottom: region that has almonds. Transducer: 1-MHz, 12.5-mm active area diameter.

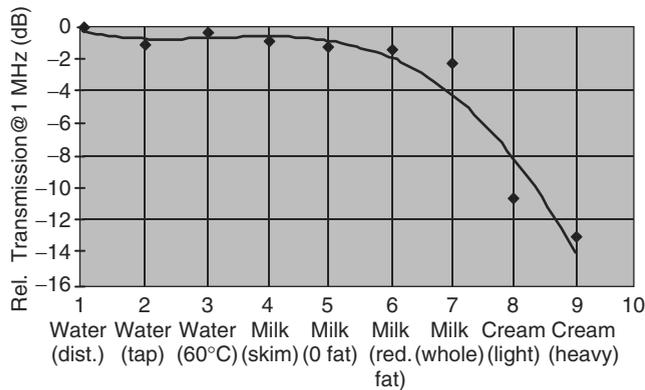


Figure 50. Estimation of fat content in various types of milk and milk products using distilled water as a reference. Ultrasound measurement here is the relative transmission of 1 MHz frequency in samples. Samples in plastic bottles were analyzed in the transmission mode and transmitting and receiving transducers were separated from the bottles by ~ 15 mm of ambient air.

and food packages has also been determined by this technique. Analogous to green ceramics and like material (Figs. 30, 33, and 34), tablets, capsules, and other powder-based pharmaceutical products have also been characterized as functions of the velocity and frequency dependence of ultrasound attenuation. This method has also been successfully applied to detect the presence or absence of liquids (milk, beverages, chemicals) in cardboard cartons and plastic and metal containers. The feasibility of detecting the absence or presence of foreign and unwanted materials in liquid containers has also been successfully demonstrated.

Medical

One of the first medical applications of NC transducers was evaluating burnt skin and bed sores in burn victims (24). By using 2-MHz transducers and a prototype portable ultrasonic pulser-receiver, many observations were made at various points on a healthy and a burnt human hand at the Burn Center, University of California, Irvine,

Figure 51. Estimation of sugar content in water. Ultrasound measurement here is the relative transmission of 1-MHz frequency in samples. Samples in plastic bottles were analyzed in the transmission mode, and transmitting and receiving transducers were separated from the bottles by ~ 15 mm of ambient air.

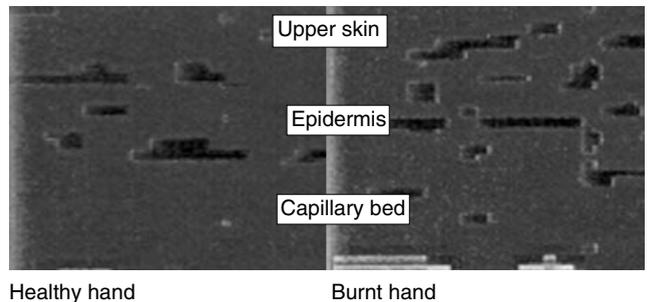
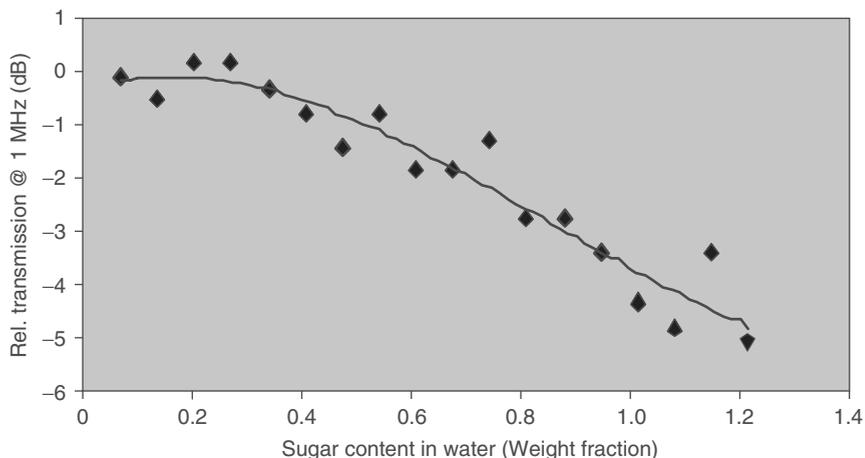


Figure 52. Noncontact ultrasound image cross section of healthy (left) and burnt (right) human hands. Note the damaged region between the epidermis and capillary bed in the burnt hand. The transducer used for this analysis was a noncontact 2-MHz in the reflection (pulse-echo) mode. These images were generated by Joie P. Jones, University of California, Irvine, and his Burn Center team.

under the direction of Joie P. Jones. The NC transducer was used in the reflection (pulse-echo) mode. The collected data were processed to create internal images of two skin conditions (Fig. 52). Detection of damage underneath the burnt skin is evident from the disruption of the interface between the epidermis and the capillary bed. Figure 53 shows the propagation of noncontact ultrasound through a human heel paving the way for bone disease (osteoporosis) diagnostics without any contact with the patient.

Very High Frequency Noncontact Ultrasonic Antennas

Because the sensitivity of NC transducers is very high, it is also possible to use them in a passive mode as “listeners” for very high frequencies. To demonstrate this, we conducted an experiment using the setup shown in Fig. 54. Here, the source of high frequency is 25-mm thick carbon steel, generated by a transducer (in contact with steel) that has a bandwidth at -6 dB from 800 kHz to 8 MHz. This transducer was excited by a single burst of a 16-volt sine wave. A noncontact transducer, nominally 3.5 MHz and 12.5-mm active area diameter, was placed 3 mm from

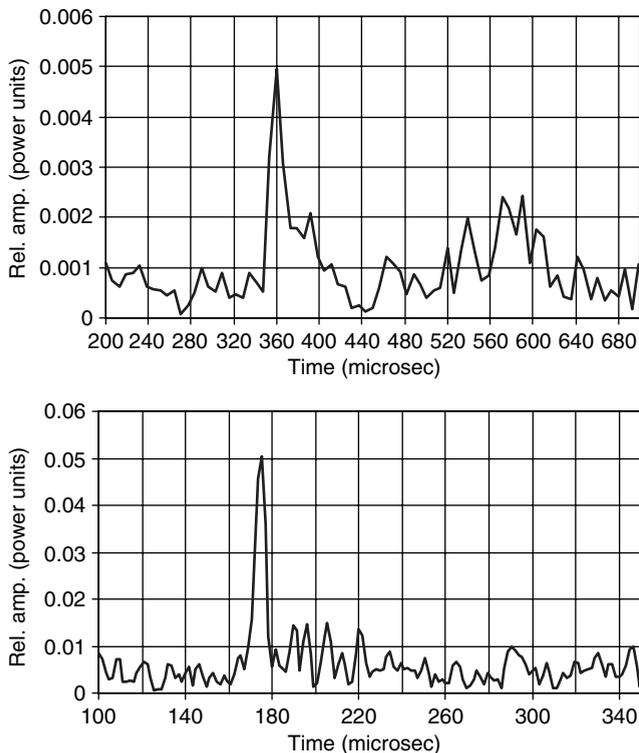


Figure 53. Noncontact ultrasound transmission through a human heel using 250-kHz (top) and 500-kHz (bottom) frequency transducers. The first peak corresponds to ultrasound transmission through air, skin, tissue, and heel bone. Other peaks are not identified.

the material surface in ambient air. The ultrasound received by this transducer was amplified by a 64-dB gain. Figure 55 shows the time and frequency domain of the ultrasound detected (heard) by the NC transducer. By sweeping the frequency across a wide range, the frequency-dependent response from the source (vibrating system) can be investigated and related to its characteristics or condition. In this mode, we successfully interrogated frequencies

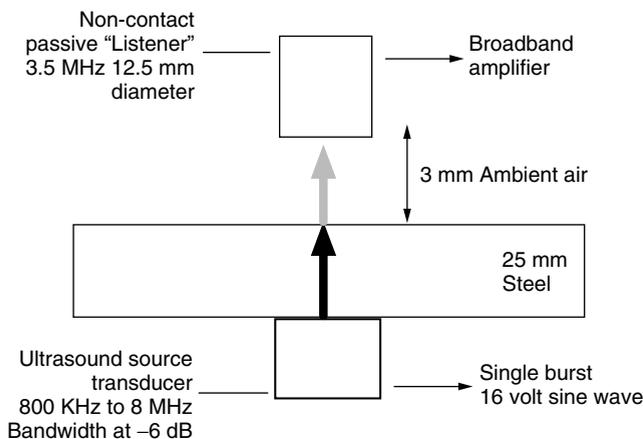


Figure 54. Experimental setup for passive operation of noncontact transducer.

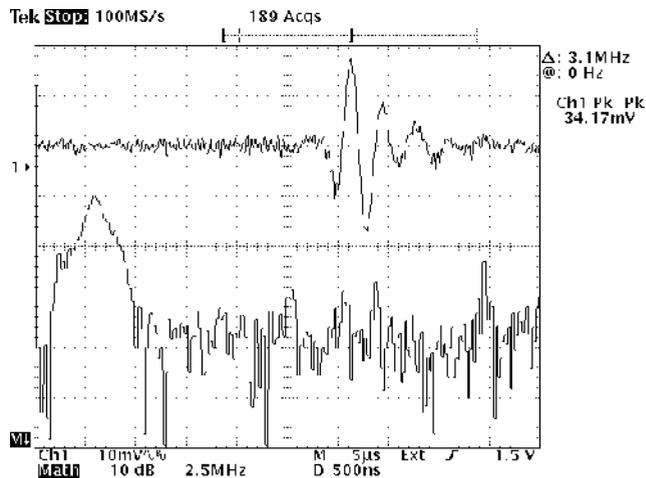


Figure 55. Time and frequency domains of ultrasound detected by noncontact transducer, per Fig. 54 setup.

as high as 7 MHz in ambient air. This opens the door to noncontact acoustic emission, acoustoultrasonics, and any other situation where detection of high frequency ultrasound is desired. Applications of the passive use of NC transducers are dynamics of vibration, materials cutting, testing of railroad, highways, bridges, runways, etc.

Other Noncontact Ultrasound Applications

Besides the applications of NCU described here, this mode can also be used for level detection; dimensional and proximity analysis; high temperature material evaluation; analysis of liquid-sensitive and hazardous material, and analysis of gases and liquids. Finally, it suffices to say that if ultrasound can be propagated through a medium or reflected from an interface, then much information about the medium and the interface can be obtained.

CONCLUSIONS

In this paper, we outlined the significance of ultrasound for nondestructive characterization of materials and for non-invasive diagnostic applications in the medical field. We have also shown the feasibility of noncontact ultrasonic measurements in the time, frequency, and image domains, analogous to other wave-based methods.

Underscoring the significance of the noncontact ultrasound mode, we presented a detailed discussion about the difficulty of achieving this mode. We have also shown that this work ultimately resulted in very high transduction noncontact transducers, thus making the noncontact ultrasound mode a reality. Applications of these transducers in industry and the medical field have been described by using documentary evidence.

We also provided an introduction to a novel ultrasonic noncontact analyzer and its applications for characterizing industrial and biomedical materials and products.

We believe that the noncontact ultrasound mode is among the most significant developments for characterizing and analyzing all states of matter. Though we have

provided selected examples of its applications, there is no doubt that the users of this technology will further enhance its use in materials quality, process control, and health care in our increasingly complex world. This advancement in the field of ultrasound and materials characterization has opened much needed and unprecedented opportunities in research and education.

ACKNOWLEDGEMENTS

The author gratefully acknowledges the assistance of M. Langron, Ultrason Laboratories, in producing the transducers used for this paper. The enthusiastic support and valuable suggestions of E. Blomme, Katholieke Hogeschool, Belgium and M. Landa, Academy of Sciences, Czech Republic, are acknowledged in kind. The work presented in this article was supported by the continuing efforts of SecondWave and Ultrason Laboratories for the advancement of industry and medical science through innovative developments in ultrasound.

BIBLIOGRAPHY

1. J. Curie and P. Curie, *Bull. no. 4 Soc. Mineral. France* **3**:90 (1880), *C.R. Acad. Sci. Paris* **91**:294 (1880).
2. Apparatus for Warning a Ship at Sea of its Nearness to Large Objects Wholly or Partially under Water, *Brit. Pat. Specification* 11,125, March 27, 1913, R.L. Richardson.
3. R.E. Green, in *Materials Analysis by Ultrasonics*, A. Vary, ed., Noyes Data, NJ, 1987, p. 6.
4. Z. Cho, J.P. Jones, and M. Singh, *Foundations of Medical Imaging*. Wiley, NY, 1993, pp. 477–486.
5. R.M. White, *J. Appl. Phys.* **34**: 3559–3567 (1963).
6. A.A. Bondarenko, Y.B. Drobot, and S.V. Kruglov, *Soviet J. NDT* **12**: 655–658 (1976).
7. H.M. Ledbetter and J.C. Moulder, *J. Acoust. Soc. Am.* **65**: 840–842 (1979).
8. A.M. Aindow, R.J. Dewhurst, S.B. Palmer, and C.B. Scruby, *NDT Int.* **17**: 329–335 (1984).
9. G.A. Allers, in *Intelligent Processing of Materials and Advanced Sensors*, H.N.G. Wadley, P.A. Parish, B.B. Rath, and S.M. Wolf, eds., Metallurgical Society, PA, 1986, pp. 17–27.
10. J.A. Brunk, Allied Signal, private communication, 1999.
11. J.A. Brunk, C.J. Valenza, and M.C. Bhardwaj, in *Acousto-Ultrasonics, Theory and Applications*, J.C. Duke, Jr., ed., Plenum Press, NY, 1988, pp. 231–238.
12. M.C. Bhardwaj and A. Bhalla, *J. Mater. Sci. Lett.* **10** (1991).
13. N. Kulkarni, B. Moudgil, and M. Bhardwaj, *Am. Ceram. Soc., Ceram. Bull.* **73**(6): (1994).
14. J.D. Fox, B.T. Khuri-Yakub, and G.S. Kino, *1983 IEEE Ultrasonics Symp.*, 1983, pp. 581–592.
15. T. Yano, M. Tone, A. Fukumoto, *IEEE Trans. UFFC* **34**(2): 222–236 (1987).
16. M.I. Haller and B.T. Khuri-Yakub, *IEEE Ultrasonics Symp.*, 1992, pp. 937–939.
17. D. Reilly and G. Hayward, *IEEE Ultrasonic Symp.*, 1991, pp. 763–766.
18. Ultrasonic Transducer for High Transduction in Gases and Method for Ultrasound NonContact Transmission into Solid Materials, US and international patents pending and in process, 1997–1999, M.C. Bhardwaj.
19. D.W. Schindel, D.A. Hutchins, L. Zou, and M. Sayer, *IEEE Trans. Ultrasonics Ferroelectric Frequency Control* **42**: 42–51 (1995).
20. I. Ladabaum, B.T. Khuri-Yakub, and D. Spoliansky, *Appl. Phys. Lett.* **68**: 7–9 (1996).
21. M. Castaings and B. Hosten, *Ultrasonics* **36**: 361–365 (1998).
22. M. Landa, M.C. Bhardwaj, and I. Neeson, Institute of Thermomechanics, Academy of Sciences of the Czech Republic, Prague, CZ, Report no. Z1266/99 (1999).
23. M.C. Bhardwaj, *Mater. Res. Innovation* **1**: 188–196 (1997).
24. J.P. Jones, D. Lee, M. Bhardwaj, V. Vanderkam, and B. Achauer, *Acoust. Imaging* **23**: (1997).
25. M.C. Bhardwaj, *Proc. Am. Ceram. Soc.* **89**: (1998).
26. T. Carneim, D.J. Green, and M.C. Bhardwaj, *Ceram. Bull.* (1999).
27. B.R. Tittmann, M.C. Bhardwaj, V. Vandervalk, and I. Neeson, *Proc. 23rd Annu. Conf. Composites Adv. Ceram. Mater. Struct.* The American Ceramic Society, Westerville, OH, 1999.
28. M.C. Bhardwaj, I. Neeson, M.E. Langron, and V. Vandervalk, *24th Annu. Conf. Composites Adv. Ceram. Mater. Struct.* The American Ceramic Society, Westerville, OH (2000).
29. R.Y. Vun, Q. Wu, M. Bhardwaj, and G. Stead, *Proc. 12th Int. Symp. Nondestructive Test. Wood*, University of Western Hungary, Sopron, Hungary, 2000.

PAINTS

SHIGENORI EGUSA*
 Japan Atomic Energy
 Research Institute
 Takasaki-shi, Gunma, Japan

INTRODUCTION

Paints are used everywhere in an industrialized society (1,2). The most important functions of paints are protection and decoration of a substrate. Paints can protect substrates against corrosion, oxidative aging, weathering, and mechanical damage and can also provide pleasant color contrasts or a lustrous appearance, hide imperfections in the substrate such as knots in wood, or enhance the beauty of the substrate by using a wood grain. In other words, paints can add to the useful life of materials and also to their attractiveness (1).

Smart paints are an innovative type of paint that has a sensor function as well as the protective and decorative functions of conventional paints. Smart paints can detect abnormal vibration of a structural material by monitoring the natural frequencies and mode shapes of the material. They can also detect damage generated in the material by monitoring the acoustic emission (AE) wave traveling from the damage location to the material surface. Vibration and AE can be monitored in real time, thus enabling health monitoring of the material even during operation.

Smart paints are used in large-scale structures such as vehicles operated at high speeds, civil infrastructures of huge mass and volume, and special facilities that contain large amounts of petroleum, nuclear fuel, and explosive substances. An accident in these facilities can be catastrophic because an enormous amount of energy stored in the form of kinetic, potential, or internal energy is released suddenly by the accident. Smart paints can possibly prevent such a disaster by warning of abnormal vibration and damage generated in a structural material. Hence, one reference goes so far as to say "Brush with disaster—Smart paint warns of impending doom" (3).

The frequency of health monitoring needed for structural materials increases steadily as age increases because the corrosion of steel and concrete progresses gradually during the service period of several decades. Smart paints can be applied to a structural material at any time before and after the construction of the structure, thus making health monitoring quite, easy even for a structure already in active service. Smart paints can make a significant contribution to increasing the service life of a structure, and consequently to saving natural resources.

*Deceased

BASIC CONCEPTS OF SMART PAINTS

The frequency range covered by vibrational measurements is the low-frequency range below ~ 20 kHz (4), whereas that covered in AE wave monitoring is the ultrasonic frequency range above ~ 20 kHz (5). Therefore, if the sensitivity of a smart paint is high enough in both frequency ranges, the paint can be used as a vibrational and AE sensor integrated into a structural material. Such a sensor function of a smart paint is analogous to the action of a sponge that discharges and soaks up water in response to the application and release of external pressure (6). In this analogy, a smart paint is a sponge that repeats the cycle of releasing and drawing an electrical charge at the natural frequency of a structural material or at a frequency of the AE wave traveling through the material.

A smart paint is applied directly to the surface of a structural material when the material is a conductor like metal or carbon fiber composite. In this case, the conducting material can be used as a bottom electrode for the smart paint. When the structural material is an insulator like concrete or ceramic, on the other hand, an electroconductive paint is first applied to the material surface, thus forming a thin conducting layer as a bottom electrode. Then, the smart paint is applied to the surface of the bottom electrode. Whether the structural material is conducting or insulating, an electroconductive paint is applied to the surface of the smart paint film, thus forming a thin conducting layer as a top electrode. Then, a high voltage is applied to the smart paint film using the top and bottom electrodes, thus making the film piezoelectrically active. This poling procedure is usually performed in air at room temperature.

Smart paints are piezoelectric composites that consist of piezoceramic and polymer phases (see Characterization of Piezoelectric Ceramic Materials; Piezoelectricity in Polymers). Thus, smart paints and piezoelectric composites have essentially the same nature with respect to many factors such as the ceramic/polymer composition, the method of preparation, the poling procedure, and the mechanical, electrical, and piezoelectric properties. An essential difference exists in that a piezoelectric composite is used as a discrete point sensor or actuator, but a smart paint is used as a continuously distributed sensor that can cover a large surface area of a structural material.

PIEZOELECTRIC COMPOSITES

Piezoceramics such as barium titanate (BaTiO_3) and lead zirconate titanate (PZT) are typical piezoelectric materials that have excellent properties such as a high electromechanical coupling coefficient and a moderate dielectric constant (7,8). Piezoceramics, however, have the problem that the high density inherent in ceramics makes the specific acoustic impedance much higher than that of water or human tissue, thus causing impedance mismatch (7). Brittleness common to all ceramics is

another drawback of piezoceramics. Piezoelectric polymers such as poly(vinylidene fluoride) (PVDF), on the other hand, do not have the problems of brittleness and impedance mismatch, and furthermore have the excellent property that they can be formed into thin, broad films. However, the electromechanical coupling coefficients and the dielectric constants of piezoelectric polymers are much lower than those of piezoceramics (8).

A solution to these problems is the previously mentioned piezoelectric composites that consist of piezoceramic and polymer phases. The polymer phase in the composites increases the composite toughness and also decreases the composite density and dielectric constant, thus solving the problems of piezoceramics and piezoelectric polymers simultaneously (9–11). The electrical and mechanical properties of piezoelectric composites are determined primarily by the fraction of the piezoceramic and polymer phases and by the properties of these constituent materials (12–14). Composite properties are affected also by the connectivity pattern of the piezoceramic and polymer phases (15–20).

COMPOSITION OF SMART PAINTS

The smart paints reported so far are piezoelectric composites made up of piezoceramic particles dispersed in a polymer matrix. The polymer matrix need not be piezoelectrically active, and hence popular polymers such as alkyd, acrylic, and epoxy resins can be used as the matrix resin. The preparation of smart paints and the application procedures are essentially the same as those of conventional paints, except for poling for a dried film of smart paint. As a result, most of the fundamental characteristics and functions of conventional paints are imparted to smart paints, thus enabling smart paints to have protective, decorative, and sensor functions simultaneously.

Smart paints can form continuous paint films covering a large surface area of a structural material. Because of the electrically insulating nature of the paint film, however, the electrical charge actually detected is only that generated in a region that has an electrode on the surface of the paint film. Therefore, if a set of separate electrodes is formed on the paint film surface, the electrical charge generated in each region can be detected and analyzed separately. This feature of smart paints enables the application of the paints as a vibrational modal sensor that can determine the natural frequencies and mode shapes of a structural material (21,22). Furthermore, this feature enables another application of smart paints as an AE sensor that can determine the damage location in a structural material quite easily without using the conventional technique based on the arrival time difference of an AE wave (5).

Paints in general can be applied to all kinds of materials such as metals, composites, concrete, and ceramics; the material surface can be flat, curved, or even irregularly shaped. Furthermore, paints can be applied and reapplied at any time, when necessary. Final dry films of paints are generally light, flexible, and tough. These excellent properties of paint in general are imparted to smart paints as well, thus giving the smart paints further useful features

as vibrational and AE sensors integrated into a structural material.

FORMATION OF SMART PAINT FILMS

Paint Preparation, Application, and Curing

Paints in general are made up of three components: pigment, binder, and volatile liquid (1,2). The volatile liquid is a solvent or a nonsolvent that provides a practical viscosity for packaging and application and does not normally become part of the dried paint film. The binder is a film-forming substance which is mostly a polymeric material such as alkyd, acrylic, or epoxy resin. The binder is used as a solution in a solvent or as a dispersion of fine particles in a nonsolvent. Such a solution or dispersion is called a vehicle. Paint pigments are finely divided, insoluble, solid particles such as titanium dioxide (TiO_2), zinc oxide (ZnO), and calcium carbonate (CaCO_3). The pigment particles are dispersed stably in the paint vehicle before application and the pigment particles are distributed uniformly throughout the binder resin in the dried paint film. The decorative functions of a paint are due, for the most part, to the pigment.

The basic components of smart paints are essentially the same as those of conventional paints, except that piezoceramics such as PZT and BaTiO_3 are used as pigments in smart paints. The piezoceramics used in the smart paints so far are PZT (23–30) and lead titanate (PbTiO_3) (23), and the binders used are acrylic resin (23), polyurethane (23), and epoxy resin (25–29). Smart paints made up of these components are prepared by essentially the same procedure as used for conventional paints. Smart paints are applied by using familiar coating tools such as brushes, rollers, or spray guns. Smart paints are also cured in the usual way in air at ambient temperature or at elevated temperatures.

Electrode Formation and Poling

A simple method for forming an electrode on the surface of a paint film is to apply an electroconductive paint by using a coating tool such as a brush or roller. A more elaborate method is to deposit a vapor of gold or aluminum onto the paint film surface (30). A screen mask technique is also effective for this purpose, especially when the electrode pattern is complicated. The main advantage of this technique is that leads as well as electrodes can be printed on the paint film surface, as shown in Fig. 1. This technique, however, has the disadvantage that it cannot be used for large structures such as airplanes, trains, or bridges.

For such large structures, an ordinary coating method by brush, roller, etc. may be the most practical for forming an electrode on the paint film surface. As a lead for the electrode, on the other hand, a thin electrical wire or tape $\sim 50 \mu\text{m}$ thick or so may be the most practical choice for a large structure because such a thin wire or tape is comparable in thickness to a paint film and hence, can be buried in the paint film or under a topcoat. Note that when smart paints are put into practical use, the electrodes and leads are covered by a topcoat, thus making the appearance exactly the same as that of conventional paints.

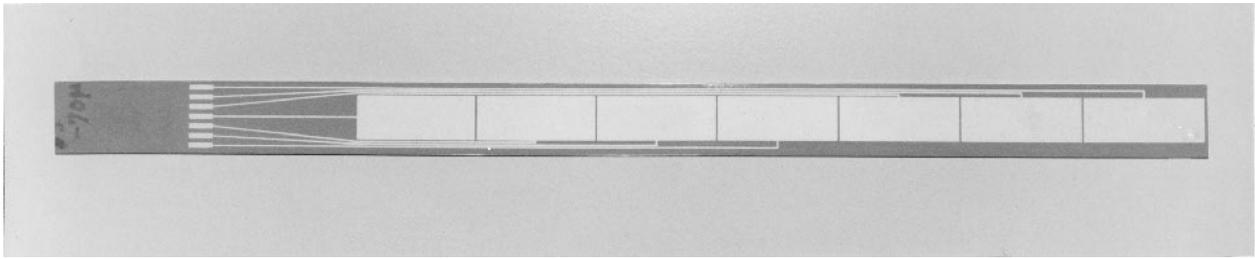


Figure 1. Electrodes and leads printed on a PZT/epoxy paint film formed on one surface of an aluminum beam. The left end of the beam where the leads come together is wrapped in an electrically insulating material. The aluminum beam is clamped at this section for vibrational measurements.

Piezoelectric composites are usually poled in an oil bath at elevated temperatures because poling at a higher temperature achieves saturation poling in a lower poling field. For smart paints, on the other hand, poling is done in air at room temperature because even room temperature poling can achieve high enough piezoelectric activity for the paint application to serve as vibrational and AE sensors integrated into a structural material (25–29).

EVALUATION OF SMART PAINT FILMS

The sensor function of smart paints relies heavily on the piezoelectric activity of the poled paint film. Usually, the activity is expressed in terms of a piezoelectric constant which is the ratio of the charge developed per unit surface area or the voltage developed per unit film thickness to the stress or strain applied externally. The charge-to-stress, voltage-to-stress, charge-to-strain, and voltage-to-strain ratios are the piezoelectric constants d , g , e , and h , respectively (7).

Piezoelectric materials are inherently anisotropic, and hence two subscripts are attached to the piezoelectric constant to describe the anisotropic properties. The first subscript is used to indicate the direction of the charge or voltage development, and this is always the film thickness direction for a piezoelectric film such as PVDF or a smart paint film. The second subscript is used to indicate the direction of the stress or strain applied externally, and this direction is any of the 1, 2, and 3 axes of the film which correspond to the length, width, and thickness directions, respectively (7).

Sensitivity as a Vibrational Sensor

When a structural material is deformed, strain is developed in all directions of the material, including the direction tangent to the material surface. This is also true when the structural material is vibrating. For a smart paint used as a vibrational sensor, therefore, one of the most important sensitivities to be evaluated is the piezoelectric constant e_{31} because this constant is the ratio of the charge per unit surface area to the strain in the direction tangent to the paint film surface.

The e_{31} constant is evaluated from vibrational measurement on a cantilever beam like that shown in Fig. 1. A typical example of the measurement is shown in Fig. 2

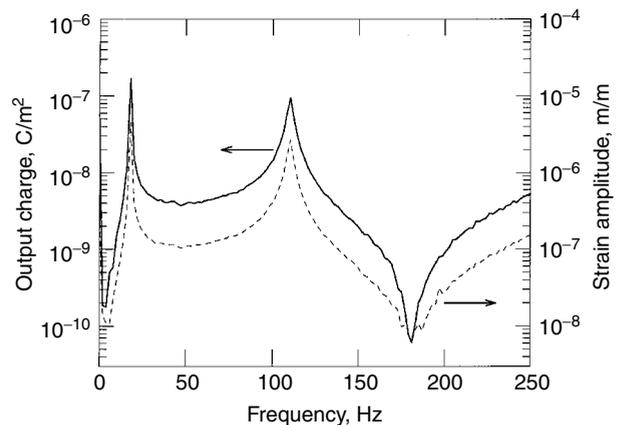


Figure 2. Frequency spectra of output signals from a PZT/epoxy paint film formed on one surface of an aluminum beam and from a strain gauge bonded to the opposite surface of the beam.

for a paint film which has the PZT/epoxy composition of 53/47 by volume and is formed on the surface of an aluminum beam 3.0 mm thick, 30 mm wide, and 460 mm long (350 mm long as a cantilever beam) (27). This example is for a 109- μm thick paint film cured at room temperature and poled at 240 kV/cm for 5 min. The spectrum shape obtained from the paint film is similar to that obtained from a strain gauge which is bonded to the opposite surface of the beam to monitor the strain developed in the direction of the cantilever length. Then, the e_{31} constant is evaluated from the charge-to-strain ratio at a natural frequency of 18 or 112 Hz.

The e_{31} constant thus evaluated depends on many factors such as the poling field, the film thickness, the cure temperature, and the PZT/epoxy composition (26,27). A typical example of the poling-field and film-thickness dependence is shown in Fig. 3 for paint films cured at room temperature that have the PZT/epoxy composition of 53/47 by volume (27). The e_{31} constant increases steadily as the poling field increases for all of the paint films shown here, and saturation poling is not achieved, even at a high poling field of ~ 150 kV/cm. The e_{31} constant obtained at a particular poling field, say, 100 kV/cm, increases as film thickness increases from 33 to 152 μm , thus exhibiting a clear film-thickness dependence. This point is further described later.

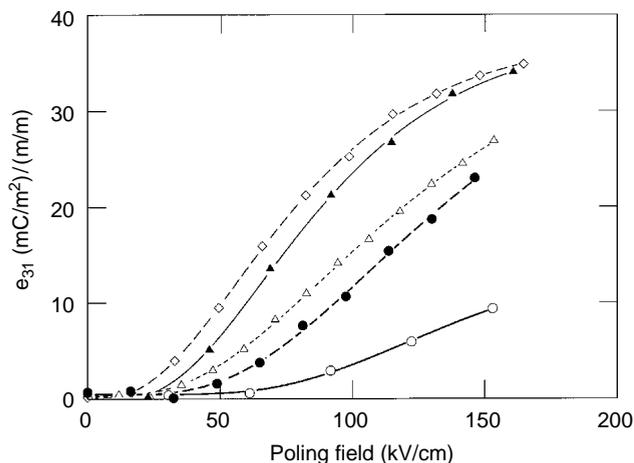


Figure 3. Plots of the piezoelectric constant e_{31} vs. the poling field for PZT/epoxy paint films cured at room temperature and evaluated as a vibrational sensor.

Sensitivity as an Acoustic Emission Sensor

In many cases, eventual failure of a structural material occurs after a certain amount of damage accumulates within the material. The generation of such damage is almost always accompanied by the emission of an AE wave, and hence the damage generated and accumulated can be detected by monitoring the AE wave (5). The AE wave is emitted in all directions, and consequently, an AE wave that arrives at the material surface and enters the smart paint film on the material surface always exists. Furthermore, an AE wave that enters the paint film nearly perpendicularly always exists. Such an AE wave develops strain in the paint film in the direction normal to the film surface because the AE wave is a compression wave in which particle motion is in the same direction as the propagation of the wave. For a smart paint used as an AE sensor, therefore, the sensitivity to be evaluated is the piezoelectric constant h_{33} because the h_{33} constant refers to the ratio of the voltage per unit film thickness to the strain in the direction normal to the paint film surface.

For a conventional AE sensor, the sensitivity s is usually given by $s = V/v_0$, where V is the output voltage of the sensor and v_0 is the velocity amplitude of AE waves (31). The strain amplitude of AE waves ϵ_0 is given by $\epsilon_0 = v_0/v$, where v is the phase velocity of AE waves. Combining these equations with $h_{33} = (V/d)/\epsilon_0$ leads to $s = h_{33}d/v$, where d is the film thickness. This equation indicates that the paint film sensitivity as an AE sensor s is independent of the frequency of AE waves and that the sensitivity increases linearly as film thickness increases. This equation also indicates that the h_{33} constant is calculated from $h_{33} = sv/d$.

The paint film sensitivity as an AE sensor is evaluated from measurement using an ultrasonic transducer to produce AE waves and a laser Doppler vibrometer to monitor the velocity amplitude of the AE waves (28). A typical example of the measurement is shown in Fig. 4 for a paint film that has the PZT/epoxy composition of 53/47 by volume and is formed on the surface of square aluminum plate

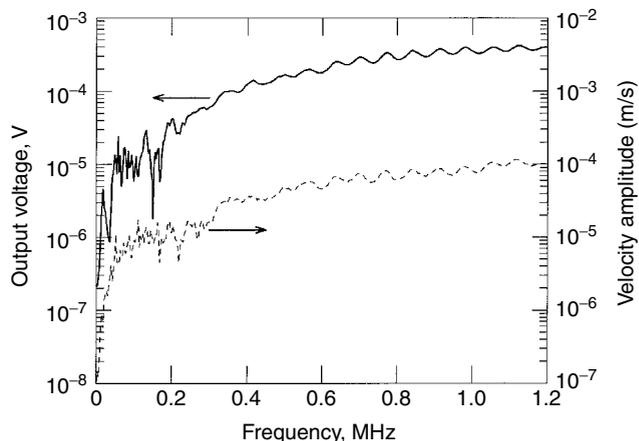


Figure 4. Frequency spectra of output signals from a PZT/epoxy paint film formed on one surface of an aluminum plate and from a laser Doppler vibrometer that monitors the velocity amplitude of AE waves.

0.2 mm thick that has 50 mm sides. This example is for a 152- μ m thick paint film cured at room temperature and poled at 184 kV/cm for 5 min. The spectral shape obtained from the paint film is similar to that obtained from the laser vibrometer in the frequency range above ~ 0.3 MHz. Such a similarity of spectral shapes reflects a nearly flat frequency response of the paint film to AE waves. Then, the paint film sensitivity as an AE sensor is evaluated from the average ratio of the output voltage of the paint film to the velocity amplitude of AE waves in the frequency range 0.3–1.0 MHz.

The paint film sensitivity thus evaluated, s can be converted into the h_{33} constant by using the relationship $h_{33} = sv/d$, where v is the phase velocity of AE waves in the PZT/epoxy paint film. The h_{33} constant calculated by using an assumed value of $v = 2850$ m/s (6) is plotted in Fig. 5 as a function of film thickness for paint films cured at room-temperature that have the PZT/epoxy composition of 53/47

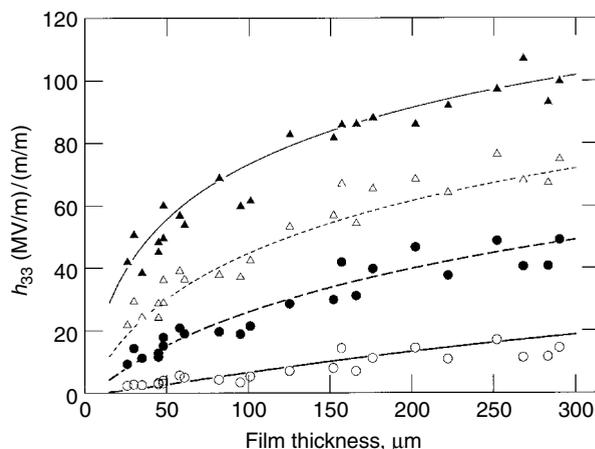


Figure 5. Plots of the piezoelectric constant h_{33} at 50 (\circ), 100 (\bullet), 150 (Δ), and 250 kV/cm (\blacktriangle) vs. film thickness for PZT/epoxy paint films cured at room temperature and evaluated as an acoustic emission sensor.

by volume (28). It is seen that the h_{33} constant obtained at a poling field of 50, 100, 150, or 250 kV/cm increases steadily as film thickness increases, thus exhibiting a clear film-thickness dependence. Such a film-thickness dependence is also observed for the e_{31} constant shown in Fig. 3.

FACTORS DETERMINING POLING BEHAVIOR OF SMART PAINT FILMS

The poling behavior of a PZT/epoxy paint film depends on the film thickness, as shown in Figs. 3 and 5. Furthermore, the poling behavior also depends on the cure temperature and the PZT/epoxy composition (26–29). Such complicated poling behavior is virtually determined by the electric field that acts on the PZT particles dispersed in the epoxy matrix. The most important factors that determine the electric field and, consequently, the poling behavior of the paint film are the electrical conductivities of the PZT particles and the epoxy matrix, the connectivity pattern of the PZT phase, and the space charge accumulated at the PZT/epoxy interface.

Electrical Conductivities of Constituent Materials

It is now well established that in poling a composite specimen made of piezoceramic particles dispersed in a polymer matrix, the electric field that acts on the ceramic particles is very low compared with that applied externally to the composite specimen (14,32). This occurs because the electrical conductivity of polymeric materials in general is much lower than that of ceramic materials, and hence the polymer matrix in the composite specimen bears almost all of the externally applied electric field at the expense of the electric field that acts on the ceramic particles. As a result, the piezoelectric activity of the ceramic/polymer composite specimen is very low, compared with a pure piezoceramic specimen poled in the same electric field. This idea explains why saturation poling is not achieved, even in a high poling field of ~ 150 kV/cm, as seen in Fig. 3. Saturation poling for a pure PZT ceramic specimen, on the other hand is achieved in a low poling field of ~ 10 kV/cm (12).

A promising solution to this problem is to increase the electrical conductivity of the polymer matrix up to that of the ceramic particles, so that the electric field distribution becomes uniform throughout the composite specimen. This can be achieved by adding a small amount of a semiconductor filler such as carbon, germanium, or silicon to the composite specimen (32). This can also be achieved by poling at a high temperature where the electrical conductivity of the polymer matrix becomes equal to that of the ceramic particles (33).

Connectivity Pattern of Ceramic Phase

Figure 6 is a scanning electron microscopy (SEM) picture that shows the internal microstructure of a paint film that has the PZT/epoxy composition of 53/47 by volume (27). It is seen that the size of PZT particles ranges from ~ 0.5 to ~ 1.5 μm , and that a substantial fraction of the PZT particles are in contact with each other, so that the PZT phase

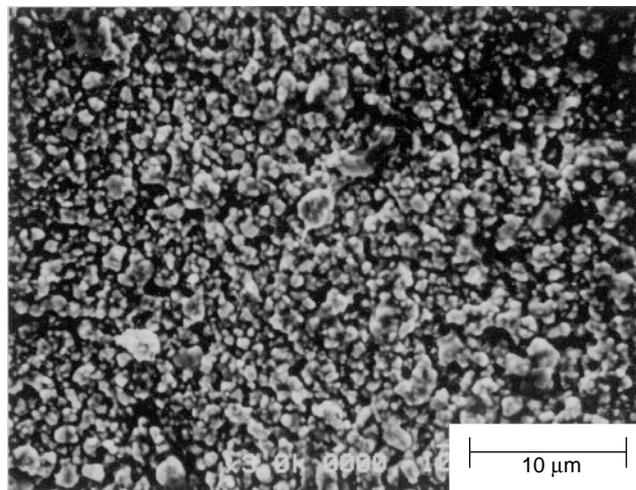


Figure 6. SEM picture of a paint film that has the PZT/epoxy composition of 53/47 by volume. This example is a 49- μm thick paint film cured at 150°C.

is practically self-connected in three dimensions. The self-connectivity of the PZT phase is one of the most important factors that determines the poling behavior of a PZT/epoxy paint film. In fact, the paint film is hardly poled when the PZT volume fraction is decreased to such a level that the PZT particles are isolated from one another by the continuous phase of the epoxy matrix (26).

Figures 3 and 5 show that the poling behavior of a PZT/epoxy paint film depends on the film thickness even when the PZT volume fraction remains constant at 53%. A SEM picture like that shown in Fig. 6, however, detects no observable difference in the PZT phase connectivity for paint films that have different thicknesses. The difference in the PZT phase connectivity is reflected much more explicitly in the current–voltage characteristic of the paint film rather than in the SEM picture, as described here.

Space Charge at the Ceramic/Polymer Interface

The current–voltage characteristic of a PZT/epoxy paint film shows that the conduction is ohmic in a low electric field, whereas in a high electric field, the space-charge-limited (SCL) conduction predominates over ohmic conduction (28). Furthermore, the current–voltage characteristic shows that the critical electric field at which the ohmic-to-SCL transition takes place decreases as the film thickness decreases. The result is that conduction during the poling process is mostly SCL for a thin film, whereas conduction is mostly ohmic for a thick film.

The SCL conduction becomes predominant when a space charge of electrons is injected into the PZT/epoxy paint film during the poling process. The space charge has a tendency to build up preferentially at the interface between the PZT and epoxy phases in the paint film (28). The space charge decreases the electric field acting on the PZT phase, and hence decreases the piezoelectric activity of the paint film obtained in a given poling field. This effect of the space charge becomes significant, particularly for a thin film, because SCL conduction becomes more

predominant as the film thickness decreases. Therefore, the film-thickness dependence of the piezoelectric constant shown in Figs. 3 and 5 is ascribed to the space charge of electrons injected into the paint film during the poling process.

The fact that the current-voltage characteristic of a PZT/epoxy paint film depends on the film thickness is closely related to the drying rate of the wet paint film. In fact, it is well known that the thickness of a wet paint film has a significant influence on the rate of solvent evaporation and, consequently, on film formation during curing (3). Thus, it is quite possible that the degree of self-connectivity of the PZT phase depends on the thickness of the dried paint film. Therefore, the drying rate of the wet paint film is another important factor that determines the poling behavior of a PZT/epoxy paint film.

TECHNIQUES FOR APPLYING SMART PAINT FILMS

Techniques for applying smart paint films as vibrational and AE sensors are essentially the same as those for a PZT ceramic or PVDF film bonded to the surface of a structural material. Theories, models, methods, and systems constructed for use of the PZT and PVDF sensors (21,22,34) can also be applied to smart paint films used as vibrational and AE sensors integrated into a structural material.

Vibrational Modal Sensor

One example of an application of smart paints is a vibrational modal sensor integrated into a structural material. As noted before, the sensitivity of the paint film used for this purpose is the e_{31} constant which is the ratio of the charge per unit surface area to the strain in the direction tangent to the paint film surface. Figure 7 shows a result of vibrational modal testing of a cantilever beam like that shown in Fig. 1 by using a PZT/epoxy paint film

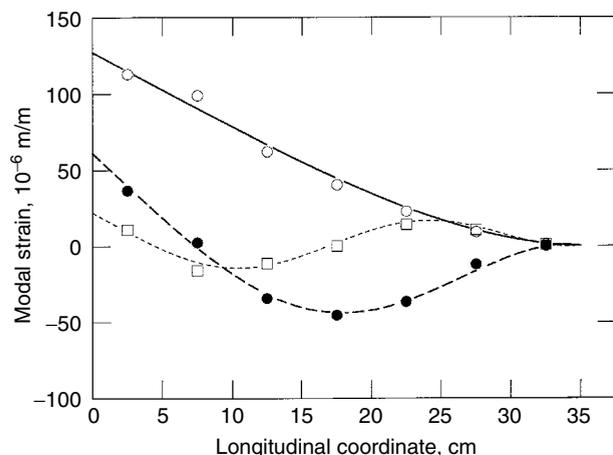


Figure 7. Modal strain shapes of a cantilever aluminum beam for the first (\circ), second (\bullet), and third modes (\square) determined by a PZT/epoxy paint film formed on the beam surface.

that has an e_{31} constant of 9.0×10^{-3} (C/m²)/(m/m) (26). A set of vibrational measurements is carried out for all of the electrodes formed on the paint film surface: an identical excitatory force is applied at a fixed point on the cantilever beam. Then, the output charge of the paint film at each electrode is converted into the strain using the e_{31} constant and is plotted against the distance from the clamped end of the beam to the center of each electrode. The modal strain shapes thus obtained are shown in Fig. 7 for the first three modes at 18, 112, and 315 Hz.

It is worth noting that the modal strain shapes shown in Fig. 7 can be converted into modal displacement shapes by $d^2\phi/dx^2 = -\varepsilon/\eta$, where ϕ is the transverse displacement of a uniform cantilever beam, x is the longitudinal coordinate of the beam, ε is the longitudinal strain in the beam surface, and η is the half-thickness of the beam (35). Modal displacement shapes determined by this equation are identical to those determined by a laser Doppler vibrometer that measures the transverse movement of the beam surface (26). Thus, smart paints offer an interesting and promising alternative to conventional sensors such as accelerometers and laser vibrometers (1).

FUTURE DIRECTIONS

Smart Paints

The highest sensitivity of smart paint films achieved so far is $e_{31} = \sim 40 \times 10^{-3}$ (C/m²)/(m/m) as a vibrational sensor and $h_{33} = \sim 100 \times 10^6$ (V/m)/(m/m) as an AE sensor, as shown in Figs. 3 and 5. For commercially available PVDF films, the sensitivity is $e_{31} = \sim 66 \times 10^{-3}$ (C/m²)/(m/m), $e_{32} = \sim 6.8 \times 10^{-3}$ (C/m²)/(m/m), and $h_{33} = \sim 50 \times 10^6$ (V/m)/(m/m), determined in essentially the same way described before for smart paint films. This indicates that the sensitivity of smart paint films is comparable to that of PVDF films. So far as sensitivity is concerned, therefore, smart paints have already reached a level suitable for practical use.

For smart paints to be put into practical use, however, the paints must meet performance requirements such as exterior durability and sensitivity stability. Exterior durability is the paint films resistance to environmental factors such as uv radiation, heat, moisture, oxygen, and ozone (2). These environmental factors can cause mechanical degradation of paint films, thus leading to the failure of the protective and decorative functions of smart paints. These environmental factors may also cause electrical degradation of paint films, thus leading to the failure of the sensor function of smart paints. Considering that smart paints are truly appreciated when used in severe and isolated environments, the evaluation of exterior durability and sensitivity stability is absolutely necessary for the paints to be put into practical use.

Smarter Paints

According to a concept of intelligent materials in Japan, the intelligence in materials is classified into three categories; intelligence from the human standpoint, intelligence inherent in materials, and intelligence at the most

primitive levels in materials (36). The intelligence from the human standpoint is a relative concept based on the value of a material and its utility in relation to all aspects of society such as economy, conservation of resources, intensiveness of information, human friendliness, reliability, harmony with the environment, and optimum life span.

Water-borne piezoelectric paints are smarter paints from the standpoint of harmony with environment (37). A paint that can spontaneously become a piezoelectric film after the usual drying process will also be a smarter paint from the standpoint of human friendliness. In fact, poling a paint film at a high voltage is dangerous work and should be avoided if possible. A feasibility study of a poling-free piezoelectric paint shows that a paint made of PVDF particles and epoxy resin does not need poling for the final dry film to be piezoelectrically active (38). At the present stage, however, the piezoelectric activity is not enough for practical use of the paint film. Studies are currently under way to increase the piezoelectric activity of the paint film.

From the standpoint of intensiveness of information, a smarter paint of the future will have a sensor function for material conditions such as vibration and damage generation and also for atmospheric variables such as temperature, pressure, moisture, and wind velocity. Such a paint resembles human skin in that the skin has a sensor function for the external stimuli imposed on the human body and also for the surrounding conditions such as temperature, humidity, wind, and rain. The ultimate goal of smart paints, therefore, should be to mimic the human skin as closely as possible.

ACKNOWLEDGMENTS

The work in smart paints by S. Egusa and N. Iwasawa was supported by the Japan Atomic Energy Research Institute through the Special Program for Fundamental Researches (1991–1994) and through REIMEI Research Resources (1998).

BIBLIOGRAPHY

- J.H. Lowell, in *Coatings*, J.I. Kroschwitz, ed., *Encyclopedia of Polymer Science and Engineering*, 2e., Wiley-Interscience, NY, 1985, Vol. 3, pp. 615–675.
- Z.W. Wicks, Jr., in *Coatings*, J.I. Kroschwitz, ed., *Encyclopedia of Polymer Science and Engineering*, 2e., Wiley-Interscience, NY, 1989, Supplement Vol. pp. 53–122.
- O. Graydon, *New Scientist*, p. 20, October 17, 1998.
- D.J. Ewins, *Modal Testing: Theory and Practice*. Research Studies Press, Taunton, 1984.
- C.B. Scruby, *J. Phys. E: Sci. Instrum.* **20**: 946–953 (1987).
- KYNAR Piezo Film Technical Manual, Pennwalt Corporation, Valley Forge, PA, 1987, p. 6.
- A.J. Moulson and J.M. Herbert, *Electroceramics*. Chapman & Hall, London, 1990, Chap. 6.
- M.V. Gandhi and B.S. Thompson, *Smart Materials and Structures*. Chapman & Hall, London, 1992, Chap. 5.
- T. Kitayama and S. Sugawara, *Proc. Gr. Inst. Electr. Comm. Eng. Jpn.*, 1972, CPM 72-17 (in Japanese).
- L.A. Pauer, *IEEE Conf. Res.*, pp. 1–5 (1973).
- W.B. Harrison, *Proc. Workshop Sonar Transducer Mater.* Naval Research Laboratories, November 1975, p. 257.
- T. Furukawa, K. Fujino, and E. Fukada, *Jpn. J. Appl. Phys.* **15**(11): 2119–2129 (1976).
- T. Furukawa, K. Ishida, and E. Fukada, *J. Appl. Phys.* **50**(7): 4904–4912 (1979).
- T. Furukawa, K. Suzuki, and M. Date, *Ferroelectrics* **68**: 33–44 (1986).
- H. Banno and S. Saito, *Jpn. J. Appl. Phys.* **22** (Supplement 22-2): 67–69 (1983).
- H. Banno, *Ferroelectrics* **50**: 3–12 (1983).
- R.E. Newnham, D.P. Skinner, and L.E. Cross, *Mater. Res. Bull.* **13**: 525–536 (1978).
- R.E. Newnham, L.J. Bowen, K.A. Klicker, and L.E. Cross, *Mater. Eng.* **2**: 93–106 (1980).
- R.E. Newnham, *Ferroelectrics* **68**: 1–32 (1986).
- R.E. Newnham and G.R. Ruschau, *J. Am. Ceram. Soc.* **74**(3): 463–480 (1991).
- C.-K. Lee and F.C. Moon, *J. Appl. Mech.* **57**: 434–441 (1990).
- S.A. Collins, D.W. Miller, and A.H. von Flotow, *Sensors for Structural Control—Applications Using Piezoelectric Polymer Film*. Space Engineering Research Center #12-90, Massachusetts Institute of Technology, Cambridge, MA, 1990.
- K.A. Hanner, A. Safari, R.E. Newnham, and J. Runt, *Ferroelectrics* **100**: 255–260 (1989).
- C.A. Rogers and S.C. Stein, *Proc. 1st Int. Conf. Intelligent Mater.* 1993, pp. 87–93.
- S. Egusa and N. Iwasawa, *Proc. 1st Int. Conf. Intelligent Mater.* 1993, pp. 101–104.
- S. Egusa and N. Iwasawa, *J. Mater. Sci.* **28**: 1667–1672 (1993).
- S. Egusa and N. Iwasawa, *Ferroelectrics* **145**: 45–60 (1993).
- S.S. Egusa and N. Iwasawa, *J. Appl. Phys.* **78**: 6060–6070 (1995).
- S. Egusa and N. Iwasawa, *J. Smart Mater. Struct.* **7**: 438–445 (1998).
- J.M. Hale and J. Tuck, A Novel Strain Transducer Using Piezoelectric Paint. *Proc. Mech. Eng.* in press.
- ASTM E1106-86, *Standard Method for Primary Calibration of Acoustic Emission Sensors*. American Society for Testing and Materials, Philadelphia, PA, 1986, pp. 489–498.
- G. Sa-Gong, A. Safari, S.J. Jang, and R.E. Newnham, *Ferroelectrics Lett.* **5**: 131–142 (1986).
- J.P. Dougherty and Y. Chen, *Proc. 2nd Int. Conf. Intelligent Mater.* 1994, pp. 462–473.
- C.A. Rogers, ME 4016, Virginia Polytechnic Institute and State University, Blacksburg, VA (private communication, 1991).
- S.H. Crandall, N.C. Dahl, and T.J. Lardner, *An Introduction to the Mechanics of Solids*. McGraw-Hill, NY, 1972, p. 628.
- T. Takagi, *Proc. Int. Workshop Intelligent Mater.*, Tsukuba, Japan, 1989, pp. 1–10.
- J.M. Hale, University of Newcastle, Newcastle, England (private communication, 1999).
- S. Egusa, 1998 REIMEI Conf., Japan Atomic Energy Research Institute, Tokai, Japan, July 14–15, 1999.

PEST CONTROL APPLICATIONS

SHERRY DRAISEY
Good Vibrations Engineering, Ltd
Nobleton, Ontario, Canada

INTRODUCTION

The smart aspects of the piezoceramic ultrasonic application being used for pest control are just beginning to evolve. Pest control, using ultrasonics, is based on developing a pressure environment which is extremely unpleasant or deadly to the pests in question. The feedback aspect of smart structure applications involves three types of sensing:

- motion sensors (designed to power up the ultrasonic device when large pest groups have been detected)
- pressure sensors (these are used in fluid media to sense if pressure levels have risen enough to generate structural instability)
- sound sensors (for antinoise generation to stop the sound from being externally transmitted) that coordinate the antinoise generation

Airborne or land pests, such as some insects, spiders, rodents, and small cats and dogs are driven away by the unpleasant sound created by the noise generated by the ceramic elements. For fluid-borne pests, the ceramic is driven to create a pressure field that includes cavitation. The release of energy from the collapse of cavitating bubbles provides the source deadly to small microorganisms.

Table 1 lists the types of pests that have been effectively deterred by ultrasonic measures. The table lists the frequency range that has been successful for these pests, as well as the approximate coverage (or flow rate) across which they are effective. The coverage is directly related to the system size and power.

The Environmental Protection Agency (EPA) has suggested that pest control devices have a deterrent effect of >60% to be considered viable.

SOUND AS A PEST DETERRENT

The control of airborne and land pests is based on generating high-frequency noise. This is done to disturb and confuse the species, making the environment generally unpleasant. The sound levels are in the range of 90+ dB at 1 meter from the source.

The concept behind ultrasonic pest control is to alter the behavior patterns of the pests to the extent that they are forced to leave the area. Some devices have been designed for operation within buildings, others for outdoors. Versions of the devices target specific pest groups (mice), and more sophisticated versions have settings that allow selecting particular pest groups.

The power supplies for the designs varies from plug-in wall units (110 or 220/240 V) to battery operated systems.

Motion sensors are used for detecting larger size pests. This reduces power consumption and eliminates unnecessary noise pollution.

Test Results

The test data presented here were provided by the Weitech company, a manufacturer of a variety of ultrasonic deterring devices designed to produce ultrasonic sound in air.

Mosquitoes. At least one company's test results of the high-frequency ultrasonic deterrent effect on mosquitoes has suggested that it does not meet the EPA suggested deterrent level.

Small Rodents. The available test results (1) for small rodents depend on the particular rodent. Two types of rodents are considered. For each test set, there were six rodents in the sample—three males and three females. They were housed in two adjoining chambers, one exposed to the ultrasonic sound (~90 dB), the other at much lower noise levels (30 to 35 dB or lower).

Two parameters are used to evaluate the influence of the ultrasound—food consumption (measurement of the daily food consumption in the treated and untreated chambers) and activity (animal track evidence in the treated and untreated chambers). Before the introduction of ultrasonic treatment, healthy mice that had good hearing (hearing test—Preyer's reflex, a reaction to loud noise) are housed in the two chambers, and their activity and food consumption levels are measured.

The effect of the ultrasonic deterrent on the Norway rat (*Rattus norvegicus*) is more pronounced than on wild house mice (*Mus musculus*). The average weight of the Norway rats in the test was 237 grams (8.4 oz). The average weight of the wild house mice was 17 grams (0.6 oz). The results are shown in Figs. 1 and 2 as an index (the ratio of the treated measurements to the total measurements). Food consumption influence is shown in black bars, and tracking activity is shown in gray.

Table 1. Pests Effectively Controlled by Ultrasonic Devices

Pest	Frequency Range	Coverage (varies with power consumption)
Dogs, cats, skunks	14–25 kHz	278.8 m ² (4000 ft ²)
Mice	26–50 kHz	46.4 sq m ² (500 ft ²)
Moths	40 kHz	5.7 m ³ (200 ft ³)
Rodents, spiders, some insects	26–42 kHz	74.3 m ² (800 ft ²)
Microorganisms	23 kHz	273.6 liters/h (60 imp. gal/h)

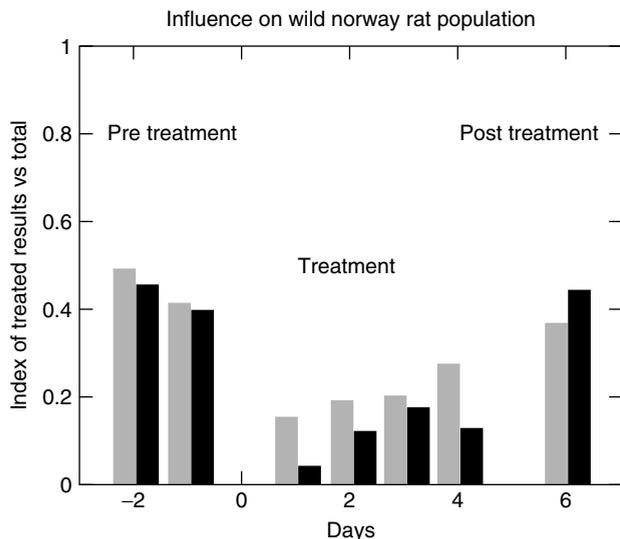


Figure 1. The influence of ultrasonic noise on the Norway rat population.

Figure 1 shows the effect of treatment on the Norway rat. Figure 2 shows the effect of the treatment on wild house mice. The influence on both populations is most significant for food consumption. The tracking activity of the wild house mice is not heavily influenced by the ultrasonic effect.

The rodents' hearing was checked before and after the testing. Only rodents that had good hearing were selected for the study. It has been postulated that the rodents might eventually become accustomed to the noise, but this was not the case. There were instances where rodents were not influenced, but this was due to hearing loss.

The sound patterns (frequency and amplitude) of four of the pace electronic pest repeller units were measured.

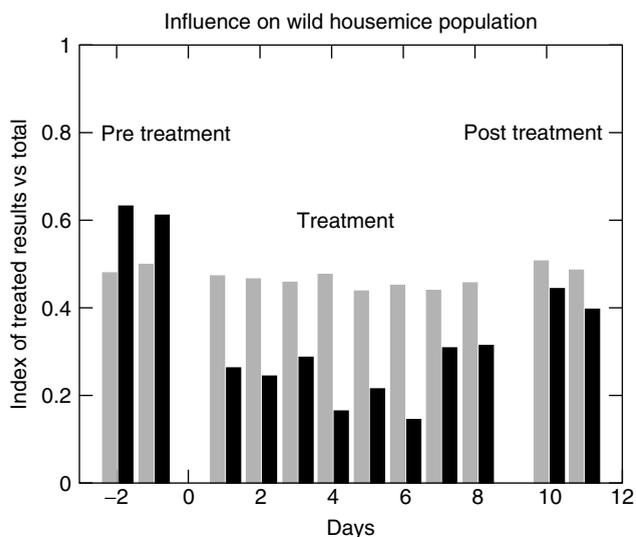


Figure 2. The influence of ultrasonic treatment on the wild house mice population.

The primary source of total sound output was at 40 kHz and above. The sound output dropped slightly at 31.5 kHz. Sound output below 20 kHz was negligible.

CAVITATION AS A DESTRUCTOR

Piezoceramic elements are commonly used to induce cavitation in fluids in biological applications for scaling instruments, but killing microorganisms is normally done by high-temperature sterilization. The erosive effect of cavitation is what is useful in removing a variety of type of scales. Cavitation is caused when the localized pressure drops below the fluid vapor pressure. This results in cavitating bubbles.

The collapse of cavitating bubbles is accompanied by a rapid release of energy. It is the collapse of the cavitating bubbles that is used to destroy microorganisms. It is not clear whether the microorganism population is immediately killed by the bubble collapse, or if the population is just weakened enough to limit its viability.

The generation of cavitation is limited to areas fairly close to the pressure/sound source. Cavitation can be applied to a large volume of fluid either by moving the source through the fluid or by moving the fluid past the source. The application described here moves the fluid past the source by pumping the volume through tubing to ensure fairly even exposure of the liquid to the pressure field. This does not sterilize the fluid, but it does eliminate a significant portion of the microorganism population.

The biological test results available indicate that cavitation does significantly reduce the population in both water and diesel fuel, but the effect varies for the types of microorganisms tested. The population reduction is of the order of 50%.

It is expected that piezoceramically induced cavitation could be used to reduce zebra mussel population in nuclear reactor water intake tubes by interfering with the zebra mussels during an early stage of their development, such as the larval stage.

The specific engineering design that follows was based on controlling microbial growth in military marine diesel tanks. These populations are currently controlled by "good housekeeping" of ships' tanks and by using environmentally harmful biocides. If an ultrasonic cavitation system were to be installed on a ship, it would be necessary to include an antinoise system to cancel the ultrasonic sound that creates the cavitation. This would be needed to minimize the likelihood that the vessel would be detected by unfriendly ships.

Engineering Application/Design

The cavitation of a fluid is induced when local pressure drops below its vapor pressure. It involves the release of relatively small amounts of energy (compared to boiling), so that though there is a temperature change in the fluid; it is small (of the order of 1–2°C, depending on exposure time and volume).

One of the well-known side effects of cavitation is its erosive effects on materials. This presents a practical problem

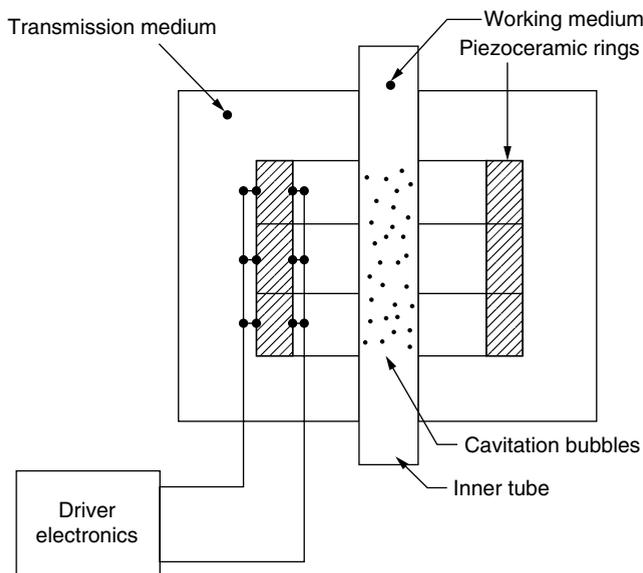


Figure 3. Schematic of cavitation concept.

in trying to use cavitation. The components used to cause the cavitation need special consideration to survive the erosive environment.

A general requirement for pest control is that it is needed for large volumes. Cavitation is a fairly local effect. To apply it to a large liquid volume, the fluid must be brought into a fairly local range. One way of achieving this is a flow-through system. The liquid is pumped through tubes that are exposed to the cavitating field. Such an arrangement could involve expenditures of significant amounts of power.

A flow-through configuration was studied analytically to achieve maximum fluid cavitation at minimum power consumption. The particular system modeled was based on a two-fluid system to avoid the electrode erosion that would be induced by cavitation. Figure 3 shows the conceptual arrangement. The fluid immediately adjacent to the electrodes is pressurized to eliminate cavitation. This fluid is used to transmit energy through a thin-walled pipe (stainless steel) into the fluid that contains the microorganism. The analytical model of the system was a piezodynamic field modeled by using finite elements. It is based on a finite element formulation of the piezoceramic elements, the physical piping structure, a liquid transmission medium, and the sound pressure field experienced by the microorganism-borne fluid (either water or diesel fuel).

The model was then test verified before applying it to a specific design.

Finite Element Formulation. The finite element method is an analytic technique for solving general field problems. It offers a number of advantages over competing methods. It can handle arbitrary geometries and both static and dynamic problems. It uses matrix numerical methods for which very efficient and general algorithms have been developed.

The special purpose FE formulation developed to handle both the fluid characteristics and the electrical input (as well as the normal structural characteristics) was based on the principles of the FE method in (2). The code modeled the structural behavior of the elements that represent the piezoelectric components, as outlined in (2, p. 22). The piezoelectric behavior was included using the approach of (3, p. 86). The fluid areas of the model were analyzed using the approach described in (2, p. 540).

The degrees of freedom of the model are the group of

- nodal displacements of the solid components,
- nodal pressures of the fluid components,
- nodal electrical potentials of the piezoelectric components, and
- the junction voltages of an external electrical circuit connected to the piezoelectric components (this latter capability was not used, though it is included for possible future use).

Then, the defining equations of the finite element approach used are

$$[A_2] \left\{ \frac{d^2 w}{dt^2} \right\} + [A_1] \left\{ \frac{dw}{dt} \right\} + [A_0] \{w\} + [A_{-1}] \int \{w\} dt + [A_{-2}] \iint \{w\} dt \cdot dt = \{b\}, \quad (1)$$

where

$$[A_2] = \begin{bmatrix} M & 0 & 0 & 0 \\ S & G & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad [A_1] = \begin{bmatrix} c & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

$$[A_0] = \begin{bmatrix} K & \frac{1}{\rho} S^T & E & 0 \\ 0 & H & 0 & 0 \\ E^T & 0 & -\epsilon \nabla^2 & 0 \\ 0 & 0 & 0 & C \end{bmatrix},$$

$$[A_{-1}] = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & R \end{bmatrix}, \quad [A_{-2}] = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & I \end{bmatrix},$$

$$\{b\} = \begin{Bmatrix} F \\ 0 \\ Q \\ Q_N \end{Bmatrix}, \quad \{w\} = \begin{Bmatrix} U \\ P \\ \Phi \\ v \end{Bmatrix}.$$

In these equations,

$$M = \int [N_s]^T \rho_s [N_s] dV_s$$

$$S = \int_S [N_f]^T \rho_f [N_s] dS_{sf}$$

$$G = \int [N_f] \frac{1}{a^2} [N_f] dV_f$$

$$c = \int [N_s]^T \mu_s [N_s] dV_s$$

$$f = \int [N_f]^T \mu_f [N_f] dV_f$$

$$K = \int [B]^T [D] [B] dV_p$$

$$E = \int [B_e]^T [\epsilon] [B_e] dV_p$$

$$H = \int [\nabla N_f]^T [\nabla N_f] dV_f$$

I = external circuit inductance

C = external circuit capacitance

R = external circuit resistance

U = solid element nodal displacements

P = fluid element nodal pressures

V = external circuit voltages

F = externally imposed force on solid element nodes

Q = externally imposed charges on piezoelectric elements

Q_N = externally imposed charges on external circuit

ϕ = piezoelectric element nodal potentials

a = speed of sound in fluid

where

$[N_s]$ = shape function matrix for solid elements

$[N_f]$ = shape function matrix for fluid elements

$[B]$ = shape function derivatives giving strain in solid elements

$[B_e]$ = derivatives of potential shape function in piezoelectric elements

ρ = mass density (subscript s for solid, f for fluid)

μ = damping (subscript s for solid, f for fluid).

The model assumed axisymmetry which was implemented as described in (2, p. 119). The elements describe the cross section of the complete unit from the centerline out, that is, that section which is rotated about the axis of symmetry to sweep out the 3-D geometry of the unit. The elements used were eight-node, isoparametric quadrilaterals, using quadratic shape functions for all fields (2-D solid displacements, fluid pressures, and electrical fields). Third-order Gaussian numerical integration was used for all element integrals. The integrals across volume are done by the usual finite element approach of integrating

across each element independently, followed by assembling the resulting equations into matrix form, as described in (2, p. 9).

Damping was included in the model by adding material damping to the fluid regions, as described in the preceding equations. Based on experimental measurements, enough damping was included to give a resonant amplification (Q factor) of 5 to 8. Two extreme conditions were used. In the first, damping was distributed across both the transmission and working media. In the second, damping was concentrated in the working medium. The first case corresponds most closely to low excitation levels, whereas the second should more closely match high excitations when cavitation is occurring. Then, the energy dissipation will be concentrated in the working medium because of the cavitation.

The model is linear. This is expected to give good results up to the point at which cavitation begins. Beyond that point, the response of the system is no longer linear because the fluid behaves effectively less stiff on the negative side of the pressure wave than on the positive side due to the formation of cavitating bubbles. In principle, this effect could be modeled using the nonlinear approaches described in (2, p. 450). This simplification was accepted because the objective was to compare alternative designs, rather than to analyze the behavior in absolute terms. It is assumed that systems that give a greater linear response will also give a greater nonlinear response. This may not be true in unusual cases, and it may not represent the effect of changes in the spatial distribution of the acoustic field in all cases (it would be expected that the "softening" nonlinearity which will occur here would tend to make the energy distribution more uniform in the system, compared to the linear case).

Figure 4 shows typical results from the model. These show the pressure distribution across the fluid cross section for 100 volt peak-peak excitation of the piezo rings for various excitation frequencies. It can be seen that the energy in the working medium in all cases is concentrated at the center. At low frequencies, only a single pressure peak occurs. At higher frequencies, when the wavelength of the sound waves in the fluid becomes comparable to the dimensions of the device, two and then three pressure peaks

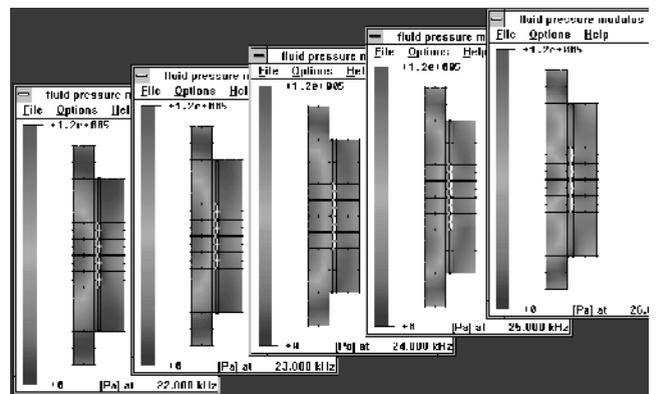


Figure 4. Finite element predictions of cavitating field.

Table 2. Finite Element Model Parameters

Parameter	Material	Dimensions
Inner tubing	Stainless steel tube ($E = 30E6$ psi)	1.5 in outer diameter 0.012 in wall thickness
Piezoceramic rings (stack of four)	PZT4	2 in diameter 0.125 in wall thickness 0.5 in height
Transmission fluid	SAE 10W30 motor oil	Density, speed of sound
Working fluid	Water or diesel fuel	Density, speed of sound

occur axially along the centerline. These observations are consistent with qualitative results. These results were obtained by suspending an aluminum foil strip in the cavitating field. Because it is known that cavitation erodes aluminum, the distribution and degree of perforation provide an indication of the cavitating intensity.

The specific parameters of the model are listed in Table 2.

Test Verification of Analytical Model. Modeling a combined electrical/piezoelectric/structural/fluid system is complex. A number of approximations and simplifications were made. For this reason, some model correlation was done in advance of prototype development (experimental data taken from breadboard unit). The FE model was done for a four-ring prototype. The experimental testing was done on a three-ring arrangement.

There were two type of measurements made for the correlation exercise, the current–voltage relationship and sound pressure measurements. The predicted and measured current versus voltage relationship for the system is shown in Figure 5. Measured values are shown at 22.7 kHz

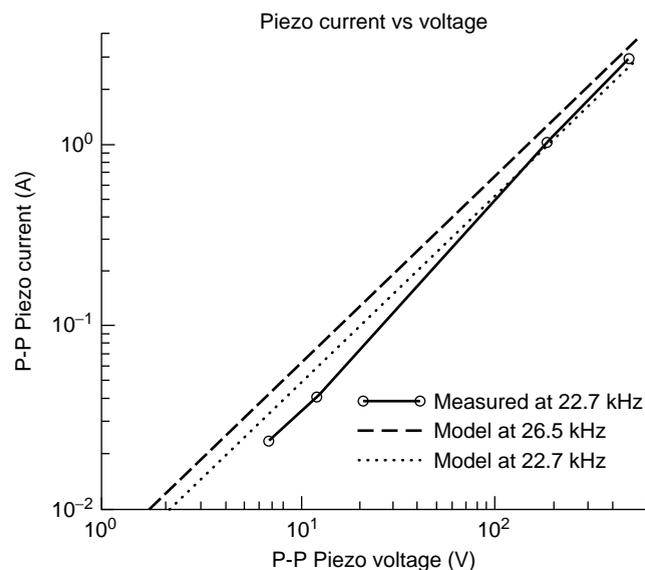


Figure 5. Measured and predicted current vs voltage.

which gives the peak piezo current. Model values are shown for both this frequency and for 26.5 kHz, which is the frequency at which the model shows peak current. It can be seen that the measured values at low voltages are about 60% of the modeled values. This is mainly due to the four rings in the model versus three in the breadboard. The sound pressure field was measured using the Specialty Engineering Associates needle hydrophone, Model SPRH-2-0500.

Figure 6 shows the response of the hydrophone at two different excitatory voltage levels, as captured on a digital storage oscilloscope. Note that the two cases were at slightly different frequencies. These frequencies correspond to the peak responses at each excitatory level. That they are different indicates nonlinearity in the model. It can be seen that the hydrophone response waveform is unsymmetrical and has pressure spikes on the positive voltage (low pressure) side. This is an indication of cavitation. It is more prominent at the higher excitatory voltage.

The model predicts that the peak pressure in the unit should be 1 kPa per volt of excitation. The transducer output should be 0.25 mV per volt of excitation. The results in Fig. 6 show a 20-mV peak-to-peak response at 130-V peak-to-peak excitation in (a) and 65 mV response at 240 V excitation, or 0.16 mV/V and 0.27 mV/V, respectively. This agreement is reasonable given the uncertainty of the hydrophone (it was being used somewhat out of its design frequency range). The model predicts that the pressure should lead the voltage by 10 to 20°, and it can be seen that this is reasonable, though the experimental measurements do not really allow testing this.

Figure 7 shows the pressure distribution measured along the centerline of the device for low voltage excitation (where the nonlinearity of the system does not confuse the results), and Fig. 8 shows the pressure distribution measured across the centerline at the midheight of the piezo rings. The hydrophone readings in these figures have been converted to acoustic pressures. The model predictions are also shown. It can be seen that the model and measured values show the same trends and the differences are 1–3 dB.

Design Studies

Outer Diameter of Transmission Medium. A design was studied to optimize the outer diameter of the transmission medium on the sound intensity in the working medium.

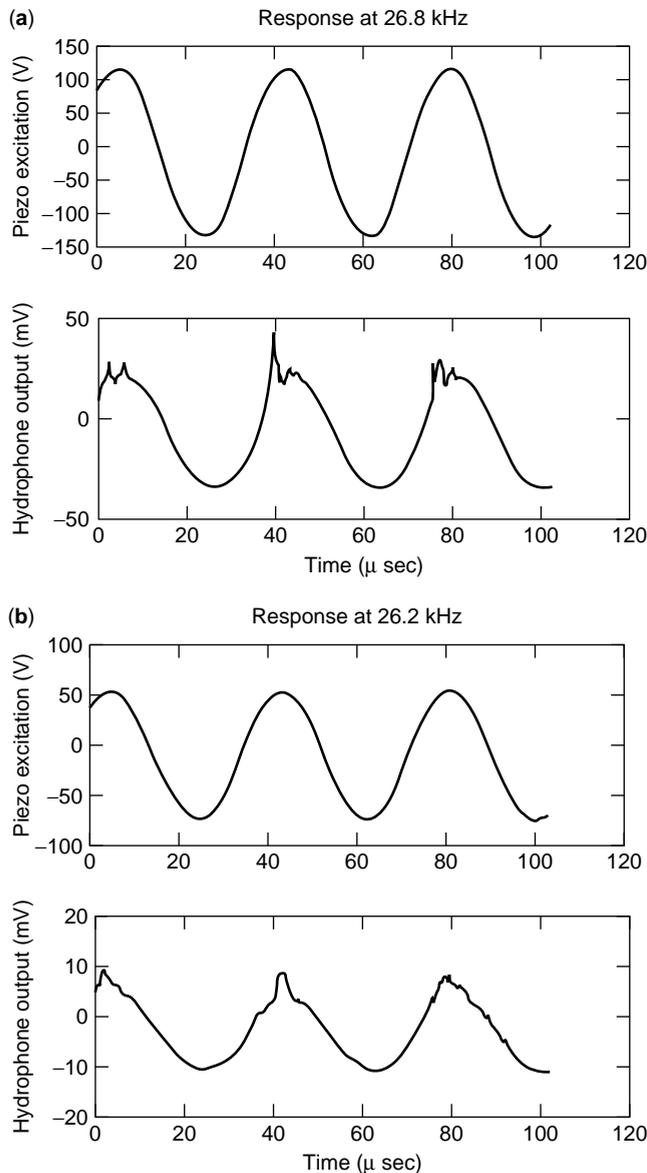


Figure 6. Hydrophone response at (a) 130 V p-p excitation; (b) 240 V p-p excitation.

The integral of acoustic pressure across the volume of the working medium was used as a performance indicator.

Two extremes of damping models were used—damping concentrated in the working medium and damping distributed over both working and transmission media. Figure 9 shows the results for both cases (as the integral of pressure vs. the outer diameter, (OD) of the transmission medium). It can be seen that when damping is concentrated in the working medium, the optimum occurs at an OD of 113 mm because the spacing between the outside of the piezo ring and the OD of the transmission medium is about one-half an acoustic wavelength. Such a condition would be expected to result in translating the high acoustic impedance condition at the rigid outer wall to a low acoustic impedance at the ring [see (8), p. 18 for an example]. This low acoustic impedance of the transmission medium

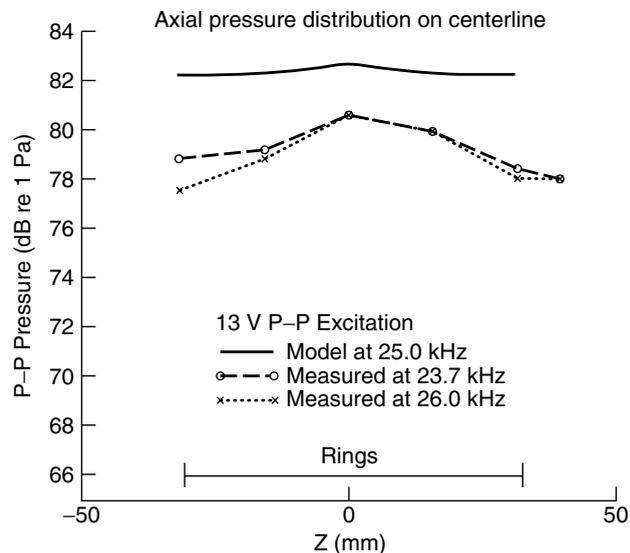


Figure 7. Acoustic pressure distribution along centerline.

at the ring is mismatched to that of the ring so that the coupling between the ring and transmission medium is poor at the outside of the ring. Little energy is launched outward from the ring, leaving more to be launched inward to the working medium.

The figure also shows that when damping is distributed across both transmission and working media, the optimum occurs at a lower OD. This may be due to the fact that when damping is included in the transmission medium, the increase in transmission medium volume, which occurs as its OD is increased, results in more energy losses in the system, thus biasing the optimum to a smaller diameter.

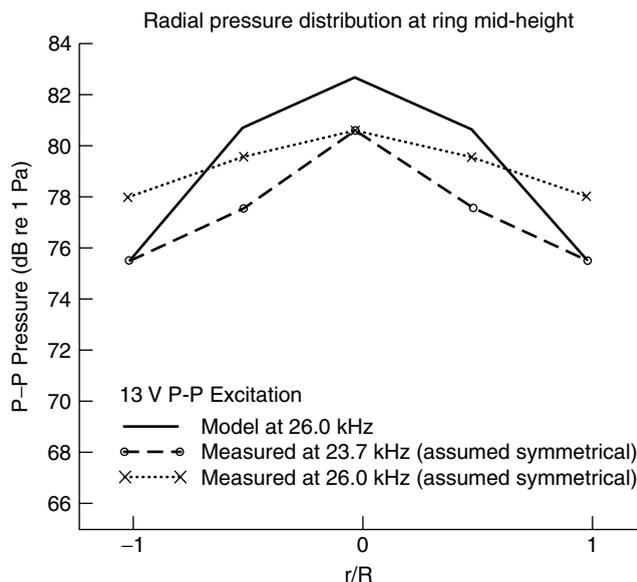


Figure 8. Acoustic pressure distribution across diameter at ring midheight.

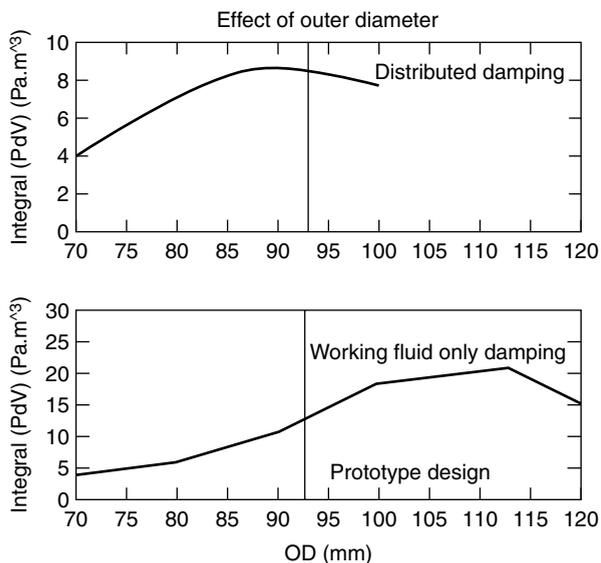


Figure 9. $\int_0^{F_f} Power_{Acoustic} vs \phi$.

Electronics Concept. Three electronics concepts were considered, and two were experimentally evaluated:

- a function generator to produce a sinusoidal (or other) waveform and a power amplifier to generate a final high-power output signal to be sent through a transformer to the piezo elements in the mechanical module
- a high-power oscillator
- a switching power supply

The first approach was used in prototype testing and development. It was not continued in the higher power, high flow-rate evaluation unit because the readily available

power amplifiers are limited in power (so would have to be ganged to drive the larger system) and the class A amplifier action used is relatively inefficient, making cooling of the electronics an issue.

The high-power oscillator was not developed because of concerns of achieving high power without instability problems.

The switching power supply was used for designing the evaluation unit. It is in line with current methods of driving high-power motors using pulse-width modulation (PWM). Digital circuitry is used to generate square waveforms. These may be duty-cycle modulated and are used to switch power MOSFET transistors on and off rapidly so that the average voltage presented to the equipment as a result of the variable duty-cycle appears sinusoidal. Such an approach is efficient because the transistors are always completely on or completely off (except during short switching transients), and they dissipate little power in either of these states. In our case, the output frequencies are too high for true PWM, but square waves can be generated at these frequencies and filtered to eliminate higher harmonics.

Figure 10 shows an electronic filtering concept evaluated by analysis. A high voltage supply that has positive and negative polarity and a 33% duty cycle is switched on and off. The fundamental frequency of the source is 25 kHz. This is followed by a three-pole low-pass filter that has a cutoff at 62.5 kHz. The output from this filter feeds a tuned circuit that represents the piezo rings (21.2-nF capacitance and a 100-ohm resistor to simulate a system Q of 3) in series with an inductance chosen to tune the circuit to the 25 kHz fundamental. This makes the driven system of this tuned circuit appear resistive at the fundamental frequency and so matches the low-pass filter's output impedance expectation. Note that no transformer is shown, though by adding a transformer between the filter and the piezo, lower voltages would exist in the left-hand

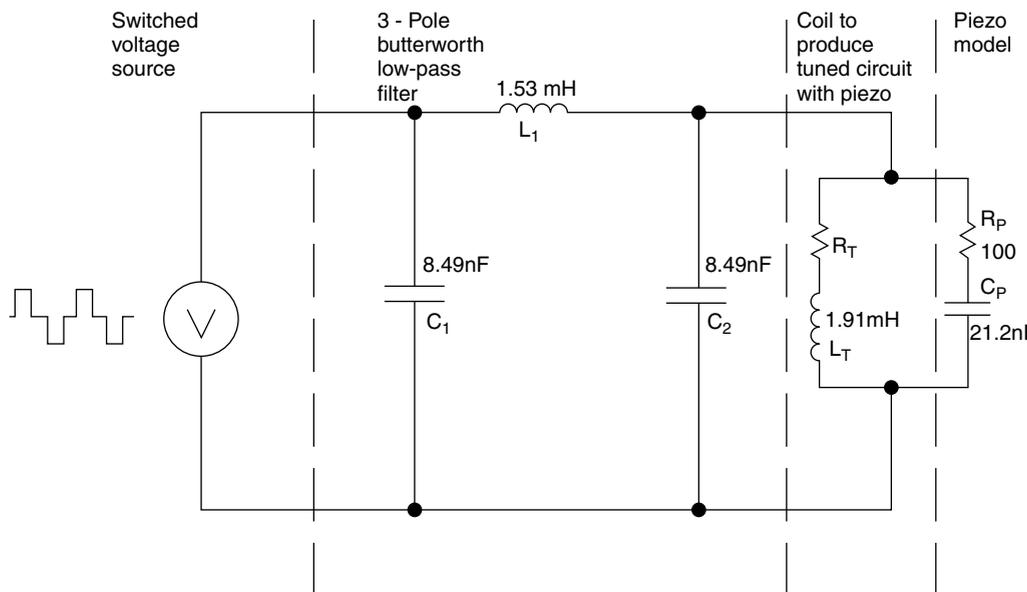


Figure 10. Electronics concept.

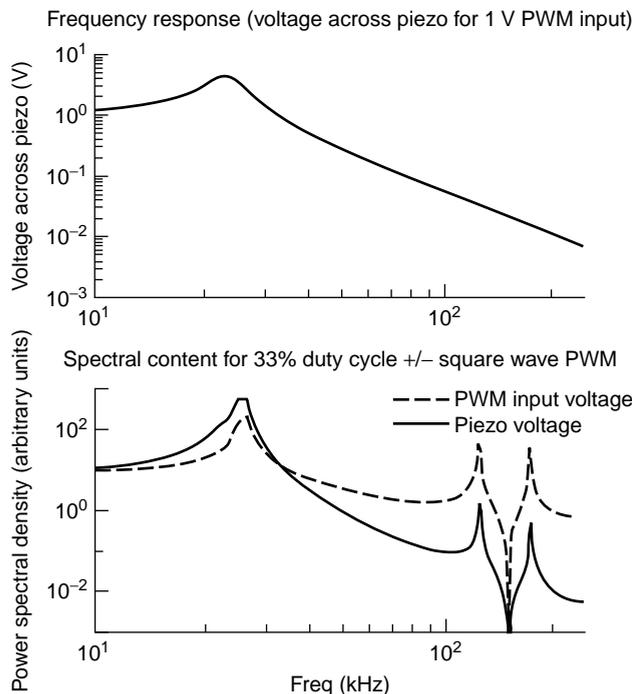


Figure 11. Frequency response function of electronics concept.

side of the circuit which would probably ease component choice.

Figure 11 shows the calculated frequency response function. It also shows the spectral content of the voltage out of the switched power supply and into the piezo. The output from the switched power supply it is assumed, is both positive and negative in the 33% duty cycle and has switching transients 25% as long as the on-time, that is, $1.67 \mu\text{s}$. Summing all power above the fundamental to 250 kHz gives a total harmonic distortion figure of 71% for the switched power supply output that has this waveform, but only 4% for the voltage across the piezo.

A breadboard of this system was built and tested. It was felt that the advantages of the switching amplifier concept outweighed its disadvantages for a production application. A commercial supplier (Instruments Inc. of San Diego CA) was found.

Implementation Issues. The thin walled stainless steel tube that contains fluid-borne microorganisms was designed to be as thin as possible to maximum the pressure transmitted through to the fluid. The thickness is limited by the pressure in the transmission medium. The thin walled tube is fairly close to buckling under the pressure of the transmission medium.

In the prototype system, there was no pressure sensor to ensure that the pressure of the transmission medium was maintained between 30–100 psi. The small temperature change ($1\text{--}2^\circ\text{C}$) that results from the excitation of the system causes the pressure to vary. The temperature change is kept to this low level by pumping the working fluid continuously past the transmission medium. During

biological evaluation of the prototype system, the pressure did drift above 100 psi. After completing of prototype testing, the system was dismantled, and it was discovered that the tubing had buckled.

The evaluation unit which was built as a follow-on to the prototype includes both a temperature and pressure sensor as part of the design. This ensures that the system will shut down before the critical pressure is exceeded. In an early version of the evaluative design (which contained 16 piezo rings, rather than the original four), the stainless steel tubing did buckle because the unsupported length of the tubing had more than doubled. Modifications of the tubing boundary conditions were made to ensure that buckling did not occur but at the same time maintained as thin a profile as possible to maximize the energy transfer to the microorganism-borne fluid.

Another significant issue that arose during early testing of the evaluative system relates to the importance of tolerancing the rings themselves. After short runs of the 16-ring stack system, failures in the rings occurred. They were failing mechanically—breaking into two pieces. The initiation of the crack seemed to be associated with a burn mark on the ring. It was postulated that the set of rings being used was not sufficiently well toleranced for roundness. The system was rebuilt using rings of improved tolerance (proved by Sensor Technologies of Collingwood, Ontario). There have been no ring failures since the system was rebuilt.

The original electronic drive for the system was based on square wave input switching. When this was implemented, switching noise was feeding back to the input, causing noise spikes that were outside the acceptable range of the microprocessor. To eliminate this problem, the signal generator was rebuilt to use sine wave excitation.

Figure 12 shows a drawing of the cavitation portion of the system. The elements of the figure are as listed in Table 3.

Effectiveness of Cavitation in Destroying Microorganisms

The effectiveness of using a cavitation field to destroy microorganisms was measured for two types of fluid hosts (water and diesel fuel) (9) and three types of microorganisms:

- *Serratia marcescens*
- *Pseudomonas aeruginosa*
- *Saccharomyces cerevisiae* (yeast)

The fitted results are shown in Fig. 13, plotted as a function of exposure time to the cavitation field. Regression analysis was used to fit the data to the following equation:

$$\log\left(\frac{\text{Irradiated}}{\text{Control}}\right) = (\text{Slope} \times \text{Time}) + \text{const.} \quad (2)$$

These test results were for microorganisms exposed to cavitation while the working medium was moving (being pumped) through the cavitation field. Earlier test results were performed while the medium was static during

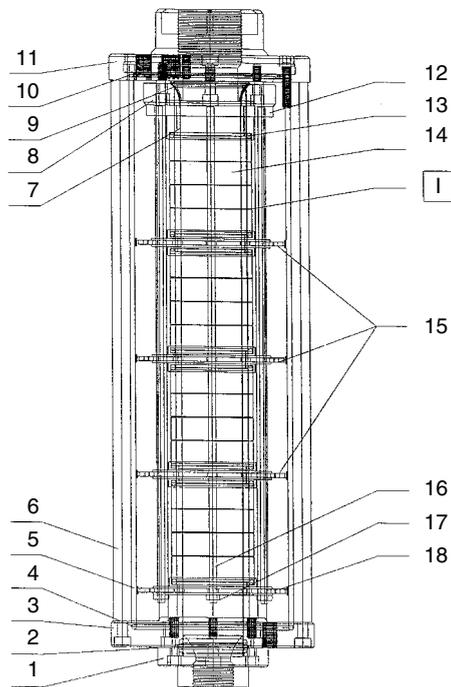


Figure 12. Cavitation unit—16 ring.

exposure to the cavitation field. The cavitation effect was more pronounced on the moving population than on the static population. It was hypothesized that the motion ensured improved distribution of the microorganisms in the cavitation field.

There were two different strains of *Pseudomonas aeruginosa* used in the study. Tests in water were done using ATCC 10145. A strain of *Pseudomonas aeruginosa* was isolated from a sample of marine diesel fuel. This strain would not survive at elevated temperatures (37°C) where the ATCC 10145 thrived.

Table 3. Parts of Cavitation Unit

Drawing Label	Part
1	Lower sealing flange
2	Hydraulic O-ring
3	Lower flange
4	Hydraulic O-ring
5	Body
6	Body assembly rods
7	Flow-through tubing
8	Supporting ring
9	Hydraulic O-ring
10	Hydraulic O-ring
11	Upper flange
12	Upper supporting ring
13	Hydraulic O-ring
14	PZT ring, 2.0 in OD
15	Middle PZT supporting ring
16	PZT Assembly rods
17	Self-locking nuts
18	Lower PZT supporting ring

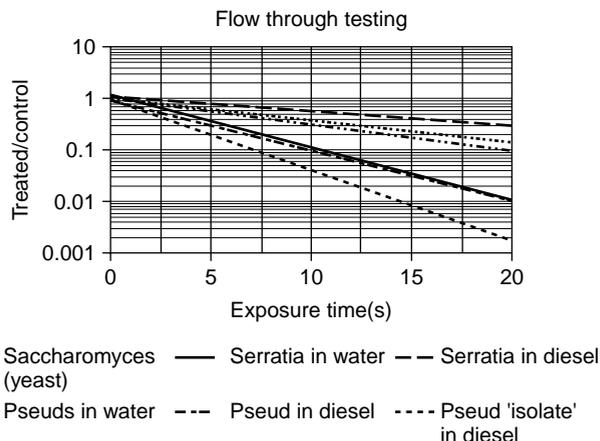


Figure 13. Biological test results.

The results were based on a flow-through testing system that involved recirculating the population to obtain the required exposure time. Figure 14 shows a schematic of the experimental facility. The contaminated working fluid was recirculated during testing. This eliminated the need for disposal of large volumes of contaminated fluid. The recirculating effect underestimates the effectiveness of the method because the population is being gradually reduced for each pass through the cavitation field.

It had been postulated that the pumping action itself might influence the microorganism population, but that effect was studied and found insignificant on either the *Serratia marcescens* or the *Pseudomonas aeruginosa*. There did seem to be a small effect on the yeast results.

An attempt was made to predict the kill efficiency of a single pass of the population through the cavitation field. Kill efficiency e is the ratio of microorganisms per unit volume of fluid killed in one pass to microorganisms present in an untreated unit volume of fluid.

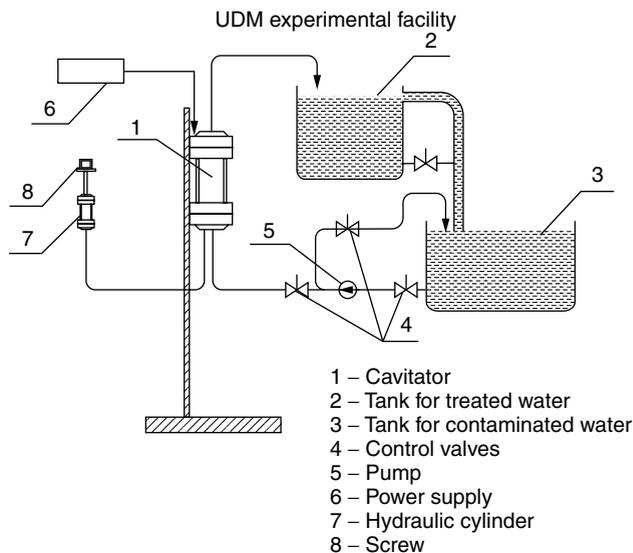


Figure 14. Schematic of flow-through experimental facility.

NOTATION

- C_0 = initial concentration (microorganism's/litre)
 C_n = concentration after n passes through cavitation field
 e = kill efficiency
 n = number of times sample passed through cavitation field
 V = volume of cavitation field
 X = holding tank volume

$$\frac{C_n}{C_0} = \left[\frac{X - e \times V}{X} \right]^n \quad (3)$$

When this equation is applied to the yeast test data obtained, the resulting kill efficiency is 0.49. When it is applied to the test results for *Pseudomonas aeruginosa* in diesel fuel, the resulting kill efficiency is 0.45. These results were based on an exposure time of 3.15 seconds in the cavitation field.

BIBLIOGRAPHY

1. A.D. Ashton. Laboratory Evaluation of Ultrasonic Devices: Weitech Electronics,
2. O.C. Zienkiewicz, *The Finite Element Method*. McGraw-Hill, NY, 1977.
3. K. Ragulskis, R. Bansevicius, R. Barauskas, and G. Kulvietis, *Vibromotors for Precision Microrobots*. Hemisphere, NY, 1988.
4. Modern Piezoelectric Ceramics, Morgan Matroc Vernitron Division, Bedford, OH, 1988.
5. J.R. Frederick, *Ultrasonic Engineering*. Wiley, NY, 1965.
6. S.S. Save, A.B. Pandit, and J.B. Joshi, *Chem. Eng. J.* **55** B67–B72 (1994).
7. A.J. Chapman, *Heat Transfer*. Macmillan, NY, 1967.
8. G.L. Goberman, *Ultrasonics: Theory and Application*. Hart P, NY, 1969.
9. S. Draisey. *Ultrasonic Destruction of Microorganisms in Shipboard Fuels: Biology Report*. Canadian National Defence Report, DREA CR 98/426.

PHOTOCROMIC AND PHOTO-THERMO-REFRACTIVE GLASSES

L.B. GLEBOV
 University of Central Florida
 Orlando, FL

INTRODUCTION

Inorganic glasses are the main transparent material, which people have long used for observation (windows in buildings, windshields in cars, eyeglasses, prisms and lenses in optical instruments), light delivery (light bulbs, projectors, lasers, optical fibers), and fine arts (crochery, bijouterie, jewelry). The ability of glasses to change coloration after exposure to sunshine was well known since the last century. A new era in glass application was started

in 1949 by S.D. Stookey's publication (12) in which recording a permanent photographic image in silicate glass was described. This two-step process of exposure to UV radiation and thermal development that resulted in a crystalline phase precipitation in the exposed areas was similar to the classical photographic process. As a result of intensive research during a long period of time, a great number of different photosensitive glasses were developed, which have found very wide application in different branches of industry and personal use. When exposed to optical radiation, these glasses (and glass ceramics) change their optical properties (absorption, refraction, or scattering) instantly or after thermal development, permanently or transiently. Among the great variety of photosensitive glasses, we emphasize only the two most widely used types.

The largest commercial application was obtained for so-called "photochromic glasses," which exhibit reversible coloration after exposure to UV or visible light and can vary their absorption depending on the illumination level. Glasses that contained small concentrations of microcrystals of silver and copper halides, proposed by Armistead and Stookey in 1965 became the most widely used for reversible coloration (13). A peculiarity of these materials is that they are produced by glassmaking technology whereas the photochromic processes occur in microcrystals distributed in the glass matrix. Several hundred original papers were dedicated to different aspects of heterogeneous photochromic glasses in those years. The vast bibliography and detailed descriptions of these heterogeneous photochromic glasses were collected in books (3,4), and therefore we will not include a list of original publications in this article.

Another type of photosensitive glass, which is beginning its application in optics and photonics right now, is "photo-thermorefractive (PTR)" glass. If this glass is exposed to UV radiation followed by heat treatment, it varies in refractive index. A phase hologram in the volume of this glass was recorded in 1990 by Glebov and coauthors (5). The feature of this process is that homogeneous glass is exposed to light and a microcrystalline phase is produced in the volume of the glass matrix by a thermodevelopment process. No books have been written on this subject. The main results concerning phase hologram recording in glasses can be found in a few original papers (5–7) and a survey (8). Similar processes of photoionization followed by thermoinduced crystallization were studied for single- and full-color photography in polychromatic glasses, as described in (1, 9–12). Thus, these references can also be used for learning the basic physical phenomena that result from irradiation and development of PTR glasses. Some basic data concerning intrinsic absorption, electronic excitation, and nonlinear photoionization in multicomponent glasses can be found in (13,14).

PHYSICAL PRINCIPLES OF PHOTOSENSITIVITY IN GLASSES

Photosensitivity is the variation in glass properties from exposure to optical radiation. Photoinduced processes can be caused by the absorption of light and consequent

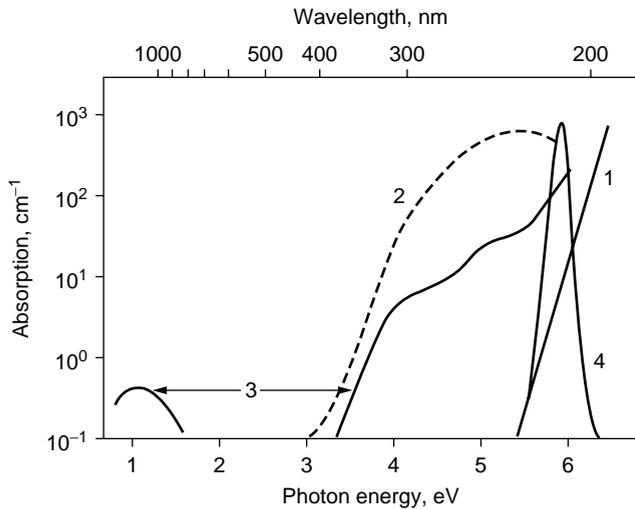


Figure 1. Absorption spectra of $25\text{Na}_2\text{O}-75\text{SiO}_2$ glass. 1: intrinsic absorption; 2 and 3: extrinsic absorption of 0.1 wt.% of Fe^{3+} and Fe^{2+} , respectively; and 4: color center generation spectrum (arbitrary units).

excitation of electrons from ground to upper levels by which these electrons can be delivered to other places (we will not consider heating and posterior melting or ablation). Absorption spectra of solids may be conventionally divided into three groups. Absorption due to electron transitions in defect-free substances of stoichiometric composition is called “intrinsic,” “basic,” or “fundamental” absorption. The absorption in atoms or molecules that are present as small additives is called “extrinsic,” or “dopant,” or “impurity” absorption. The absorption by defects in the host substance created by chemical or physical effects is called “induced,” or “additional,” or “defect” absorption.

The absorption spectra of widespread alkali silicate glass, which is the basis of the majority of technical glasses, are presented in Fig. 1. Intrinsic absorption (curve 1) is in the range of 210 nm (6 eV) and exhibits an exponential dependence of the absorption coefficient on photon energy (or wave number). This absorption is caused by basic structural units of silicate glass (Si–O–Na), which are called L centers. An example of extrinsic absorption in $25\text{Na}_2\text{O}-75\text{SiO}_2$ glass is shown by curves 2 and 3 for ferric (Fe^{3+}) and ferrous (Fe^{2+}) ions, which determine the actual absorption of commercial silicate glasses in the near IR, visible, and near UV spectral regions. Induced absorption produced by UV and γ radiation (Fig. 2) is caused by ionization in the glass matrix and further trapping of electrons and holes at different glass matrix defects. The presence of different dopants and impurities results additional induced absorption bands. Extrinsic absorption can be caused by additional ions distributed in the glass matrix and also by bigger units, for example, microcrystals. The absorption spectra of borosilicate glass doped with copper and chlorine, which has undergone heat treatment, are shown in Fig. 3. Instead of absorption of copper ions in the glass in the far UV region, a narrow absorption peak near 380 nm (3.25 eV) is seen in these spectra, which corresponds to excitons in CuCl crystals precipitated in the glass matrix as the result of heat treatment. Induced absorption can

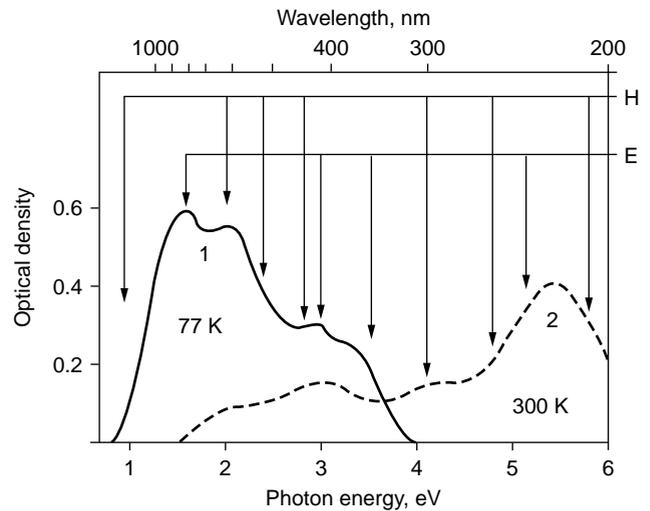


Figure 2. Induced absorption spectra of $25\text{Na}_2\text{O}-75\text{SiO}_2$ glass. 1: exposure to UV at 77 K; 2: γ irradiation at 300 K. Arrows show the positions of the absorption bands of electron (E) and hole (H) color centers.

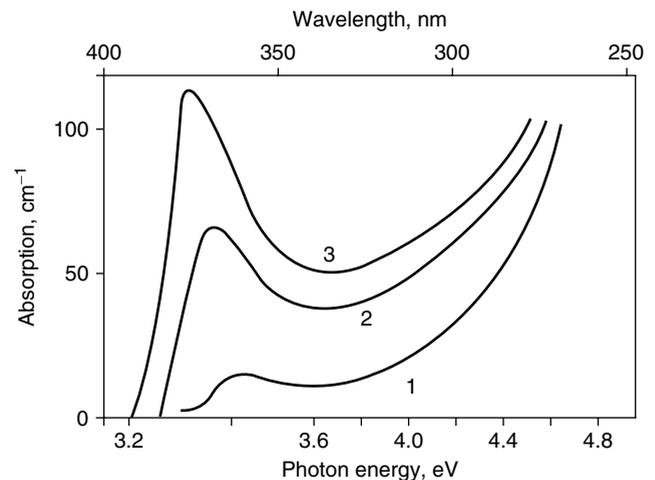


Figure 3. Absorption spectra of borosilicate glass doped with copper and chlorine after 2 hours of treatment at $T(^{\circ}\text{C})$: (1) 550, (2) 600, (3) 650.

also be produced by relatively big particles. Photoinduced precipitation of microcrystals of such metals as gold, silver, and copper causes additional absorption, usually called colloidal coloration.

Glass exposure to radiation whose photon energy is more than the intrinsic absorption edge (curve 1 in Fig. 1) causes photoionization in the glass matrix followed by the generation of both electron and hole color centers. The dependence of the induced absorption on the photon energy (or wavelength) is called the color center generation spectrum or the spectrum of photosensitivity (curve 4 in Fig. 1). Photoionization in the glass matrix (generation of both electron and hole centers) is impossible if the photon energy of the exciting radiation is less than a bandgap, which is determined by the position of the intrinsic absorption (curve 1 in Fig. 1). In other words, the long wavelength

edge of the color center generation spectrum (curve 4 in Fig. 1) coincides with the intrinsic absorption edge (curve 1 in Fig. 1).

The photosensitivity spectrum can be shifted to the long wavelength side if the glass is doped with some ions in a lower valence state, and the dopant's excited level is placed above the threshold of the charge carrier's mobility. In this case, a mobile electron can be trapped either by defect at an intrinsic electron center formation or by another dopant, that is, to recharge the activators. The depth of the dopant ground level in $\text{Na}_2\text{O}-3\text{SiO}_2$ glass is 5.2 eV for Fe^{2+} , 5.0 eV for Tb^{3+} , and 3.6 eV for Ce^{3+} . Comparison of these values with curve 3 in Fig. 1 shows that the ionization threshold of Fe^{2+} corresponds to the long wavelength edge of the absorption band whose maximum is at 6.5 eV (191 nm). Excitation using smaller photon energy causes tunnel ionization whose efficiency is about one to two orders of magnitude less than that of over-barrier ionization. The thresholds of tunnel ionization of dopants in $\text{Na}_2\text{O}-3\text{SiO}_2$ glass are 3.5 eV for Fe^{2+} , 3.1 eV for Tb^{3+} , and 3.1 eV for Ce^{3+} . Referring Fig. 1, one can see that the tunnel ionization of Fe^{2+} is obtained at an excitation of the long wavelength bands whose peaks are at 5.1 and 4.4 eV (243 and 282 nm) up to 3.5 eV (350 nm). Unlike intrinsic ionization that inevitably produces electron and hole centers, the only hole center generated from the excitation of dopant absorption bands is the same (but oxidized) dopant ion. All newly created centers are electron centers (either intrinsic or extrinsic).

The other way to shift photosensitivity to the long wavelength side is to use nonlinear ionization produced by powerful optical irradiation. In silicate glass exposed to pulsed radiation whose photon energy is more than half of the bandgap ($h\nu > 3$ eV, $\lambda < 400$ nm) and whose irradiance is more than 1 MW/cm^2 , both electron and hole color centers appear as a result of two-photon ionization in the glass matrix. The final concentration of color centers is determined by equilibrium between two-photon generation and single-photon bleaching of color centers.

INDUCED COLORATION OF REVERSIBLE PHOTOCROMIC GLASSES

Generally, the term photochromism may be treated as any variation of color induced by optical radiation, but usually people use a narrower definition, which excludes irreversible color changes. So, photochromism is a reversible variation in color (i.e., of the absorption spectrum or spectrum of attenuation) of a material under optical radiation that relaxes when exposure stops. Naturally, when experimental conditions are changed, for example, a temperature change, the magnitude of the photochromic effect can vary (even to complete disappearance). Therefore, we shall call a photochromic material one that, under specified operating conditions, becomes colored by optical radiation and restores its transparency after radiation ceases.

Relaxation of induced absorption after illumination ceases is usually caused by thermal fading of color centers, which are not stable at a given temperature. This is the most important feature of photochromic materials

because reversibility of the photochromic effect means the absence of any stable induced centers generated by illumination. A great number of electron and hole color centers in silicate glasses produced by UV radiation (Fig. 2) leads to fatigue because of the progressive accumulation of stable color centers. This is the reason that these glasses are not used as photochromic materials, although photochromism was discovered in cerium-doped, reduced silicate glasses. Glasses doped with microcrystals of silver and copper halides (Fig. 3) show complete reversibility of coloration at room temperature and therefore have the widest commercial application.

The main feature of photochromic glasses, variable optical density both observed during exposure and upon its cessation, has to be taken into account to determine characteristics such as integral and spectral sensitivity, darkening degree and rate, thermal fading, and optical bleaching rates. Let us define the main concepts required for photochromic material characterization. Light absorption (or, more exactly, light attenuation or losses, that is the sum of absorption and scattering) is characterized by the transmittance, $\tau = I_{tr}/I_0$ (where I_{tr} and I_0 are the intensities of transmitted and incident light, respectively), or the optical density, $D = -\log_{10} \tau$. The optical density of a sample before irradiation (original absorption, clear glass) is D_0 (Fig. 4). The optical density of the sample at the moment exposure ceases (induced absorption, dark glass) is D_{exp} . The optical density in t seconds of the thermal fading process (induced absorption, partially relaxed glass) is D_t . The spectral dependences of τ_0 and D_0 are the transmission or absorption spectra of clear glass. The spectral dependences of τ_{exp} and D_{exp} are the transmission or absorption spectra of dark glass. Glass has a gray color if the absorption (transmission) spectrum is flat in the visible region. A brown color means that the absorption in the blue region is greater than that in the red region.

The dependences of D_{exp} and D_t on the time of illumination or aging are the kinetics of coloration and relaxation, respectively (Fig. 4). D_{exp} increases when the exposure time increases and comes to the equilibrium level D_e

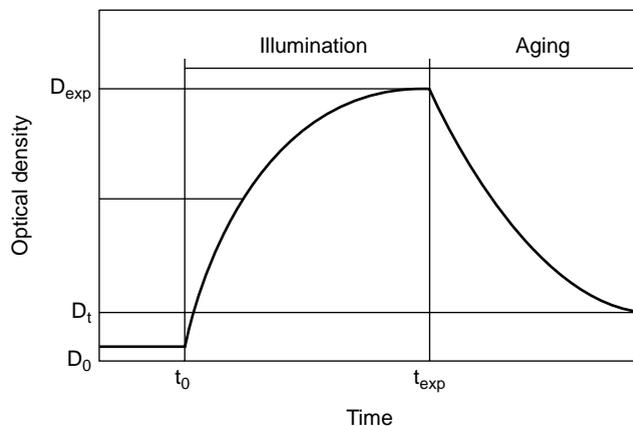


Figure 4. Kinetics of photochromic glass darkening under illumination and fading in the aging process. D_0 , D_{exp} , and D_t are the optical densities of clear, dark, and relaxed glass, respectively.

when the rate of color center generation is equal to the rate of thermal fading. The criterion of relaxation characterizes the degree of thermal fading in a certain time after illumination ceases:

$$K_{\text{rel}} = \frac{D_{\text{exp}} - D_t}{D_{\text{exp}} - D_0} \quad (1)$$

The value of that time interval should be selected on the basis of the practical applications of a photochromic glass. Thus, for photochromic lenses used as sunglasses, a time interval of 180 s is recommended. From Eq. (12), it is obvious that, if a glass has faded completely in that time, $K_{\text{rel}} = 1$. Contrariwise, if the induced absorption has not reduced at all in that time, $K_{\text{rel}} = 0$. Now, there are photochromic glasses whose K_{rel} vary in the entire range from zero to about one. K_{rel} for a particular glass can be changed by temperature variation.

An important parameter is the spectral sensitivity of a photochromic material, the dependence of the saturated photoinduced optical density (D_e) on the photon energy of the exciting radiation. This dependence is called the color center generation spectrum. The absorption edge of photochromic glass determines the position of the color center generation spectrum because photosensitive crystals absorb exactly in that region (compare curves 1 and 2 in Fig. 5). The short wavelength edge of the color center generation spectrum is connected with the decrease of the thickness of the layer containing color centers, that is due to the increase of the glass absorption coefficient. The long wavelength edge is caused by a decrease in the absorption and in the efficiency of photosensitive center ionization. These photosensitive centers are usually copper centers in silver halide crystals or excitons in a crystalline phase of copper chloride. Owing to that, the position of the maximum in the color center formation spectrum does not coincide with that of any maximum in the photochromic glass absorption spectrum. Moreover, its position is determined by the spectral shape of the photochromic glass absorption edge,

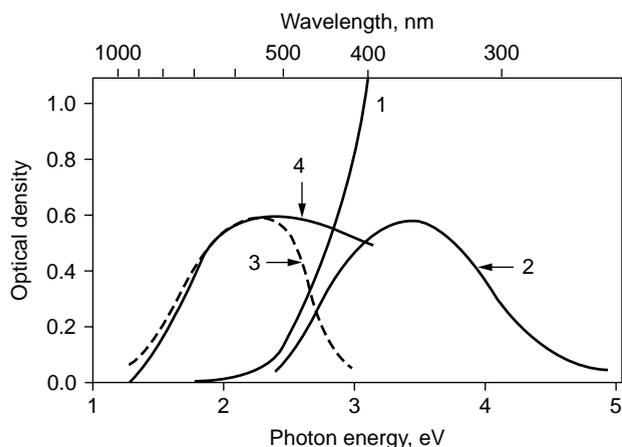


Figure 5. Spectra of glass doped with AgCl(Br). Absorption of original glass (12) and color centers (3), color center generation (13) and bleaching (4) efficiency. Sample thickness 5 mm.

is a function of the sample thickness, and drifts to the short wavelength side as the thickness decreases. The absorption spectrum of an exposed glass doped with AgCl microcrystals is presented in Fig. 5, curve 3. This absorption represents a wide band in the visible spectral range. The spectral shape of this band is usually ascribed to precipitation of colloidal silver particles on the surface of halide microcrystals. Curve 4 in Fig. 5 shows that excitation of the absorption band of color centers destroys these centers and causes optical bleaching. Thus, optical bleaching by visible light is a process additional to thermal fading, which accelerates the relaxation of darkened silver halide photochromic glass.

The photosensitivity of photochromic glasses doped with CuCl can be shifted from the UV region to the long wavelength side. Virgin photochromic glass is photosensitive only to UV irradiation and cannot be darkened by visible light. Excitation of glasses doped with CuCl that are exposed to UV radiation does not produce optical bleaching, as shown in Fig. 5 (curve 4) for silver halide glasses. On the contrary, initial additional absorption (induced by UV radiation) can be intensified by additional exposure to visible and even IR radiation having photon energy much below the ionization threshold of copper centers. Note that the power density of long wavelength irradiation must be high enough to produce this intensification. It is shown in Fig. 6 that the spectra of additional absorption produced in this glass after irradiation at various wavelengths are the same. Consequently, this long wavelength sensitivity results from generating new color centers by exciting the same color centers. Therefore this process is called “cooperative breeding of color centers.”

The mechanism of two-photon cooperative breeding is as follows. Initial exposure to UV radiation causes ionization

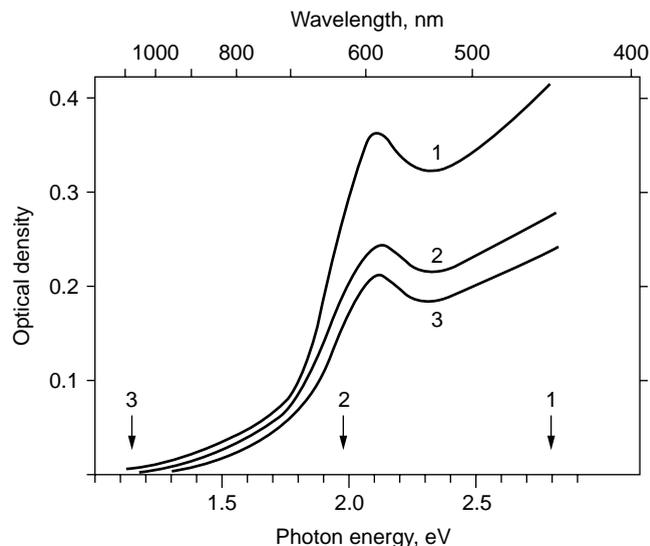


Figure 6. Spectra of induced absorption in copper halide photochromic glass (thickness 5 mm) after exposure to radiation at different wavelengths: (12) 440 nm (2.78 eV), (13) 633 nm (1.96 eV), and (3) 1060 nm (1.17 eV).

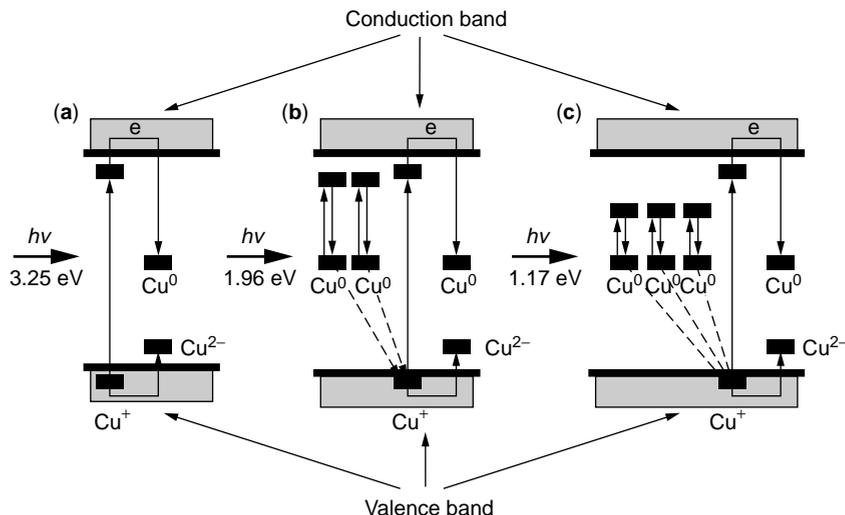


Figure 7. Energy diagram of the first stage of photochromic glass coloration at (a) short wavelength coloration, (b) two-photon cooperative breeding, and (c) three-photon cooperative breeding.

of a photosensitive center (Cu^+) and generates electrons and hole centers (Cu^{2+}). Then released electrons produce color centers by reducing copper (Cu^+) or silver (Ag^+) ions. The initial concentration of color centers (Fig. 7a) is determined by the number of UV-ionized photosensitive centers. This concentration can be rather small and even invisible to the naked eye. Linear absorption of two photons of visible light by two color centers causes a transition of these centers to excited states (Fig. 7b). Further, these centers simultaneously transfer the accumulated energy to the photosensitive centers (Cu^+) and return to their ground states. An excited photosensitive center releases an electron and converts to its ionized state in the same manner as after linear excitation, as illustrated in Fig. 7a. The released electron is trapped by an acceptor, converts to a reduced state (Cu^0), and this is a first stage in generating a new color center. Thus, the number of color centers increases after each cycle. This means that induced absorption increases in the process of exciting previously induced color centers without altering the spectrum of the induced absorption. The efficiency of this nonlinear process is proportional to the squared intensity of the exciting long wavelength radiation.

The coloration caused by exposure to pulsed IR radiation can be explained similarly to the three-photon cooperative breeding of color centers (Fig. 7c). The latter process obeys the cubical dependence of efficiency on the intensity of the exciting radiation. There are several important features of cooperative breeding of color centers. The first is a very high level of additional absorption because photosensitivity in this case is not connected with the sharp absorption edge of glass (Fig. 5) and a thick slab can be homogeneously colored. The second is the opportunity of localizing colored spots in arbitrary places of the bulk glass. The spots are produced by focusing the exciting beam because photosensitivity is proportional to the squared or cubical intensity of the exciting radiation and therefore, is concentrated near the focal plane. The third is an opportunity to store a latent image produced by UV radiation that can be revealed by photodevelopment.

HETEROGENEOUS PHOTOCROMIC GLASSES

Photochromic glasses co-doped with silver and copper halides are heterogeneous materials. They represent two-phase systems that consist of a vitreous host and dispersed photosensitive microcrystals. This is important because microcrystals show a reversible photochromic effect without fatigue. However, in a two-phase system, light attenuation is caused by absorption of each phase and also by scattering produced by the difference between the refractive indexes of the crystalline and vitreous components. Therefore, the parameters of the crystalline phase should be chosen to prevent strong scattering. The size of the particle of most photosensitive microcrystals, whose refractive index is about 2, should be no more than 10–20 nm to keep scattering below the level of acceptability for optical applications.

The main approach to producing dispersed microcrystals in a vitreous host is crystalline phase growth as a result of host glass heat treatment at temperatures from 500–700°C, depending on host composition. These temperatures correspond to a viscosity range from 10^{10} – 10^{13} poise. To secure crystalline phase precipitation, special requirements are applied to the host glass. First, this glass should be an oversaturated solution of the photosensitive phase (silver and copper halides) that allows effective diffusion of these components in the temperature range mentioned. Second, the solubility of the photosensitive components must drop quickly when cooling to allow the homogeneous glass to melt at high temperature and the crystalline phase to precipitate in the secondary heat treatment process. The last is usually connected with phase separation (immiscibility) and altered coordination of different components in the host glass.

The best glass, which satisfies the requirements mentioned before, is alkaline borosilicate glass. This glass matrix is the basis for almost all commercial photochromic glasses manufactured by a number of companies in different countries. Halides (Cl, Br, I) of silver and copper are photosensitive components, which are added to the batch.

Cations such as Mg, Ca, Ba, Zn, Cd, Al, and Pb, or anions such as P and S are used by different companies as additions to modify technical and end use properties. These compositional changes lead to variations in photosensitivity, the criterion of relaxation, and induced absorption spectra. Photochromic glasses can be divided into two large groups: silver halide glasses that have small concentrations of copper, which usually exhibit faster relaxation and lower sensitivity and copper halide glasses that have small concentration of silver, which exhibit slower relaxation and higher sensitivity. In silver halide glasses, small additions of copper are a sensitizer.

The traditional schedule for photosensitive phase creation, "bottom-to-top," consists of four stages: melting, rough annealing and cooling to room temperature, additional heat treatment (roasting), and final annealing. Final annealing is necessary for stress relaxation because crystalline phase precipitation occurs at temperatures above the glass transition temperature. The other method of sensitization is "top-to-bottom," which is used for mass production because of heat energy saving. In the latter, the glass casting cools down to roasting temperature but not to room temperature. It requires the other schedule (time and temperature) because the most effective growth of nucleation centers occurs at temperatures below the roasting temperature.

OPTICAL WAVEGUIDES IN PHOTOCHROMIC GLASSES

The largest commercial application of photochromic glasses is for sunglasses. Tens of millions of photochromic lenses are produced worldwide each year for this purpose. However, the alkaline borosilicate origin of photochromic glasses allows some other applications in modern optics and photonics. It is well known that these glasses are suitable for ion exchange and, consequently, planar and channel waveguides can be created on this glass. Besides that, the mildly sloping dependence of photochromic glass viscosity on temperature allows creating of optical fibers. The optical properties of photochromic waveguides compared with bulk photochromic glasses are unusual because of structural transformations in the ion-exchanged layers or in the drawn fibers and the peculiarities of light propagation in waveguides. An important feature of ion-exchanged glass is incompleteness of structural relaxation. The exchange of ions that have different radii creates stresses in glass. These stresses produce strong differences between the refractive indexes of waveguide modes that are orthogonally polarized (birefringence). Compression of silver halide photochromic glass after substituting Na^+ by K^+ at temperatures below the glass transition temperature reaches 1 GPa and produces birefringence up to 20% of the total refractive index variation, as shown in Fig. 8.

Exposure of waveguides in photochromic glasses to UV radiation produces reversible coloration. This means that ion-exchange treatment does not destroy the photosensitive crystalline phase and this technology is available for photosensitive waveguide fabrication. However, parameters of coloration and relaxation of photochromic waveguides are different compared to bulk glass. For silver

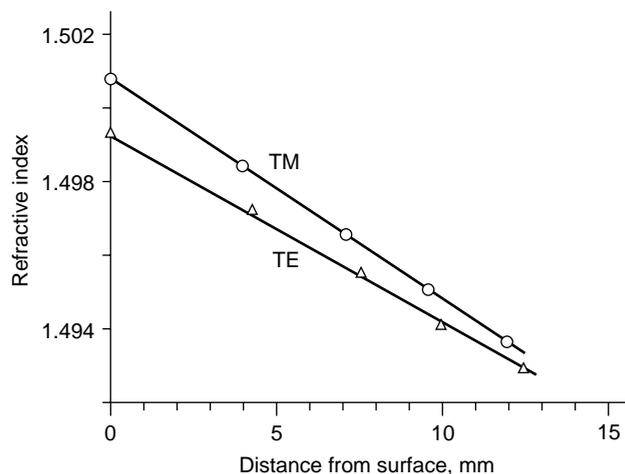


Figure 8. Refractive index profiles of photochromic glass after $\text{Na}_{\text{glass}}\text{-K}_{\text{melt}}$ ion exchange. TE or TM polarizations mean electric or magnetic field oriented along the surface, respectively.

halide glasses, the criterion of relaxation in waveguides is more than that in bulk glass. This means that relaxation in waveguides occurs faster. For copper halide glasses, relaxation in the waveguide was not detected, which means that the coloration of these waveguides is stable. There is a difference in photosensitivity between different waveguide modes. Modes with low numbers propagate near the surface and have lower sensitivity than modes that have a large number and propagate in deep layers. This difference is caused by copper (which is a sensitizer) depletion in the surface layer as result of copper exchange for potassium or other ions. This phenomenon can be used for mode selection.

The other feature of photochromic waveguides is anisotropy of photosensitivity and induced coloration. This phenomenon is connected with ion-exchange stresses. Dichroism (the difference between induced absorption for orthogonal polarizations) is proportional to birefringence in a waveguide. It is important to note that photosensitive microcrystals are plastic or melted at the temperatures of ion exchange. Therefore, dichroism is determined by stresses and also by orientation of liquid drops of the photosensitive phase caused by ion-exchange stresses.

The discrete structure of light propagation in photosensitive planar waveguides gives one more opportunity for multiplexing by mode selection. If a mode in such a waveguide (Mode #1 in Fig. 9) is excited by actinic radiation, the waveguide becomes colored. The spatial profile of induced absorption is determined by the spatial profile of the exciting modes intensity. As a result, a sort of distributed absorbing mask will be formed in the waveguide whose absorption profile is similar to that of the intensity distribution of actinic radiation in the waveguide. Consequently, losses for mode #1 increase after excitation of this mode by actinic radiation. The attenuation of other modes is determined by overlapping of their fields by the distributed mask, that is, by the field of the mode that induced this absorption. Because field profiles for the modes that have different numbers essentially differ from each other

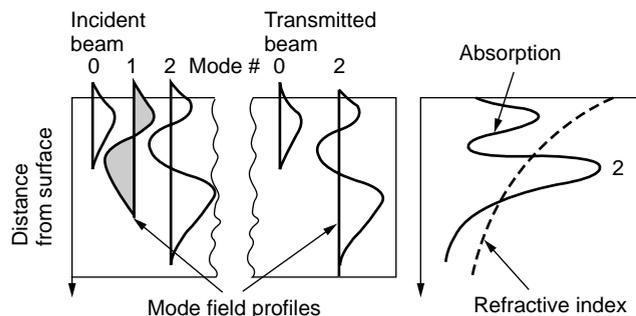


Figure 9. Sketch of a waveguide mode selector. The darkened profile corresponds to the exposed mode, which produces a similar profile of photoinduced absorption and prevents propagation of this mode.

(Fig. 9), the losses for different modes should be significantly different. An example of a mode spectrum of a planar waveguide excited by actinic radiation in the TE_0 mode is shown in Fig. 10. A mode selection of about 10 dB/cm can be reached without special effort in planar waveguides on commercial photochromic glasses. The problem of mask bleaching can be solved by using probe radiation at longer wavelengths, where bleaching is not effective, or using, as described earlier, cooperating breeding of color centers for writing by high-power radiation.

Optical fibers were drawn from photochromic glasses. It was found that thermal treatment of these fibers produces photochromic properties. Fiber plates were made from photochromic glass as a core and a transparent optical glass as a cladding, or vice versa. High contrast was obtained in this fiber element compared to bulk photochromic glass plate. This feature of photochromic fiber plate is determined by gradual leakage of actinic radiation from transparent glass to photochromic glass. This effect increases the length of the interaction of actinic radiation with photochromic glass and, consequently, increases dramatically the induced absorption and possible contrast of a photochromic attenuator.

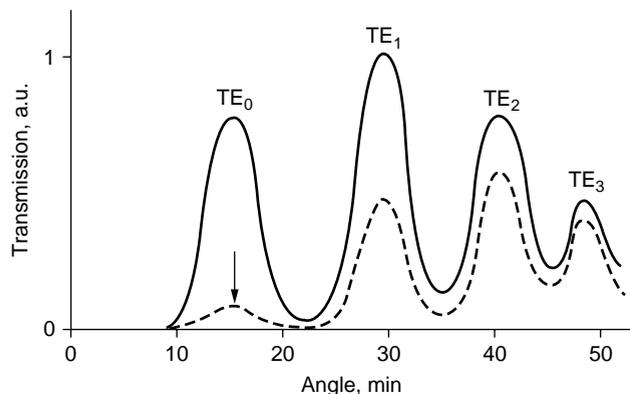


Figure 10. Effect of exposure to powerful excitation of the fundamental mode (shown by arrow) on the dependence of photochromic waveguide transmission on the angle of incidence onto the input coupler prism (spectrum of waveguide modes). Solid lines before exposure, dashed lines after exposure.

INDUCED REFRACTION THROUGH IRREVERSIBLE PHOTOINDUCED CRYSTALLIZATION

It is clear that photochromic glasses can be used for recording information. Actually some photos and holograms were recorded in these glasses but no great success was obtained because of small contrast in photography and small diffraction efficiency in holography. For highly efficient holography, it is necessary to produce variation in the refractive index but not in the absorption coefficient. The refractive index in glasses, where color centers are induced by radiation, can vary for very small values, less than 10^{-6} . This is not enough for efficient diffraction. Recent discovery of a strong photoinduced refractive index variation in Ge-doped silica opened a new very promising approach for efficient Bragg grating recording in optical fibers. Another approach, which allows an increase of sensitivity of several orders of magnitude compared to Ge-doped silica and avoids interaction between writing and diffracted beams, is based on a two-step process of exposure and development in multicomponent silicate glasses doped with fluorine, silver, and cerium.

Phase volume holograms of high diffraction efficiency were produced in lithium aluminum silicate and sodium zinc aluminum silicate glasses doped with silver and cerium by exposure to UV radiation followed by thermal treatment. Diffraction was caused by a difference in refractive indexes in exposed (enriched by microcrystals) and unexposed (original glass) areas. This phenomenon is called the "photo-thermorefractive" process. Glasses that possess these properties are called "photo-thermorefractive" (PTR) glasses. This two-step process (exposure and thermal development that leads to crystallization) was used earlier to record a translucent image in glass due to light scattering caused by a difference between the refractive indexes of the precipitated crystalline phase and the glass matrix. Later, colored images were recorded in similar glasses by photothermal precipitation of a number of complex crystals of different compositions, sizes, and shapes.

The sequence of processes, which occurs in these glasses and produces coloration, follows (Fig. 11). The first step is exposure of the glass to UV radiation, which ionizes a cerium ion. The electrons released from cerium are then trapped by a silver ion. As a result, silver is converted from a positive ion to a neutral atom. This second stage corresponds to latent image formation, and no significant changes in optical properties of glass occur, except light coloration in near UV and blue regions.

The next step in the process is obtained by thermal development at elevated temperatures. The high diffusion coefficient of silver atoms in silicate glasses leads to the creation of tiny silver crystals at temperatures from 450–500°C. A number of silver clusters arise in exposed regions of the glass after aging at these elevated temperatures. This is the third stage of the process. Further, these silver particles serve as the nucleation centers for sodium and fluorine ion precipitation. Cubic sodium fluoride crystal growth occurs at temperatures from 500–550°C because the PTR glass composition is an oversaturated solution of these components. This is the last step, which finishes the photo-thermorefractive process. Further heat treatment

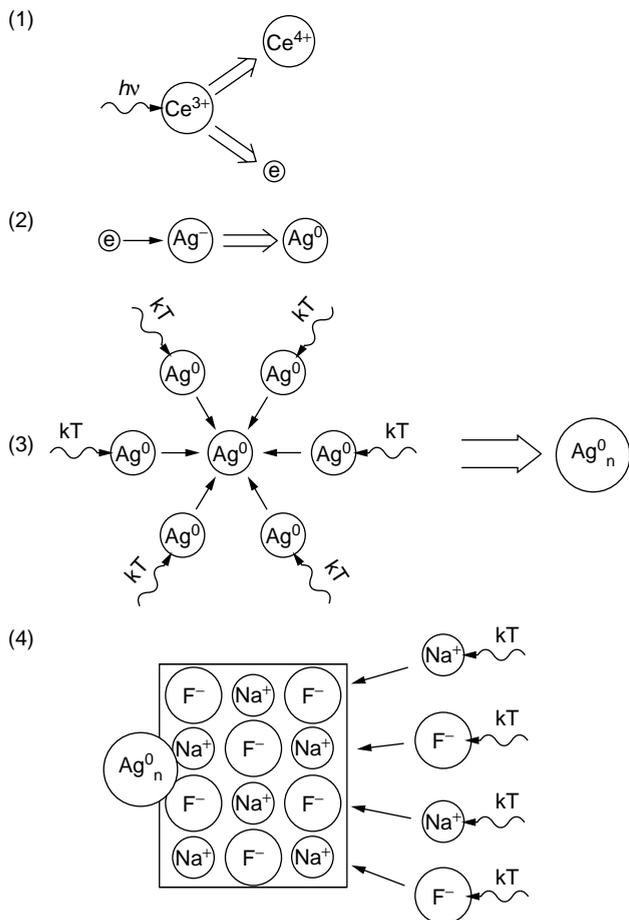


Figure 11. Stages of the photo-thermorefractive process.

leads to the growth of elongated pyramidal complex Na, Ag-F, Br crystals on the surface of cubic NaF crystals. This mixture of crystals can produce an opal coloration in large crystal sizes or a yellow coloration caused by colloidal silver precipitated on the interfaces of dielectric crystals. A second exposure to UV followed by a second heat treatment produces a different coloration because of metallic silver reduction on the surfaces of the dielectric pyramids. The final resulting coloration depends on the size and aspect ratio of these silver particles. These two last steps are used for photography because strong scattering does not allow using them in holography.

A refractive index decrease of about 5×10^{-4} occurs in the areas of glasses exposed to nitrogen laser radiation at 337 nm. The refractive index of NaF in the red spectral region is $n_{NaF} = 1.32$ compared to the refractive index of PTR glass $n_{PTR} = 1.49$. The small value of the refractive index change is due to the small volume fraction of the precipitated crystalline phase, which produces no scattering in the exposed volume. However, it is sufficient to result in highly efficient Bragg grating recording in samples more than several hundreds of microns thick. This photo-thermoinduced refraction is stable up to 400°C. The photosensitivity is in the range of several tens of mJ/cm^2 at wavelengths in the absorption band region of Ce^{3+} , which has a maximum

near 300 nm and a long wavelength tail up to 400 nm. This means that several commercial lasers such as N_2 , Ar, and He-Cd, can be used for recording. Once developed, holograms in PTR glass are not destroyed by further exposure to visible or UV radiation.

PHOTO-THERMOREFRACTIVE GLASS

The composition (mol. %) of PTR glass which was used for hologram recording is $15Na_2O-5ZnO-4Al_2O_3-70SiO_2-5NaF-1KBr-0.01Ag_2O-0.01CeO_2$. Absorption spectra of PTR glasses are presented in Fig. 12. Figure 12a shows the UV part of the absorption spectrum. One can see the wide absorption band of Ce^{3+} that has a maximum at 305 nm. The short wavelength absorption in the region $\lambda < 270$ nm is due to several components, such as Ce^{4+} , Ag^+ , Br^- , and Fe^{3+} . The short wavelength edge, at which writing radiation is attenuated by two times in the recording medium (optical density about 0.3), is placed at 330 nm for a 1-cm thick plate and at 265 nm for a 1-mm thick plate. The range of photosensitivity of this glass is from 280–360 nm.

Absorption of PTR glass is less than 0.01 cm^{-1} in the visible and near IR regions, which is close to the limit of measurements, and therefore it is not shown in Fig. 12. One can see in Fig. 12b that detectable absorption occurs at wavelengths higher than 2700 nm. Absorption in this spectral region is usually ascribed to different vibrations of hydroxyl groups in the glass network and reaches several cm^{-1} in regular silicate glasses. Hydroxyl absorption in fluorine-containing PTR glass is lower compared to similar fluorine-free silicate glass. This phenomenon is caused by high volatilization of HF molecules, which can result from the interaction of fluorine and hydrogen in the glass melting process. This decrease of IR absorption in PTR glass results in an opportunity for PTR use in the middle IR region up to 4300 nm for 1-mm thick specimens.

Additional absorption of PTR glass under UV exposure that is used in hologram recording in this glass is shown in Fig. 12c, curve 1. Detectable photoinduced absorption is seen only in the UV region. Even at the recording wavelength, this absorption is less than 0.1 cm^{-1} and cannot impact the recording process significantly. The small tail of the induced absorption spectrum in the blue region can be distinguished by the naked eye as a slight yellow coloration of the exposed area. Thermodevelopment causes colloidal silver and sodium fluoride precipitation in the glass matrix. Fluoride crystals are colorless and can result in scattering if the size of the crystals is too large (more than 100 nm). A shoulder near 450 nm in the additional absorption spectrum after thermal treatment in Fig. 12c (curve 2) is ascribed to silver particles in glass matrix. One can see that the visible additional absorption does not exceed 0.3 cm^{-1} and 0.03 cm^{-1} in the blue and red regions, respectively. This means that losses in this region do not exceed a few percent for a 1-mm thick plate. Additional absorption in the whole IR region is not detectable and therefore is not shown in Fig. 12c. Consequently, this glass can be used successfully at all wavelengths important for lasers and optical communication in the visible and near IR regions.

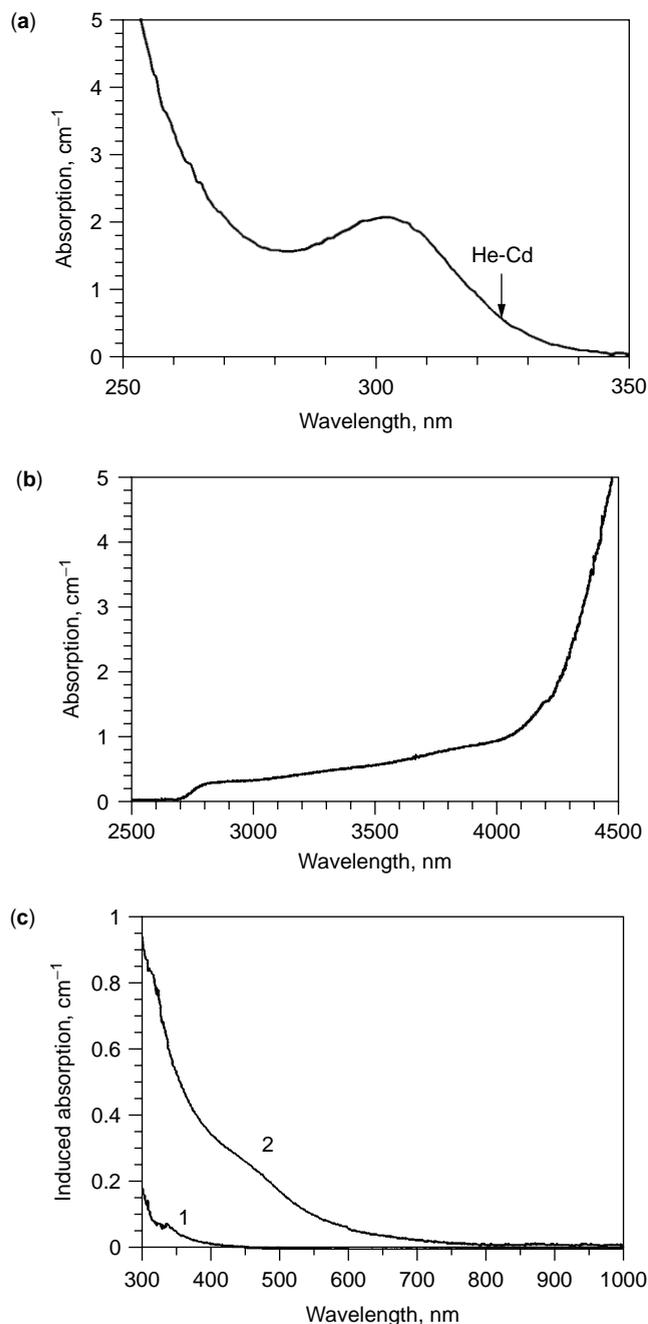


Figure 12. Absorption spectra of PTR glass: (a) and (b) original glass in the UV and IR spectral regions, (c) induced absorption after exposure to 325 nm for 400 mJ/cm² (12) and consequent thermal development for 1 hour at 520°C (13). Arrow shows the position of the wavelength of the writing He–Cd laser.

Optical microscopy of exposed and developed samples used for induced absorption measurements has shown optical inhomogeneities in the exposed region. The structure of these inhomogeneities appears as a series of parallel, continuous, aligned filaments whose widths are tens of microns oriented in the direction of light propagation in the glass sample. These microscopic features are caused by structures whose different refractive indexes arise in glass

processing (phase structures). It is proved that these phase patterns are not an intrinsic feature of PTR glass but are caused by various defects of the sample bulk and surfaces. Some additional patterns were found in micrographs; they are combinations of different rings and fringes. It was found that they are recordings of the interference patterns produced by matching propagating beams to beams consequently reflected from the back and front surfaces of different elements in the optical setup. Diffraction of the exciting beam on different apertures produces systems of straight or curved fringes that have variable periods dependent on the shape and position of the aperture. It is necessary to make special adjustments to eliminate these interference and diffraction patterns in the plane of the recording to avoid these parasitic structures. Therefore, the homogeneity of the photosensitive medium (including surface and volume defects) and the writing beam (including interference and diffraction patterns of low visibility) must be tested to avoid undesirable losses.

The pattern of probe radiation transmitted through exposed area consists of the zero and first orders of diffraction but exhibit some rings. The diameters and positions of these rings on the screen depend on the incident angle of the probe beam and on the feature of the writing pattern. The origin of these rings follows. Each medium causes scattering of propagating light. Therefore, even for single beams propagating in a photosensitive medium, one can observe an interference pattern produced by matching the original and scattered beams. In this case, the probe beam used for hologram reading should be scattered twice. The first time is regular scattering by the medium. The second time is scattering produced by a hologram of scattered light recorded together with the main hologram. This hologram can be completely reconstructed only by the reading beam of the same wavelength and direction as the writing beam. When the wavelengths or the directions of the writing and reading beams are different, the whole hologram of scattered light cannot be read out because its wavefronts are not planar. At each angle of incidence, the reading beam can read only that part of the hologram, for which Bragg conditions are satisfied. Because the angular diagram of scattering has cylindrical symmetry, this part should be a ring. All phase defects mentioned (filaments, fringes, and rings) appear in all materials but they are visible well in PTR glass because of the high homogeneity and transparency of this material.

BRAGG GRATINGS IN PTR GLASS

The dependence of the absolute diffraction efficiency of Bragg gratings recorded in PTR glasses in the thermal treatment period is shown in Fig. 13. The specimen exposed for 400 mJ/cm² has undergone consecutive thermal treatments for 10–15 minutes each at 520°C and in intervals between, was cooled down to room temperature for diffractive efficiency measurements. The absolute diffraction efficiency is

$$\eta_A = \frac{I_1}{(1 - \rho)^2 I_L}, \quad (2)$$

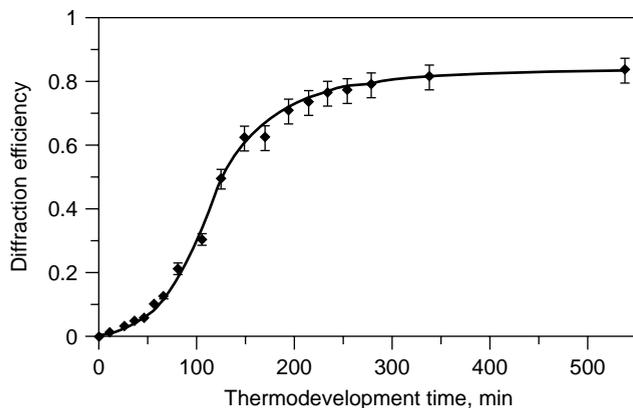


Figure 13. Effect of the period of thermal treatment on the absolute diffraction efficiency of a Bragg grating in PTR glass. Exposure 400 mJ/cm^2 at 325 nm , spatial frequency 600 mm^{-1} . Development at 520°C . Specimen thickness 1.42 mm .

where I_L and I_1 are the intensities of the incident and diffracted beams, respectively. The reflection coefficient (ρ) is calculated by the Fresnel formula $\rho = (n - 1/n + 1)^2$.

The dependence of diffraction efficiency versus development time has an inflection point at the beginning of the process and is saturated at the 85% level after long heat treatment. Note that this multiple heat treatment is not the same as a regular development for one or several hours because this procedure includes multiple heating and cooling. However, the curve in Fig. 13 shows a tendency for the diffraction efficiency to approach a high value after some exposure at elevated temperature.

The growth of diffraction efficiency in increasing periods of thermal development is obviously caused by refractive index changes that result from crystalline phase precipitation. Figure 14 shows the dependence of the refractive index on the thermal treatment period. This photothermoinduced refractive index was calculated from

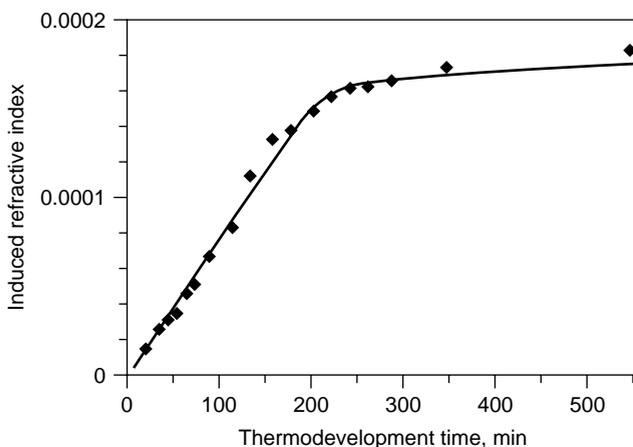


Figure 14. Effect of the period of thermal treatment on the induced refractive index. Exposure 400 mJ/cm^2 at 325 nm , spatial frequency 600 mm^{-1} . Development at 520°C . Specimen thickness 1.42 mm .

Kogelnik's equation:

$$\delta n = \frac{\lambda \cos \Theta \arcsin(\sqrt{\eta_R})}{\pi d}, \quad (3)$$

where λ is the wavelength of the reading beam, Θ is the Bragg angle, and d is the thickness of the specimen. The linear dependence of induced refractive index on the thermal treatment period is present in Fig. 14. The function $\delta n(t)$ shows no inflection point compared to $DE(t)$ (Fig. 13). The linear dependence of $\delta n(t)$ up to the value of 0.00015 allows writing high efficiency holograms in glass plates more than several hundreds of microns thick. The optical quality of inorganic glass allows using plates up to several centimeters thick. The saturation of the diffraction efficiency in Fig. 13 corresponds to the refractive index saturation at about 0.00017 in Fig. 14. No oscillations of diffraction efficiency were recorded in this experiment in long development periods up to 13 hours. This means that no significant result exceeding π for the induced phase was obtained and, consequently, no additional refractive index growth occurred.

The effect of the spatial frequency of the interference pattern on the diffraction efficiency of the grating in PTR glasses is shown in Fig. 15. This was measured in a thin sample of 1.65 mm in a transmittance configuration when writing (325 nm) and reading (633 nm) beams were directed from the same side of the glass plate. This configuration allows spatial frequency variations below 2500 mm^{-1} . Exposure or development of gratings was not optimized for different spatial frequencies. No significant dependence of diffraction efficiency on special frequency can be observed in the region from $300\text{--}2500 \text{ mm}^{-1}$ in Fig. 15. The absence of a drop in the frequency response at low frequencies is a feature of the PTR process, which requires transport of species in the glass matrix to build single crystals (tens of nanometers) and does not require transport of species between exposed and unexposed areas, as is necessary in photorefractive crystals. The

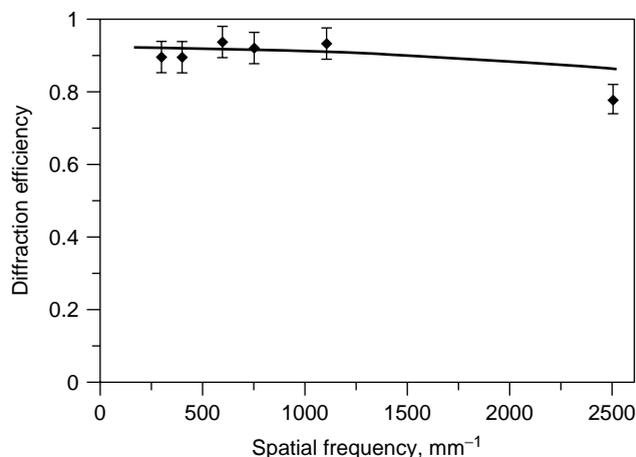


Figure 15. Dependence of the absolute diffraction efficiency on the spatial frequency of the grating. Exposure 600 mJ/cm^2 at 325 nm , development 90 min . at 520°C . Specimen thickness 1.65 mm .

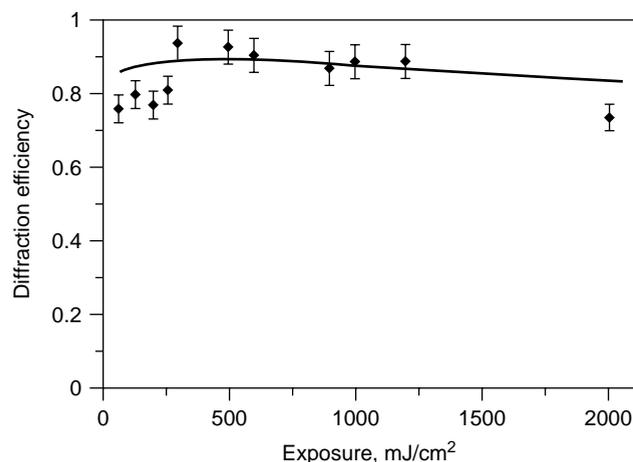


Figure 16. Maximum absolute diffraction efficiencies of Bragg gratings in PTR glasses for different exposures to the radiation of a He-Cd laser at 325 nm.

absence of a drop at high spatial frequencies means that no fringe smearing occurs in the developed interferogram and, consequently, no detectable diffusion of components at distances comparable with the half-period of the gratings studied (up to 200 nm) occurs in PTR glass during thermal processing. These data show that diffusion of glass components in the development process cannot affect the saturation in Fig. 14, which was observed for gratings that have a spatial period of 1600 nm. The lack of drop in the amplitude-frequency response at low frequencies (Fig. 15) is an advantage of PTR glasses compared to photorefractive crystals; this results in a distinct opportunity to design holographic optical elements that have very small diffraction angles.

An interesting consequence of the low level of induced losses (Fig. 12c, curve 2) is the rather low sensitivity of PTR-grating diffraction efficiency on exposure because underexposure can be compensated for by overdevelopment, and vice versa. Figure 16 illustrates this feature of PTR glass. In this figure, the best diffraction efficiencies for specimens of different thickness from different melts, which had undergone different development procedures, are plotted versus exposure to the radiation of a He-Cd laser. A high absolute diffraction efficiency of 80% and more is observed in Fig. 16 for exposures that ranged between 50 mJ/cm² and 5 J/cm².

SUMMARY

Photochromic glasses that have completely reversible coloration are made of borosilicate glasses doped with microcrystals of copper and silver halides. These glasses are sensitive to near UV radiation. Photosensitivity can be extended to visible and near IR regions by cooperative breeding of color centers. Induced coloration is a wide band that covers the whole visible region. Photocontrolled waveguides can be fabricated in photochromic glasses. These waveguides can serve as attenuators and mode selectors. Photo-thermorefractive glasses that have irreversible

photoinduced refraction are aluminosilicate glasses doped with silver, cerium, and fluorine. These glasses are sensitive to near UV radiation. Their photosensitivity is comparable with the best organic and inorganic materials, it allows wide variations of exposure because of image amplification in the thermal development process, and it has high diffraction efficiency and high transparency from the UV to the IR region.

BIBLIOGRAPHY

1. S.D. Stookey, *Ind. Eng. Chem.* **41**: 856–861 (1949).
2. US Pat. 3, 208, 860, 1965, W.H. Armistead and S.D. Stookey.
3. R.J. Araujo and N.F. Borrelli, in *Optical Properties of Glass*, D.R. Uhlmann and N.J. Kreidl, eds., Westerville, OH, 1991: 125.
4. A.V. Dotsenko, L.B. Glebov, and V.A. Tsekhomskii, *Physics and Chemistry of Photochromic Glasses*. CRC, Boca Raton, FL, 1997.
5. L.B. Glebov, N.V. Nikonorov, E.I. Panysheva, G.T. Petrovskii, V.V. Savvin, I.V. Tunianova, and V.A. Tsekhomskii, *Sov. Phys. Dokl.* **35**: 878 (1990).
6. L.B. Glebov, N.V. Nikonorov, E.I. Panysheva, G.T. Petrovskii, V.V. Savvin, I.V. Tunimanova, and V.A. Tsekhomskii, *Opt. Spectrosc.* **73**: 237 (1992).
7. O.M. Efimov, L.B. Glebov, L.N. Glebova, K.C. Richardson, and V.I. Smirnov, *Appl. Opt.* in press.
8. L.B. Glebov, *Glass Sci. Technol. (Glastechnische Berichte)*, in press.
9. S.D. Stookey, G.H. Beall, and J.E. Pierson, *J. Appl. Phys.* **49**: 5114–5123 (1978).
10. N.F. Borrelli, J.B. Chodak, D.A. Nolan, and T.P. Seward, *J. Opt. Soc. Am.* **69**: 1514–1519 (1979).
11. A.V. Dotsenko, A.M. Efremov, V.K. Zakharov, E.I. Panysheva, and I.V. Tunimanova, *Fiz. i Khim. Stekla* **11**: 592–595 (1985) (in Russian).
12. E.I. Panysheva, I.V. Tunimanova, and V.A. Tsekhomskii, *Glass Phys. Chem.* **17**: 543–549 (1991).
13. V.I. Arbuzov, *Glass Phys. Chem.* **22**: 477–489 (1996).
14. L.B. Glebov, O.M. Efimov, A.M. Mekryukov, and Yu.A. Matveev, *J. Opt. Technol.* **62**: 780–785 (1995).

PIEZOELECTRICITY IN POLYMERS

ALEKSANDRA VINOGRADOV
Montana State University
Bozeman, MT

INTRODUCTION

The diverse group of “smart” piezoelectric materials is distinguished by their ability to react actively to changing stimuli as a result of converting mechanical to electrical energy and vice versa. Synthetic piezoelectric polymers, an integral part of the “smart” materials group, exhibit a type of behavior that is often compared with biological reactions involving transformations of the sensed information into the desired response. Due to such

special qualities, piezoelectric polymers have been increasingly used in a rapidly expanding range of applications. At present, these materials continue to offer unprecedented design opportunities, leading to the belief that the industry is on the verge of major technological breakthroughs.

PIEZOELECTRICITY: AN OVERVIEW

Piezoelectricity is a material property that is observed as an electric charge or voltage produced by applied mechanical forces or, conversely, as mechanical deformation that is caused by an applied electric field. These piezoelectric effects have been defined, respectively, as “direct” and “converse.” The latter classification provides a convenient basis for reference purposes, although it is clear that both phenomena have the same physical origin.

Rapid progress in piezoelectric investigations was made at the beginning of the twentieth century after Pierre and Jacques Curie discovered the direct piezoelectric effect in tourmaline crystals in 1880. Subsequently, piezoelectric effects were observed and studied in other crystals, such as quartz, zinblende and Rochelle salt, providing enhanced understanding of the piezoelectric phenomenon and leading to new discoveries of piezoelectric effects in a variety of materials. In the 1940s, research efforts were particularly focused on the piezoelectric response of ferroelectric polycrystalline ceramics, including lead zirconate titanate (PZT), lithium niobate, and barium titanate. For several decades, and, increasingly, toward the mid-1960s, piezoelectricity was investigated as a common property of biopolymers, including natural biological materials that form the structures of plants, animals, and humans. Since 1969, when the strong piezoelectric effect in polyvinylidene fluoride (PVDF) was first discovered by Kawai, attention has been attracted to the piezoelectric properties of synthetic polymers. At present, the traditional group of smart materials involving piezoelectric crystals, ceramics, and polymers is expanding as a new generation of laminated composites that have embedded piezoelectric elements has recently emerged. The history of scientific developments in the dynamic and growing field of smart materials has been reviewed in (1–3).

In phenomenological terms, piezoelectricity is described as coupling between a quasi-static electric field and dynamic mechanical motion. Typically, the direct and converse piezoelectric effects have been treated as reversible. Respectively, the constitutive equations of linear piezoelectricity are based on the principle of energy conservation. The piezoelectric constitutive law can be presented in several alternative forms. One of the formulations is given by

$$\begin{aligned} [\varepsilon] &= [\mathbf{C}][\sigma] + [\mathbf{d}]^T[E], \\ [D] &= [\mathbf{d}][\sigma] + [\mathbf{e}][E], \end{aligned} \quad (1)$$

where $[\sigma]$ and $[\varepsilon]$ denote, respectively, stress and strain tensors that satisfy the condition of symmetry, that is, $\sigma_{ij} = \sigma_{ji}$, and $\varepsilon_{kl} = \varepsilon_{lk}$ ($i \neq j$, $k \neq l$); $[D]$ and $[E]$ denote, respectively, the electric flux density and the electric field;

$[\mathbf{C}]$ is the elastic compliance matrix whose components satisfy the condition $c_{ijkl} = c_{ijlk} = c_{jikl} = c_{klij}$; $[\mathbf{d}]$ is the matrix of piezoelectric coefficients $d_{ijk} = d_{ikj}$; $[\mathbf{d}]^T$ is the transpose of $[\mathbf{d}]$; and $[\mathbf{e}]$ represents the dielectric permittivity matrix whose components $e_{ij} = e_{ji}$ ($i \neq j$, $k \neq l$), $i, j, k, l = 1, 2, 3$. Other forms of the linear piezoelectric constitutive equations are given in (4).

In the general case of fully populated matrices $[\mathbf{C}]$, $[\mathbf{d}]$, and $[\mathbf{e}]$, the electromechanical properties of an anisotropic piezoelectric continuum are defined by 21 independent elastic constants, 18 piezoelectric coefficients, and 6 dielectric constants. However, the actual number of parameters required to characterize the properties of various piezoelectric materials is less than the total of 45. The structure and content of the matrices $[\mathbf{C}]$, $[\mathbf{d}]$, and $[\mathbf{e}]$ depend on the type of material microstructure. The anisotropic properties of piezoelectric crystals and, respectively, the composition of the matrices $[\mathbf{C}]$, $[\mathbf{d}]$, and $[\mathbf{e}]$ are determined by the type of symmetry in the crystal lattice. Because only those crystals that possess no center of symmetry on the atomic scale tend to exhibit piezoelectric effects, only 20 out of 32 crystallographic classes of crystals are piezoelectric. Specific characteristics of various groups of piezoelectric crystals and ceramics, their classification, and properties have been considered in (1,4,5). The material properties of piezoelectric polymers are discussed in detail in the following sections.

It is important to note that the theory of linear piezoelectricity is based on the assumptions of infinitesimal deformations, linear stress–strain relations, and stationary electric fields with respect to an inertial reference frame. Attempts have been made to develop more general nonlinear piezoelectric material models that take into account the effects of higher order electromechanical couplings, such as electrostriction, nonlinear strain-displacement relations, and the material response to large driving voltages. Research efforts in this regard have been reviewed (4,6). A systematic account of anelastic properties of piezoelectric polymers has been given in (7).

SYNTHETIC PIEZOELECTRIC POLYMERS

The diverse group of piezoelectric materials includes a variety of synthetic polymers such as polypropylene, polystyrene, and poly(methyl methacrylate); semicrystalline polyamides such as nylon-11; and amorphous polymers such as vinyl acetate. However, piezoelectric effects in these materials are relatively weak, often unstable, and are considered of limited practical significance. Strong piezoelectricity has been observed only in the synthetic polymer poly(vinylidene fluoride) (PVDF or PVF₂) and PVDF copolymers.

Poly(vinylidene fluoride) is a semicrystalline polymer whose typical crystallinity is approximately 50%. The amorphous phase of the polymer has the properties of a supercooled liquid. The glass transition temperature of the polymer is about -50°C . The molecular structure of poly(vinylidene fluoride) consists of the repeated monomer unit $-\text{CF}_2-\text{CH}_2-$. The atoms are covalently bonded, forming long molecular chains. Because the hydrogen atoms are

positively charged and the fluoride atoms are negatively charged with respect to the carbon atoms, PVDF is inherently polar. However, the net polar moment of the material in its original state is zero due to the random orientation of the individual crystallites.

Permanent dipole polarization of PVDF is obtained through a technological process that involves stretching and polarizing extruded thin sheets of the polymer. Stretching aligns molecular chains in the stretch direction. An applied electric field of up to 100 kV/mm at an elevated, typically, 103°C temperature causes permanent polarization that is maintained after the material cools to room temperature. Sessler (8) provides an overview of polymer polarization methods. In general, it has been observed that polarization in PVDF depends on a number of factors, including polarizing temperature, polarizing time, polarizing process, electrode conditions, and the morphology of the material.

Typically, PVDF is produced in thin films whose thicknesses range from 9 to 800 μm (10^{-6} m). A thin layer of nickel, silver, or copper is deposited on both film surfaces to provide electrical conductivity when an electric field is applied, or to allow measuring the charge induced by mechanical deformation.

ELECTROMECHANICAL PROPERTIES OF PVDF

Since the discovery of piezoelectric effects in PVDF (9), the properties of this material have been studied by many investigators. Research accomplishments in this subject area have been reviewed in (8,10,11).

Typically, the piezoelectric properties of PVDF are determined within the framework of linear piezoelectric theory. An expanded form of the constitutive law defined by Eqs. (1) is formulated for piezoelectric polymers as

$$\begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{22} \\ \varepsilon_{33} \\ \varepsilon_{23} \\ \varepsilon_{31} \\ \varepsilon_{12} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & c_{13} & 0 & 0 & 0 \\ c_{12} & c_{22} & c_{23} & 0 & 0 & 0 \\ c_{13} & c_{23} & c_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & c_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & c_{55} & 0 \\ 0 & 0 & 0 & 0 & 0 & c_{66} \end{bmatrix} \begin{bmatrix} \sigma_{11} \\ \sigma_{22} \\ \sigma_{33} \\ \sigma_{23} \\ \sigma_{31} \\ \sigma_{12} \end{bmatrix} + \begin{bmatrix} 0 & 0 & d_{31} \\ 0 & 0 & d_{32} \\ 0 & 0 & d_{33} \\ 0 & d_{24} & 0 \\ d_{15} & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} E_1 \\ E_2 \\ E_3 \end{bmatrix} \quad (2)$$

$$\begin{bmatrix} D_1 \\ D_2 \\ D_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & d_{15} & 0 \\ 0 & 0 & 0 & d_{24} & 0 & 0 \\ d_{31} & d_{32} & d_{33} & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \sigma_{11} \\ \sigma_{22} \\ \sigma_{33} \\ \sigma_{23} \\ \sigma_{31} \\ \sigma_{12} \end{bmatrix} + \begin{bmatrix} e_{11} & 0 & 0 \\ 0 & e_{22} & 0 \\ 0 & 0 & e_{33} \end{bmatrix} \begin{bmatrix} E_1 \\ E_2 \\ E_3 \end{bmatrix}$$

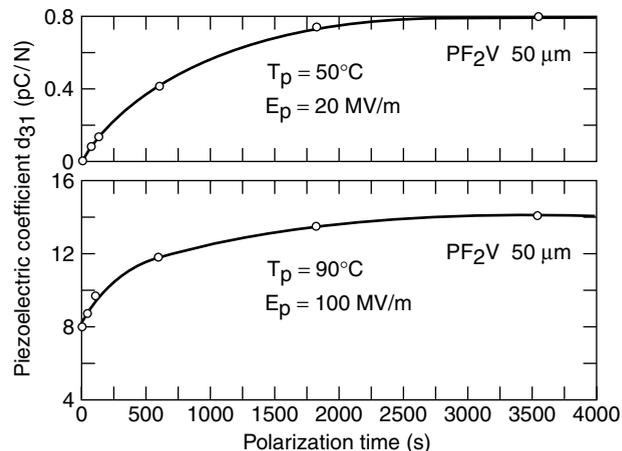


Figure 1. Dependence of the coefficient d_{31} of PVDF on polarization time t_p (12).

The latter equations are formulated in an orthogonal coordinate system 1-2-3, so that axes 1 and 2 are considered in the plane of a PVDF film, whereas axis 3 is normal to the film surface. Axes 1 and 2 are, respectively, parallel and normal to the orientation of the polymer's aligned molecular chains.

According to Eqs. (2), coupling of the electromechanical material properties of PVDF is characterized by five piezoelectric coefficients contained in the matrix $[d]$. The most important coefficients that determine the magnitude of piezoelectric effects are the coefficients d_{3j} , ($j = 1, 2, 3$). Sometimes, the hydrostatic coefficient, $d_h = d_{31} + d_{32} + d_{33}$ that determines the electric charge generated by hydrostatic pressure is used to represent the degree of piezoelectric effects in a material.

The values of the piezoelectric coefficients of PVDF depend on the polarization conditions in terms of the polarization time t_p , polarization temperature T_p , and polarizing field strength E_p (12). In particular, the dependence of the coefficient d_{31} on t_p , T_p , and E_p for a PVDF thin film stretched at a 4:1 ratio, is illustrated in Figs. 1–3.

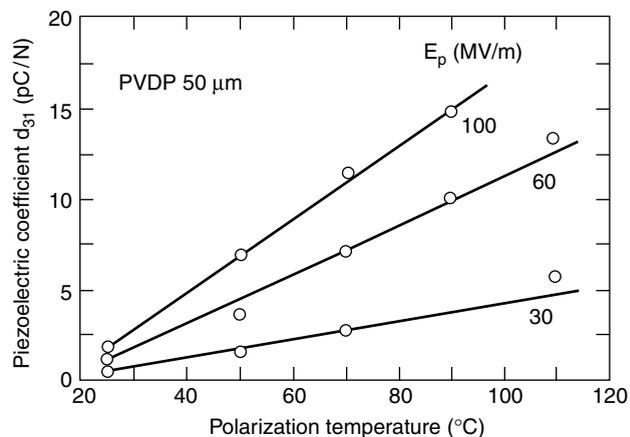


Figure 2. Dependence of coefficient d_{31} of PVDF on polarization temperature T_p (12).

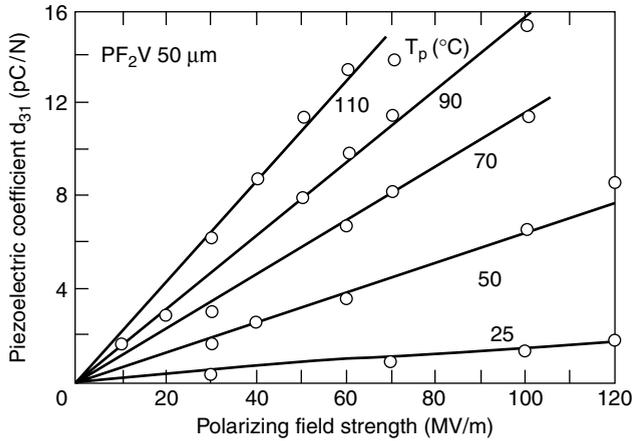


Figure 3. Dependence of coefficient d_{31} of PVDF on polarizing field strength E_p (12).

A number of experimental techniques have been developed to determine the values of the piezoelectric coefficients of PVDF. In particular, the response of 20- μm thick PVDF films has been studied under the conditions of superimposed static and sinusoidal loads (13). The electric charge resulting from the mechanical loading has been measured for various values of the static load and at various temperatures; the amplitude (0.15 N) and frequency (15 Hz) of the dynamic load remained unchanged. It has been determined that the piezoelectric coefficient d_{31} of PVDF strongly depends on temperature, particularly, in the range from -40 to -50°C , close to the glass transition temperature T_g . A similar dependence of coefficient d_{31} on temperature has been observed in (14).

The electromechanical response of PVDF as a function of temperature has been studied in (15) using the piezoelectric resonance method. By applying an alternating stress in the material directions 1, 2, and 3 and using polarization measurements along axis 3, it has been determined that $d_{31}, d_{32} > 0$, and $d_{33} < 0$. In addition, it has been observed that the piezoelectric coefficients of PVDF tend to increase with temperature, as illustrated in Fig. 4.

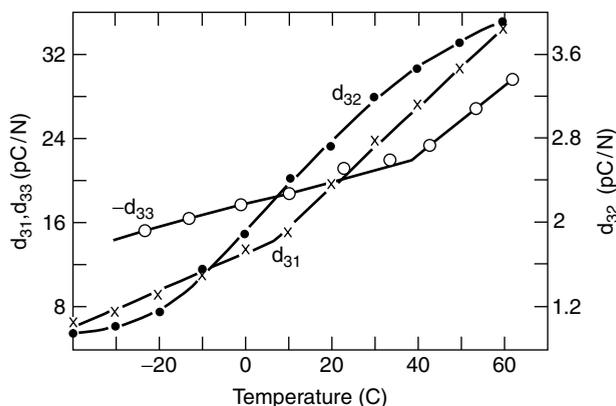


Figure 4. Piezoelectric coefficients d_{31} , d_{32} , and d_{33} of PVDF as functions of temperature (15).

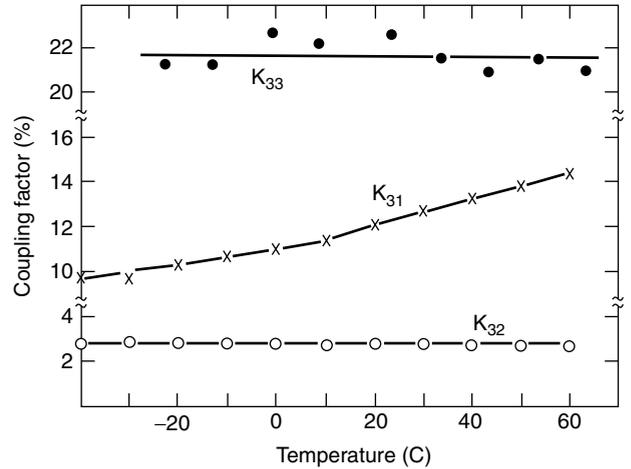


Figure 5. Electromechanical coupling factors k_{31} , k_{32} , and k_{33} of PVDF as functions of temperature (15).

The piezoelectric properties of PVDF have been also characterized in terms of electromechanical coupling factors k_{31} , k_{32} , and k_{33} . These coefficients represent the ratios between the dissipated and input energies in the respective material directions. It has been determined (15) that the electromechanical coupling factor k_{31} of PVDF tends to increase with temperature, whereas k_{32} and k_{33} remain temperature insensitive. These results are illustrated in Fig. 5.

The shear piezoelectric properties of uniaxially oriented PVDF films have been studied in (16). It has been observed that polarization of PVDF samples is linearly proportional to applied shear stresses. It has been determined that the values of the piezoelectric coefficients d_{15} and d_{24} range from -13 pC/N^{-1} to -27 pC/N^{-1} and from -23 pC/N^{-1} to -38 pC/N^{-1} , respectively.

The mechanical properties of PVDF have been defined by the constitutive equations of linear elasticity in the form of a generalized Hooke's law. For orthotropic materials, the coefficients of the compliance matrix $[C]$ in Eqs. (2) can be represented such that

$$\begin{aligned} c_{11} &= 1/Y_1, c_{22} = 1/Y_2, c_{33} = 1/Y_3, c_{44} = 1/2G_{23}, \\ c_{55} &= 1/2G_{31}, c_{66} = 1/2G_{12} \\ c_{12} &= -\nu_{12}/Y_1 = -\nu_{21}/Y_2, c_{13} = -\nu_{13}/Y_1 = -\nu_{31}/Y_3, \text{ and} \\ c_{23} &= -\nu_{23}/Y_2 = \nu_{32}/Y_3 \end{aligned} \quad (3)$$

where Y_1 , Y_2 , and Y_3 are the elastic moduli in directions 1, 2, and 3, respectively; G_{12} , G_{31} , and G_{23} denote the shear moduli; and ν_{12} , ν_{23} , and ν_{31} are Poisson ratios whose first index indicates the direction of contraction or expansion and the second indicates the direction of force action. Note that due to the symmetry of the compliance matrix $[C]$, the mechanical properties of PVDF thin films are characterized by nine independent elastic constants.

The elastic response of PVDF has been studied in (14,15,17–21). It has been observed that the experimental values of the elastic moduli Y_1 and Y_2 have been consistently very close. This result has been often interpreted

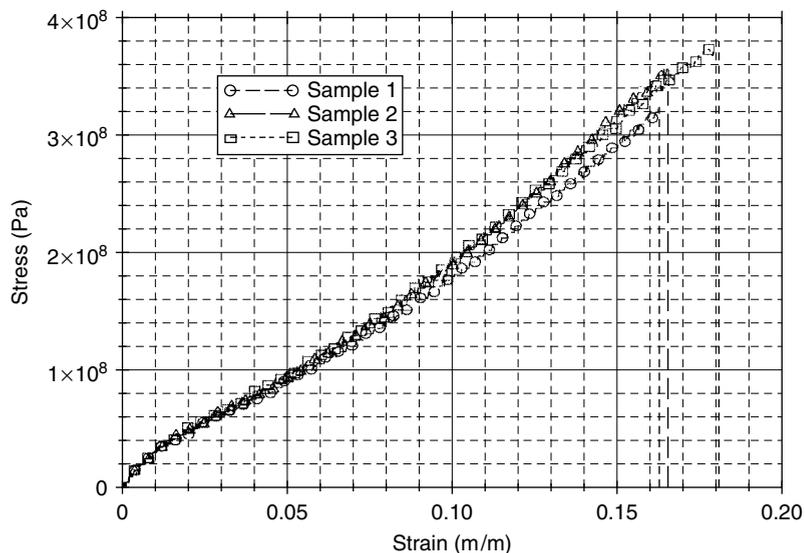


Figure 6. Stress–strain response of PVDF (direction 1) (20).

as evidence that the mechanical properties of PVDF are isotropic. However, it has been demonstrated in (18–20) that PVDF thin films exhibit significantly different responses, depending on the orientation of the aligned molecular chains. In the latter studies, 28- μm PVDF samples were tested under displacement controlled experimental conditions at a strain rate of 1.27 cm/min. The respective stress–strain diagrams for both in-plane material directions of PVDF are given in Figs. 6 and 7.

It is clear that the mechanical properties of PVDF thin films strongly depend on the orientation of the polymer's molecular chains aligned in the stretch direction. The diagram in Fig. 6 demonstrates that the stress–strain response of the material in the direction of the aligned molecular chains (direction 1) is characterized by a continuous increase of stresses that culminates in sudden failure. This type of response is typical for brittle materials. In

contrast, the stress–strain diagram in Fig. 7 for the material direction normal to the alignment of molecular chains (direction 2) is characteristic of ductile material behavior that involves an increase in stresses up to a certain maximum value and a following sharp decrease of load-carrying capacity.

Besides the observed differences in the stress–strain behavior, the ultimate stresses $(\sigma_u)_i$ and ultimate strains $(\varepsilon_u)_i$ ($i = 1, 2$) in the respective in-plane material directions of PVDF have considerably different values: $(\sigma_u)_1 = 3.5 \times 10^8$ Pa, and $(\sigma_u)_2 = 5 \times 10^7$ Pa.

The Poisson ratio for uniaxially stretched PVDF films has been measured experimentally in (21). Material samples were subjected to uniaxial tension in the direction of the aligned molecular chains. The values of the Poisson ratios ν_{31} and ν_{21} were obtained by measuring the respective deformations in the thickness and width directions of

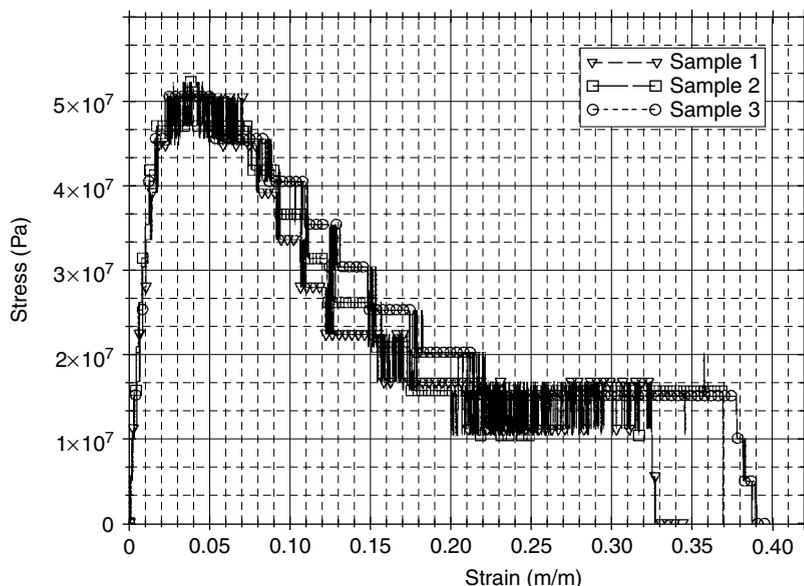


Figure 7. Stress–strain response of PVDF (direction 2) (20).

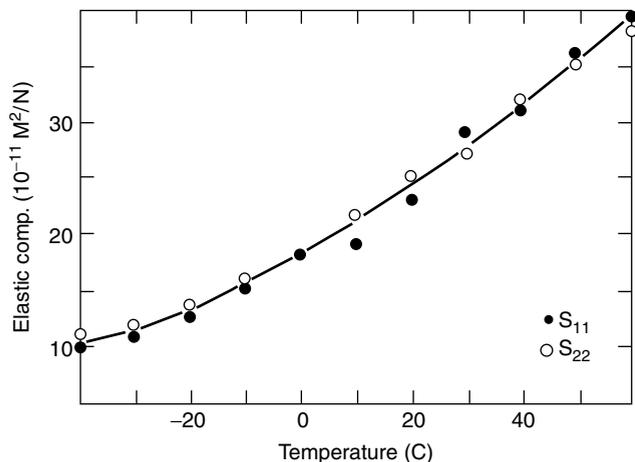


Figure 8. Elastic compliances of PVDF as functions of temperature (15).

the samples. It has been determined that $\nu_{21} \sim 0.1$ and $\nu_{31} \sim 0.8$. It is important to note that the value of ν_{31} exceeds 0.5, the theoretical maximum possible value of the Poisson ratio for isotropic elastic materials. This result indicates that PVDF thin films are highly anisotropic.

Experimental studies (14,15,22–24) indicate that the elastic properties of PVDF are temperature-dependent. In particular, according to the results reported in (15), the elastic compliances of PVDF increase with temperature. The yield stress and yield strain of PVDF are also temperature-dependent (22). These results are illustrated in Figs. 8, 9, and 10.

Due to the fact that the electromechanical response of PVDF depends on a number of factors, including polarization conditions, stress/strain rates, temperature, and hydrostatic pressure, the reported data for the values of the piezoelectric and elastic constants of the polymer appear to involve certain inconsistencies. Nevertheless, it is possible to identify the typical values of the electromechanical characteristics of PVDF such as summarized in Table 1.

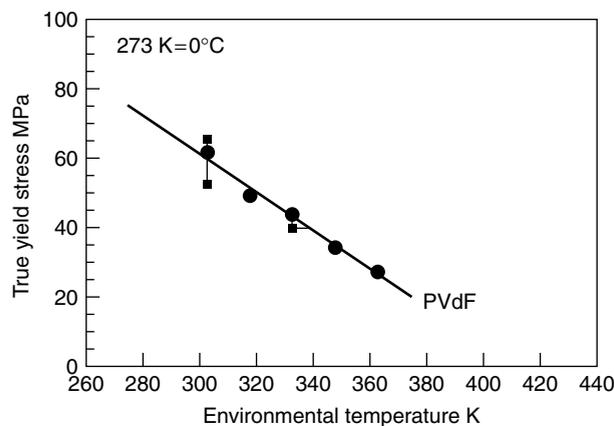


Figure 9. Temperature dependence of the true yield stress of PVDF (22).

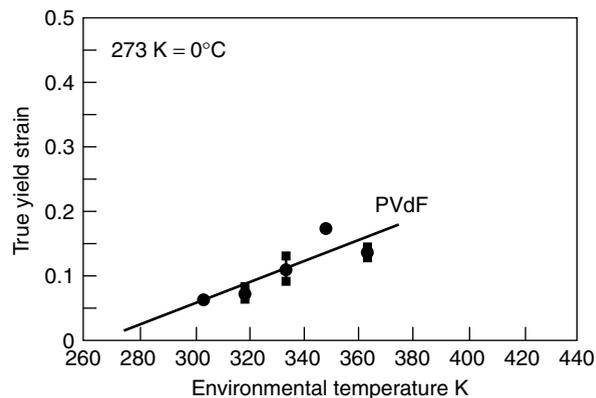


Figure 10. Temperature dependence of the true yield strain of PVDF (22).

NONLINEAR AND TIME-DEPENDENT EFFECTS

The constitutive law of linear piezoelectricity in the form of Eqs. (1) tends to neglect energy dissipation, time-dependent effects, and various nonlinearities in the electromechanical response of piezoelectric materials. However, there is consistent experimental evidence that these assumptions have certain limitations. It has been observed that, in general, all piezoelectric materials exhibit nonlinear effects, as well as dielectric and mechanical energy losses, although to different degrees. Thus, energy losses in piezoelectric crystals and ceramics are negligible (26,27), whereas in piezoelectric polymers such effects are of practical significance (28).

One study demonstrates strong nonlinear dependence of the transverse piezoelectric response of PVDF on the applied stress (29). It has been observed that the piezoelectric coefficient d_{32} of 22- μm uniaxially oriented PVDF films becomes negative under large stresses. This effect appeared reversible upon unloading but tended to repeat itself in subsequent loading–unloading cycles.

Under cyclic conditions, piezoelectric polymers exhibit energy losses observed from hysteresis loops formed by the electric displacement D as a function of electric field E (10, 30–34). Furukawa et al. (30) subjected 20- μm thick PVDF films to high sinusoidal electric fields whose amplitudes ranged from 40 to 120 MV/m in the frequency range of 10^{-4} – 10^{-2} Hz at temperatures between -100 and 100°C . These experiments demonstrated a strong dependence of D on temperature and on the amplitude and frequency of the electric field. At sufficiently high electric fields, D – E hysteresis loops have been observed, even in the temperature range below the glass transition temperature of the polymer. The D – E response of PVDF samples at different temperatures is illustrated in Fig. 11.

D – E hysteresis loops similar to those shown in Fig. 11 have been obtained for PVDF copolymers, vinylidene fluoride-trifluoroethylene (VDF-TrFE), and vinylidene fluoride-tetrafluoroethylene (VDF-TFE) (35–37). Similarly, the piezoelectric coefficients of PVDF and its copolymers have demonstrated hysteresis under variable electric fields (38–41).

Table 1. Electromechanical Characteristics of PVDF

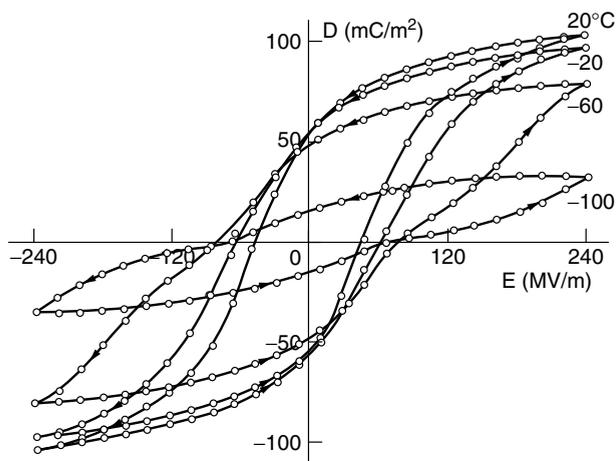
Property	Value
Range of film thicknesses	9–800 μm (10^{-6} m)
Mass density, ρ	1.78×10^3 kg/m ³
Water absorption	0.02%
Operating temperature range	–40 to 80°C
Glass transition temperature, T_g	–60 to –20°C
Melting temperature, T_m	170–178°C
Maximum operating voltage	30 V/ μm (750 V/mil)
Breakdown voltage	100 V/ μm (2000 V/mil)
Capacitance	380 pF/cm ² for 28- μm films at 1 kHz
Piezoelectric coefficients	$d_{31} = 21.4 \times 10^{-12}$ C/N; $d_{32} = 2.3 \times 10^{-12}$ C/N; $d_{33} = -31 \times 10^{-12}$ C/N; $d_{24} = -35 \times 10^{-12}$ C/N; $d_{15} = -27 \times 10^{-12}$ C/N
Electromechanical coupling factors	$k_{31} = 12\%$; $k_{32} = 3\%$; $k_{33} = 19\%$ at 1 kHz
Permittivity	$(106\text{--}113) \times 10^{12}$ F/m
Relative permittivity	12–13
Young's moduli	$Y_1 = 2.56 \times 10^9$ Pa; $Y_2 = 2.6 \times 10^9$ Pa
Yield stress	$(\sigma_y)_1 = 4.5 \times 10^7$ Pa; $(\sigma_y)_2 = 3.9 \times 10^7$ Pa
Yield strain	$(\epsilon_y)_1 = 1.8\%$; $(\epsilon_y)_2 = 1.4\%$;
Ultimate stress	$(\sigma_u)_1 = 3.5 \times 10^8$ Pa; $(\sigma_u)_2 = 5 \times 10^7$ Pa
Ultimate strain	$(\epsilon_u)_1 = 16.9\%$; $(\epsilon_u)_2 = 2.5\%$
Dielectric loss factor	0.015–0.25
Mechanical loss tangent	0.10

One of the most obvious indicators of the time-dependent behavior of piezoelectric polymers is their tendency to undergo piezoelectric, dielectric, and mechanical relaxation. As an example, relaxation of the piezoelectric coefficient d_{31} of PVDF thin films stretched at a 5:1 ratio and polarized for 2 hours at a polarization temperature $T_p = 85^\circ\text{C}$ and electric field $E_p = 50$ MV/m is illustrated in Fig. 12.

The relaxation properties of PVDF attracted attention in the early 1960s (42), and since that time, the relaxation phenomenon in piezoelectric polymers has been studied extensively (43–47). Typically, piezoelectric, dielectric, and mechanical relaxation of piezoelectric polymers is characterized by the complex coefficients

$$d_{ijk} = d_{ijk}' - id_{ijk}'',$$

$$e_{ij} = e_{ij}' - ie_{ij}'',$$

**Figure 11.** D – E hysteresis loops of PVDF at various temperatures (30).

and

$$c_{ijkl} = c_{ijkl}' - ic_{ijkl}'', \quad (4)$$

that determine energy losses in terms of the respective loss tangents such as

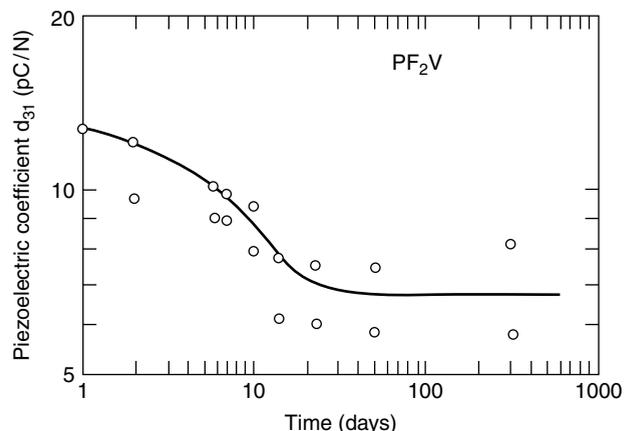
$$(\tan \delta_{ijkl})_m = c_{ijkl}''/c_{ijkl}',$$

$$(\tan \delta_{ijk})_p = d_{ijk}''/d_{ijk}',$$

and

$$(\tan \delta_{ij})_d = e_{ij}''/e_{ij}'. \quad (5)$$

The experimental values of the dielectric complex coefficients for PVDF and P(VDF-TrFE) are summarized in (43). In general, it has been observed that the dielectric complex coefficients of PVDF and its copolymers depend nonlinearly on temperature and frequency (45–47). The dielectric loss tangent for PVDF films is reportedly in the

**Figure 12.** Piezoelectric coefficient d_{31} of PVDF film (12).

range of 0.015 to 0.25, depending on the experimental conditions (25,42,44).

Time-dependent mechanical properties of PVDF have been studied in (18–20) based on a series of quasi-static creep tests and dynamic mechanical tests of 28- μm thick, commercially produced PVDF thin films. In the experiments, PVDF samples were tested in two in-plane material directions, parallel (direction 1) and perpendicular (direction 2) to the polymer's aligned molecular chains. Creep experiments were performed at 10 different stress levels under sustained loading conditions at room temperature. Strain measurements were taken by a linear variable differential transformer (LVDT).

The time-dependent response of PVDF thin films has been described by the constitutive equations of linear viscoelasticity in either of two alternative forms (48,49):

$$\varepsilon(t) = \sigma(0)C_i(t) + \int_0^t C_i(t-\tau) \frac{d\sigma}{d\tau} d\tau, \quad (6)$$

$$\sigma(t) = \varepsilon(0)Y_i(t) + \int_0^t Y_i(t-\tau) \frac{d\varepsilon}{d\tau} d\tau, \quad i = 1, 2, \quad (7)$$

where the functions $C_i(t)$ and $Y_i(t)$ ($i = 1, 2$) denote, respectively, the creep compliances and relaxation moduli of the polymer in both material directions.

The creep compliances of PVDF were determined by using the method of dynamic mechanical testing and analysis (DMTA), commonly employed for testing and characterizing polymers and polymer matrix composites (50).

The dynamic experimental program described in (18–20) was implemented by subjecting PVDF samples to sustained tensile stresses at several levels below σ_{Y_i} ($i = 1, 2$) that had a superimposed sinusoidal strain $\varepsilon = \varepsilon^\circ \sin(\omega t)$. The response of the material was measured in terms of the respective stress $\sigma = \sigma^\circ \sin(\omega t + \delta)$, and the phase angle δ represented the loss of mechanical energy. Dynamic tests were performed in each material direction for 20 different values of frequency in the range from 1 to 50 Hz at ambient temperature. The viscoelastic properties of the polymer in both directions were characterized in terms of the respective storage and loss moduli, Y'_i and Y''_i , and storage and loss compliances, C'_i and C''_i ($i = 1, 2$).

The experimental frequency range was expanded on the basis of the temperature-frequency correspondence principle that provides a correlation between the viscoelastic material characteristics at a base temperature T_0 and the respective material characteristics at a different temperature T , such that

$$Y'(\omega, T_0) = Y'(\omega a_T, T_0), \quad (8)$$

where a_T denotes the shift factor (48).

The shift factor a_T for PVDF thin films was determined by repeating the entire set of dynamic experiments at 13 different temperature levels in the range from 25.5 to 81°C for the material direction 1 and at 10 different temperature levels in the range from 24.4 to 81.1°C for the material direction 2. On this basis, the loss and storage relaxation moduli and the loss and storage creep compliances of PVDF were determined. As an example, the storage moduli

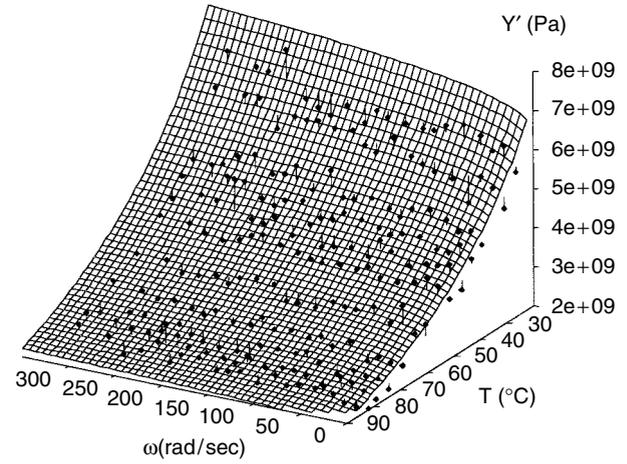


Figure 13. Storage modulus of PVDF (direction 1) (18).

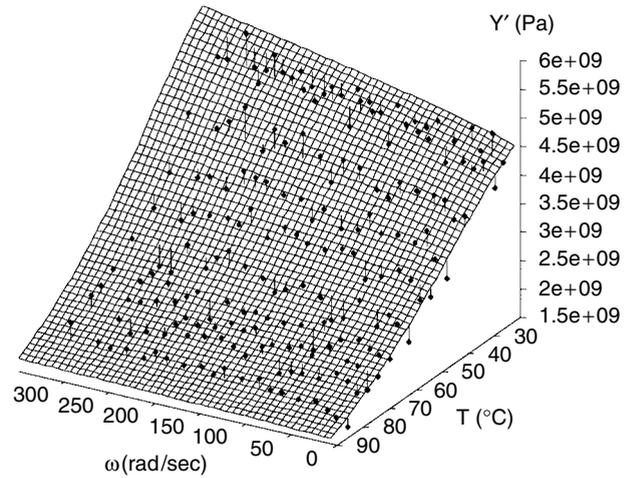


Figure 14. Storage modulus of PVDF (direction 2) (18).

of PVDF as functions of frequency and temperature are shown in Figs. 13 and 14 in directions 1 and 2, respectively.

Using the approximate inverse Fourier transformation method (48), the creep compliances $C_i(t)$ ($i = 1, 2$) of PVDF were obtained numerically. An analytical approximation of $C_i(t)$ ($i = 1, 2$) was obtained as a power law

$$C_i(t) = a_i + b_i t^{\alpha_i}, \quad i = 1, 2, \quad (9)$$

where the coefficients a_i , b_i , and α_i were determined for the material direction 1 as

$$a_1 = 3.206 \times 10^{-10}, b_1 = 5.018 \times 10^{-11}, \alpha_1 = 0.107, \quad (10)$$

within the range ($0.0032 < t < 1.91 \times 10^6$) s. and, for the material direction 2, as

$$a_2 = 3.514 \times 10^{-10}, b_2 = 0.111 \times 10^{-11}, \alpha_2 = 0.085, \quad (11)$$

within the range ($0.0032 < t < 4.001 \times 10^4$) s. The respective relaxation functions of PVDF were obtained (51), as illustrated in Figs. 15 and 16.

It is important to note that, the time-dependent mechanical response of PVDF films is independent of stress only

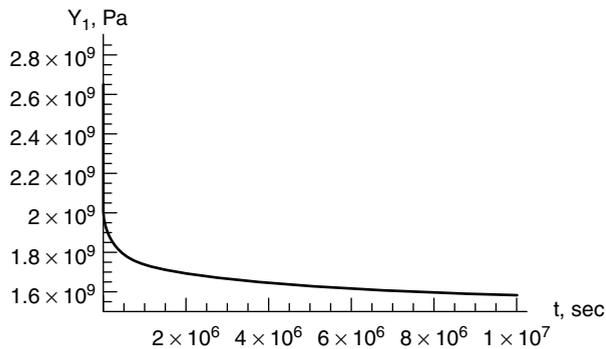


Figure 15. Relaxation function of PVDF (direction 1) (51).

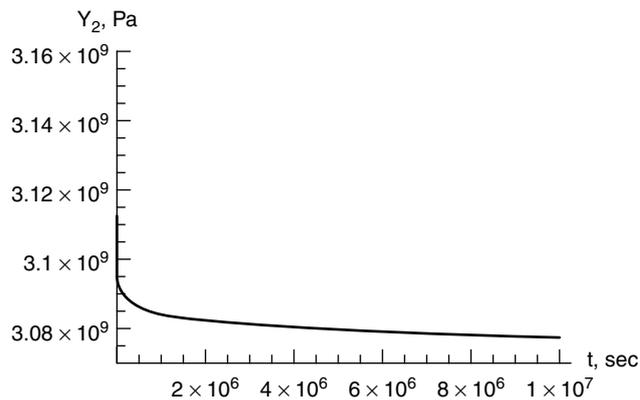


Figure 16. Relaxation function of PVDF (direction 2) (51).

at stress levels below 57% of the yield stress $(\sigma_Y)_1$ in the material direction 1 and below 76% of $(\sigma_Y)_2$ in the material direction 2 (18). Beyond these limits, the creep and relaxation behavior of PVDF is nonlinear.

It is important to note that, as determined by Holloway (18), the time-dependent mechanical response of PVDF films is stress independent only at the stress levels below 57% of the yield stress $(\sigma_Y)_1$ in the material direction 1 and below 76% of $(\sigma_Y)_2$ in the material direction 2. Beyond these limits, the creep and relaxation behavior of PVDF is nonlinear.

It has been shown (52) that the mechanical and dielectric relaxation responses of piezoelectric solids are interrelated. The same molecular relaxation mechanisms that give rise to mechanical relaxation also give rise to dielectric relaxation. It has been suggested (53) that the inherent coupling between the time-dependent electrical and mechanical properties of piezoelectric materials can be described by the constitutive equations in a form similar to the constitutive law of linear viscoelasticity:

$$\varepsilon_{ij} = \int_0^t C_{ijkl}(t-\tau) \frac{d\sigma_{kl}}{d\tau} d\tau + \int_0^t d_{kij}(t-\tau) \frac{dE_k}{d\tau} d\tau, \quad (12)$$

$$D_i(t) = \int_0^t d_{ijk}(t-\tau) \frac{d\sigma_{jk}}{d\tau} d\tau + \int_0^t e_{ij}(t-\tau) \frac{dE_j}{d\tau} d\tau. \quad (13)$$

Combined nonlinear and time-dependent effects in the behavior of piezoelectric polymers were studied in (54), based on measurements of the stress relaxation modulus of uniaxially stretched 30- μm thick PVDF films in the temperature range from 30 to 100°C. It was observed that, for temperatures below 40°C, the behavior of PVDF samples in the draw direction followed the typical relaxation pattern. However, at elevated temperatures ranging from 50 to 100°C, PVDF samples demonstrated an inverse behavior, an increase of the relaxation modulus after a certain period of continuous decrease. For example, at 100°C, the relaxation modulus of PVDF decreased for about 1 min, reached a minimum, and, subsequently, increased from about 1.14 GPa to 2.97 GPa during a period of 36 hours. This behavior of PVDF, identified as “inverse stress relaxation,” was accompanied by a gradual decay of the piezoelectric properties of the material.

To date, significant progress has been made in the development of various constitutive models of nonlinear piezoelectricity (7,55–57). However, many aspects of experimentally observed nonlinear and time-dependent phenomena that characterize the behavior of piezoelectric polymers still remain unexplained.

APPLICATIONS

Piezoelectric polymers have been increasingly integrated in structural design as active elements that can sense and respond intelligently to external stimuli. On this basis, a new generation of so-called “smart structures” or “smart material systems” has emerged that can detect changes in loading or environmental conditions, decide rationally on a set of respective actions, and carry out these actions in a controlled manner. A broad range of applications using such functions include active vibration damping, acoustic suppression, damage detection, shape and position control of compliant structures, and self-inspection of structural integrity. Systematic reviews (3,25,58,59) provide a consistent account of modern technological developments in the field of smart material systems.

Material selection in designing smart material systems involves considerations of such factors as the maximum achievable strain, stiffness, spatial resolution, frequency bandwidth, and temperature sensitivity. Traditionally, piezoelectric ceramics such as lead zirconate titanate (PZT) and barium titanate (BaTiO_3) have played a leading role in many applications due to their dielectric strength and stable electromechanical properties at high temperatures up to 400°C. However, the potentials of piezoelectric ceramics are limited because these materials are brittle, somewhat heavy, and are difficult to scale to larger applications. In this regard, piezoelectric polymers offer definite advantages because they are light, flexible, easy to shape, and can be bonded to almost any surface. The attractive properties of PVDF and PVDF copolymers include stable response characteristics in a wide frequency range up to 10^9 Hz, low acoustic impedance; a high degree of resistance to impact, and resistance to moisture absorption, intense ultraviolet, and nuclear radiation. However, the effectiveness of piezoelectric polymers tends to decrease in

low-frequency applications, and their use is limited to a temperature range not exceeding 100°C.

The key factor that defines the application range of piezoelectric polymers is their use in the design of transducers, sensors, and actuators. An impressive array of PVDF applications in transducers, including loudspeakers, optical scanners, light deflectors, and variable aperture diaphragms, modulators for fiber optics, pyroelectric detectors, and capacitors has been delineated in (60). PVDF based transducers are particularly effective in high-frequency applications such as acoustics, ultrasonics, and nondestructive material evaluation (61). Examples of such applications include medical diagnostics (62), marine fouling prevention (63), acoustic microscopy (64), and damage detection in fibrous composite materials (65).

Piezoelectric sensors have been typically used for strain measurements through the readings of voltage, rates of voltage change, or the frequency spectra of the signal generated by the sensor. Various commercial applications of piezoelectric film sensors include contact switches, musical instruments, and vibrational sensing devices (66). Because they react to temperature changes, PVDF-based motion sensors are widely used in energy management systems to control room lights, appliance displays, and HVAC equipment (67). PVDF thin films have recently been employed in more advanced sensing technologies, for example, tactile sensors that recognize objects with a high degree of precision and also temperature and pressure sensors that can replicate the functions of human skin (68).

At present, a number of PVDF-based discrete piezoelectric actuators have been developed for different modes of operation. Traditionally, bending mode actuators have been designed in the form of bimorph beams consisting of two piezoelectric layers of opposite polarity bonded together. When an electric field is applied, one of the layers expands while another contracts, producing bending deformation. As shown by Wang and Cross (69), piezoelectric bimorph actuators in the form of cantilever beams can generate considerable tip displacements, although they tend to produce low forces. Thicker films and multilayer designs can expand the range of forces produced but reduce the respective displacements. A three-layer piezoelectric actuator with hysteresis has been analyzed and tested in (70).

The effectiveness of bimorph piezoelectric actuators has been enhanced by shaping them into a curvilinear configuration. As an example, a semicircular bimorph C-block actuator was proposed in (71) and analyzed in (72). Similarly to the beam bending actuator, the C-block design uses the response of two bonded piezoelectric layers actuated by equal and opposite electric fields. As shown in (72), the C-block actuator produces an increased stroke or force output, especially when combined in series or in parallel, to form larger actuator architectures.

Concurrently, a double curvature piezoelectric actuator for vibration control in microgravity environments has been proposed in (73,74). This low force PVDF actuator uses the bimorph design concept and can be produced in a variety of sizes, depending on the performance requirements. The efficiency of the actuator can be enhanced by introducing multilayer configurations and by creating more complex architectures, as in the C-block design.

The bimorph configuration involving PVDF thin films has been effectively incorporated (75,76) into the design of flexible mirror systems to control their shape and, consequently, provide the required precision of the optical surface. A similar concept was implemented (77) in designing large-scale deployable membrane mirrors for space exploration telescopes.

In general, integration of the electromechanical properties of piezoelectric polymers into structural design provides the capability of controlling the mechanical characteristics of structures in terms of stiffness or damping, or modifying the structural response in terms of position or velocity. This type of built-in structural intelligence has been particularly effective in applications involving vibration control and damping enhancement of flexible structural elements.

Piezoelectricity has played a major role in the development of various "passive" and "active" vibration control strategies. Passive damping involves converting mechanical strain energy into electrical energy, which is subsequently dissipated by a simple resistive element. Active vibrational control is achieved through interactive functions of three main components: a sensor that identifies the present state of the structure, a cognitive interpretation and decision system that controls and optimizes, and an actuator that modifies the response of the structure. Active control of sound radiation is based on a conceptually analogous approach.

Typically, piezoelectric sensors and actuators used in vibration control applications are either bonded to the surface or embedded within the structure as a patch, a continuous single layer, or multiple material layers. Spatial arrangements of segmented piezoelectric elements can be optimized to achieve the desired effects.

The response of intelligent structures that have integrated piezoelectric sensors and actuators has been studied extensively. In particular, the effectiveness of passive piezoelectric damping has been examined in (78). Various problems of active and passive vibration control of smart beams using bonded and embedded piezoelectric sensors and actuators have been investigated in (79–85). The linear vibration theory of piezoelectric plates has been developed in (86). An exact solution has been derived for piezothermoelastic orthotropic flat panels subjected to external pressure, and thermal and electrostatic excitations (87). The vibration response of simply supported elastic rectangular plates excited by two-dimensional patch actuators bonded to the plate surface has been investigated in (88). Active vibration control of plates using patches composed of a viscoelastic damping layer sandwiched between two piezoelectric layers has been studied in (89). A similar approach was adopted in (90) to control the vibrations of cylindrical shells actively. The potentials of vibrational control of cylindrical shells using curved piezoelectric actuators have been studied in (91). The problem of active noise control of an elastic panel harmonically excited by multiple piezoelectric actuators has been analyzed in (92).

As an integral part of these efforts, a number of studies specifically focused on using piezoelectric polymers for active vibration and noise control of structural components. Thus, the problem of vibration control of beams

using distributed PVDF sensors has been examined in (93,94). The performance of shaped PVDF modal sensors employed to control specific vibrational modes of rectangular plates under steady-state harmonic excitations has been studied in (95). The results obtained demonstrated the high effectiveness of PVDF sensors in active vibration control applications in both a resonant and off-resonant range of frequencies. A similar conclusion was reached (96) based on a study of shaped PVDF sensors for active structural acoustic control of rectangular plates. One theoretical and experimental investigation (97) concerns the performance of distributed PVDF sensors and actuators that can distinguish between bending and torsional vibration modes of rectangular plates. The results of the study indicate that PVDF thin films can be effectively used in microactuator devices as well as in modal control applications of larger continuous structures. Studies (98) demonstrated the effectiveness of distributed piezopolymer actuators for active control of sound fields radiated from composite structures in acoustic control applications. Similarly, the high effectiveness of PVDF-based distributed sensors and actuators used for active vibration control of flexible manipulators was observed (99).

Piezoelectric polymers represent a group of primary candidates for shape and positional control of flexible structures in weight-sensitive applications. Examples of such applications are smart skin for airborne structures (100) and highly compliant smart material systems for space applications (101). The latter comprise a diverse range of ultra-lightweight structures such as solar sails, deployable membrane mirrors, atmospheric balloons, antennae, and reflectors. By incorporating the capabilities of piezoelectric polymers in structural design, it is possible to enhance the performance of such structures by actively controlling their shape and stability.

At present, a large class of smart composite structures has been developed that combines the traditional advantages of laminated composites and the adaptive capabilities of piezoelectric materials. These structures offer numerous technological benefits. However, they exhibit complex electromechanical behavior that depends on a combination of many factors, such as the individual properties of active and passive constituents, specific material lay-ups, interfacial conditions, and the effects of damage evolution processes. These and other related issues have attracted considerable research interest.

Various theories and analyses of laminated composite structures that have integrated self-sensing, control, and diagnostic functions have been developed. In particular, the coupled mechanical, electrical, and thermal response of piezoelectric composite beams has been analyzed in (102). The development of the general theory of piezoelectric composite plates involves the efforts of many (103–109). The response of curvilinear piezoelectric composite structures has been analyzed in (110–113).

A general finite element formulation for analyzing distributed thermopiezoelectric sensors and actuators as elements of intelligent structures has been proposed in (114). A nonlinear three-dimensional constitutive theory of anisotropic piezoelectro-thermoviscoelasticity for

nonhomogeneous layered media has been developed in (115).

An innovative technique for damage diagnostics of laminated composites by an integrated sensor-actuator system in the form of a thin flexible Stanford Multi-Actuator-Receiver Transduction (SMART) layer embedded within laminated structures has been described in (116). Applications of the method involve identifying the location and force of the unknown external impact, estimating the extent of the impact damage, and monitoring the cure conditions of composites.

Advanced technological developments in intelligent material systems consistently stimulate the search for smart materials that have novel or improved characteristics. Currently, attention is focused on the effectiveness of composite materials that combine the superior piezoelectric properties of ceramics and the compliance and flexibility of various polymers. Piezoelectric composites of this type are typically produced by integrating ceramic fibers or particles of lead zirconate titanate (PZT) or calcium modified lead titanate (PbTiO_3) within a polymer matrix. Fibrous composites have been referred to as 1–3 composites because the fibers have unidirectional orientation, whereas the particulate composites, known as 0–3 composites, are isotropic. Fibrous piezoelectric composites usually have better piezoelectric properties; however, their fabrication processes are complex. Particulate composites can be produced as thin films and have the advantage of being less expensive. The performance and properties of 1–3 and 0–3 piezoelectric ceramic/polymer composites have been investigated in (117–120). In general, potential applications of active polymeric composite materials (APCM) have been discussed in (121).

CONCLUDING REMARKS

At present, the field of intelligent material systems is expanding at an unprecedented rate. The guiding principles behind this progress are structural efficiency, functionality, precision, and durability. Superior adaptive capabilities and other attractive qualities of piezoelectric polymers determine their increasingly leading role in the design of intelligent structures whose applications range from aerospace, construction and transportation to physics and life sciences.

It is clear that effective implementation of piezoelectric polymer systems directly depends on the degree to which their behavior and properties are understood. To date, despite considerable progress, material characterization of piezoelectric polymer films is far from complete. Challenges arise due to the sensitivity of the polymers to variations in fabrication and temperature conditions, time-dependent effects, and material nonlinearities. The matter is complicated by the technological necessity of providing electrical conductivity by depositing metallic surface layers. Effectively, piezoelectric polymers represent a composite material, whose response depends strongly on the thickness and properties of the individual constituents. Due to these factors, stable experimental conditions and highly precise measurements are required to characterize

the properties of piezoelectric polymers accurately in the practical range of their operating conditions.

In the immediate future, continuing progress in the field of smart materials will depend on the intensity of research efforts directed toward the development of piezoelectric polymer systems that have enhanced adaptive capabilities, formulation of advanced theoretical models, and implementation of innovative testing methodologies. On this basis, the unprecedented opportunities offered by the new generation of intelligent materials will continue to stimulate further technological progress and, ultimately, contribute to the betterment of humanity.

BIBLIOGRAPHY

1. W.G. Cady, *Piezoelectricity. An Introduction to the Theory and Applications of Electromechanical Phenomena in Crystals*. Dover, NY, 1964, Vols. 1 and 2.
2. P.E. Dunn and S.H. Carr, *A (MRS) Bull.* **2**: 22–31 (1989).
3. S.S. Rao and M. Sunar, *Appl. Mech. Rev.*, **47**(4): 113–123 (1994).
4. T. Ikeda, *Fundamentals of Piezoelectricity*. Oxford Sci. Oxford, 1990.
5. J. Zelenka, *Piezoelectric Resonators And Their Applications*. Elsevier, NY, 1986.
6. H.F. Tiersten, *J. Appl. Phys.* **74**(5): 3389–3393 (1993).
7. N.G. McGrum, B.E. Reid, and G. Williams, *Anelastic and Dielectric Effects in Polymeric Solids*. Dover, NY, 1991.
8. G.M. Sessler, in *Polymeric Electrets, Electrical Properties of Polymers*, D.A. Seanor, ed., Academic Press, NY, 1982, pp. 241–284.
9. H. Kawai, *Jpn. J. Appl. Phys.* **8**: 975–976 (1969).
10. M.G. Broadhurst and G.T. Davis, *Ferroelectrics* **60**: 3–13 (1984).
11. R.G. Kepler and R.A. Anderson, *Ferroelectric Polymers, Advances in Physics*, **41**(10). Taylor and Francis, Washington, 1992, pp. 1–57.
12. B. Hilezer and J. Malecki, *Electrets*. Elsevier, Amsterdam, 1986.
13. B.R. Hahn, *J. Appl. Phys.* **57**(4): 1294–1298 (1985).
14. H. Ohigashi, *J. Appl. Phys.* **47**(3): 949–955 (1976).
15. P. Destruel, F.S. Rojas, D. Tougne, and Hoang-The-Giam, *J. Appl. Phys.* **56**(11): 3298–3303 (1984).
16. E.L. Nix and I.M. Ward, *Ferroelectrics* **67**: 137–141 (1986).
17. R.G. Kepler and R.A. Anderson, *J. Appl. Phys.* **49**(8): 4490–4494 (1978).
18. F.C. Holloway, M.S. Dissertation, Montana State University-Bozeman, 1997.
19. A.M. Vinogradov, and F. Holloway, *J. Adv. Mater.* **29**(1): 11–17 (1997).
20. A.M. Vinogradov and F. Holloway, *Ferroelectrics* **226**: 169–181 (1999).
21. S. Tasaka and S. Miyata, *Ferroelectrics* **32**(1): 17–23 (1981).
22. T. Aya and T. Nakayama, *JSME Int. J.* **40**(3): 343–348 (1997).
23. T.T. Wang, J.M. Herbert, and A.M. Glass, *The Applications of Ferroelectric Polymers*. Chapman and Hall, NY, 1988.
24. K. Tashiro, H. Tadokoro, and M. Kobayashi, *Ferroelectrics* **32**: 167–175 (1981).
25. M.V. Gandhi and B.S. Thompson, *Smart Materials And Structures*. Chapman & Hall, London, 1992.
26. B. Jaffe and W. R. Cook, *Piezoelectric Ceramics*. Academic Press, London, 1971.
27. D.A. Berlincourt, D.R. Curran, and B. Jaffe, in *Piezoelectric and Piezomagnetic Materials and Their Function in Transducers, Physical Acoustics*, Vol. 1A, E.P. Mason, ed., Academic Press, NY, 1964, pp. 169–270.
28. W.P. Leung and K.K. Yung, *J. Appl. Phys.* **50**(12): 8031–8033 (1979).
29. B.R. Hahn, *J. Appl. Phys.* **57**(4): 1294–1298 (1985).
30. T. Furukawa, M. Date, and E. Fukada, *J. Appl. Phys.* **51**(2): 1135–1141 (1980).
31. J.C. Hicks and T.E. Jones, *J. Appl. Phys.* **32**(1/4): 119–126 (1981).
32. T. Takahashi, M. Date, and E. Fukada, *J. Appl. Phys.* **32**(1/4): 73–77 (1981).
33. D.M. Gookin, E.W. Jacobs, and J.C. Hicks, *Ferroelectrics* **57**: 89–98 (1984).
34. T. Furukawa and N. Seo, *Jpn. J. Appl. Phys.* **57**(4): 675–680 (1990).
35. T. Furukawa, G.E. Johnson, H.E. Bair, Y. Tajitsu, A. Chiba, and E. Fukada, *Ferroelectrics* **32**: 61–67 (1981).
36. S. Tasaka and S. Miyata, *J. Appl. Phys.* **57**(3): 906–910, 1985.
37. G.T. Davis, M.G. Broadhurst, A.J. Lovinger, and T. Furukawa, *Ferroelectrics* **57**: 73–84 (1984).
38. M. Oshiki and E. Fukada, *J. Mater. Sci.* **10**: 1–6 (1975).
39. M. Tamura, K. Ogasawara, N. Ono, and S. Hagiwara, *J. Appl. Phys.* **45**(9): 3768–3771 (1974).
40. J.I. Scheinbeim, C.H. Yoon, K.D. Pae, and B.A. Newman, *J. Appl. Phys.* **51**(10): 5156–5160 (1980).
41. B. Rezvani and J.G. Linvill, *Appl. Phys. Lett.* **34**(12): 828–830 (1979).
42. T. Wentink, Jr., *J. Appl. Phys.* **32**(6): 1063–1064 (1961).
43. H. Wang, Q.M. Zhang, and L.E. Cross, *J. Appl. Phys.* **74**(5): 3394–3398 (1993).
44. K. Koga and H. Ohigashi, *J. Appl. Phys.* **59**(6): 2142–2151 (1986).
45. S. Osaki and T. Kotaka, *Ferroelectrics* **32**: 1–11 (1981).
46. H. Arisawa, O. Yano, and Y. Wada, *Ferroelectrics* **32**: 39–41 (1981).
47. N. Koizumi, J. Hagino, and Y. Murata, *J. Appl. Phys.* **32**: 141–147 (1981).
48. J.D. Ferry, *Viscoelastic Properties of Polymers*. 3ed, Wiley, NY, 1980.
49. R.M. Christensen, *Theory of Viscoelasticity. An Introduction*. Academic Press, NY, 1971.
50. K.P. Menard, *Dynamic Mechanical Analysis. A Practical Introduction*. CRC, NY, 1999.
51. A.M. Vinogradov, Proc. SPIE 3667, 1999, pp. 711–718.
52. R. Lakes, *IEEE Trans. Sonics Ultrasonics* **SU-27**(4): 208–213 (1980).
53. R. Lakes, *Viscoelastic Solids*. CRC, Boca Raton, 1998.
54. T.T. Wang, *J. Appl. Phys.* **53**(3): 1828–1829, 1982.
55. R.A. Toupin, *J. Rational Mech. Anal.* **5**(6): 849–915, 1956.
56. H.F. Tiersten, *J. Acoust. Soc. Am.* **70**(6): 1567–1576 (1981).
57. G.A. Maugin, in *Nonlinear Electromechanical Effects and Applications*, Series in Theoretical and Applied Mechanics, R.K.T. Hsieh, ed., World Scientific, Singapore, 1985.

58. E.F. Crawley, *AIAA J.* **32**(8): 1689–1699 (1994).
59. M. Sunar and S.S. Rao, *Appl. Mech. Rev.* **52**(1): 1–16 (1999).
60. M.A. Marcus, in *Electroresponsive Molecular and Polymeric Systems*, Vol. 1, T.A. Skotheim, ed., Marcel Dekker, NY, 1988.
61. Y. Bar-Cohen, T. Xue, and S.-S. Lih, *NDTnet* **1**(9): (1996). <http://www.ndt.net/article/yousi/yosi.htm>
62. K. Kobayashi and T. Yasuda, *Ferroelectrics* **32**: 181–184 (1981).
63. M. Latour and P.V. Murphy, *Ferroelectrics* **32**: 33–37 (1981).
64. S. Smolorz and W. Grill, *Res. Nondestructive Evaluation* **7**(4): 195–201 (1995).
65. R. Stiffler and E.G. Henneke, II, *Mater. Evaluation* **41**: 956–960 (1982).
66. Piezo Film Sensors Technical Manual, AMP, 1993.
67. D. Halvorsen, *Mach. Design* 47–51 (July 1994).
68. C.A. Rogers, D.K. Barker, and C.A. Jaeger, *Proc. U.S. Army Res. Office Workshop Smart Mater. Struct. Math. Issues*, VPI, Blacksburg, VA, Sept. 15–16, 1988, pp. 17–28.
69. Q.-M. Wang and L.E. Cross, *Ferroelectrics* **215**: 187–213 (1998).
70. T.S. Low and W. Guo, *J. Micromechanical Systems* **4**(4): 230–237 (1995).
71. D. Brei, *Proc. Symp. Micro-Mechanical Syst. ASME Winter Annu. Meet.*, Chicago, IL, 1994, Vol. 2, pp. 717–723.
72. C.E. Seeley, A. Chattopadhyay, and D. Brei, *AIAA J.* **34**(1): 123–128 (1996).
73. V.H. Schmidt, D. Brandt, F. Holloway, A. Vinogradov, and D. Rosenberg, *Proc. 10th IEEE Int. Symp. Appl. Ferroelectrics 1*, East Brunswick, NJ, August 18–21, 1996.
74. G. Bohannan, H. Schmidt, D. Brandt, and M. Mooibroek, *Ferroelectrics* **224**: 211–217 (1999).
75. T. Sato, H. Ishida, and O. Ikeda, *Appl. Opt.* **19**(9): 1430–1434 (1980).
76. T. Sato, H. Ishikawa, and O. Ikeda, *Appl. Opt.* **21**(20): 3664–3668 (1982).
77. J.W. Martin, J.A. Main, and G.C. Nelson, *Proc. 1998 ASME Int. Mech. Eng. Cong.*, Nov. 15–20, 1998, Anaheim, CA, pp. 217–223.
78. A.R. Ramachandran, Q.C. Xu, L.E. Cross, and R.E. Newnham, *Proc. First Joint U.S./Jpn. Conf. Adaptive Struct.*, Nov. 13–15, 1990, Maui, Hawaii, pp. 525–538.
79. E.F. Crawley and J. de Luis, *AIAA J.* **25**(10): 1373–1385 (1987).
80. M.-H.H. Shen, *Smart Mater. Struct.* **3**: 439–447 (1994).
81. W.B. Wu and M.J. Tzeng, *Proc. First Joint U.S./Jpn. Conf. Adaptive Struct.* Maui, Hawaii, Nov. 13–15, 1990, pp. 882–914.
82. R. Chandra and I. Chopra, *AIAA J.* **31**(9): 1692–1701 (1993).
83. G.L. Ghiringhelli, P. Masarati, and P. Mantegazza, *J. Intelligent Mater. Syst. Struct.* **8**(10): 842–858 (1997).
84. A.M. Baz and J.-J. Ro, *Shock and Vibration* **2**(1): 33–42 (1995).
85. A. Baz, *J. Vibration Acoust. Trans. ASME* **119**(4): 166–172 (1997).
86. H.F. Tiersten, *Linear Piezoelectric Plate Vibrations*. Plenum Press, NY, 1969.
87. G.P. Dube, S. Kapuria and P.C. Dumir, *Int. J. Mech. Sci.* **38**(11): 1161–1177 (1996).
88. E.K. Dimitriadis, C.R. Fuller, and C.A. Rogers, *J. Vibrations Acoust. Trans. ASME* **113**: 100–107, 1991.
89. A. Baz and J. Ro, *Smart Mater. Struct.* **5**: 272–280 (1996).
90. A. Baz and T. Chen, *35th AIAA Aerosp. Sci. Meet. Exhibit*, Reno, NV, Jan. 6–9, 1997, pp. 168–180.
91. V.R. Sonti and J.D. Jones, *AIAA J.* **34**(5): 1034–1040, 1996.
92. B.-T. Wang, C.R. Fuller, and E.K. Dimitriadis, *AIAA J.* **29**(11): 1802–1809 (1991).
93. S. Miller and J. Hubbard, *Proc. 6th VPI&SU/AIAA Symp. Dynamics Control Large Struct.*, Blacksburg, VA, 1987, pp. 375–390.
94. T. Bailey and J.E. Hubbard, *J. Guidance Control Dynamics* **8**: 605–611 (1985).
95. Y. Gu, R.L. Clark, C.R. Fuller, and A.C. Zander, *J. Vibration Acoust.* **116**: 303–308 (1994).
96. C. Guigou, A. Berry, and F. Charette, *ASME Proc.* 1994, AD-Vol. 45/MD-Vol. 54, pp. 247–255.
97. C.-K. Lee and F.C. Moon, *J. Acoust. Soc. Am.* **85**(6): 2432–2439, (1989).
98. S.J. Kim and K.W. Yoon, *J. Sound Vibration* **202**(4): 461–476 (1997).
99. H.S. Tzou, *J. Robotic Syst.* **6**(6): 745–767 (1989).
100. Vi. Varadan and VA. Varadan, *ECCM Smart Composites Workshop*, Sept. 21–22, 1993, Bordeaux, France, 1993, pp. 17–22.
101. C.H. Jenkins, V.D. Kalanovic, S.M. Faisal, K. Padmanabhan, and M. Tampi, *1998 ASME Int. Mech. Eng. Congr.*, AD-Vol. 57/MD-Vol. 83, pp.177–188.
102. H.-J. Lee and D.A. Saravanos, *AIAA J.* **34**(6): 1231–1237 (1996).
103. D.A. Saravanos, P.R. Heyliger, and D.A. Hopkins, *Int. J. Solids Struct.* **34**(1): 359–378 (1997).
104. Y.Y. Tang and K. Xu, *J. Thermal Stresses* **18**: 87–104 (1995).
105. K. Xu, A.K. Noor, and Y.Y. Tang, *Comput. Methods Appl. Mech. Eng.* **126**: 355–371 (1995).
106. W. Sun, *SPIE* **1917**: 497–507 (1993).
107. H.S. Tzou and R. Ye, *J. Vibration Acoust.* **116**(4): 489–495 (1994).
108. W.S. Anders and C.A. Rogers, *Proc. 1st Joint U.S./Jpn. Conf. Adaptive Struct.*, Nov. 13–15, 1990, Maui, Hawaii, pp. 285–303.
109. W. Shields, J. Ro, and A. Baz, *Smart Mater. Struct.* **7**: 1–11 (1998).
110. H.S. Tsou and M. Garde, *J. Sound Vibration* **132**(3): 433–450 (1989).
111. H.S. Tzou and Y. Bao, *Proc. 1994 Int. Mech. Eng. Cong.*, Chicago IL, Nov. 6–11, 1994, ASME DE-Vol. 75, pp. 149–157.
112. P.H. Larson and J.R. Winson, *Proc. 8th Tech. Conf. Am. Soc. Composites*, Oct. 19–21, 1993, Cleveland OH, pp. 141–150.
113. P.H. Larson, Ph.D. Thesis, University of Delaware, 1994.
114. S.S. Rao and M. Sunar, *AIAA J.* **31**(7): 1280–1286 (1993).
115. H.H. Hilton, J.R. Vinson, and S. Yi, *Proc. 11th Int. Conf. Composite Mater.* July 14–18, 1997, Gold Coast, Australia, Vol. VI, pp. 444–453.
116. M. Lin and F.-K. Chang, *Mater. Today* **2**(2): 18–22 (1999).
117. H.L.W. Chan and J. Unsworth, *J. Appl. Phys.* **65**(4): 1754–1758 (1989).
118. J.A. Chilton, G.M. Garner, R.W. Whatmore, and F.W. Ainger, *Ferroelectrics* **109**: 217–222 (1990).
119. S.P. Marra, K.T. Ramesh, and A.S. Doudlas, *Composite Sci. Technol.* **59**: 2163–2173 (1999).
120. S.P. Marra, K.T. Ramesh, and A.S. Doudlas, *Smart Mater. Struct.* **8**: 57–63 (1999).
121. L.S. Pinchuk and A.V. Makarevich, *Mech. Composite Mater.* **31**(3): 211–215 (1995).

POLY(*p*-PHENYLENEVINYLENE)

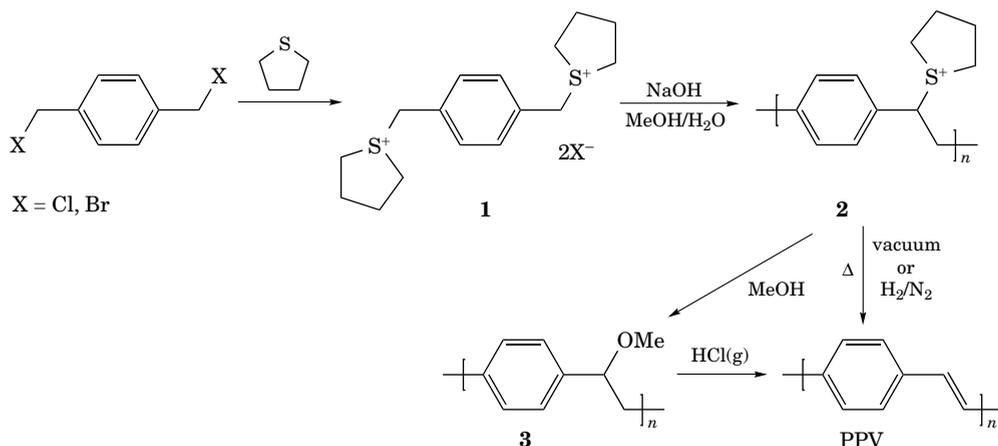
MICHAEL O. WOLF
The University of British Columbia
Vancouver, British Columbia
Canada

INTRODUCTION

Poly(*p*-phenylenevinylene) (PPV) and its derivatives are polymers that have been widely studied due to their potential applications in optoelectronic devices. PPV is a conjugated polymer whose backbone consists of alternating single and double bonds. Many conjugated polymers are known and exhibit remarkably high electrical conductivities when oxidatively or reductively doped (1). Much of the early research on PPV focused on the relatively disappointing properties of the doped material; however, interest in this material was reawakened in 1990 when Friend and co-workers in Cambridge discovered that films of undoped PPV could be used as the emitting layer in organic

insoluble and intractable materials. Therefore, an important consideration in all preparative routes that are used to prepare high molecular weight material, is the solubility of the growing polymer chain. One of the most successful approaches to high molecular weight (10,000–100,000) PPV is the Wessling or sulfonium precursor route which proceeds via a soluble precursor polymer that is subsequently thermally converted to fully conjugated PPV. Other methods are available but generally produce low molecular weight material.

Sulfonium Precursor Route. In this route, polymerization of the bis-sulfonium salt 1 with base yields a soluble polyelectrolyte 2 (Scheme 1) (8,9). This intermediate may then be purified, processed, and finally thermally converted to PPV. Both the nature of the sulfide used in the sulfonium salt and the counterion affect the conditions required in the preparation, as well as the molecular weight and structure of the resulting polymer (9,10). A modified sulfonium precursor route has also been developed, in which the soluble methoxy-substituted polymer 3 is converted to PPV in the presence of HCl gas (11).



Scheme 1

electroluminescent (EL) devices (2). This discovery stimulated intense research in the area, including many fundamental studies of the properties of PPV and its derivatives, and extensive academic and industrial interest in the applications of the materials.

This article provides an overview of the methods by which PPV and some of its derivatives may be prepared, the physical and electronic properties of these materials, and the applications that are being explored. The reader who seeks a deeper and more detailed understanding of this fascinating material is referred to several excellent and comprehensive reviews which have been published on the synthesis, properties, and applications of PPV (3–7).

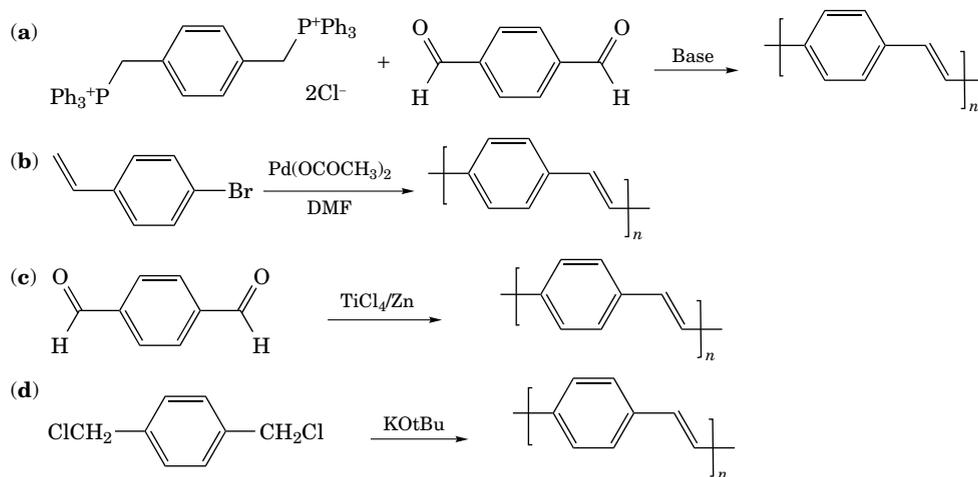
METHODS OF PREPARATION

Unsubstituted PPV

Many methods have been devised to prepare PPV for fundamental and applied studies. Due to its rigid conjugated backbone, unsubstituted PPV and even short oligomers are

The mechanism of polymerization has been the subject of some debate in the literature; both radical and anionic mechanisms are proposed. The presence of oxygen during polymerization results in lower molecular weight polymer consistent with a radical mechanism (8). A *p*-xylylene or *p*-benzoquinodimethane intermediate is postulated and has been observed spectroscopically. Other studies have suggested that this intermediate polymerizes via an anionic mechanism (12,13). Despite these conflicting results, the polymerization is typically carried out in the absence of oxygen, and the thermal conversion step is done either in vacuum or under an inert or forming gas atmosphere.

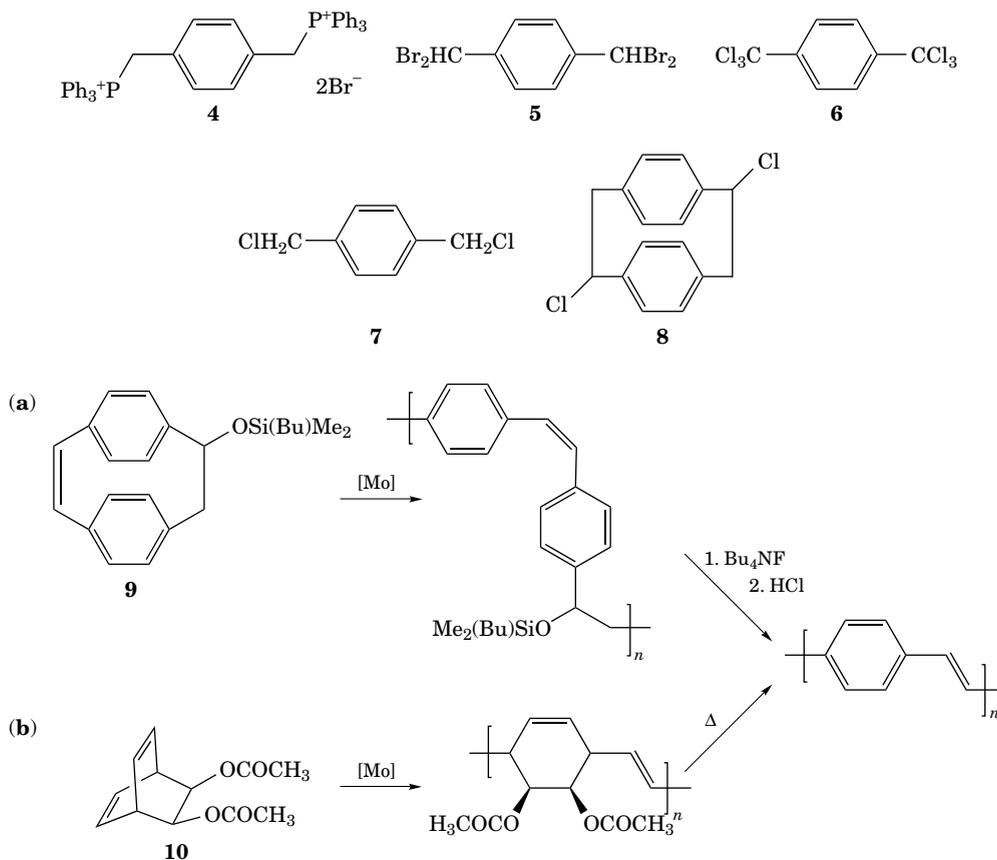
Other Routes. Other synthetic routes to PPV include the Wittig reaction (Scheme 2a) (14), the Pd-catalyzed Heck reaction (Scheme 2b) (15), and the McMurry coupling (Scheme 2c) (16). Reaction of α, α' -dichloro-*p*-xylene with potassium *tert*-butoxide yields PPV (Scheme 2d); this procedure, discovered by Gilch and Wheelwright, is referred to as the Gilch route (17). All of these methods yield PPV directly from soluble monomers and thus produce primarily low molecular weight oligomers.



Scheme 2

Electropolymerization is a convenient method for preparing insoluble conjugated polymers such as PPV because it yields thin films directly on an electrode surface. Such a film may then be directly used in an application that requires a conducting contact, such as electroluminescence. Several methods have been reported for the electrochemical preparation of PPV films. Electrochemical reduction of 4–6, it has been shown, yields PPV films on metal and indium tin oxide (ITO) electrodes (18–20). Another direct route to thin films is chemical vapor deposition (CVD) from precursors such as 7 or 8 (21,22).

Ring-opening metathesis polymerization (ROMP) has been used to prepare unsubstituted PPV via a soluble precursor. Substituted paracyclophan-1-ene 9 (23) and norbornadiene derivatives 10 (24) have been used as monomers, where $\text{Mo}(\text{NAr})(\text{CHCMe}_2\text{Ph})[\text{OCMe}(\text{CF}_3)_2]_2$ ($\text{Ar} = 2,6\text{-diisopropylphenyl}$) was used as the catalyst in the polymerization (Scheme 3). This route proceeds via a living polymerization and consequently yields a relatively narrow molecular weight distribution.

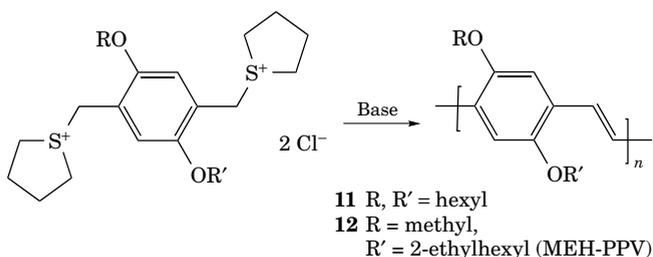


Scheme 3

Substituted PPV

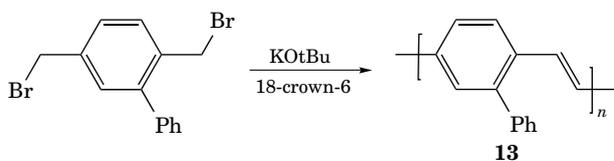
To increase the solubility and processibility of PPV in the conjugated form, substituents such as alkoxy or phenyl groups have been added to the backbone structure. In addition to enhancing the solubility, these substituents also change the electronic properties of the polymer via both inductive and conjugative effects. An additional benefit of soluble derivatives of PPV is that techniques for polymer characterization that require soluble material may be used and provide direct information about molecular weights and higher order structure.

Sulfonium Precursor Route. The Wessling route has also been used to produce soluble derivatives from monomers that contain solubilizing substituents on the phenyl ring. For example, dialkoxy substituted monomers yield **11** which is soluble in organic solvents such as chloroform and chlorobenzene (**25**), as well as poly[2-((2-ethylhexyl)oxy-5-methoxy-*p*-phenylene)vinylene] (MEH-PPV) (**12**) (Scheme 4) (**26**). The branched side chains in MEH-PPV improve the solubility of this derivative versus unbranched analogs, and this polymer is one of the most popular for electroluminescent applications.

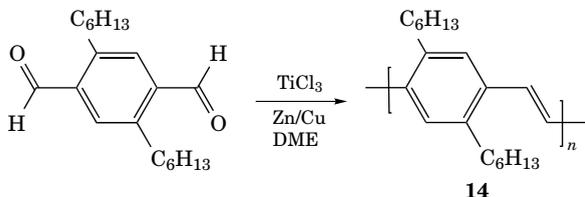


Scheme 4

Other Routes. Substituted PPV derivatives have also been prepared by several other routes. When the resulting polymers are soluble, these methods are often successful in preparing high molecular weight material. The Gilch route has been used to prepare phenyl substituted polymers (**13**) (Scheme 5) (**27**), and the McMurry coupling has been used to yield a dihexyl substituted polymer (**14**) that is soluble in a range of organic solvents (Scheme 6) (**28**).



Scheme 5



Scheme 6

Copolymers

A number of studies have focused on the preparation of copolymers that contain fully conjugated backbones, as well as those that contain both conjugated and nonconjugated blocks. This work has been motivated by the desire to prepare materials whose range of electronic properties can easily be tuned by variations in the proportion and nature of the monomers used in the copolymer synthesis. In addition, local variations in the π - π^* energy gap can be introduced in this way which, it has been shown, result in dramatic improvements in the performance of these copolymers in electroluminescent devices.

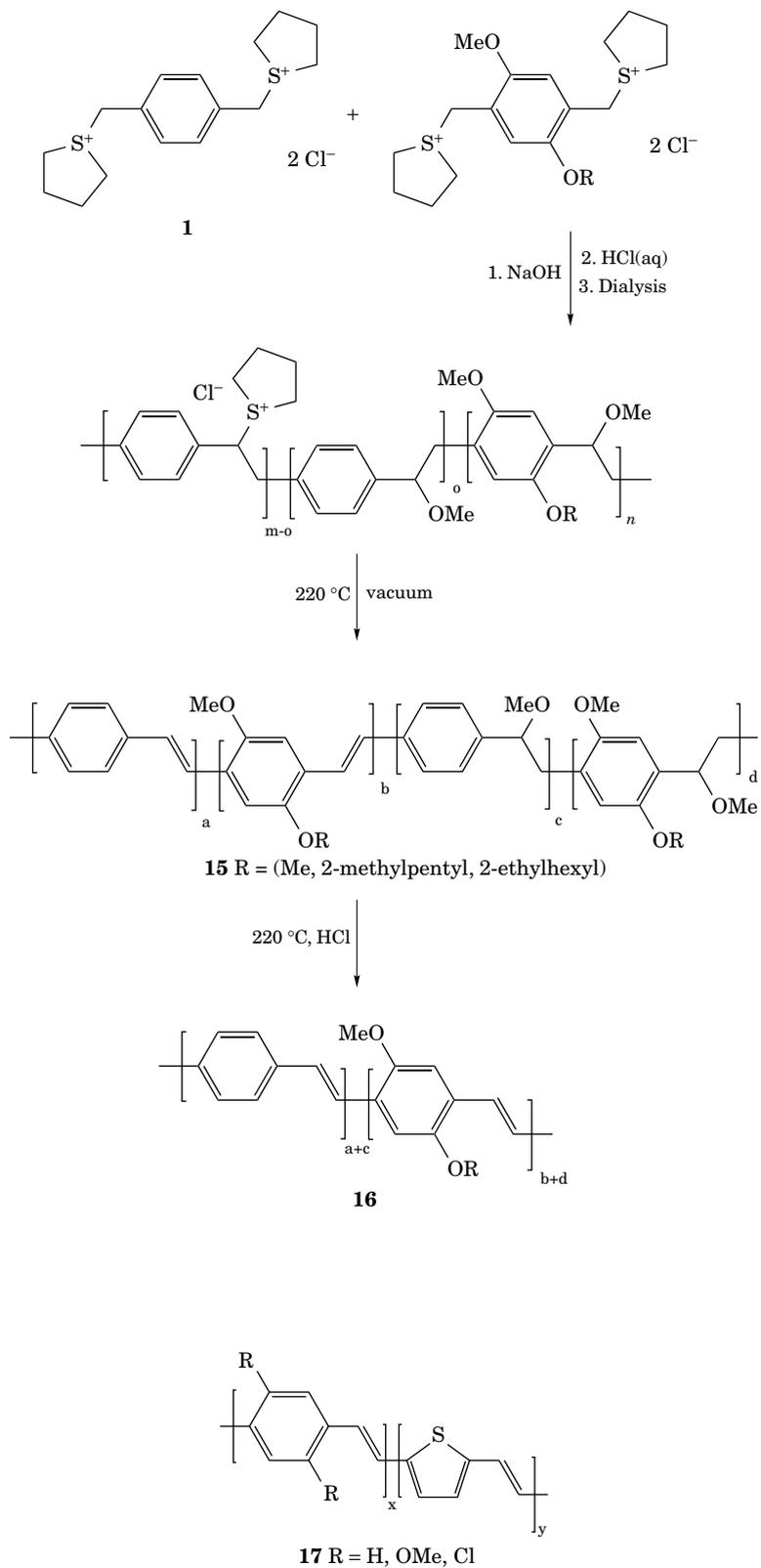
The sulfonium precursor route has been used to prepare copolymers by using various proportions of different monomers in the synthesis. This method has yielded both partially conjugated (**15**) and fully conjugated polymers (**16**) (Scheme 7) (**29,30**), as well as copolymers that contain other aromatic groups in the backbone in addition to phenylenevinylene moieties such as in copolymer **17** (**31**).

The Wittig reaction has also been successfully used to prepare soluble copolymers using substituents such as alkoxy groups on the backbone (**18**) (**32**), as well as CF_3 substituted copolymers that contain flexible, nonconjugated spacers (**19**) (**33**). Heck chemistry has been applied to prepare copolymers such as **20** (**34**).

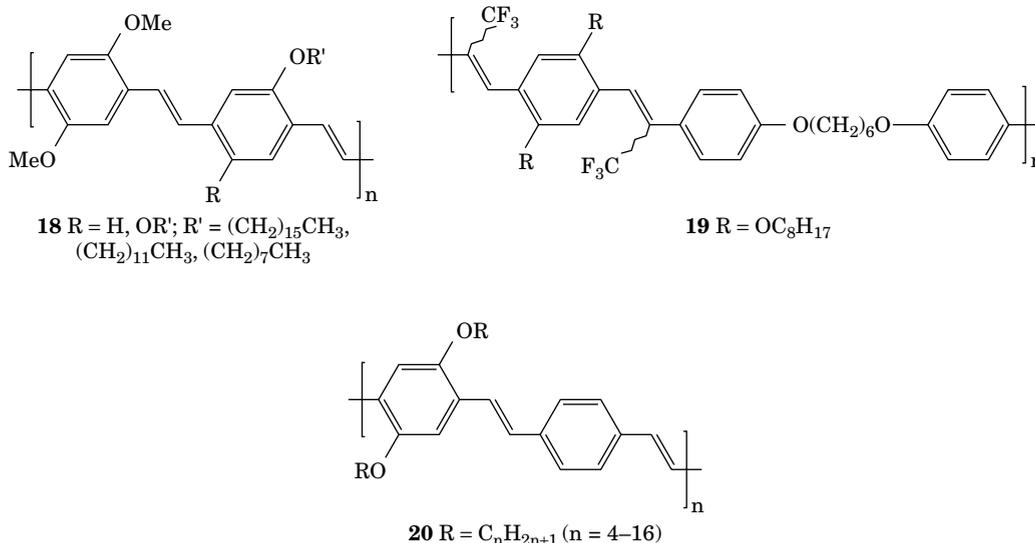
Composites and Blends

The mechanical and optoelectronic properties of PPV and its derivatives may be optimized for specific applications by using polymer blends and composite materials. Blends of MEH-PPV and polyethylene have been used to enhance the degree significantly to which the conjugated polymer chains are aligned by stretching (**35**). The density of entanglements in gels of polyethylene is much lower than in spin-cast polymer solutions, and this effect remains after removing the solvent. This allows tensile drawing of such blends to large draw ratios (>200), which causes the conjugated polymers to align to a degree normally expected in single crystals. Charge transport through blends can also be controlled by using a host polymer that has specific properties. Blue emitters have been prepared from blends of poly(*p*-phenylphenylenevinylene) (PPPV) in poly(9-vinylcarbazole) (PVK), a hole-transporting polymer (**36**). In addition to enhancing the processibility of PPPV, PVK blue-shifts the electroluminescence, enhances hole-transport, and increases the probability of radiative recombination due to the dilution effect.

Composites of PPV and silica (SiO_2) and vanadium oxide (V_2O_5) have been prepared for use in nonlinear optical applications (**37-39**). These composites are prepared via sol-gel processing methods, and allow combining the superior nonlinear optical properties of the conjugated polymers with the very low optical losses found in inorganic glasses. Composite films of insulating SiO_2 , TiO_2 , and Al_2O_3 nanoparticles and MEH-PPV were prepared and result in more efficient charge injection and transport in electroluminescent devices formed from them, as well as enhanced emission intensities (**40,41**). Photovoltaic and time-resolved microwave conductivity measurements were also used to study nanocrystalline TiO_2 /PPV composites; these studies show that excitons generated in the polymer are



Scheme 7



dissociated at the polymer/TiO₂ interface and the electrons are transferred to the nanocrystals (42).

Composites of PPV in films of the polyelectrolyte Nafion have been synthesized by electrostatically binding the dicationic monomer 1 to the film, followed by treatment with base and thermal conversion (43). Ordered nanocomposites of PPV have been synthesized from mixtures of polymerizable lyotropic liquid crystals using PPV precursors (44). Photopolymerization of the host results in a hexagonal architecture, from which thin films and fibers can be fabricated. A significant enhancement in the photoluminescence of the composite relative to PPV was found.

PROPERTIES

Absorption and Emission

Films of PPV show three absorption bands whose maxima are at 6.12, 5.06, and 3.09 eV (45). The two higher energy bands are attributed to localized molecular states, whereas the lower energy band is due to a delocalized electronic excitation. The emission spectrum of PPV upon excitation at 355 nm is resolved into three lines whose spacing of 0.16 eV is due to vibronic coupling (46). Migration of the excited state to the longest conjugation length segments in the polymer appears to occur before radiative decay because smearing out of the vibrational fine structure is expected if emission from a distribution of sites within the polymer occurs. The photoluminescent efficiency varies between 5 and 25%, depending on the synthetic route used and the conversion conditions (47).

Two descriptions of the excited state have been applied to organic semiconductors, the exciton and the band model. The appropriate model depends on the extent of coupling between sites; strong coupling yields uncorrelated electrons and holes, and weak coupling favors correlated electron-hole pairs (excitons). Time-resolved fluorescence and polarized fluorescence experiments suggest that the

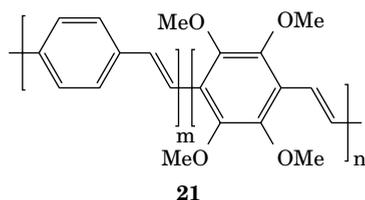
exciton model is appropriate in PPV (48). Rothberg and co-workers examined the relative effects of interchain versus intrachain excitations in MEH-PPV by comparing excited state lifetimes and quantum yields in films to those of dilute solutions (49). They observed significantly lower quantum yields for emission in films and attributed this to the formation of nonemissive interchain excitons, that are not formed in dilute solution. They also concluded that film morphology can play a significant role in the photophysical behavior of PPV (50). The presence of trace oxygen in the conversion of precursor polymer to PPV reduces the photoluminescence of the resulting material (51). It was shown that this is correlated with the formation of carbonyl groups in the polymer backbone and can be prevented by carrying out the conversion in a reducing atmosphere.

Photoconductivity

Photoconductivity in PPV was first measured three decades ago (52). A later study revealed low dark conductivities for PPV films (<10⁻¹⁵ S/cm) but significant photoconductivity upon irradiation at 440 nm. Significant conductivity was also found upon irradiation of the film in the near-infrared region, despite insignificant optical absorption in this region. This was attributed to a charge-transfer mechanism that involves trace oxygen (53). Oriented PPV films showed enhancements in photoconductivity in light polarized parallel to the direction in which the films were stretched (54). Transient photoconductivity measurements have also been used to address the question of the nature of the charge carriers in MEH-PPV films. The exciton model predicts strong dependence of photoconductivity on temperature, and this is observed for films thicker than 120 nm. In thinner samples, both steady-state and fast time-resolved photoconductivity measurements demonstrate that photoconductivity is independent of temperature. These results are inconsistent with the exciton model (55).

Doping and Electrical Conductivity

Pristine PPV films are insulators at room temperature; however, exposure to oxygen causes an increase in conductivity to 10^{-11} S/cm, an effect attributed to reversible doping where the oxygen acts as an electron acceptor (56). Irreversible doping of PPV films with strong oxidants such as FeCl_3 or H_2SO_4 produces black films whose conductivities are very high relative to the pristine material (57). Films doped with sulfuric acid showed conductivities of $\sim 10^4$ S/cm that were only weakly temperature dependent, indicating metallic behavior. The conductivity of films doped with FeCl_3 was slightly lower (10^3 S/cm) and decreased with temperature. The doped films are stable in oxygen but are moisture sensitive. Copolymers of PPV derivatives that contain electron-donating groups, such as poly(1,4-phenylenevinylene-*co*-2,3,5,6-tetramethoxy-1,4-phenylenevinylene) (21), can also be doped with weaker oxidants such as I_2 to give materials whose conductivities are as high as 7×10^{-2} S/cm (58).



Microstructure and Liquid Crystallinity

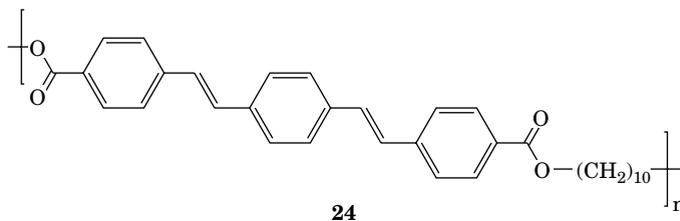
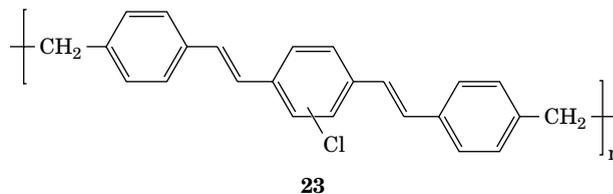
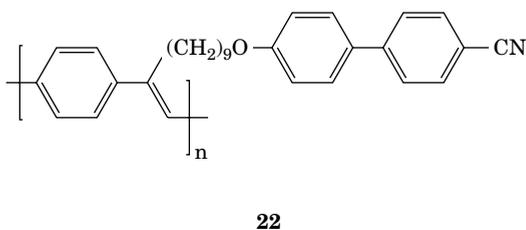
The degree of structural order in PPV films depends greatly on the method of preparation. The degree of broadening in electron diffraction patterns has been used to assess the extent of ordering in different PPV samples (59). High-resolution transmission electron microscopy revealed, crystalline regions of approximately 7 nm in oriented PPV films, and these crystallites are retained upon doping the films with H_2SO_4 (60). Substituents also affect the structural order of PPV derivatives. Methoxy substituents allow chains to interlock and

form a more ordered three-dimensional structure, whereas methyl groups distort the backbone due to repulsive interactions between the backbone and side groups, resulting in a more disordered material (61). X-ray diffraction studies on MEH-PPV films cast from different solvents have shown that the chain orientation is anisotropic and the aromatic rings are oriented predominantly parallel to the film plane (62). The crystallinity of films cast from THF is highest, and the crystalline domains are largest compared to films prepared from other solvents.

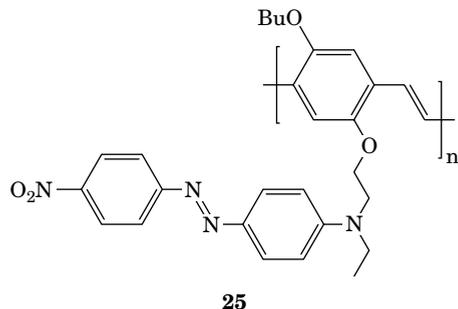
Liquid crystallinity has been observed in derivatives of PPV that bear mesogenic substituents, as well as in copolymers that contain phenylenevinylene segments in the main chain. Copolymers in which some of the phenyl rings have alkoxy side chains (20) exhibit a nematic liquid crystalline phase which has been characterized by polarized microscopy and differential scanning calorimetry (34). The temperature range between the melting point and the nematic–isotropic phase transition, it was found, depends on the length of the alkoxy group. A PPV derivative (22) that bears the well-known cyanobiphenyl mesogen as a side chain has both nematic and smectic mesophases (63). This polymer was oriented by rubbing a film with a Teflon stick, and a significant degree of orientation was observed by polarized UV/visible and IR spectroscopies. Several examples of main-chain liquid crystalline polymers that contain phenylenevinylene moieties bridged by saturated linkers are known. The thermotropic polymer 23 was prepared by using a Wittig procedure, and it was found that it melts anisotropically (64). A related main-chain polymer (24) has a mesophase that exists between 218 and 275°C (65).

Nonlinear Optical Properties

For many optical signal processing applications, it is desirable for materials to have large optical nonlinearities and fast response times. For example, third-order nonlinear optical (NLO) properties result in laser pulse induced refractive index changes that occur on the femtosecond timescale. These changes could be exploited in fast optical switches. Conjugated polymers are expected



to be good candidates for such applications due to the delocalization of charge in the polymer backbone (66).



PPV has many of the characteristics desired in a NLO material, including good transparency, high π -electron density, and optical quality films that may be oriented and ordered. PPV has a third-order nonlinear optical susceptibility ($\chi^{(3)}$) of 7.8×10^{-12} esu (67), whereas $\chi^{(3)}$ for a substituted derivative, poly(2,5-dimethoxy-*p*-phenylenevinylene), is 5.4×10^{-11} esu at $1.85 \mu\text{m}$ (68). The higher value for the dimethoxy-substituted derivative may be due to more extended conjugation in this material. An alternative strategy which has been investigated is to introduce a NLO-active moiety pendent to a PPV backbone. For example, Disperse Red 1 has been tethered in this way (25), and the resulting polymer has a $\chi^{(3)}$ value of 2.5×10^{-12} esu (69).

APPLICATIONS

Photovoltaics

Heterojunctions between conjugated polymers and films of electron acceptors behave as rectifying p-n junctions and may be used in photovoltaic devices. Such junctions have been prepared by vacuum evaporation of n-type buckminsterfullerene (C_{60}) onto spin-cast films of p-type MEH-PPV on ITO-glass substrates (Fig. 1) (70). These devices behave as rectifiers in the dark and pass a photocurrent when

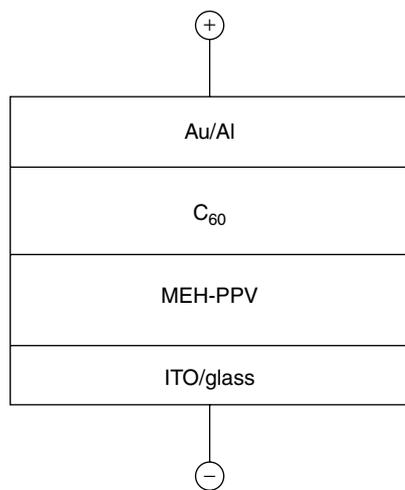


Figure 1. Schematic of a MEH-PPV photovoltaic cell.

illuminated by visible light. The open circuit voltage (V_{oc}) saturates at 0.53 V and has a fill factor of 0.48 and a power conversion efficiency of 0.04%. The Cambridge group also reported photovoltaic devices in which a heterojunction between bis(phenethylimido)perylene and PPV is sandwiched between ITO and Al. These devices had a somewhat greater fill factor (0.6), and V_{oc} approached 1 V; the quantum yield was 6% (electrons per incident photon) (71). A related approach involves using a blend of MEH-PPV and bis(phenethylimido)perylene that gives a fill factor of 0.27 at an open circuit voltage of 0.58 V; however, improvements in these devices are limited by the poor solubility of the bis(phenethylimido)perylene (72).

Photovoltaic cells have also been constructed from light-emitting electrochemical cells (LECs; see later) (73). In these devices, a phase-separated blend of MEH-PPV and cyano-PPV is used along with a solid electrolyte that consists of a mixture of polyethylene oxide (PEO) and LiCF_3SO_3 . Sandwich photovoltaic cells using Al and ITO as the electrode contacts were doped using a prebias of 3 V and resulted in a V_{oc} of 1.0 V and a power conversion efficiency of 0.1%, assuming a fill factor of 0.25. The built-in potential is determined by the chemical potential difference between the p-doped and n-doped layers, rather than the work function of the electrodes; thus, air-stable electrodes can be used in these cells.

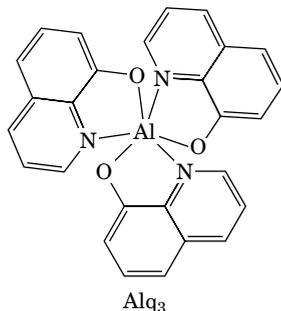
Optical Memory

Data have been permanently stored in films of PPV derivatives by irradiating films of the sulfonium precursor polymers using either a Xe arc lamp or Ar ion laser (488 nm) (74). Subsequently heating the films resulted in the formation of colored, conjugated films only in regions that were not irradiated. Photochemical scission of the polymer chains leaves a water-soluble residue which is readily removed by rinsing the heat-treated films in water. This process may also be used for lithographic patterning of PPV films onto substrates.

Light-Emitting Devices

In the late 1980s, Tang and VanSlyke reported electroluminescent devices that used thin films of 8-hydroxyquinoline aluminum (Alq_3), as the emitting material (75). They distinguished these devices from those based on conventional inorganic semiconductors by calling them "organic," despite the fact that the emissive compound is actually an inorganic coordination complex. The devices consisted of a layer of a hole-transporting aromatic amine on an ITO electrode, 600 Å of the luminescent Alq_3 , and a Mg/Ag electrode. The devices behaved like rectifiers and emitted light whose peak intensity at 550 nm had a forward bias of as little as 2.5 V. In a subsequent publication, Tang and VanSlyke showed that doping the Alq_3 layer with other highly fluorescent molecules, such as coumarin 540, increases the electroluminescent efficiency and allows tuning the color from blue-green to orange-red (76).

In 1990, the discovery that PPV could be used as the emitter in an electroluminescent device was reported by Friend and co-workers at the Cavendish Laboratory in



Cambridge (2). The devices consisted of a PPV film, prepared by using the sulfonium precursor route, sandwiched between an indium tin oxide (ITO) and an Al electrode; green-yellow light was emitted under a forward bias of 14 V and the quantum efficiency was up to 0.05%. Shortly after this initial publication, Braun and Heeger demonstrated that MEH-PPV could also be used to fabricate EL devices in which the polymer was directly cast from solution in the conjugated form. They used both indium and calcium cathodes and observed visible light at 4 V forward bias using a calcium cathode whose efficiency was 1%.

Device Operation. Single-layer devices consist of an electroluminescent layer sandwiched between an electron-injecting cathode (usually a low work-function material such as Ca or Al) and a hole-injecting anode (most commonly the transparent conductor indium tin oxide (ITO) on glass).

The operation of the device under forward bias may be understood by using a simple band diagram (Fig. 2a). The anode and cathode materials are chosen to provide low

barriers to electron and hole injection by matching the valence and conduction band energies of the polymer to the electrode work functions. A disadvantage of the single-layer device configuration is that charge recombination often occurs close to the cathode because most EL polymers are better hole conductors than they are electron conductors. The metal electrodes can quench excitons in close proximity, thus reducing EL efficiency.

An approach that has been used successfully to move the emitting zone away from the electrodes is constructing two-layer cells in which recombination occurs at the interface between the two organic layers (Fig. 2b). Here, the material in the layer adjacent to the cathode is selected for high electron mobility but hole mobility lower than the EL polymer which is located adjacent to the anode. Thus, electrons and holes are readily injected into the adjacent layers that contact the respective electrodes and accumulate at the interface between the two layers. A number of materials have been exploited for use as the electron transporting layer (ETL), including 2-(4-biphenyl)-5-(4-*tert*-butylphenyl)-1,3,4-oxadiazole (butyl PBD) 26 (77). Two-layer devices constructed from PPV as the emitting layer and butyl PBD dispersed in poly(methylmethacrylate) (PMMA) as the ETL showed a 10-fold improvement in efficiency relative to analogous single-layer devices constructed only from PPV. Polymers that contain oxadiazole moieties pendent from the backbone and in the main chain, have been synthesized and tested as ETLs and also improve external quantum efficiencies in two-layer devices (78). The EL efficiency is temperature independent in these devices, suggesting that charge injection from both electrodes is well balanced.

Hole injection from the anode can be improved by using a hole-transport layer that functions by improving

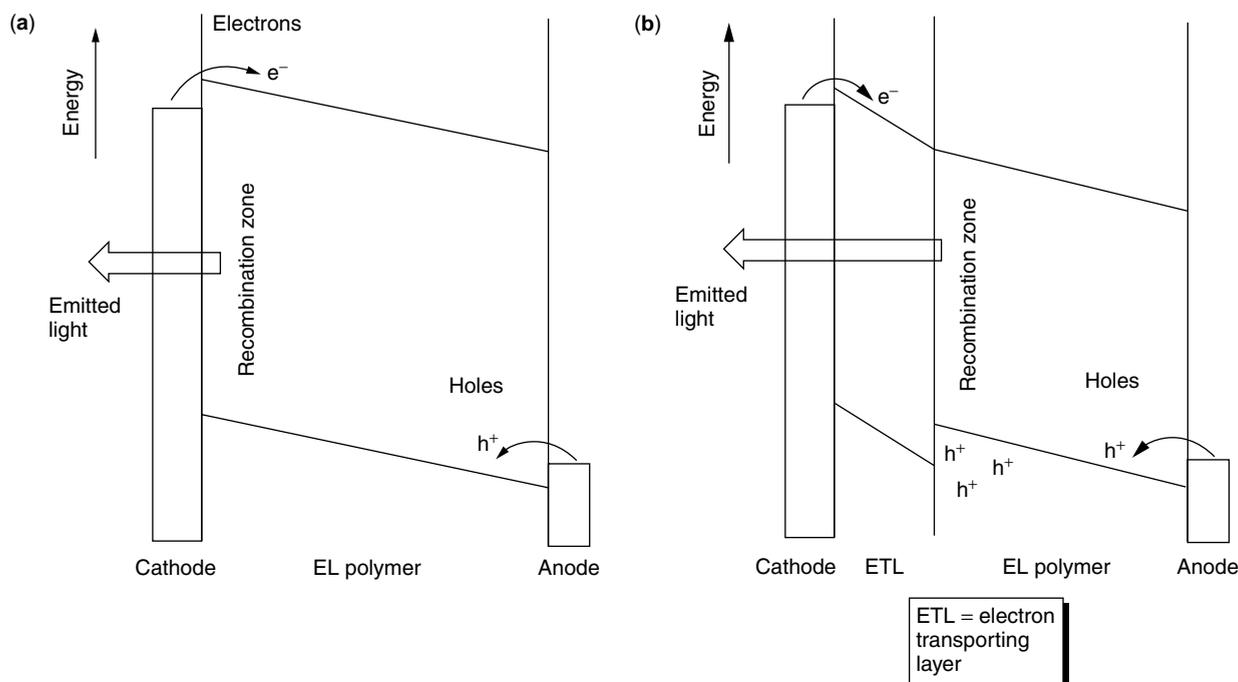
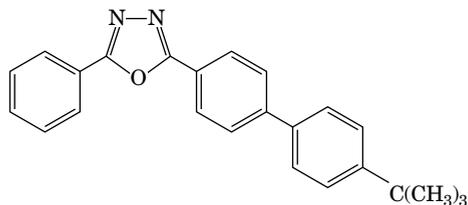
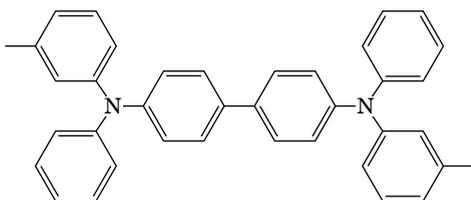


Figure 2. Band diagram for (a) single-layer and (b) two-layer polymer EL devices.

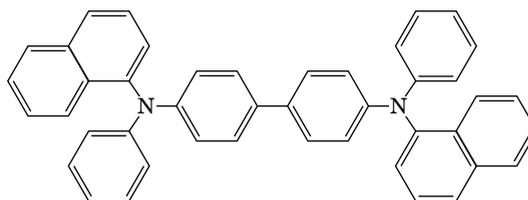


26

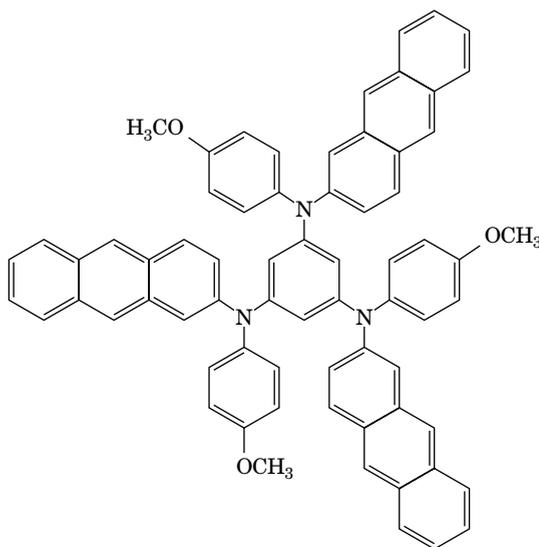
hole injection into the electroluminescent layer and maximizing recombination of injected charges. Optimally, these materials should possess high glass transition temperatures (T_g) and low ionization potentials. A large variety of triaryl amines have been investigated for use as hole-transport materials including *N,N'*-diphenyl-*N,N'*-di-*m*-tolyl-1,1'-biphenyl-4,4'-diamine (TPD) 27 (79), *N,N'*-1-naphthyl-*N,N'*-diphenyl-1,1'-biphenyl-4,4'-diamine (NPD) 28 (80), 1,3,5-tris(2'-anthracyl-4''-methoxyphenylamino)benzene 29 (81), and the atropisomeric *cis*- and *trans*-5,11-dihydro-5,11-di-1-naphthylindolo [3,2-*b*]carbazole 30 (82).



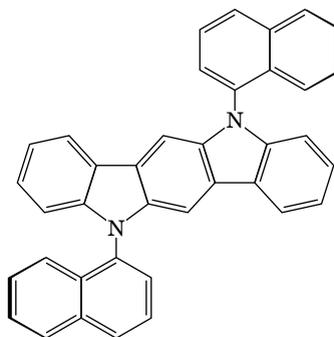
27



28



29

*trans*-30

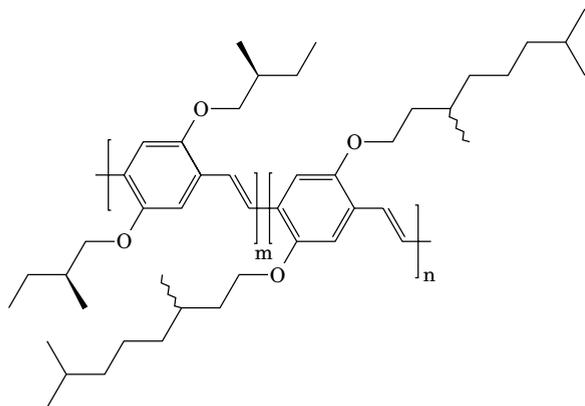
The brightness of EL polymer devices varies, depending on the device configuration and the polymer used. Uniax Corporation has reported ITO/PANI-CSA/MEH-PPV/Ca (PANI-CSA = polyaniline-camphor sulfonic acid) devices whose brightness is as high as 10,000 cd/m² for bias voltages >5 V, more than twice the brightness of a fluorescent lamp (83). Device brightness can be increased by using short drive pulses, that permit high current operation (84).

Luminous efficiencies as high as 0.8 cd/A can be achieved this way using current densities of 1 kA/cm² in single-layer PPV-based devices. Poly(3,4-ethylenedioxythiophene) (PEDOT) doped with polystyrene sulfonic acid (PSS) has also been investigated as a polymeric anode and it has been shown, results in increases in efficiency, brightness, and lifetime for devices based on MEH-PPV (85). When PPV is thermally converted by using a PEDOT/PSS overlayer, an interface of large surface area results due to interpenetration of the two layers, possibly contributing to the improvement in quantum efficiency measured for these devices (86).

An application for which polymer-based EL devices may be uniquely suited is as flexible emitting materials. Such devices were first reported in 1992 by workers at Uniax, using a thin layer of PANI-CSA on flexible poly(ethylene terephthalate) (PET) as the anode, an MEH-PPV film as the emitter, and vacuum-deposited Ca as the cathode (87). The PANI-CSA film is light green, but the absorption is low in the spectral region where MEH-PPV emits (500–700 nm), so emission is readily observed through the anode.

Doped silicon has also been used as a hole-injecting electrode material (88). Devices consisted of a MEH-PPV film on a thin SiO₂ layer atop a heavily p- or n-doped Si substrate. The top contact was a semitransparent Ca or Au electrode, and light emission was observed through this layer. The insulating SiO₂ layer allows matching the electrode Fermi level to the conduction band of the polymer, thus permitting the devices to operate in reverse bias using

a high work-function cathode such as Au. The Cambridge group also investigated the effect of separating the polymer layer from the Al cathode by a thin SiO₂ layer. This results in an increase in PL quantum efficiency because the emission zone is brought away from the metal interface (89).



31

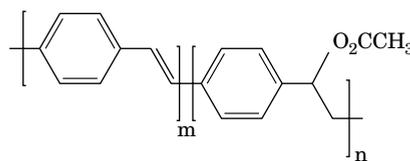
Circularly polarized electroluminescence (CPEL) has been obtained from a single-layer device in which the emitter is a PPV derivative that has chiral pendent groups (31) (90). The CPEL is due to the inherent chirality of the polymer; polymers that contain racemic side groups did not show any effect, so it is believed that aggregation of the polymer chains in the solid state, enhances the magnitude of the effect.

The external efficiencies of electroluminescent devices based on PPV derivatives are typically between 0.2 and 4% (4), comparable to the efficiencies of inorganic EL devices. In addition to using electron- and hole-transporting layers to enhance efficiencies, other strategies have also been used. Son and co-workers reported polymers in which cis linkages were intentionally introduced into the PPV backbone (91). These defects prevent the polymer chains from packing efficiently in the solid state, thus producing amorphous PPV films. Both single- and two-layer devices prepared from this material had significantly higher efficiencies than comparable devices fabricated from PPV synthesized in the conventional fashion (high trans content). This improvement in efficiency may result from better chain separation in the amorphous material.

Degradation of polymer EL devices that leads to limited lifetimes is a serious problem, particularly for commercial applications that typically demand lifetimes of >10,000 hours. The low work-function metals used as cathode materials (Al, Ca, In) are air and moisture sensitive to varying degrees. Encapsulation to prevent intrusion of air and water is necessary for practical devices. It has also been shown that the polymer layer is susceptible to degradation by singlet oxygen, which can form from oxygen impurities in the film by energy transfer from a nonradiative exciton in the polymer (92). Strict removal of oxygen in the film preparation process thus results in longer device lifetimes. Oxidation of MEH-PPV, it has been shown, results in reduced fluorescence due to the formation of carbonyl groups, along with reduced carrier mobility as a consequence of chain scission. Localized shorts also form in these devices, which are initially isolated by melting of the

cathode; however, eventually these coalesce and cause device failure (93).

Although most studies have been carried out on PPV prepared by solution processing, EL devices have also been made from PPV prepared by chemical vapor deposition (94). The advantage of this approach is that many of the side reactions that typically occur during solution processing are eliminated, and this method is also compatible with existing methods for processing inorganic LEDs. Single-layer devices prepared this way had turn-on voltages as low as 4.5 V and brightness of 20 cd/m³.



32

It has also been demonstrated that modifications of the PPV structure affect the performance of EL devices (95). A PPV copolymer that bears acetate side groups (32) which disrupt the conjugation in the backbone has been used to fabricate encapsulated single-layer devices using Al cathodes and ITO anodes which operated continuously in air for 12,000 hours without degradation. The devices were encapsulated by gluing a glass slide over the top of the device using epoxy resin.

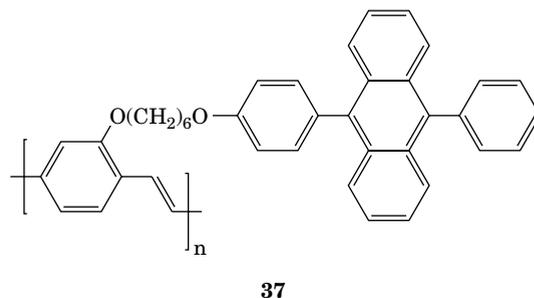
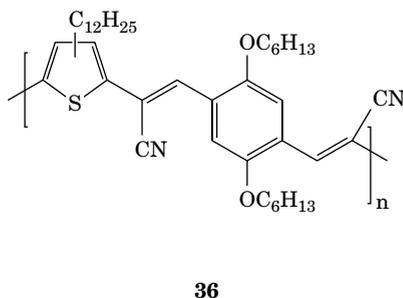
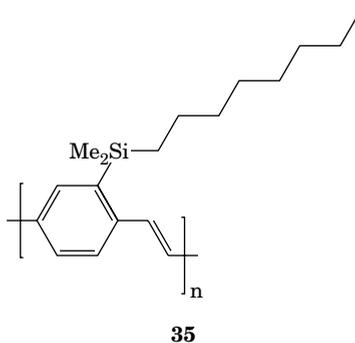
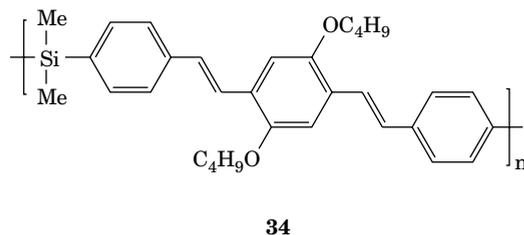
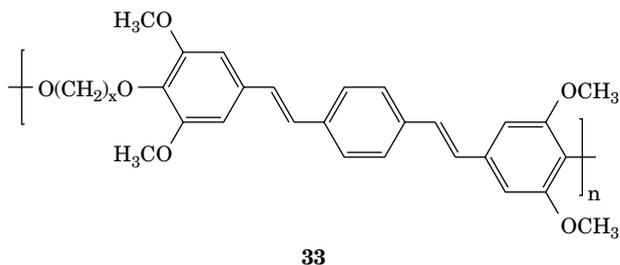
For many display applications, it is useful to pattern the emitting area of a device. Several successful approaches have been developed in this vein, including ink-jet printing and solvent-assisted micromolding. In the ink-jet approach, a conventional printer was modified to spray a solution of PEDOT onto defined areas of an ITO electrode (96). The whole area was then coated by a MEH-PPV film, and a Ca cathode was deposited to complete the device. Emission occurs only in the areas defined by the PEDOT layer due to the enhanced charge injection from this polymer relative to the ITO layer. Typical pixel sizes ranged from 180 to 400 μm, and the density of the dots could be used to control the brightness of emission from a given area. Solvent-assisted micromolding involves using methanol to wet a polydimethylsiloxane mold and allowing a solution of PPV precursor polymer to wick into the recessed regions of the mold (97). The mold is initially prepared by casting and curing a prepolymer against a lithographically prepared relief structure. When the methanol evaporates, the mold is removed, leaving the PPV precursor polymer patterned on the surface. Curing yields PPV, and the devices are completed by depositing Ca or Al. Emitting features as small as 800 nm can be prepared by this method.

Color Range and Tuning. Variations in polymer structure allow tuning the emitted color across the entire visible spectrum. Polymers such as 16 in which the conjugation is broken due to the presence of methoxy groups show blue-shifted electroluminescence relative to unsubstituted PPV and an emission maximum at 508 nm (98). Blue emission was also obtained from a copolymer containing short methoxy-substituted phenylenevinylene segments (33)

(99). Single-layer devices that contain this polymer sandwiched between Al and ITO have a maximum at 465 nm in the EL spectrum. A PPV derivative that has Si groups in the backbone (34) emits blue light in both single and two-layer device configurations (100). Green light was obtained from multilayer devices using poly(2-methyloctylsilyl-1,4-phenylenevinylene) (DMOS-PPV) (35) as the emitter combined with an electron-conducting and hole-blocking layer of either butyl PBD or a polymeric oxadiazole derivative (101). A low-band-gap, cyano-substituted thienylene phenylenevinylene copolymer 36 emits in the near-IR region. The maximum in the EL spectrum is at 740 nm, and emission tails down to 1000 nm (102).

These structures were prepared by alternate absorption of PPV precursor polymer and negatively charged CdSe particles onto an ITO substrate. Electroluminescence in these devices, arises exclusively through charge recombination on the CdSe particles and has a maximum intensity at 650 nm. The PPV, it is believed, acts only to transport charge, thus significantly reducing the turn-on voltage in these devices compared to those using insulating poly(allylaminehydrochloride) as the host.

Polymer-based light-emitting devices in which the emitted color may be tuned by varying the bias voltage have been fabricated. In one approach, the emitting layer consists of a mixture of poly(2,5-dioctyl-*p*-phenylenevinylene)



Several groups have developed polymer LEDs that emit light across a broad wavelength range, resulting in white light. Photoluminescence from polymer 37 arises from both the pendent diphenylanthracene units and the polymer backbone, thus covering much of the visible spectrum (103). However, electroluminescence from single-layer devices that contain this polymer is significantly red-shifted relative to the photoluminescence that emits green-red light apparently only from the main chain. Another approach that has produced white electroluminescence involves CdSe/PPV multilayer structures (104).

(ROPPV-8) and Alq₃ sandwiched between a Mg:In cathode and an ITO anode (Fig. 3a) (105). The electroluminescence from this device varies from orange to greenish-yellow as the bias voltage is increased from 15 to 22 V. Increasing the voltage increases the proportion of emitted light from the Alq₃. Devices in which ROPP-8 and Alq₃ are in separate layers emit only from the polymer layer, regardless of the applied voltage. Another approach uses a three-layer device, consisting of PPV, perfluoropropylated poly(*p*-phenylene) (F-PPP), and poly(3-dodecylthiophene) (PDT) (Fig. 3b) (106). The F-PPP acts as an

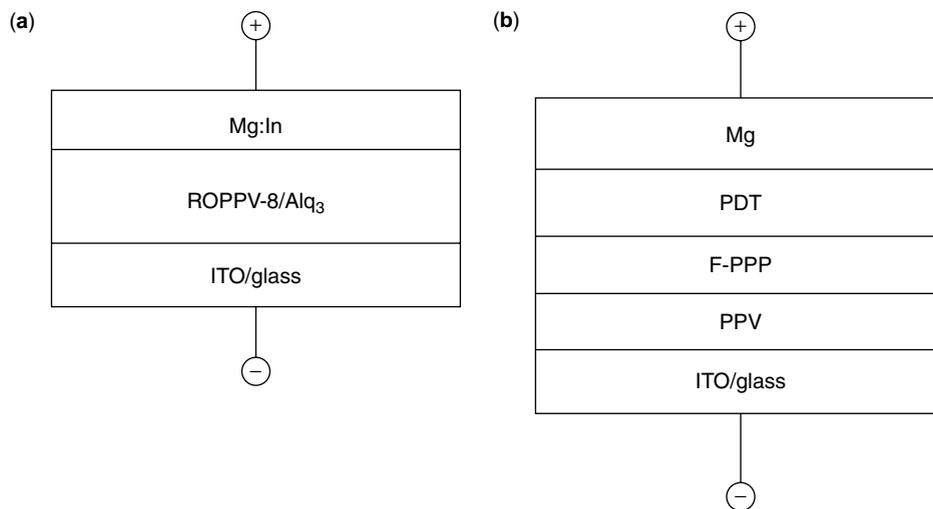


Figure 3. Schematics of tunable color light-emitting devices.

electron-blocking layer and confines emission only to the side of the device to which a negative bias is applied. Thus, green light is emitted when the negative bias is on the PPV side, and red light is emitted when it is applied to the PDT side.

Self-Assembled Devices. Rubner and co-workers at MIT developed a self-assembly approach to constructing PPV-containing heterostructures for use in EL devices (107). The heterostructures are built up by alternately dipping a substrate in solutions of the sulfonium PPV precursor polymer and anionic polyelectrolytes such as the sodium salts of poly(styrenesulfonate) (PSS) or poly(methacrylic acid) (PMA). The layers are then heated to convert the PPV to the fully conjugated form. This method allows very precise control of thickness, as well as manipulation of the nature of the surface (cationic or anionic) in contact with the anode and cathode. Device efficiency is very sensitive to the chemical nature of the interface; the PMA/Al interface consistently yielded efficiencies greater by a factor of 2 than devices that have a PPV/Al interface.

Light-Emitting Electrochemical Cells

Recently, Heeger and co-workers developed light-emitting devices in which a blend of a semiconducting, luminescent polymer and a solid-state electrolyte is used in the emitting layer (108). These devices function as solid-state electrochemical cells, and charge injection results in forming a p-i-n junction through electrochemical doping. Many of the requirements of conventional polymer LEDs, such as a match between the electrode work function and the polymer valence and conduction bands, are circumvented in such light-emitting electrochemical cells (LECs).

Polymer LECs, it has been demonstrated, function in a dynamic-junction mode, in which ions move under the external bias to create the junction. These devices degrade when the drive voltage is beyond the window of electrochemical stability. When polyethylene oxide (PEO) is used

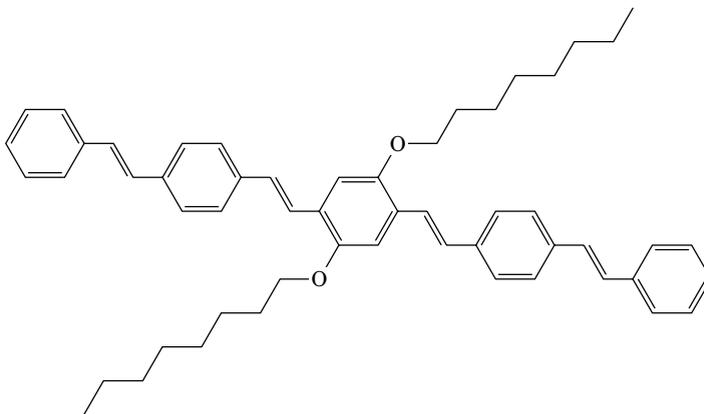
as the ion transport medium, the devices can operate in a frozen-junction mode at 100 K. Response times in this mode are comparable to those of polymer LEDs. An improvement on this approach allows the frozen junction to function at room temperature (109). In these single-layer devices, a PPV derivative such as poly(2-butyl-5-(2-ethyl(hexyl))-1,2-phenylenevinylene) was blended with an electrolyte containing a crown ether and lithium triflate. The devices were activated by applying a small voltage while heating to 60–80°C, creating the p-i-n junction, and functioned at room temperature at a turn-on voltage of 2.1 V and an external EL efficiency of 2–3%.

Stimulated Emission and Lasers

In 1996, Friend and co-workers demonstrated optically driven lasing from thin films of PPV sandwiched between a distributed Bragg reflector and an Ag mirror (110). Excitation was by a Nd:YAG laser, frequency-tripled to 355 nm. Emission occurs in three cavity modes, equally distributed at low excitation energy, but the 545 nm emission dominates at higher excitation energy, indicating that the device is lasing. In the same year, the Santa Barbara group reported laser emission from solutions and films of MEH-PPV and TiO₂ nanocrystals. The nanoparticles multiply scatter photons in the composite material, and the scattering length exceeds the gain length (111). Stimulated emission has also been observed from crystalline films of an oligomer of phenylenevinylene substituted with octyloxy groups (34) (112).

FUTURE CONSIDERATIONS

The last 10 years have seen an explosion of research interest in PPV and its derivatives, driven by the potential of these materials as emitters in EL devices. Companies such as Philips, Uniax (DuPont Displays), Cambridge Display Technology, and Covion are engaged in bringing display applications for low-information content (cell phone and



38

alphanumeric displays) based on organic emitters to market, and in late 1999, Pioneer Electronics had already begun selling a monochrome display for use in vehicles. Applications for higher information content, such as computer and television displays, may be developed in the near future.

Although many of the problems which are of concern in optoelectronic applications of organic materials have been overcome, there are still many challenges ahead in the field. Many of the polymers that have been made recently have not yet been investigated in detail, and more work is needed to optimize them for specific applications. Fundamental questions also remain, such as the exact nature of the charge carriers in PPV films, and further studies of the photophysical behavior of these polymers are required.

Perhaps the most exciting prospect which has arisen from the intense research activity on PPV is the realization that organic materials and polymers can be used in real-world optoelectronic applications. Hopefully, this will stimulate efforts to make and study more novel materials and to find new applications for them. In particular, any integration of conventional and molecular electronics will require many advances. The success of PPV and other organic EL materials has demonstrated that research in this field is a worthwhile enterprise, and the development of entirely new classes of materials that have unknown properties may be on the horizon.

BIBLIOGRAPHY

1. T. Skotheim, R.L. Elsenbaumer, and J.R. Reynolds, eds., *Handbook of Conducting Polymers*. Marcel Dekker, NY, 1998.
2. J.H. Burroughes, D.D.C. Bradley, A.R. Brown, R.N. Marks, K. Mackay, R.H. Friend, P.L. Burns, and A.B. Holmes, *Nature* **347**: 539–541 (1990).
3. F.R. Denton, III and P.M. Lahti, *Plast. Eng.* **49**: 61–102 (1998).
4. A. Kraft, A.C. Grimsdale, and A.B. Holmes, *Angew. Chem. Int. Ed.* **37**: 403–428 (1998).
5. J. Salbeck, *Ber. Bunsenges. Phys. Chem.* **100**: 1667–1677 (1996).
6. J.R. Sheats, H. Antoniadis, M. Hueschen, W. Leonard, J. Miller, R. Moon, D. Roitman, and A. Stocking, *Science* **273**: 884–888 (1996).
7. J.R. Sheats, Y.-L. Chang, D.B. Roitman, and A. Stocking, *Acc. Chem. Res.* **32**: 193–200 (1999).
8. R.A. Wessling, *J. Polym. Sci. Polym. Symp.* **72**: 55–66 (1985).
9. R.W. Lenz, C.C. Han, J. Stenger-Smith, and F.E. Karasz, *J. Polym. Sci. Part A Polym. Chem.* **26**: 3241–3249 (1988).
10. R.O. Garay, U. Baier, C. Bubeck, and K. Muellen, *Adv. Mater.* **5**: 561–564 (1993).
11. P.L. Burn, D.D.C. Bradley, R.H. Friend, D.A. Halliday, A.B. Holmes, R.W. Jackson, and A. Kraft, *J. Chem. Soc. Perkin Trans. 1*: 3225–3231 (1992).
12. P.M. Lahti, D.A. Modarelli, F.R. Denton, III, R.W. Lenz, and F.E. Karasz, *J. Am. Chem. Soc.* **110**: 7258–7259 (1988).
13. B.R. Cho, M.S. Han, Y.S. Suh, K.J. Oh, and S.J. Jeon, *J. Chem. Soc. Chem. Commun.* 564–566 (1993).
14. R.N. McDonald and T.W. Campbell, *J. Am. Chem. Soc.* **82**: 4669–4671 (1969).
15. W. Heitz, W. Bruegging, L. Freund, M. Gailberger, A. Greiner, H. Jung, U. Kampschulte, N. Niessner, and F. Osan, *Makromol. Chem.* **189**: 119–127 (1988).
16. A.W. Cooke and K.B. Wagener, *Macromolecules* **24**: 1404–1407 (1991).
17. H.G. Gilch and W.L. Wheelwright, *J. Polym. Sci. Part A-1* **4**: 1337–1349 (1966).
18. H. Nishihara, M. Tateishi, K. Aramaki, T. Ohsawa, and O. Kimura, *Chem. Lett.* 539–542 (1987).
19. W.-P. Chang, W.-T. Whang, and P.-W. Lin, *Polymer* **37**: 1513–1518 (1996).
20. M. Tateishi, H. Nishihara, and K. Aramaki, *Chem. Lett.* 1727–1728 (1987).
21. E.G.J. Staring, D. Braun, G.L.J.A. Rikken, R.J.C.E. Demandt, Y.A.R.R. Kessener, M. Bouwmans, and D. Broer, *Synth. Met.* **67**: 71–75 (1994).
22. S. Iwatsuki, M. Kubo, and T. Kumeuchi, *Chem. Lett.* 1071–1074 (1991).
23. Y.-J. Miao and G.C. Bazan, *J. Am. Chem. Soc.* **116**: 9379–9380 (1994).
24. V.P. Conticello, D.L. Gin, and R.H. Grubbs, *J. Am. Chem. Soc.* **114**: 9708–9710 (1992).
25. S.H. Askari, S.D. Rughooputh, and F. Wudl, *Synth. Met.* **29**: E129–E134 (1989).

26. F. Wudl, P.M. Allemand, G. Srdanov, Z. Ni, and D. McBranch, *ACS Symp. Ser.* **455**: 683–686 (1991).
27. C. Zhang, H. von Seggern, K. Pakbaz, B. Kraabel, H.W. Schmidt, and A.J. Heeger, *Synth. Met.* **62**: 35–40 (1994).
28. M. Rehahn and A.-D. Schlüter, *Makromol. Chem. Rapid Commun.* **11**: 375–379 (1990).
29. P.L. Burn, A. Kraft, D.R. Baigent, D.D.C. Bradley, A.R. Brown, R.H. Friend, R.W. Gymer, A.B. Holmes, and R.W. Jackson, *J. Am. Chem. Soc.* **115**: 10117–10124 (1993).
30. P.L. Burn, A.B. Holmes, A. Kraft, D.D.C. Bradley, A.R. Brown, R.H. Friend, and R.W. Gymer, *Nature* **356**: 47–49 (1992).
31. R.M. Gregorius, P.M. Lahti, and F.E. Karasz, *Macromolecules* **25**: 6664–6669 (1992).
32. G. Kossmehl and M. Samandari, *Makromol. Chem.* **186**: 1565–1574 (1985).
33. A. Lux, A.B. Holmes, R. Cervini, J.E. Davies, S.C. Moratti, J. Gruener, F. Cacialli, and R.H. Friend, *Synth. Met.* **84**: 293–294 (1997).
34. Z. Bao, Y. Chen, R. Cai, and L. Yu, *Macromolecules* **26**: 5281–5286 (1993).
35. T.W. Hagler, K. Pakbaz, K.F. Voss, and A.J. Heeger, *Phys. Rev. B Condens. Matter* **44**: 8652–8666 (1991).
36. C. Zhang, H. von Seggern, K. Pakbaz, B. Kraabel, H.-W. Schmidt, and A.J. Heeger, *Synth. Met.* **62**: 35–40 (1994).
37. Y. Pang, M. Samoc, and P.N. Prasad, *J. Chem. Phys.* **94**: 5282–5290 (1991).
38. C.J. Wung, Y. Pang, P.N. Prasad, and F.E. Karasz, *Polymer* **32**: 605–608 (1991).
39. C.J. Wung, W.M. Wijekoon, and P.N. Prasad, *Polymer* **34**: 11174–11178 (1993).
40. L. Bozano, S.E. Tuttle, S.A. Carter, and P.J. Brock, *Appl. Phys. Lett.* **73**: 3911–3913 (1998).
41. S.A. Carter, J.C. Scott, and P.J. Brock, *Appl. Phys. Lett.* **71**: 1145–1147 (1997).
42. J.S. Salafsky, W.H. Lubberhuizen, and R.E.I. Schropp, *Chem. Phys. Lett.* **290**: 297–303 (1998).
43. T.A.E. Loss, C.W. Rogers, and M.O. Wolf, *Can. J. Chem.* **76**: 1554–1558 (1998).
44. R.C. Smith, W.M. Fischer, and D.L. Gin, *J. Am. Chem. Soc.* **119**: 4092–4093 (1997).
45. J. Obrzut and F.E. Karasz, *J. Chem. Phys.* **87**: 2349–2358 (1987).
46. R.H. Friend, D.D.C. Bradley, and P.D. Townsend, *J. Phys. D Appl. Phys.* **13**: 1367–1384 (1987).
47. M. Meier, E. Buchwald, S. Karg, P. Posch, M. Greczmiel, P. Stroehriegl, and W. Riess, *Synth. Met.* **76**: 95–99 (1996).
48. U. Rauscher, H. Bäessler, D.D.C. Bradley, and M. Hennecke, *Phys. Rev. B Condens. Matter* **42**: 9830–9836 (1990).
49. M. Yan, L.J. Rothberg, E.W. Kwock, and T.M. Miller, *Phys. Rev. Lett.* **75**: 1992–1995 (1995).
50. L.J. Rothberg, M. Yan, E.W. Kwock, T.M. Miller, M.E. Galvin, S. Son, and F. Papadimitrakopoulos, *IEEE Trans. Electron Devices* **44**: 1258–1262 (1997).
51. F. Papadimitrakopoulos, K. Konstadinidis, T.M. Miller, R. Opila, E.A. Chandross, and M.E. Galvin, *Chem. Mater.* **6**: 1563–1568 (1994).
52. H.-H. Horhold and J. Opfermann, *Makromol. Chem.* **131**: 105–132 (1970).
53. S. Tokito, T. Tsutsui, R. Tanaka, and S. Saito, *Jpn. J. Appl. Phys.* **25**: L680–L681 (1986).
54. T. Takiguchi, D.H. Park, H. Ueno, K. Yoshino, and R. Sugimoto, *Synth. Met.* **17**: 657–662 (1987).
55. D. Moses, J. Wang, G. Yu, and A.J. Heeger, *Phys. Rev. Lett.* **80**: 2685–2688 (1998).
56. S. Tokito, T. Tsutsui, S. Saito, and R. Tanaka, *Polym. Commun.* **27**: 333–335 (1986).
57. M. Ahlskog, M. Reghu, T. Noguchi, and T. Ohnishi, *Synth. Met.* **89**: 11–15 (1997).
58. J.-I. Jin, C.-K. Park, and H.K. Shim, *Macromolecules* **26**: 1799–1804 (1993).
59. J.H.F. Martens, D.A. Halliday, E.A. Marseglia, D.D.C. Bradley, R.H. Friend, P.L. Burn, and A.B. Holmes, *Synth. Met.* **55**: 434–439 (1993).
60. M.A. Masse, D.C. Martin, E.L. Thomas, F.E. Karasz, and J.H. Petermann, *J. Mater. Sci.* **25**: 311–320 (1990).
61. J.H.F. Martens, E.A. Marseglia, D.D.C. Bradley, R.H. Friend, P.L. Burn, and A.B. Holmes, *Synth. Met.* **55**: 449–453 (1993).
62. C.Y. Yang, F. Hide, M.A. Diaz-Garcia, A.J. Heeger, and Y. Cao, *Polymer* **39**: 2299–2304 (1998).
63. B. Winkler, S. Tasch, E. Zojer, M. Ungerank, G. Leising, and F. Stelzer, *Synth. Met.* **83**: 177–180 (1996).
64. W. Memeger, *Macromolecules* **22**: 1577–1588 (1989).
65. M. Suzuki, J.-C. Lim, and T. Segusa, *Macromolecules* **23**: 1574–1579 (1990).
66. A. Samoc, M. Samoc, M. Woodruff, and B. Luther-Davies, *Plast. Eng.* **49**: 373–436 (1998).
67. T. Kaino, K. Kubodera, S. Tomaru, T. Kurihara, S. Saito, T. Tsutsui, and S. Tokito, *Electron. Lett.* **23**: 1095–1097 (1987).
68. T. Kaino, H. Kobayashi, K. Kubodera, T. Kurihara, S. Saito, T. Tsutsui, and S. Tokito, *Appl. Phys. Lett.* **54**: 1619–1621 (1989).
69. C.-B. Yoon and H.-K. Shim, *J. Mater. Chem.* **8**: 913–917 (1998).
70. N.S. Sacriciftci, D. Braun, C. Zhang, V.I. Srdanov, A.J. Heeger, G. Stucky, and F. Wudl, *Appl. Phys. Lett.* **62**: 585–587 (1993).
71. J.J.M. Halls and R.H. Friend, *Synth. Met.* **85**: 1307–1308 (1997).
72. J.J. Dittmer, K. Petritsch, E.A. Marseglia, R.H. Friend, H. Rost, and A.B. Holmes, *Synth. Met.* **102**: 879–880 (1999).
73. J. Gao, G. Yu, and A.J. Heeger, *Adv. Mater.* **10**: 692–695 (1998).
74. K. Yoshino, T. Kuwabara, T. Iwasa, T. Kawai, and M. Onoda, *Jpn. J. Appl. Phys.* **29**: L1514–L1516 (1990).
75. C.W. Tang and S.A. VanSlyke, *Appl. Phys. Lett.* **51**: 913–915 (1987).
76. C.W. Tang, S.A. VanSlyke, and C.H. Chen, *J. Appl. Phys.* **65**: 3610–3616 (1989).
77. A.R. Brown, D.D.C. Bradley, J.H. Burroughes, R.H. Friend, N.C. Greenham, P.L. Burn, A.B. Holmes, and A. Kraft, *Appl. Phys. Lett.* **61**: 2793–2795 (1992).
78. M. Meier, E. Buchwald, S. Karg, P. Posch, M. Greczmiel, P. Stroehriegl, and W. Riess, *Synth. Met.* **76**: 95–99 (1996).
79. K. Naito and A. Miura, *J. Phys. Chem.* **97**: 6240–6248 (1993).
80. D.R. O'Brien, P.E. Burrows, S.R. Forrest, B.E. Koene, D.E. Loy, and M.E. Thompson, *Adv. Mater.* **10**: 1108–1112 (1998).

81. M. Thelakkat and H.-W. Schmidt, *Adv. Mater.* **10**: 219–223 (1998).
82. N.-X. Hu, S. Xie, Z. Popavic, B. Ong, A.-M. Hor, and S. Wang, *J. Am. Chem. Soc.* **121**: 5097–5098 (1999).
83. G. Yu, *Synth. Met.* **80**: 143–150 (1996).
84. N. Tessler, N.T. Harrison, and R.H. Friend, *Adv. Mater.* **10**: 64–68 (1998).
85. S.A. Carter, M. Angelopoulos, S. Karg, P. J. Brock, and J.C. Scott, *Appl. Phys. Lett.* **70**: 2067–2069 (1997).
86. A.C. Arias, M. Granström, D.S. Thomas, K. Petritsch, and R.H. Friend, *Phys. Rev. B* **60**: 1854–1860 (1999).
87. G. Gustafsson, Y. Cao, G.M. Treacy, F. Klavetter, N. Colaneri, and A.J. Heeger, *Nature* (London) **357**: 477–479 (1992).
88. I.D. Parker and H.H. Kim, *Appl. Phys. Lett.* **64**: 1774–1776 (1994).
89. H. Becker, S.E. Burns, and R.H. Friend, *Phys. Rev. B Condens. Matter* **56**: 1893–1905 (1997).
90. E. Peeters, M.P.T. Christiaans, R.A.J. Janssen, H.F.M. Schoo, H.P.J.M. Dekkers, and E.W. Meijer, *J. Am. Chem. Soc.* **119**: 9909–9910 (1997).
91. S. Son, A. Dodabalapu, A.J. Lovinger, and M.E. Galvin, *Science* **269**: 376–378 (1995).
92. B.H. Cumpston, I.D. Parker, and K.F. Jensen, *J. Appl. Phys.* **81**: 3716–3720 (1997).
93. J.C. Scott, J.H. Kaufman, P.J. Brock, R. DiPietro, J. Salem, and J.A. Goitia, *J. Appl. Phys.* **79**: 2745–2751 (1996).
94. K.M. Vaeth and K.F. Jensen, *Appl. Phys. Lett.* **71**: 2091–2093 (1997).
95. J.C. Carter, I. Grizzi, S.K. Heeks, D.J. Lacey, S.G. Latham, P.G. May, O. Ruiz de los Panos, K. Pichler, C.R. Towns, and H.F. Wittmann, *Appl. Phys. Lett.* **71**: 34–36 (1997).
96. J.M. Bharathan and Y. Yang, *Appl. Phys. Lett.* **72**: 2660–2662 (1998).
97. J.A. Rogers, Z. Bao, and L. Dhar, *Appl. Phys. Lett.* **73**: 294–296 (1998).
98. P.L. Burn, A.B. Holmes, A. Kraft, D.D.C. Bradley, A.R. Brown, and R.H. Friend, *J. Chem. Soc. Chem. Commun.* 32–34 (1992).
99. I. Sokolik, Z. Yang, F.E. Karasz, and D.C. Morton, *J. Appl. Phys.* **74**: 3584–3586 (1993).
100. F. Garten, A. Hilberer, F. Cacialli, E. Esselink, Y. Van Dam, B. Schlattmann, R.H. Friend, T.M. Klapwijk, and G. Hadziioannou, *Adv. Mater.* **9**: 127–131 (1997).
101. S.T. Kim, D.H. Hwang, X.C. Li, J. Gruener, R.H. Friend, A.B. Holmes, and H.K. Shim, *Adv. Mater.* **8**: 979–982 (1996).
102. D.R. Baigent, P.J. Hamer, R.H. Friend, S.C. Moratti, and A.B. Holmes, *Synth. Met.* **71**: 2175–2176 (1995).
103. S.J. Chung, J.I. Jin, and K.K. Kim, *Adv. Mater.* **9**: 551–554 (1997).
104. M. Gao, B. Richter, and S. Kirstein, *Adv. Mater.* **9**: 802–805 (1997).
105. M. Uchida, Y. Ohmori, T. Noguchi, T. Ohnishi, and K. Yoshino, *Jpn. J. Appl. Phys.* **32**: L921–L924 (1993).
106. M. Hamaguchi and K. Yoshino, *Appl. Phys. Lett.* **69**: 143–145 (1996).
107. O. Onitsuka, A.C. Fou, M. Ferreira, B.R. Hsieh, and M.F. Rubner, *J. Appl. Phys.* **80**: 4067–4071 (1996).
108. Q. Pei, G. Yu, C. Zhang, Y. Yang, and A.J. Heeger, *Science* 1086–1088 (1995).
109. G. Yu, Y. Cao, M. Andersson, J. Gao, and A.J. Heeger, *Adv. Mater.* **10**: 385–388 (1998).
110. N. Tessler, G.J. Denton, and R.H. Friend, *Nature* **382**: 695–697 (1996).
111. F. Hide, B.J. Schwatz, M.A. Diaz-Garcia, and A.J. Heeger, *Chem. Phys. Lett.* **256**: 424–430 (1996).
112. H.-J. Brouwer, V.V. Krasnikov, T.-A. Pham, R.E. Gill, and G. Hadziioannou, *Appl. Phys. Lett.* **73**: 708–710 (1998).

POLY(VINYLLIDENE FLUORIDE) (PVDF) AND ITS COPOLYMERS

Q.M. ZHANG
VIVEK BHARTI
GEORGE KAVARNOS
Pennsylvania State University
University Park, PA

INTRODUCTION

Poly(vinylidene fluoride) (PVDF) and the family of P(VDF-TrFE) [TrFE = trifluoroethylene] copolymers are arguably the best-known examples of a class of high-performance polymers noted for their remarkable piezoelectric and ferroelectric properties (1–4). In 1969, Kawii discovered the exceptional piezoelectric behavior of PVDF, which at that time, was the highest among the known synthetic polymers (5). After more than 30 years of study and development, the piezoelectricity and electromechanical properties of PVDF and its copolymers have been improved markedly. Today this class of polymer still possesses the highest electromechanical responses over a broad temperature range among known synthetic organic materials. Further, when considered along with their easy conformability, flexibility, robustness, and lightness, it is not surprising that electroactive polymers continue to be the focus of interest of the designers of high-performance electromechanical devices (4). When PVDF is stretched and poled in a strong electric field, it exhibits piezoelectricity (3). In its piezoelectric form, PVDF finds use in transducer devices requiring the interconversion of mechanical and electrical energy. Piezoelectric PVDF can be fabricated and used in a variety of sensors and actuators such as artificial muscles and organs, medical imaging, blood-flow monitors, microphones, smart skins, underwater acoustic transducers, seismic monitors, fluid pumps and valves, surface acoustic wave devices, robots, and tactile sensing devices (4–11). P(VDF-TrFE) copolymers display similar and in some cases even superior properties (4,12).

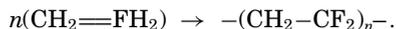
Even before the discovery of the high piezoelectricity in PVDF, the existence of its ferroelectric nature was postulated by Lando et al. based on the crystal unit cell structure. This hypothesis was confirmed about 10 years later (13–17). Many organic substances in fact exhibit a key ferroelectric property that is called polarization hysteresis (18–20). The copolymer of PVDF with TrFE and tetrafluoroethylene (TFE), however, is the only polymeric system that shows both a well-defined polarization hysteresis loop

and a transition to a paraelectric phase with increased temperature.

SYNTHETIC PATHWAYS AND MOLECULAR AND CRYSTAL STRUCTURES

Synthetic Pathways and Morphology

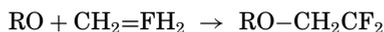
PVDF is synthesized by the successive addition of a vinylidene fluoride monomer units (21).



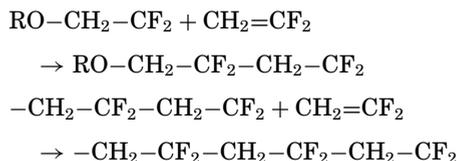
The reaction can be initiated by a free radical that might be formed by the thermal decomposition of, say, benzoyl peroxide:



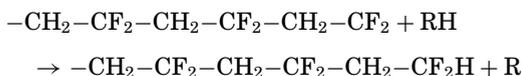
The generated free radical then reacts with a vinylidene fluoride molecule in an initiation step:



Thus begins the process of growing a polymer chain, a series of steps called chain propagation:



In this sequence, the CF_2 “head” of the VDF monomer is shown to add to the CH_2 “tail” in a “head-to-tail” addition. Propagation may also, however, involve head-to-head (CF_2 to CF_2), tail-to-tail (CH_2 to CH_2), and tail-to-head (CH_2 to CF_2) addition. Chain termination of the growing polymer chains may occur by the combination of two growing chains or by hydrogen atom abstraction from a good hydrogen donor such as isopropyl alcohol (RH):



The synthesis of PVDF is normally carried out by emulsion polymerization where a dispersing medium such as water dissipates the high heat of the polymerization reaction (21).

The copolymer P(VDF-TrFE) is synthesized in a similar manner by the copolymerization of trifluoroethylene and vinylidene (22). Here the growing chain can add to either monomer, for example

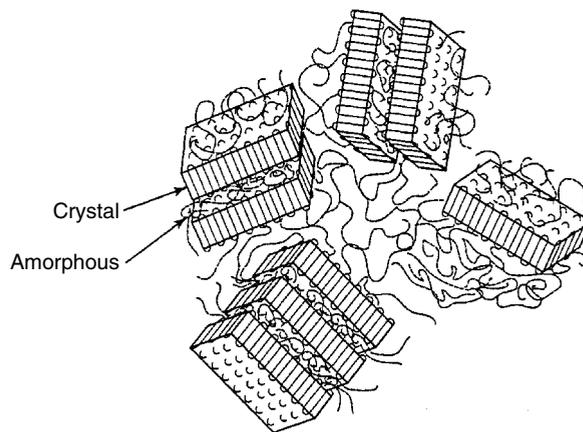
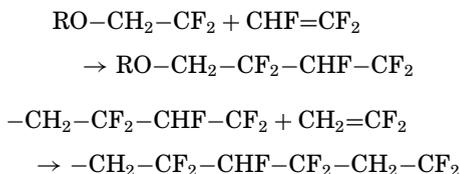


Figure 1. Polymers such as PVDF consist of crystal and amorphous regions. This figure depicts the crystalline regions as stacks of folded lamellae.

The mole ratios of VDF in P(VDF-TrFE) copolymers that display the most striking ferroelectric properties range from about 50 to over 90 mol%. On the basis of these ratios, TrFE is much more likely to be adjacent to VDF than itself in a polymer chain. The relative proportions of the monomers in the copolymers prescribe the ferroelectric properties of the copolymers, as we shall show in later sections. Furthermore, although head-to-tail bonding prevails in the polymer chains, there may be some tail-to-tail bonding ($-\text{CH}_2-\text{CF}_2-\text{CF}_2-\text{CH}_2-$) as well head-to-head ($-\text{CF}_2-\text{CH}_2-\text{CH}_2-\text{CF}_2-$). The percentage of these defects generally is only a few mole% for most PVDF compositions. It is most important to note that the monomer sequences as well as the number of head-to-head or tail-to-tail linkages give rise to defects, which ultimately may influence the material properties of PVDF and P(VDF-TrFE).

PVDF and P(VDF-TrFE) are semicrystalline polymers (24). They are comprised of ordered regions of monomer units (crystallites) surrounded by an amorphous sea of scrambled, spaghettilike chains, as shown in Fig. 1 (1,2). From many SEM studies, it has been estimated that these crystallites have thickness of about 10 to 20 nm along the polymer chain direction and extend to several microns in the other directions (23). The degree of crystallinity as well as the orientation of the crystallites can be controlled by various processing techniques. Before discussing these techniques, we present a survey of the major crystalline forms of PVDF and P(VDF-TrFE).

Molecular Conformations

PVDF and P(VDF-TrFE) are polymorphic in the sense that they may exist in several crystal forms (24). In each form, the chains are packed within crystal lattices in specific conformations. The crystal structures are described by the conformations of the chains (as a series of *trans* (T) or *gauche* (G) linkages), by the orientation of these chain sequences about the chain axis (parallel or antiparallel), and by the relative directions of adjacent chains as up-up (same direction) or down-down (opposite direction). To visualize conformations along a PVDF chain (Fig. 2), a single

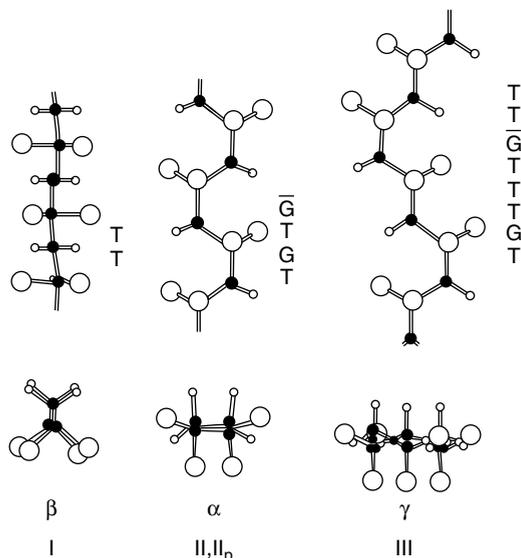


Figure 2. Extended chain segments of an all-*trans*, TGT \bar{G} , and TTTGTTT \bar{G} conformations.

carbon-carbon bond within one monomer unit can be projected perpendicular to the plane of the page (Fig. 3). One carbon atom (of an adjacent monomer) and two hydrogen or fluorine atoms are then connected to the front carbon atom by bonds shown as solid lines; likewise, three atoms are connected to the back carbon atoms by bonds shown as dotted lines. If the front carbon atom with its three atoms is rotated together while keeping the back atoms stationary, the steric energy of the structure changes because of repulsions between the atoms on the front and back carbons. The angle of rotation is $\varphi = 0$ when the four carbon atoms in figure all lie in the same plane. In this structure, the substituted atoms are eclipsed and engaged in strong mutual repulsion. As φ is varied to 60° , the repulsions decrease as the atoms on the front and back move away from one

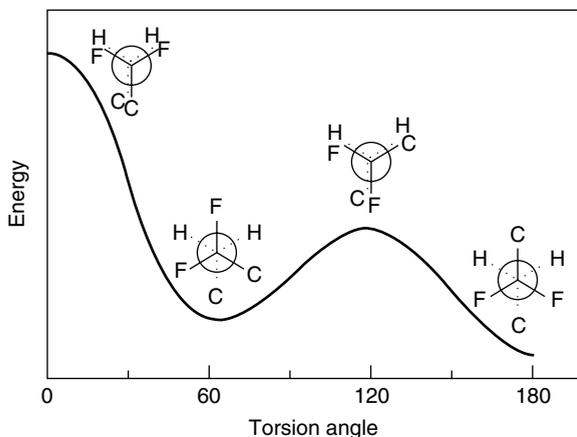


Figure 3. Interactions between fluorine and hydrogen atoms on adjacent carbon atoms in PVDF polymer chains lead to changes in the potential energy as the bond connecting the carbon atoms is rotated.

another. The energy increases again and then decreases as φ reaches 120° . Finally, going to 180° , the energy reaches a minimum value. The structures at $\varphi = 60^\circ$ and 240° are referred to as *gauche* G and \bar{G} conformations. At $\varphi = 180^\circ$, the structure is in a *trans* conformation. The energy of the *trans* conformation is the lowest of all the structures as the substituent atoms are farthest apart. The energies of the two *gauche* conformations are equal and higher than the *trans* structure. When PVDF or P(VDF-TrFE) polymers crystallize during annealing, the chains pack as ordered arrays within distinct crystal structures. The structures and lattice dimensions of these crystal structures are determined by the existence of *trans* and *gauche* conformations, whose energies are governed by the number of fluorine atoms substituted on adjacent carbon atoms in the chain.

Crystal Structures

There are four major crystalline forms of PVDF (24). In form I, which is also known as the β phase, two chains in an all-*trans* planar zigzag conformation are packed into individual orthorhombic unit cells having lattice dimensions of $a = 8.58 \text{ \AA}$, $b = 4.90 \text{ \AA}$, and the chain direction or fiber axis $c = 2.56 \text{ \AA}$ (Fig. 4) (25). The space group symmetry of each unit cell in $Cm2m$. It has been suggested that a more accurate model has CF_2 groups deflected by about 7° in opposite directions in a planar zigzag conformation with a fiber axis of $2c$ or 5.12 \AA (25). It will be noted from Fig. 4 that

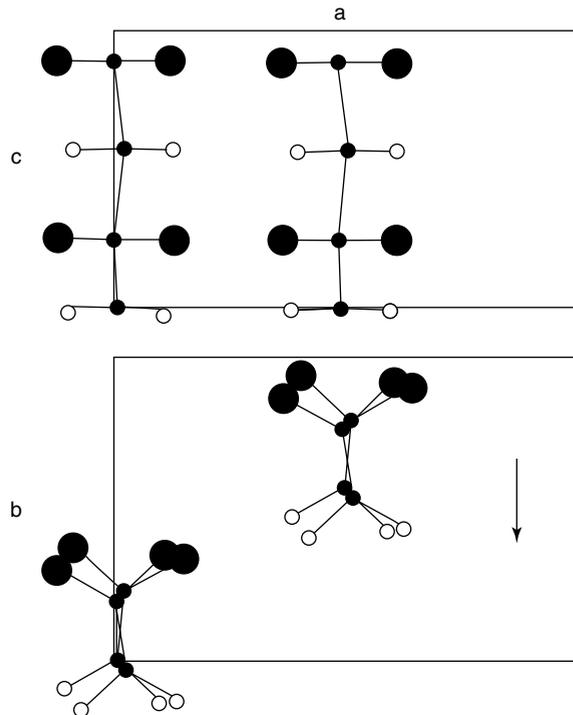


Figure 4. Projection of form I (β -phase) of PVDF. The conformation of the packed chains is all-*trans*. The chains are so aligned that the carbon-fluorine dipoles are in the same directions along the b -axis. The small and large filled circles represent carbon and fluorine atoms, respectively; the small open circles represent hydrogen atoms.

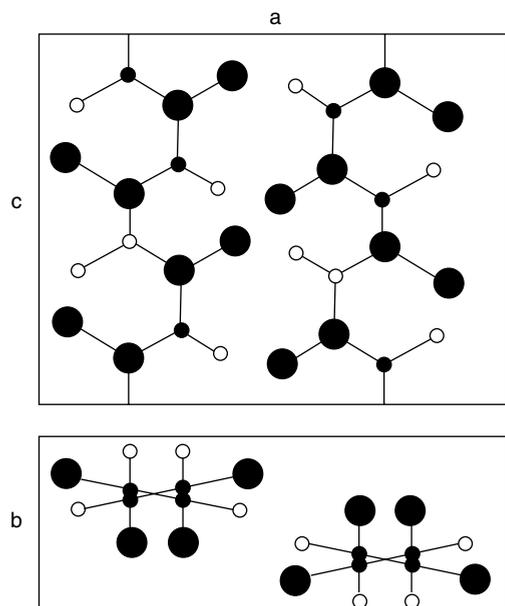


Figure 5. Projection of form II (α -phase) of PVDF. The conformation of TGT \bar{G} . The chains are packed in an antiparallel in two directions about the chain axis. The small and large filled circles represent carbon and fluorine atoms, respectively; the small open circles represent hydrogen atoms.

in the all-*trans* conformation, the fluorine atoms are positioned on one-side of the unit cell resulting in a net dipole moment. The form I unit cell is quite polar having a net dipole of 2.1 debye. As the structure of the unit cell of the form I crystal satisfies the symmetry requirement of a piezoelectric crystal, meaning that the crystal belongs to a noncentrosymmetric class, this is the form of PVDF that is responsible for its piezoelectric properties.

In form II, or the α -phase, the chain conformations are represented as a sequence of alternating *trans* and *gauche* sequences, or TGT \bar{G} (Fig. 5) (26). Each unit cell containing two chains is orthorhombic with lattice parameters $a = 4.96 \text{ \AA}$, $b = 9.64 \text{ \AA}$, and $c = 4.62 \text{ \AA}$. In the α -phase, adjacent chains are packed such that the dipole moments of the individual carbon-fluorine bonds are aligned perpendicular to the chain direction, canceling one another out. The directions of the chains consist of a statistical average of up-up and up-down orientations. A form IV (δ -phase) has been identified where the chains have the same conformations as in the α form but the carbon-fluorine bonds are aligned in one direction around the chain direction

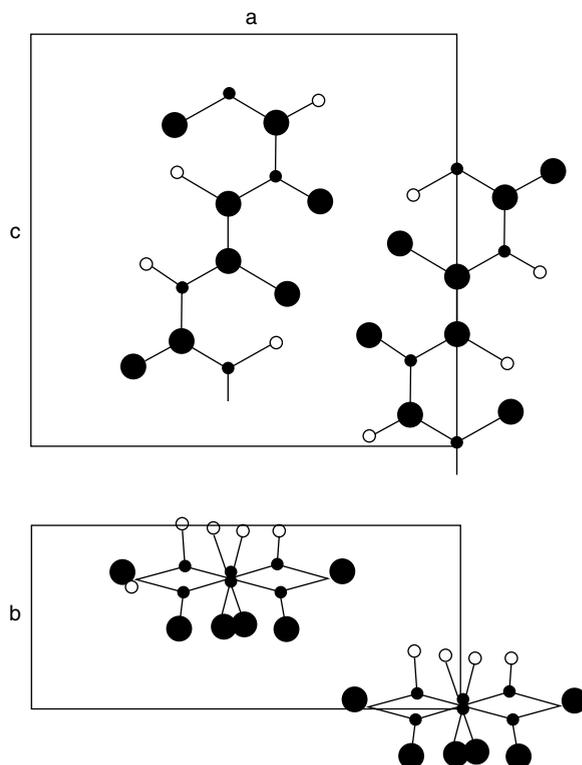


Figure 6. A projection of form III (γ -phase) of PVDF. The chain conformation is TTTGTTT \bar{G} . The small and large filled circles represent carbon and fluorine atoms, respectively; the small open circles represent hydrogen atoms.

resulting in a net dipole (27), and the crystal lattice parameters are identical to form II. In form III (γ -phase) crystals, the chain conformations are TTTGTTT \bar{G} , and the crystal lattice is monoclinic with lattice parameters $a = 4.96 \text{ \AA}$, $b = 9.67 \text{ \AA}$, $c = 9.20 \text{ \AA}$, and $\beta = 93^\circ$ (28,29). The alignment of the form III chains perpendicular to the chain axis is in one direction, resulting in a polar cell (Fig. 6).

In the P(VDF-TrFE) copolymers, the form I and form II crystal lattices are expanded structures in the directions perpendicular to the molecular chain (Table 1). The expansion of the lattices accommodates the presence of a greater number of substituted fluorine atoms, since fluorine has a van der Waals (vdW) radius of 1.35 \AA , which is to be compared with the vdW radius of hydrogen of 1.2 \AA (31). The interchain spacing from (110,200) reflection in the form I crystal as a function of the VDF content for P(VDF-TrFE) copolymer is presented in Fig. 7 (2,32,33). In the form I

Table 1. Experimental Lattice Dimensions and Angles of Form I (β) of PVDF and P(VDF-TrFE) 75/25 and 50/50 mol% Copolymers

VDF Content (mol%)	Lattice Dimensions (\AA)			Lattice Angles (deg)		
	a	b	c	α	β	γ
100 ^a	8.58	4.91	2.56	90	90	90
75 ^b	8.86	4.62	2.56	90	90	90
50 ^b	9.12	5.25	2.55	90.3	—	—

^aRef. (25).

^bRef. (30).

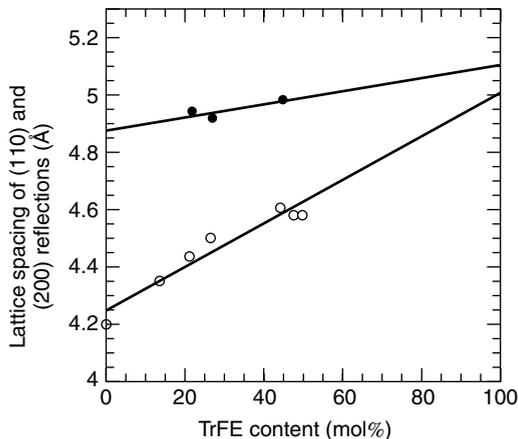


Figure 7. Variation of interchain lattice spacing with TrFE content in P(VDF-TrFE) copolymer films. Open circles are from the ferroelectric β -phase and black dots are from the paraelectric phase.

structure of copolymers, it is to be noted that the net dipole of the polar cell is reduced by the opposing orientation of the “third” carbon–fluorine dipole on the TrFE monomer unit.

There have been several studies that have attempted to use molecular modeling to simulate the crystal structures as well as mechanical properties and polarizations of PVDF and P(VDF-TrFE) crystals (34–39). The primary focus of these studies have been on developing algorithms to predict the structures and properties of PVDF and P(VDF-TrFE). For the most part, these simulations have utilized molecular mechanics force fields to calculate the crystal lattice dimensions, polarizations, and compliances, and have yielded values in close agreement with results obtained from X-ray crystallographic studies. For several model structures of P(VDF-TrFE) crystals, the increase in lattice dimensions with an increase in fluorine contents has been correctly predicted (37). For these same structures, the trends in the compliances and piezoelectric constants were calculated as a function of the molecular percentage of VDF and found to follow experimental measurements (38).

PROCESSING AND FABRICATION

There are two distinctly different ordering processes for PVDF and P(VDF-TrFE) polymers: crystalline ordering and dipolar ordering. Both orderings can be influenced to a large extent by sample processing conditions. Furthermore, one can also control the dipolar ordering by introducing defect structures in the crystalline phase. As a consequence, the ferroelectric and electromechanical responses will depend substantially on the sample treatment conditions, such as annealing, quenching, mechanical drawing, irradiation, and doping (24,39). It should be pointed out that because of the slow kinetics of the various polymer transitional processes and the high-energy barrier of the transformation between the different crystalline forms, metastable phases can be formed and be present for long periods of time even though these phases are not the most

thermodynamically favored. This phenomenon leads to the formation of different crystalline phases by making use of different sample processing conditions (24).

The eventual choice of the processing steps has much to do with the design, geometry, and size of the device that is required. In preparing PVDF or P(VDF-TrFE) films or thicker dimensions for certain applications, the synthesized polymer in the form of pellets or powder is heated to produce a melt or is dissolved in an organic solvent (40). For both PVDF and P(VDF-TrFE), these processing procedures will dictate the form that will predominate, which in turn is influenced by the energies of the crystalline forms as well as the energy barriers involved in producing these forms (41). The crystal phase that predominates during processing is also controlled by the number of fluorine atoms in the polymer chains. The increasing number of fluorine atoms enhances unfavorable van der Waals repulsive forces within the chains as well as between adjacent chains. These interactions force the energies of packed α -phase chains with TGT \bar{G} bonding, where repulsions between closely positioned fluorine atoms are enhanced, to increase significantly relative to the β -forms.

This picture helps to explain why melts of P(VDF-TrFE) crystallize directly into the electroactive polar β -phase. Not only are the crystal energies of β -phase P(VDF-TrFE) lowered, but the energy barriers to these states are reduced. In contrast, when PVDF is annealed and slowly cooled from the melt or cast from solutions of an organic solvent such as methyl ethyl ketone (MEK), formation of the α -phase rather than the β -phase is favored. Although the crystal lattice energy of the β -phase unit cell is slightly lower than that of the α -phase cell, molecular dynamical calculations have suggested that formation of the α -phase PVDF is kinetically favored; that is, the barriers for formation of the all-*trans* phase prevent crystallization directly into this phase from the melt. When films of PVDF α -phase are mechanically stretched or drawn, the TGT \bar{G} chains convert into the polar all-*trans* β -phase. This happens to be the basis for the drawing procedure followed in most commercial processes for the preparation of piezoelectric PVDF films. Quenching of melted PVDF under high pressure or casting PVDF from hexamethylphosphoric triamide solutions will also result in the direct formation of polar β -phase PVDF. Form III (γ -phase) PVDF can be produced by casting from dimethylsulfoxide. Polar form II (δ -phase) can be produced by poling α -phase PVDF in strong electric fields.

In the form I β -phase crystallites that predominate after PVDF is stretched or when the P(VDF-TrFE) copolymer is cooled from the melt, there are ferroelectric domains in crystallites that are polar but are nonetheless orientated in all crystallographically allowed directions. Furthermore, these crystallites are randomly oriented within the film. This accounts for the absence of any piezoelectric activity unless the films are poled. To be made piezoelectric, the domains must be oriented in a strong electric field called the “poling field.” Poling can be accomplished by electroding the polymer surfaces with a metal, followed by application of a strong electric field to orient the crystallites. An alternative method of poling is the use of a corona discharge where a corona charge is injected into the polymer from a needle electrode placed a centimeter or two from the

polymer film. In corona poling, no electroding is required as in the case of direct field poling.

Thus, the choice of PVDF or P(VDF-TrFE) for a specific application will in many cases be determined by the size and complexity of the desired device. P(VDF-TrFE) is normally the choice material for thick and unusual shapes in that no stretching is required to convert the polymer into its ferroelectric form prior to the poling step. Casting of polymers dissolved in organic solvents onto substrates can produce films of P(VDF-TrFE) that can be subsequently annealed. P(VDF-TrFE) can also be molded or extruded to produce larger dimensions or unusual shapes. Unlike P(VDF-TrFE) copolymers, which do not have to be mechanically drawn to induce ferroelectric β -phase, PVDF films have to be oriented by stretching in one direction (uniaxial orientation) or in two directions (biaxial orientation). Multilayering of thin polymer films is another technique for fabricating thicker elements. Multilayered and electroded polymer films that are subsequently wired in parallel have been proposed to reduce the electrical requirements for devices such as sonar drivers (7).

POLARIZATION RESPONSES AND PHASE TRANSITIONS

Phase Diagrams: An Overview

With knowledge of the major crystalline phases of PVDF and its copolymers, it is useful to review its phase diagram showing the interconversions between the various crystalline forms where the relative proportions of the monomer units are varied. The phase diagram of PVDF and P(VDF-TrFE) polymers shows a ferroelectric-paraelectric transition that signals a change from a ferroelectric (polar) phase to a paraelectric (nonpolar) phase (Fig. 8). The ferroelectric-paraelectric (F-P) transition temperature increases with vinylidene fluoride mole fraction content. Below the F-P transition, the crystal is best represented as an ordered form I (β) structure with long sequences of all-*trans* bonds. As the temperature of the crystals rises and goes through the F-P transition, an increasing number of *gauche* bonds is introduced into the ordered all-*trans* structure. As a result, the crystal lattice dimensions enlarge, as we have seen above, and the crystal regions tend toward disorder, leading to the formation of the *paraelectric* phase containing a random mixing of TG, $T\bar{G}$, TTTG, and TTT \bar{G} . Eventually, at higher temperatures the paraelectric phase passes through the melt transition.

One should note from Fig. 8 that PVDF and P(VDF-TrFE) copolymers with high VDF concentrations do not appear to possess distinct F-P transitions; rather, it appears that melting takes place before a F-P transition. However, it must be mentioned that even in the ferroelectric phase, conformational defects can be introduced as the temperature of the polymers is raised. These defects are introduced so subtly that they may not be apparent in thermal studies such as differential scanning calorimetry (DSC). Moreover, in P(VDF-TrFE), in addition to a low-temperature (LT) phase, where the chain conformation is predominantly all-*trans*, a cooled (CL) phase has been identified (38). Structural analysis indicated that the most probable structure of the so-called CL phase is a mixture of two disordered

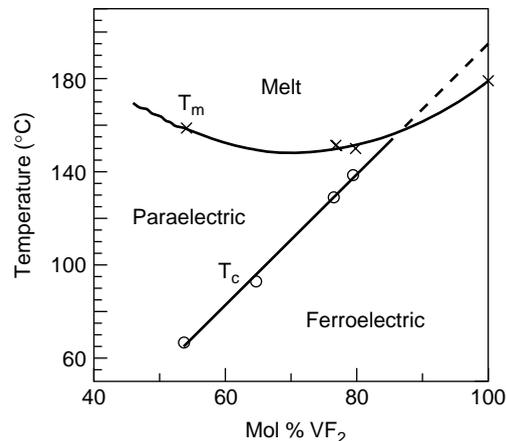


Figure 8. Phase diagram for VDF/TrFE copolymer (x: crystal melting point, o: ferroelectric to paraelectric transition point).

crystalline phases, one *trans* planar and the other 3/1 helical (frozen-in high-temperature phase) (42–45). Because of this “frozen-in” disorder, copolymers with VDF contents below 50 mol% lose ferroelectricity. Accordingly, there is no clear phase transition signal as shown in the phase diagram.

Polarization Responses and Phase Transitions

In this section, we examine the ferroelectric nature of PVDF and P(VDF-TrFE) in greater detail. As ferroelectric materials, PVDF and its copolymers with TrFE and TFE exhibit well-defined polarization hysteresis loops. Figure 9(a) shows the polarization loop measured on a P(VDF-TrFE) 68/32 mol% copolymer stretched film for which, at low cyclic electric fields (<55 MV/m), the polarization loop is nearly linear with very little hysteresis. At field amplitudes exceeding 55 MV/m, a well-defined polarization loop appears. The coercive field E_c (the field when $P = 0$) and remanent polarization P_r (the polarization when $E = 0$) do not change appreciably with the applied field amplitude. In contrast, unstretched 50/50 mol% copolymer films display very little hysteresis at fields below 25 MV/m. As the field amplitude increases beyond that, the polarization loop gradually expands where both E_c and P_r increase with the applied cyclic field amplitude, as shown in Fig. 9(b). Eventually, when the field amplitude exceeds 100 MV/m, E_c and P_r , which are now saturated, define the maximum coercive field and remanent polarization. Moreover, even at fields far below E_c measured from the saturated loop, polarization often is switchable (2). Such behavior is related to the nucleation process in the polarization switching, which will be discussed below (46,47). Because of the semicrystalline nature of the polymer, the magnitudes of P_r and E_c are particularly sensitive to sample preparation conditions. For instance, Fig. 10 illustrates that both crystallinity and remanent polarization for stretched 68/32 mol% copolymer increase with the sample annealing temperature.

The crystallization process in the copolymer is also influenced by the mobility as well as the stereoirregularity of the polymer chains (48). For PVDF homopolymer, the

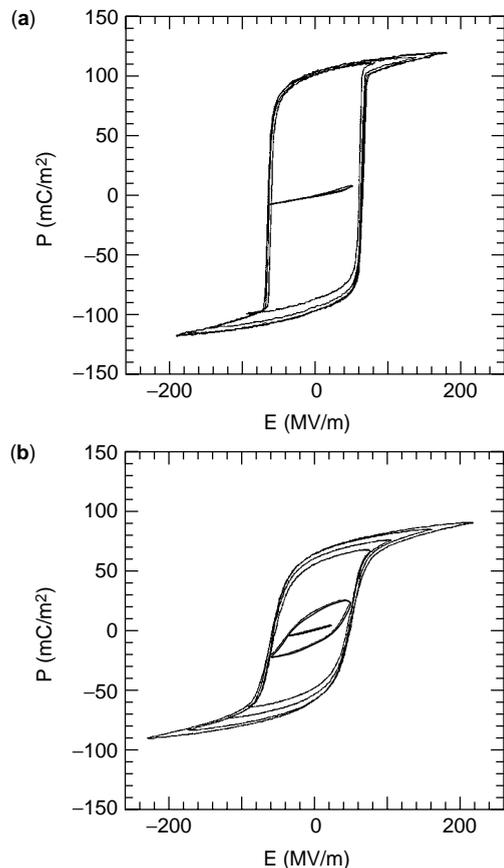


Figure 9. Polarization hysteresis loops measured at 10 Hz under different electric fields for unstretched (a) solution cast 68/32 mol%, and (b) 50/50 mol% extruded P(VDF-TrFE) copolymer films.

crystallinity is at about 50% (49). With increased TrFE content, because of the larger size of the TrFE monomer unit, the lattice spacing between the polymer chain expands and thus facilitates and enhances the crystallization process. Hence, in the copolymer of 75/25 mol%, the degree of crystallinity can reach 90%. In contrast, in polymer with higher TrFE content, the stereoirregularity introduced by

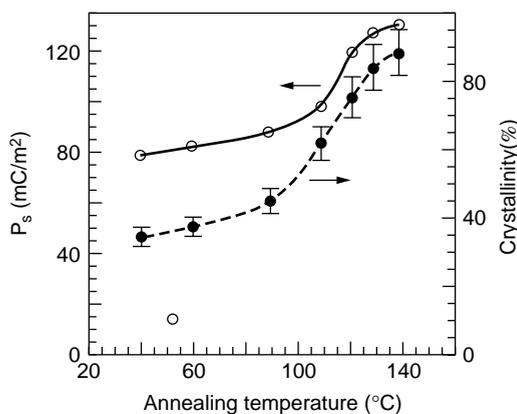


Figure 10. Effect of annealing temperature on the remanent polarization and crystallinity measured for extruded unstretched P(VDF-TrFE) 68/32 mol% copolymer films.

TrFE reduces the stability of the crystalline phase with respect to the amorphous and leads to a lowering of the crystallinity. The dependence of the crystallinity with VDF content is depicted in Fig. 11(a). Similarly, the dependence of the remanent polarization and coercive field on VDF content is illustrated in Fig. 11(b).

One of the interesting features of the data in Fig. 9(a) and (b) is that the coercive field observed here is far below that estimated from thermodynamic theory. In a recent experiment, S. Ducharme et al. (50) demonstrated that for P(VDF-TrFE) 70/30 mol% copolymer, the intrinsic coercive field at room temperature should exceed 0.5 GV/m, in substantial agreement with the thermodynamic theory that predicts a value about 10 times higher than that measured from the polarization hysteresis loop in bulk samples (50). The major reason for this difference is that polarization switching in most ferroelectric materials is through the nucleation of new domains at interfaces and surfaces. Nucleation and growth of nuclei under field are processes that typically require much lower electric fields (46,47). By taking the ultrathin film approach where the film thickness is below the critical nucleation size, S. Ducharme et al. achieved switching fields approaching values predicted from the thermodynamic theory by suppressing the nucleation process in polarization switching (50).

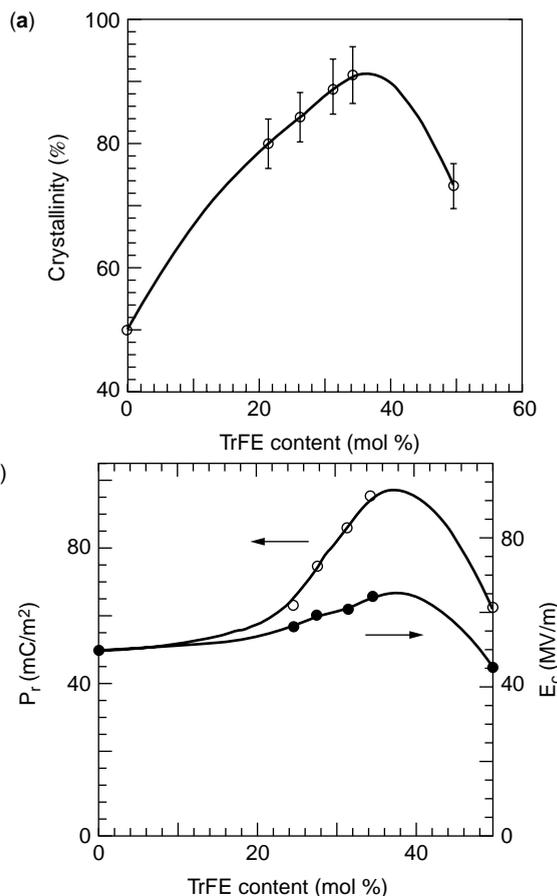


Figure 11. Change in (a) crystallinity and (b) remanent polarization P_r and coercive field E_c with TrFE content in P(VDF-TrFE) copolymer films.

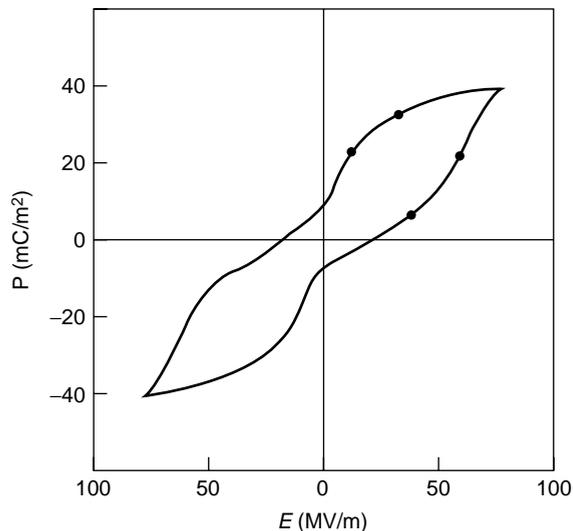


Figure 12. Polarization hysteresis loop of P(VDF-TrFE) 37/63 mol% copolymer film.

Besides the crystallinity and remanent polarization, the F–P transition temperature for a given P(VDF-TrFE) copolymer can also be varied by carefully selecting conditions for sample processing (51–53). For instance, J. Green et al. demonstrated that for P(VDF-TrFE) 60/40 mol% copolymer, the F–P transition can occur in the temperature range from 65 to 85°C (52). The sample processing methods used in this investigation included melt crystallization as well as crystallization and recrystallization from cast solutions of dimethylacetamide (DMA), dimethylformamide (DMF), cyclohexanone. The authors attributed this behavior to variations in the TrFE concentration in the crystalline phase for samples prepared under different conditions. Such variations lead to different F–P transition temperatures. But it is also quite possible that the polar ordering in the crystalline region is affected by these differing processing conditions.

Interestingly, for copolymers at compositions below about 50 mol% VDF, there is no clear transition signal with temperature, and a double polarization hysteresis loop is observed as illustrated in Fig. 12 (54,55). The latter observation seems to be quite similar to those observed in an antiferroelectric phase of a ceramic system where a ferroelectric state can be induced under a high electric field (56). It should be noted that for copolymers with VDF below 60 mol%, the cooled low (CL) temperature phase mentioned earlier was observed where the X-ray data show a double peak structure. In Fig. 13, this is near the angular position of (110,200) reflections, which are not characteristic of a ferroelectric phase (32,42,43). Structural analysis has indicated that the most probable structure of this phase is a mixture of two disordered crystalline phases, one *trans* planar and the other 3/1 helical (frozen high temperature phase) (42–45). For copolymers at VDF content higher than 50 mol%, this phase can be converted into the ferroelectric β -phase by drawing or poling (Fig. 13). However, for copolymers with VDF below 50 mol%, this CL phase will persist even after the poling field is removed, resulting in the observed double hysteresis polarization loop. It

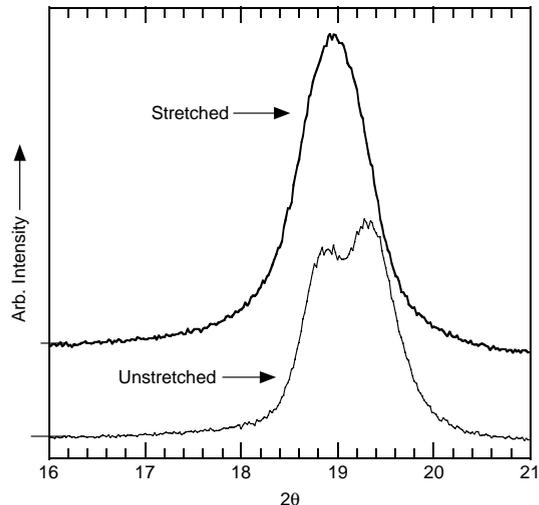


Figure 13. X-ray diffraction scans of stretched and unstretched P(VDF-TrFE) 50/50 mol% copolymer films.

should be noted that an antiferroelectric material should satisfy two criteria: (1) it must have a unit cell with anti-parallel arrangement of the dipoles, and (2) the dipoles can be switched under external fields to yield a ferroelectric phase (46).

Ferroelectricity is a collective phenomenon that involves long-range interactions in the system that include dipolar interactions in PVDF and its copolymer. One may expect that there will be a finite size effect; that is, the ferroelectricity in a small size system can be very different from that in a large bulk system. One such system, which is of great theoretical and technical interest, embodies polymer thin films. Using spin-casting techniques, Ohigash et al. prepared P(VDF-TrFE) copolymer films and showed that even at a film of 60 nm, the ferroelectric polarization switching can still be observed, but the coercive field becomes much higher than that in bulk sample (23,57). Shown in Fig. 14 is the polarization loop measured at 10 Hz for spin cast 75/25 mol% copolymer film at 65 nm thickness.

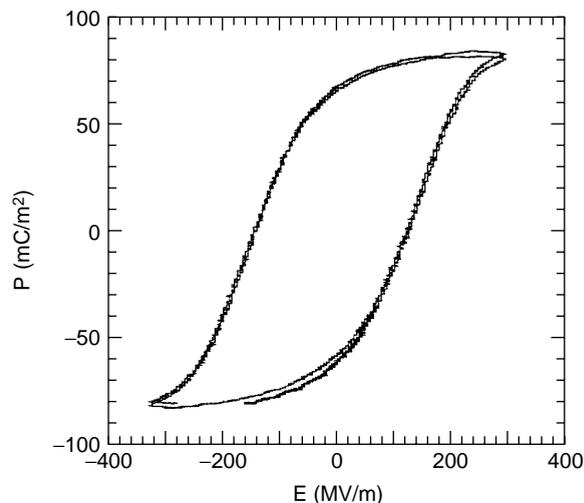


Figure 14. Polarization hysteresis loop measured at 10 Hz for 65 nm thick spin cast P(VDF-TrFE) 75/25 mol% copolymer film.

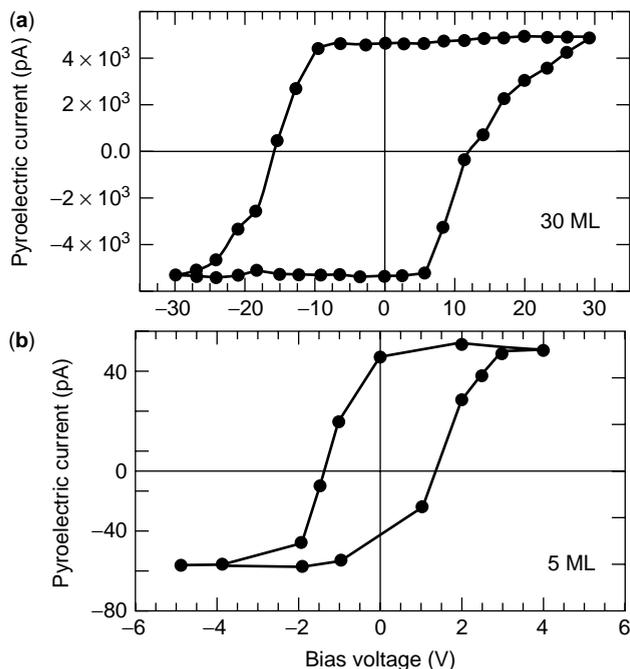


Figure 15. Polarization hysteresis loops measured at room temperature by pyroelectric technique of P(VDF-TrFE) 70/30 mol% copolymer films with different thickness: (a) 30 ML and (b) 5 ML.

More recently, ultrathin Langmuir-Blodgett (L-B) copolymer films were prepared by a Russian group (58). In this study it was shown that ferroelectricity persists even in films as thin as five molecular layers. In Fig. 15, the polarization hysteresis loops measured for these films are displayed (59). In addition, a surface ferroelectric transition was observed. The results here raise an interesting issue that concerns the role of the interchain dipolar interaction in controlling the ferroelectric behavior in this class of materials.

ELECTROMECHANICAL PROPERTIES IN NORMAL FERROELECTRIC PVDF AND ITS COPOLYMERS

In the ferroelectric phase, the copolymer is in the all-*trans* conformation, resulting in a unit cell with a large lattice constant along the polymer chain direction and a smaller unit cell dimensions perpendicular to the chain. In the paraelectric phase, however, where the crystallites adopt conformations containing a mixture of *trans* and *gauche* bonding, the unit cell dimension along the chain direction is significantly shortened while the cell dimension perpendicular to the chain expands. As a result, when the copolymer goes through the F-P transition, there is a large lattice constant change as depicted in Fig. 16 for the 65/35 mol% copolymer (31). In fact, this phenomenon has been observed for copolymers exhibiting F-P transitions. The corresponding lattice strain change through the transition perpendicular to the polymer chain is presented in Fig. 7 with VDF content. All of these results indicate that strains up to 10% can be achieved as the polymer goes through the transition.

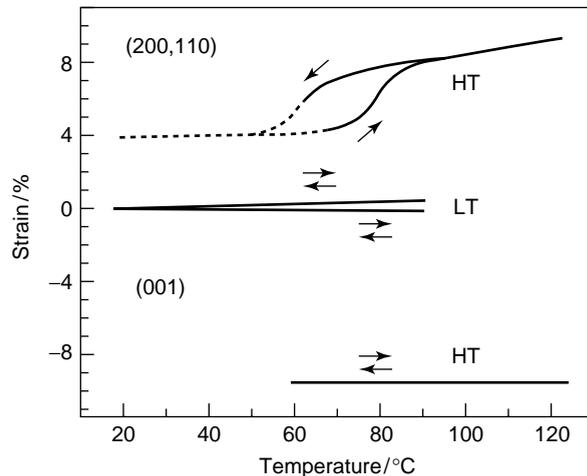


Figure 16. Lattice strain along polymer chain direction [(001)] and perpendicular to the chain [(200), (110)] obtained from the X-ray data for P(VDF-TrFE) 65/35 mol% copolymer.

The results presented indicate a strong coupling between the polarization and strain in P(VDF-TrFE) copolymer, which is the most attractive feature of this polymer. For PVDF and its copolymers, this coupling is electrostriction in nature where the strain S_i ($i = 1$ to 6) is directly proportional to the induced polarization P_j ($j = 1$ to 3) in the material (60),

$$S_i = Q_{ij}P_j^2. \quad (1)$$

For the poled P(VDF-TrFE) copolymer, the polarization is along the 3-direction (P_3). However, in the polarization switching process, it has been observed that Eq. (1) does not adequately describe the strain-polarization relationship. This is primarily due to the fact that the polarization switching is through the domain wall motion. With this in mind, it must be stated that understanding and establishing relationships between strain and polarization in the domain wall motion continue to draw intense research interest (61). One notes, for example, the polarization switching loop for 65/35 mol% copolymer and the corresponding strain changes in Fig. 17. As can be seen, there is very little strain change as the polarization switches from *A* to *B*. Early FT-IR and X-ray studies demonstrated that the switching is primarily through successive 60° domain wall motions (62–67). Owing to the pseudo-hexagonal symmetry of the unit cell in the directions perpendicular to the polymer chain, the 60° domain wall motions are not expected to generate very high strains. Therefore, to achieve a high strain response in ferroelectric materials through the polarization switching mechanism, efforts should be expended to suppress the domain wall motions that do not generate much strain response. As also shown in Fig. 17, a linear relationship of strain versus P^2 is observed in the nonswitching part of the polarization response.

Most of the electromechanical applications that use PVDF and its copolymers exploit their piezoelectric properties. For ferroelectric PVDF and its copolymers, the samples as prepared usually have negligible piezoelectric

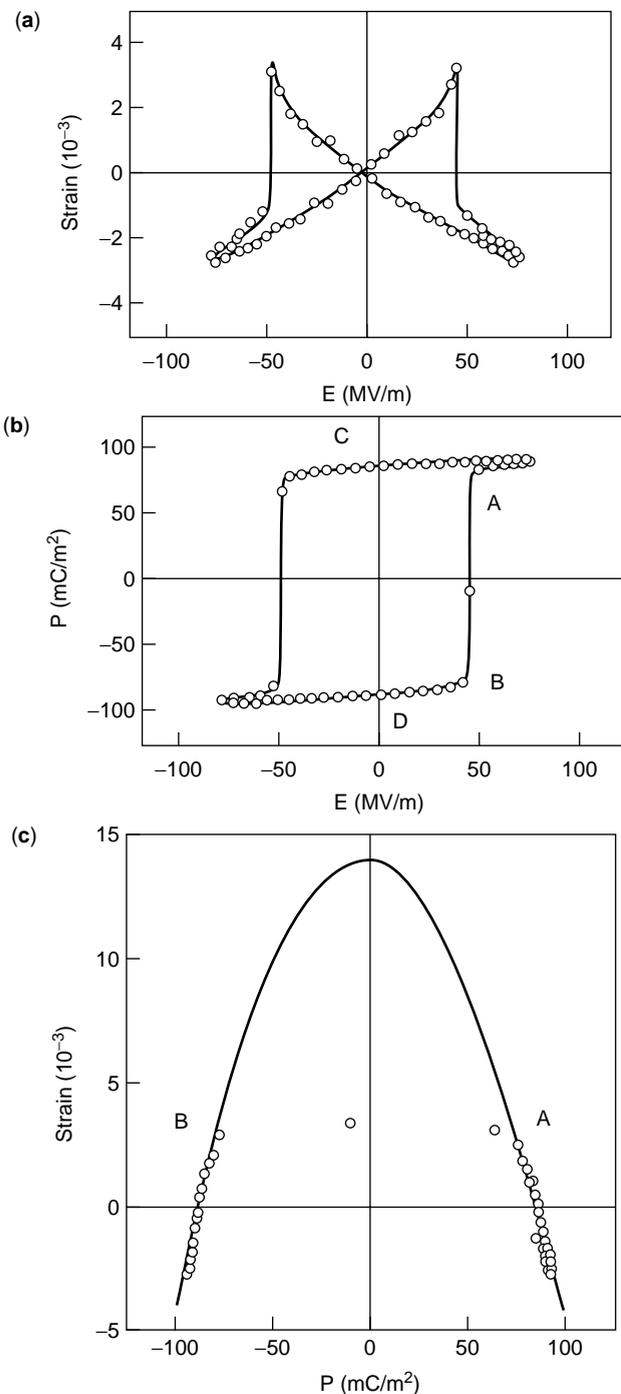


Figure 17. (a) Electric field induced strain; (b) polarization versus electric field; (c) strain versus the square of polarization of unstretched P(VDF-TrFE) 65/35 mol% copolymer films.

responses due to the fact that the polarization at each domain orientates randomly, implying that the net sample polarization is zero. In order to establish the piezoelectric state, the polymers can be poled under a field several times higher than the coercive field to induce a stable remanent polarization (at point C or D of the polarization loop in Fig. 17). For poled copolymers under a weak field (a field several times smaller than the coercive field), the strain and polarization (or field E) relationship is piezoelectric, namely a linear relationship. In addition, in the

piezoelectric state, an applied stress σ can also induce a polarization change. These responses can be written as

$$S = dE, \quad (2a)$$

$$D = d\sigma, \quad (2b)$$

where d is the piezoelectric coefficient and $D = \varepsilon_0 E + P$ is the electric displacement, ε_0 is the vacuum permittivity ($= 8.85 \times 10^{-12}$ F/m). From the Maxwell relation, it can be shown that the coefficient d in the two equations is the same. In the literature, the effect in Eq. (2a) is referred as the converse piezoelectric effect (for actuation) and in Eq. (2b) as the direct piezoelectric effect (for sensing). Adding the linear elastic (Hook's law) and dielectric relations to Eq. (2) and writing out it in the full tensor form, one obtains

$$S_{ij} = d_{kij} E_k + s_{ijkl}^E \sigma_{kl}, \quad (3a)$$

$$D_i = \varepsilon_{ik}^T E_k + d_{ikl} \sigma_{kl}, \quad (3b)$$

where s_{ijkl}^E is the elastic compliance and ε_{ik}^T is the dielectric permittivity (68,69). The superscripts E and T refer to the conditions under which these quantities are measured. That is, compliance is measured under a constant electric field and dielectric constant under a constant stress. Owing to the electromechanical coupling in a piezoelectric material, the elastic compliance under constant electric field can be quite different from that under constant charge. Using a compressed matrix (Voigt notation), Eq. (3) can be simplified and ij or kl is replaced by p or q as

$$\begin{aligned} 11 \rightarrow 1, \quad 22 \rightarrow 2, \quad 33 \rightarrow 3, \quad 23 \text{ or } 32 \rightarrow 4, \\ 31 \text{ or } 13 \rightarrow 5, \quad \text{and} \quad 12 \text{ or } 21 \rightarrow 6. \end{aligned}$$

Equation (3) is the complete constitutive equation for a piezoelectric material. In view of the symmetry relations that apply for specific piezoelectric materials, the number of independent coefficients in the equation can be reduced. For example, for an unstretched and poled piezoelectric P(VDF-TrFE) copolymer, which possesses a point group symmetry ∞m , the piezoelectric coefficient, the dielectric permittivity, and elastic compliance matrices are (68)

$$\begin{pmatrix} 0 & 0 & 0 & 0 & d_{15} & 0 \\ 0 & 0 & 0 & d_{15} & 0 & 0 \\ d_{31} & d_{31} & d_{33} & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} K_{11} & 0 & 0 \\ 0 & K_{11} & 0 \\ 0 & 0 & K_{33} \end{pmatrix},$$

$$\begin{pmatrix} s_{11} & s_{12} & s_{13} & 0 & 0 & 0 \\ s_{12} & s_{11} & s_{13} & 0 & 0 & 0 \\ s_{13} & s_{13} & s_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & s_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & s_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & s_{66} [= 2(s_{11} - s_{12})] \end{pmatrix}. \quad (4)$$

In many cases, processing conditions such as extrusion and stretching can induce anisotropy (preferred orientation of the polymer chain along the stretching or extrusion direction) in the plane perpendicular to the poling direction. As a consequence, the macroscopic symmetry of the

poled polymer is 2 mm, for which the nonzero matrix elements of the piezoelectric constitutive equation can be written as (68)

$$\begin{pmatrix} 0 & 0 & 0 & 0 & d_{15} & 0 \\ 0 & 0 & 0 & d_{24} & 0 & 0 \\ d_{31} & d_{32} & d_{33} & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} K_{11} & 0 & 0 \\ 0 & K_{22} & 0 \\ 0 & 0 & K_{33} \end{pmatrix}, \begin{pmatrix} s_{11} & s_{12} & s_{13} & 0 & 0 & 0 \\ s_{12} & s_{22} & s_{23} & 0 & 0 & 0 \\ s_{13} & s_{23} & s_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & s_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & s_{55} & 0 \\ 0 & 0 & 0 & 0 & 0 & s_{66} \end{pmatrix}. \quad (5)$$

For electromechanical application, the electromechanical coupling factor k , which measures the ability of a material to interconvert electrical and mechanical energy, is one of the fundamental parameters (69),

$$k^2 = \frac{\text{Converted mechanical energy}}{\text{Input electric energy}} \quad (6a)$$

or

$$k^2 = \frac{\text{Converted electrical energy}}{\text{Input mechanic energy}}. \quad (6b)$$

It should be noted that in electromechanical applications, an electric field is applied along a certain direction, Eq. (3), and the electromechanical actuation along the same direction or other direction is used. For example, if a polymer actuator can be made with the electric field along the 3-direction, Eq. (3), the actuation along the same direction is used. In this case, the coupling factor is the longitudinal electromechanical coupling factor k_{33} , which is related to the parameters in Eq. (3) as (69)

$$k_{33}^2 = \frac{d_{33}^2}{(\epsilon_{33}^T s_{33}^E)} \quad (7)$$

The coupling factor can also be related to the material coefficient measured under different conditions. For example, the elastic compliance s_{33}^E is related to s_{33}^D (measured under constant charge or an open circuit condition) as (69)

$$s_{33}^D = (1 - k_{33}^2) s_{33}^E \quad (8)$$

Therefore, a polymer with a large coupling factor will see a large difference in the elastic compliance when used under different external electric conditions.

The quasi-static and resonance methods are among several techniques recommended by IEEE to determine the coefficients in the piezoelectric constitutive equations as well as the coupling factors (69). For determining piezoelectric coefficients, a convenient means is to use the dilatometer technique where the strain change induced by an applied electric field under stress-free conditions can be measured, and the ratio of the strain to field yields the piezoelectric coefficient (70). A stress signal σ can also be applied to a sample. In this case, the charge output Q between the electrodes, which are shorted together, is measured to obtain the piezoelectric coefficient (71). One of the difficulties with the latter method is applying a uniaxial

stress uniformly to a sample without introducing other complications such as stress components along other directions. To obtain accurate results using the stress-induced charge method, one may also have to take into account the deformation of the sample under stress. This is accomplished by introducing a correction term to the piezoelectric coefficient as pointed out by Dvey-Aharon and Taylor (72),

$$d = \frac{\partial P}{\partial \sigma} \Big|_E = \frac{1}{A} \frac{\partial Q}{\partial \sigma} - \frac{P}{A} \frac{\partial A}{\partial \sigma} \quad (9)$$

where A is the sample surface area. For PVDF and its copolymers at room temperature, it can be estimated that the contribution from the deformation is less than 10%. For a soft piezoelectric polymer, this correction term can be quite high. On the other hand, when comparing the measured results from the converse and direct piezoelectric effects, it can be shown from the Maxwell relation that (73,74)

$$\frac{1}{A} \frac{\partial Q}{\partial \sigma} \Big|_E = \frac{\partial S}{\partial E} \Big|_\sigma \quad (10)$$

Various coefficients in the piezoelectric constitutive equation for PVDF and P(VDF-TrFE) 75/25 copolymer are listed in Table 2, where the data are taken from several sources (70,75). In most polymers, the response to external stimulus often shows strong relaxation behavior (76); hence, the coefficients in the piezoelectric constitutive equation are frequency dependent and complex, that is (77).

$$d_{ij}^* = d'_{ij} - jd''_{ij}, \quad s_{ij}^* = s'_{ij} - js''_{ij}, \quad \text{and} \quad \epsilon_{ij}^* = \epsilon'_{ij} - j\epsilon''_{ij}. \quad (11)$$

In Table 2, both the real and imaginary parts of these coefficients for P(VDF-TrFE) 75/25 mol% copolymer are listed. Apparently, the imaginary part for the copolymer cannot be neglected.

The temperature dependence of the piezoelectric and dielectric constant of 75/25 copolymer is presented in Fig. 18 and 19. Such temperature dependence behavior, where the piezoelectric coefficient and dielectric constant

Table 2. Piezoelectric, Dielectric, and Elastic Properties of PVDF and P(VDF-TrFE) 75/25 mol%

Material Parameter	PVDF	P(VDF-TrFE) 75/25	
		Real	Imaginary
d_{31} (pC/N)	28	10.7	0.18
d_{32} (pC/N)	4	10.1	0.19
d_{33} (pC/N)	-35	-33.5	-0.65
d_{15} (pC/N)	—	-36.3	-0.32
s_{11} (10^{-10} Pa $^{-1}$)	3.65	3.32	0.1
s_{22} (10^{-10} Pa $^{-1}$)	4.24	3.24	0.07
s_{33} (10^{-10} Pa $^{-1}$)	4.72	3.00	0.07
s_{12} (10^{-10} Pa $^{-1}$)	-1.10	-1.44	-0.036
s_{13} (10^{-10} Pa $^{-1}$)	-2.09	-0.89	-0.022
s_{23} (10^{-10} Pa $^{-1}$)	-1.92	-0.86	-0.022
ϵ_{33}/ϵ_0	15	7.9	0.09
k_{33}	—	0.23	—
k_{13}	0.13	0.07	—
k_t	0.144	0.196	—

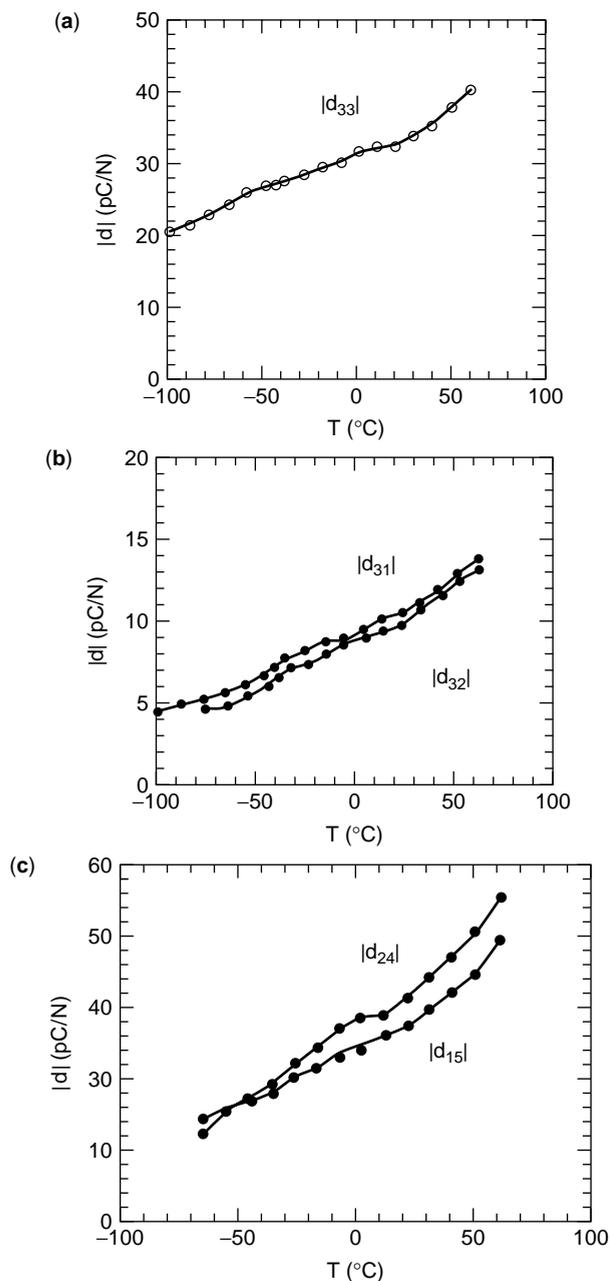


Figure 18. Piezoelectric strain coefficients as a function of temperature measured for P(VDF-TrFE) 75/25 mol% copolymer films: (a) $|d_{33}|$ (b) $|d_{31}|$ and $|d_{32}|$, and (c) $|d_{24}|$ and $|d_{15}|$.

increase as the temperature is increased toward the F-P transition temperature, is a common phenomenon for all ferroelectric-based piezoelectric materials including many commercially available piezoceramics (47,78,79).

In semicrystalline copolymers, the piezoelectric properties can often be improved by raising the degree of crystallinity, given that the piezoelectric response is mainly from the crystalline region. For example, using high-pressure crystallization to increase the crystallinity and improve the quality of the crystalline phase, Ohigashi et al. reported an improved thickness coupling factor for 80/20 copolymer to near 0.3 ($k_t = 0.3$) (80). More recently, a

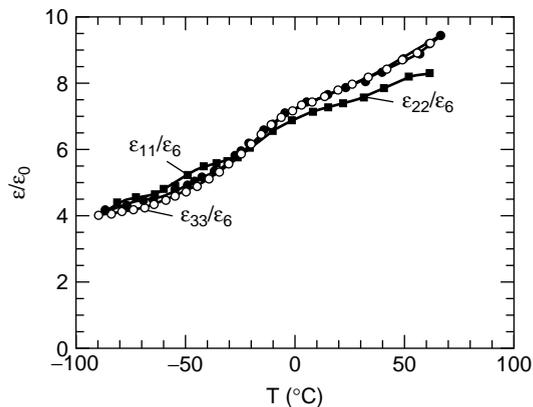


Figure 19. Dielectric constant (1 kHz) as a function of temperature for P(VDF-TrFE) 75/25 mol% copolymer films.

relatively large size single-crystal P(VDF-TrFE) 75/25 mol% copolymer was grown having a room temperature $d_{33} = -38$ pm/V and coupling factor $k_{33} = 0.33$, values that to now represent the best piezoelectric performance of known piezoelectric and ferroelectric polymers (81). Figures 20 to 22 show the temperature dependence of the

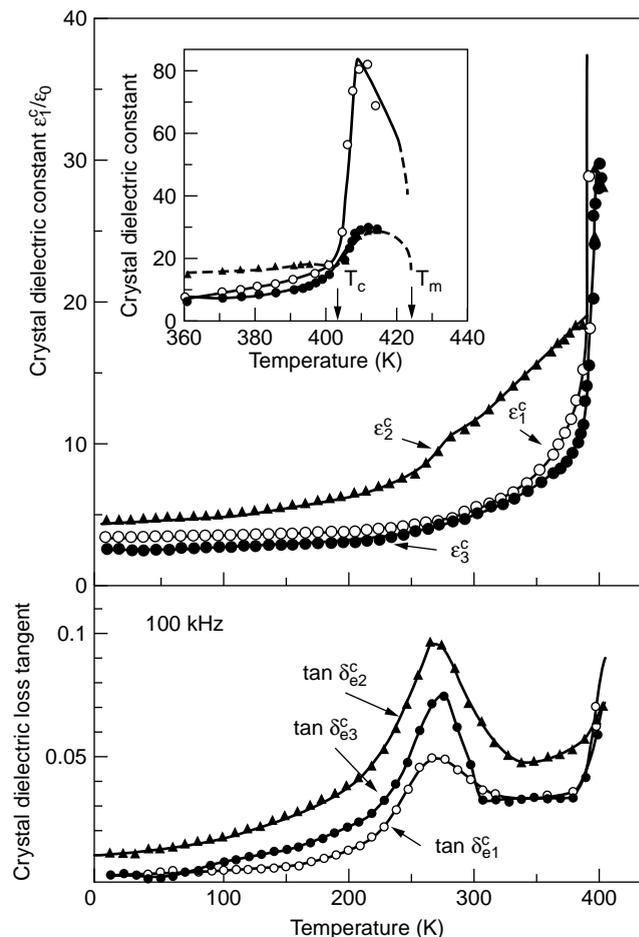


Figure 20. Temperature dependence of dielectric constants (ϵ_1^c , ϵ_2^c , ϵ_3^c) and dielectric tangent losses ($\tan \delta_{e1}^c$, $\tan \delta_{e2}^c$, $\tan \delta_{e3}^c$) of P(VDF-TrFE) single crystals.

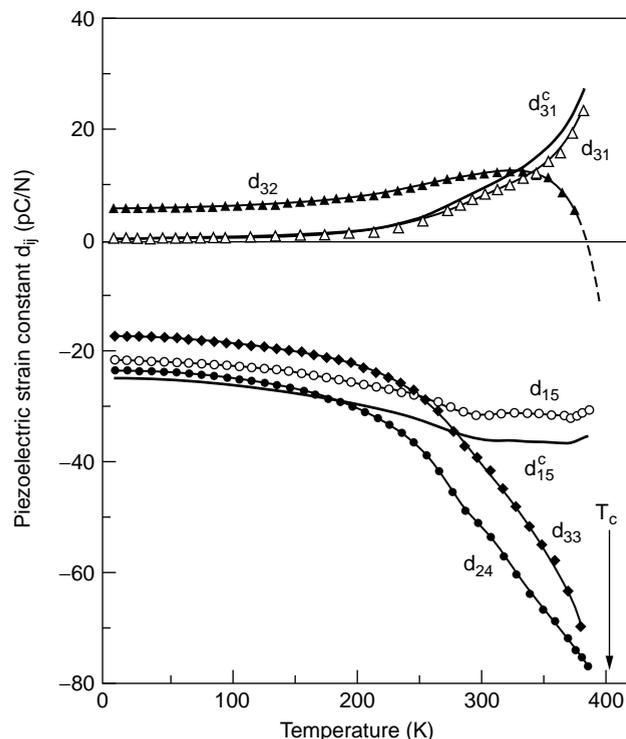


Figure 21. Temperature dependence of piezoelectric coefficients d_{31} , d_{32} , d_{33} , d_{15} , and d_{24} of P(VDF-TrFE) single crystal films and d_{31}^c and d_{15}^c for single crystal of P(VDF-TrFE) 75/25 mol% copolymer.

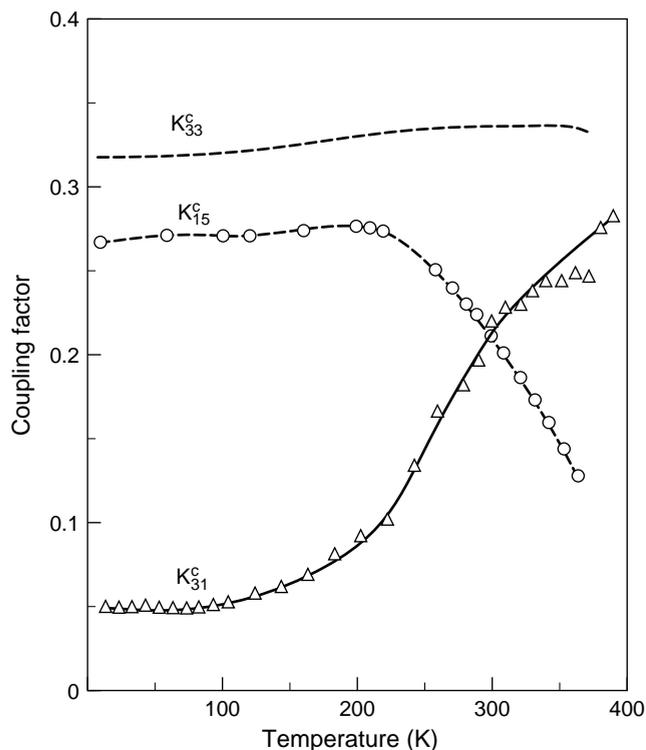


Figure 22. Temperature dependence of electromechanical coupling factors k_{31}^c , k_{33}^c , and k_{15}^c for P(VDF-TrFE) 75/25 mol% single crystals.

dielectric and piezoelectric properties and the electro-mechanical coupling factors for a copolymer single crystal.

RELAXOR FERROELECTRIC BEHAVIOR AND ELECTROSTRICTIVE RESPONSE IN P(VDF-TrFE) BASED COPOLYMERS

In addition to operating a ferroelectric polymer deep inside the ferroelectric phase (at temperatures far below the F–P transition temperature) where the piezoelectric response is the main source of the electromechanical response, one could also make use of the phase transition phenomenon where large changes in the properties can be realized. As has been shown above, the lattice strains in P(VDF-TrFE) copolymers experience large changes through the F–P transition, changes which far surpass those achievable from the piezoelectric effect. Besides the large strain response at F–P transition, there is another interesting feature associated with the F–P transition—the possibility that a very large electromechanical coupling factor ($k \sim 1$) can be obtained near a first-order F–P transition temperature (82).

Above the F–P transition temperature, a ferroelectric state can be induced by imposing an external field as schematically depicted in Fig. 23 (78,82). The first-order field-induced transition region terminates at a critical point. As shown by an earlier study based on thermodynamic theory, the coupling factor can approach 1 (near 100% energy conversion efficiency) at temperatures above the first-order transition and near the critical point (82). This is a general result for first-order F–P phase transformations despite the fact that the curves in Fig. 23 are from the parameters applicable to BaTiO_3 . For P(VDF-TrFE) copolymers, it has been shown that at compositions with VDF content $>60\%$ mol% the F–P transition is first order. Still the temperature range in which the electric field can induce the phase transition from nonpolar to polar phases depends strongly on the material. For most of the inorganic ferroelectrics, this temperature range is relatively narrow. For instance, the range is about 8°C for BaTiO_3 (as approximately measured by the temperature range between F–P transition and critical temperature) (47). For P(VDF-TrFE) copolymers, however, it has been reported that this temperature range exceeds 50°C as reported by a recent study on Langmuir-Blodgett film of P(VDF-TrFE) 70/30 mol% (83).

These results suggest that one may be able to improve the electromechanical response of P(VDF-TrFE) copolymer significantly by operating the polymer near the F–P transition. However, there are several issues associated with the first-order F–P transition in P(VDF-TrFE) copolymer that have to be addressed. As has been shown in the phase diagram (Fig. 5), the F–P transition in all P(VDF-TrFE) compositions occurs at temperatures higher than room temperature. The transition is also relatively sharp (over a relatively narrow temperature range). In addition, undesirably large hysteresis has been observed for the copolymers at the first order F–P transition. Therefore to make use of the unique opportunities near the first-order F–P transition in P(VDF-TrFE) copolymer systems, the copolymer should be modified as to broaden the phase transition

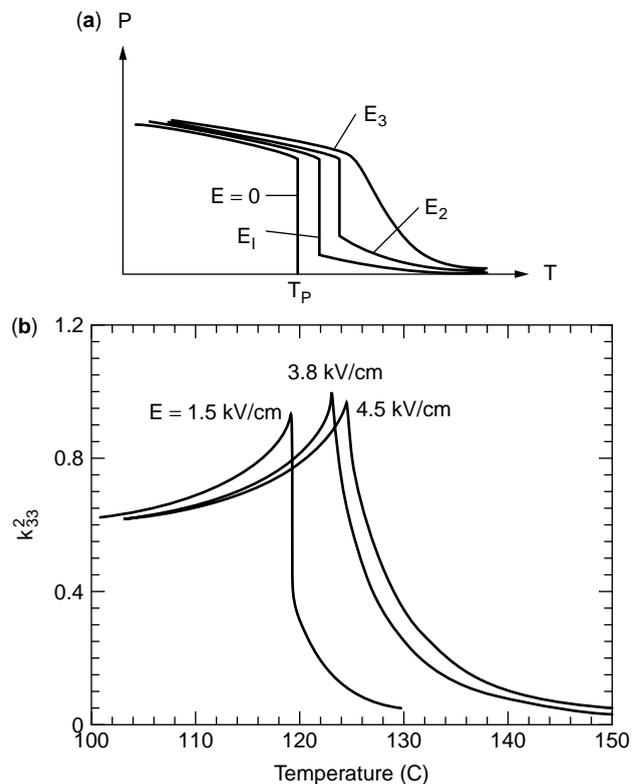


Figure 23. (a) Schematic of polarization as a function of temperature for a first-order ferroelectric-paraelectric transition under different dc bias field. The first-order polar to nonpolar transition terminates at a critical point. (b) electromechanical coupling coefficient (k_{33}) as a function of temperature under different dc bias field near a first-order F-P phase transition. The results are derived based on Landau-Devonshire theory using the parameter obtained on BaTiO₃.

region as well as to move it to room temperature while minimizing hysteresis.

To modify the phase transition behavior in P(VDF-TrFE) copolymer, one approach is to make use of high-energy irradiation. This has been demonstrated in several earlier studies. For example, Lovinger found that high-energy electron irradiation can convert the ferroelectric phase at room temperature to resemble a macroscopically paraelectric phase (84). Subsequently, Odajima et al. and Daudin et al. found that the sharp dielectric constant peak from the F-P transition can be broadened markedly and moved to near room temperature by irradiation (85,86). More recently, Zhang et al. showed that by high-energy electron irradiation, the normal ferroelectric P(VDF-TrFE) copolymers can be converted into a relaxor ferroelectric with high electrostrictive strains (87–90).

Figure 24 compares the polarization loops of P(VDF-TrFE) 50/50 mol% copolymer before and after irradiation (40 Mrad of 2.5 MeV electrons at 120°C). As shown, the irradiation effectively eliminates the room temperature polarization hysteresis. The dielectric data of the irradiated copolymer are presented in Fig. 25, where clearly the broad room temperature peak moves to higher temperature as the measuring frequency increases. In addition, it

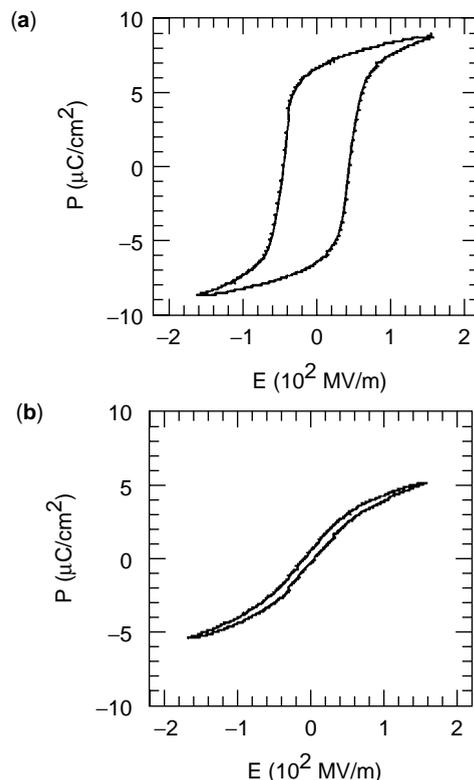


Figure 24. Polarization hysteresis loop of P(VDF-TrFE) 50/50 mol% copolymer film measured at room temperature (a) before and (b) after irradiation with electrons of 2.5 MeV energy for 40 Mrad dose at 120°C.

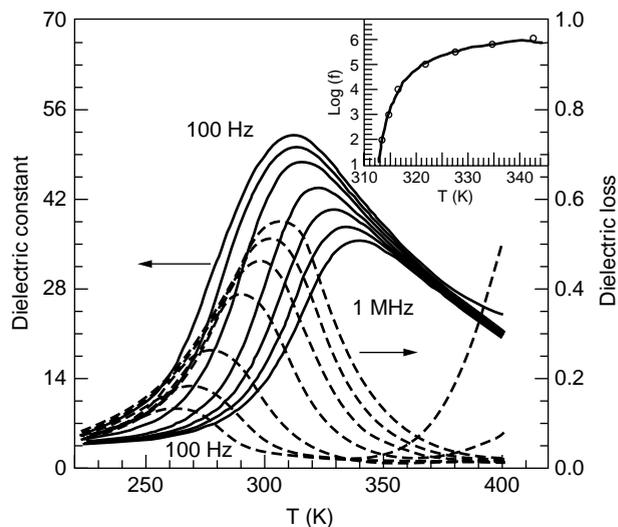


Figure 25. The dielectric constant (solid lines) and dielectric loss (dashed lines) as a function of temperature of P(VDF-TrFE) 50/50 mol% copolymer films irradiated with 40 Mrad dose at 120°C using 2.55 MeV electrons. The inset shows the fitting of the Vogel-Fulcher law. Here the solid line is the fit and the circles are the data points.

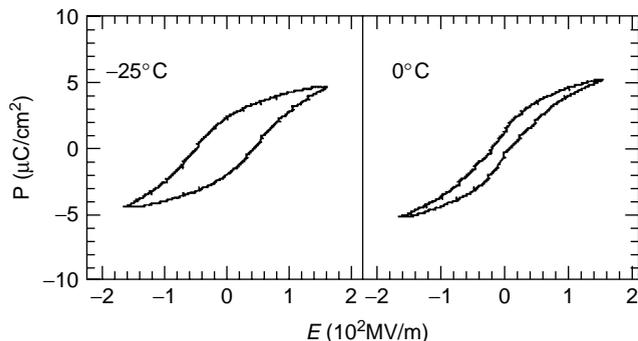


Figure 26. Gradual increase in remanent polarization and hysteresis with lowering the temperature measured for P(VDF-TrFE) 50/50 mol% copolymer films irradiated with 40 Mrad dose at 120°C using 2.55 MeV electrons.

was found that the peak temperature T_m is related to the frequency f by the Vogel-Folcher (V-F) relation, as shown in the insert of Fig. 25,

$$f = f_0 \exp\left(\frac{-U}{k(T_m - T_f)}\right), \quad (12)$$

where T_f can be regarded as a freezing temperature of the system (91).

Polarization loop measurements also reveal that as the temperature is lowered, the irradiated copolymer shows a gradual increase of the polarization hysteresis, Fig. 26. Figure 27 compares the remanent polarization for the copolymer before and after irradiation. Evidently, P_r changes with temperature rather gradually after the irradiation. The peak of the derivative of P_r with respect to temperature is very close to T_f from the V-F law fitting (305 K

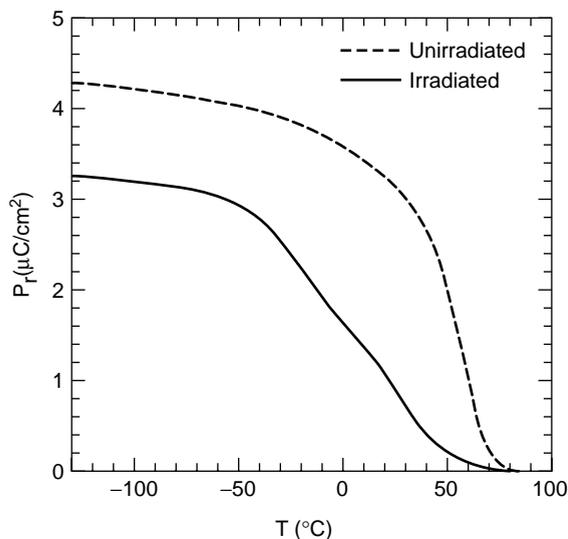


Figure 27. Comparison of change in remanent polarization (P_r) with temperature for P(VDF-TrFE) 50/50 mol% copolymer films before (dashed lines) and after (solid line) irradiating with 2.5 MeV electrons for 40 Mrad dose at 120°C.

versus $T_f = 307$ K). This result suggests a dipolar freezing process in the material. All of these features are reminiscent of those observed in ceramic $\text{Pb}(\text{Mg}_{1/3}\text{Nb}_{2/3})\text{O}_3$, which is the best-known relaxor ferroelectric in inorganic systems (92,93). Since the polarization response in P(VDF-TrFE) copolymer is from the dipoles, the irradiated copolymer should be an interesting system to study the relaxor ferroelectric behavior in a dipolar system.

The irradiated copolymer has been shown to possess a high electrostrictive strain. Figure 28(a) shows the longitudinal strain (S_3) for the copolymer 68/32 mol% after irradiation: a strain of -5% is achieved under a field of 150 MV/m (94). The plot of S_3 versus P^2 yields a straight line, indicating the response is electrostrictive in nature, $S_3 = Q_{33}P_3^2$, Fig. 28(b). For the irradiated copolymer, Q_{33} is found in the range between -4 to $-15 \text{ m}^4/\text{C}^2$, depending on the sample processing conditions. The strain response does not change appreciably with temperature as suggested by the plot in Fig. 28(c).

Of special interest is the finding that in P(VDF-TrFE) copolymers, large anisotropy in the strain responses exists along and perpendicular to the chain direction, as can be deduced from the change in the lattice parameters between the polar and nonpolar phases. Therefore, the transverse strain can be tuned over a large range by varying the film processing conditions (95,96). For unstretched films, the transverse strain is relatively small ($\sim +1\%$ level at ~ 100 MV/m), while the amplitude ratio between the transverse strain and longitudinal strain is less than 0.33. This feature is attractive for devices utilizing the longitudinal strain such as ultrasonic transducers in the thickness mode, and actuators and sensors making use of the longitudinal electromechanical responses of the material. For example, with a very weak transverse electromechanical response in comparison with the longitudinal, one can significantly reduce the influence of the lateral modes on the thickness resonance and improve the performance of the thickness transducer. On the other hand, for stretched films, a large transverse strain (S_1) along the stretching direction can be achieved as shown in Fig. 29, where the transverse strain about $+3.5\%$, as observed in the irradiated copolymer under an electric field of 110 MV/m. It is also found that for stretched films, the amplitude of the transverse strain along the stretching direction is comparable to the longitudinal strain, while the amplitude of the transverse strain in direction perpendicular to the stretching direction (S_2) is much less ($\sim 1\%$ at 100 MV/m).

It is also of interest that the strain along the thickness direction (parallel to the electric field) is always negative for P(VDF-TrFE) copolymers regardless of the sample processing conditions. In fact, this is a general feature for a system in which the polarization response originates from the dipolar interaction and is true for all polymeric piezoelectric and electrostrictive responses (97,98). The sign of the strains perpendicular to the applied field direction will depend on the sample processing conditions. For the anisotropically stretched films discussed here, the electric induced strain along the stretching direction, which is perpendicular to the applied field, is positive, whereas in the direction perpendicular to both stretching and applied field directions, the strain is negative. For unstretched samples

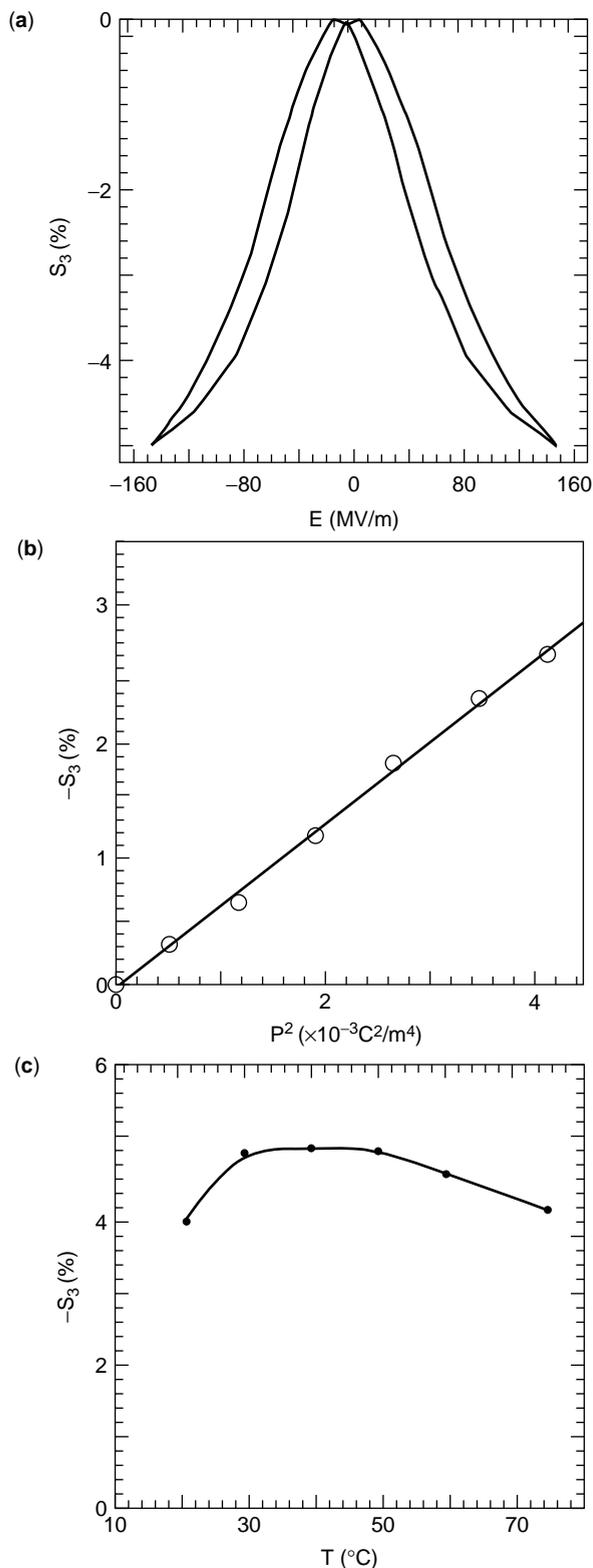


Figure 28. (a) Electric field induced strain along the thickness direction (longitudinal strain, S_3) versus electric field measured at room temperature and 1 Hz, (b) change in longitudinal strain (S_3) with square of polarization (P), and (c) temperature dependence of longitudinal strain induced under 14 MV/m and 1 Hz driving electric field, of unstretched P(VDF-TrFE) 68/32 mol% copolymer films irradiated at 105 $^{\circ}\text{C}$ with 70 Mrad dose using 1 MeV electrons.

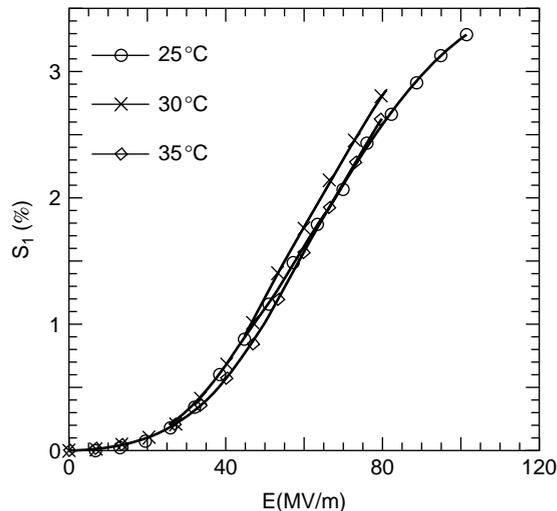


Figure 29. Transverse strain along the stretching direction (S_1) as a function of driving electric field at different temperatures measured for stretched P(VDF-TrFE) 68/32 mol% copolymer films irradiated at 100 $^{\circ}\text{C}$ with 70 Mrad dose using 1.2 MeV electrons.

that are isotropic in the plane perpendicular to the applied field, the strain component in the plane is an average of the strains along the chain (positive) and perpendicular to the chain (negative) and is in general positive.

For electrostrictive materials, the electromechanical coupling factor (k_{ij}) has been derived by Hom et al. based on the consideration of electrical and mechanical energies generated in the material under external field (99):

$$k_{3j}^2 = \frac{kS_j^2}{s_{ij}^D \left[P_E \ln \left(\frac{P_S + P_E}{P_S - P_E} \right) + P_S \ln \left(1 - \left(\frac{P_E}{P_S} \right)^2 \right) \right]}, \quad (13)$$

where $j = 1$ or 3 correspond to the transverse or longitudinal direction (e.g., k_{31} , is the transverse coupling factor) and s_{ij}^D is the elastic compliance under constant polarization, S_j and P_E are the strain and polarization responses, respectively, for the material under an electric field of E . The coupling factor depends on E , the electric field level. In Eq. (13), it is assumed that the polarization-field (P - E) relationship follows approximately

$$|P_E| = P_S \tanh(k|E|), \quad (14)$$

where P_S is the saturation polarization and k is a constant. It is found that Eq. (14) describes the P - E relationship of the irradiated copolymers studied here quite well (94).

The electromechanical coupling factors for the irradiated copolymers have been determined based on the data on the field-induced strain, the elastic modulus (Fig. 30), and polarization. Presented in Fig. 31 are k_{33} for the unstretched sample and k_{31} for the stretched sample along the drawing direction. Near room temperature and under an electric field of 80 MV/m, k_{33} can reach more than 0.3, which is comparable to that obtained in a single-crystal P(VDF-TrFE) copolymer (81). More interestingly, k_{31} of 0.45 can be obtained in a stretched copolymer, which is

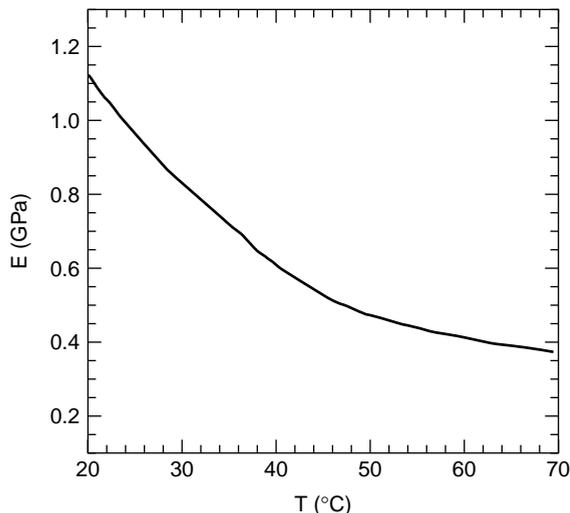


Figure 30. Temperature dependence of elastic modulus measured along the stretching direction for stretched P(VDF-TrFE) 68/32 mol% copolymer films irradiated at 100°C with 70 Mrad dose using 1.2 MeV electrons.

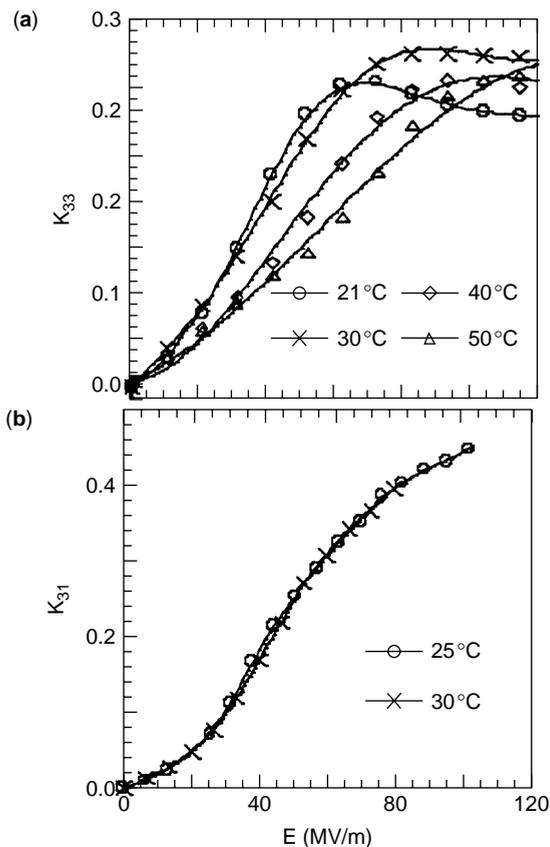


Figure 31. Dependence of electromechanical coupling coefficients on the applied electric field: (a) k_{33} for extruded unstretched P(VDF-TrFE) 68/32 mol% copolymer films irradiated at 105°C with 70 Mrad dose using 1 MeV electrons and (b) k_{31} for stretched P(VDF-TrFE) 68/32 mol% copolymer films irradiated with 70 Mrad dose using 1.2 MeV electrons at 100°C.

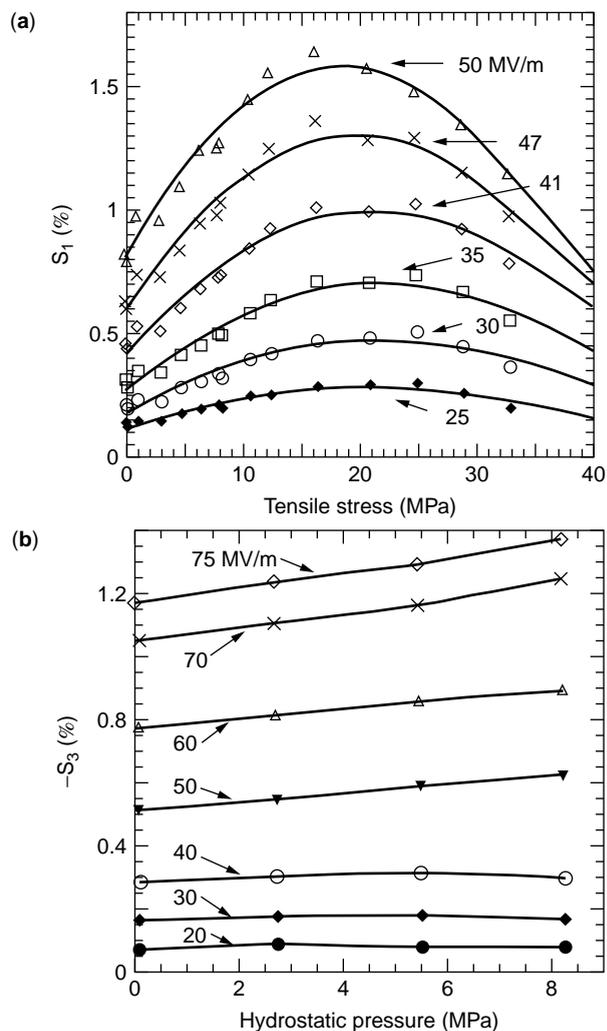


Figure 32. Effect of (a) tensile stress on transverse strains (S_1) for stretched film and (b) hydrostatic pressure on longitudinal strain (S_3) for unstretched film at room temperature under different driving electric fields. The sample used here is P(VDF-TrFE) 65/35 mol% copolymer film irradiated at 95°C with 60 Mrad dose using 2.55 MeV electrons.

much higher than values measured in unirradiated P(VDF-TrFE) copolymers.

For a polymer, there is always a concern about the electromechanical response under high mechanical load; that is, whether the material can maintain high strain levels when subject to high external stresses. Figure 32(a) depicts the transverse strain of stretched and irradiated 65/35 copolymer under a tensile stress along the stretching direction and the longitudinal strain of unstretched and irradiated 65/35 copolymer under hydrostatic pressure (100,101). As can be seen from the figure, under a constant electric field, the transverse strain increases initially with the load and reaches a maximum at the tensile stress of about 20 MPa. Upon a further increase of the load, the field-induced strain is reduced. One important feature revealed by the data is that even under a tensile stress of 45 MPa, the strain generated is still nearly the same as

that without load, indicating that the material has a very high load capability. Shown in Fig. 32(b) is the longitudinal strain under hydrostatic pressure. At low electric fields, the strain does not change much with pressure, while at high fields it shows increase with pressure.

The results demonstrate that the electrostrictive P(VDF-TrFE) copolymer has a relatively high load capability. The observed change in the strain with load can be understood based on the consideration of the electrostrictive coupling in this relaxor ferroelectric material as has been considered and discussed in (100,101).

CONCLUDING REMARKS

A large number of studies are concerned with the electromechanical properties of PVDF and P(VDF-TrFE) polymers, including both the piezoelectric responses from polymers with semicrystalline and single-crystal forms and electrostrictive responses from the newly developed high-energy irradiated copolymers. This article has consolidated these studies and emphasized the different polarization responses in ferroelectric polymers such as polarization switching, phase transformation, and pure dielectric response. Optimizing the electromechanical responses from each type of polarization responses is a fruitful area of research. By proper polymer engineering, the electromechanical properties can be improved substantially as demonstrated in the high-energy irradiated copolymers.

The discussion has included the syntheses, stereochemistry, and major crystal structures as well as their interesting morphologies, phase diagrams and phase transitions. From a practical perspective, it should be quite evident that knowledge of their macromolecular properties and structures is quite desirable to successfully exploit their piezoelectric and electrostrictive properties. In particular, the molecular conformation, crystal structures, and polymer morphology can be controlled at the molecular and mesoscopic levels, and this can be accomplished by varying the composition and electroprocessing conditions, as well as utilizing defect modification. As a result, the properties of PVDF and its copolymer depend substantially on these conditions. Although traditional PVDF and the P(VDF-TrFE) polymers have been used in the piezoelectric mode, recent evidence was presented that demonstrates a remarkable enhancement in the strain of P(VDF-TrFE) films after exposure to high-energy irradiation, which involves electrostriction. Further study in this direction is certainly merited if only to identify alternative techniques to generate electrostrictive polymer films and other avenues to achieve high electromechanical effects.

BIBLIOGRAPHY

1. A.J. Lovinger, *Science* **220**: 111 (1983).
2. T. Furukawa, *Phase Transitions* **18**(2): 14 (1989).
3. H.S. Nalwa, ed. *Ferroelectric Polymers*. Dekker, New York, 1995.
4. T.T. Wang, J.M. Herbert, and A.M. Glass, eds. *Applications of Ferroelectric Polymers*. Blackie and Son, Glasgow, 1988.
5. H. Kawai, *Jpn. J. Appl. Phys.* **8**: 975 (1969).
6. M.A. Marcus. Fifth Int. Mg. on Ferroelectricity. Pennsylvania State University, Aug. 17–21, 1981.
7. J. F. Lindberg, *Mater. Res. Soc. Symp. Proc.* **459**: 509 (1997).
8. J. Powers. In T.T. Wang, J.M. Herbert, and A.M. Glass, eds., *Application of Ferroelectric Polymers*. Blackie and Son, Glasgow, 1988, chap. 6.
9. N. Murayama, K. Nakamura, H. Obara, and M. Segawa. The Strong Piezoelectricity in Polyvinylidene Fluoride Ultrasonics **15**, (1976).
10. A.J. Cleaver and P. Pantelis. *Piezoelectric Poly(vinylidene fluoride) Films for Use in Telecommunications, Plastics in Telecommunications III*. Plastics and Rubber Institute, London, Sept. 15–17, 1982, p. 32.1.
11. H.R. Gallantree. *IEEE proc.* **130**: 219 (1983).
12. P.M. Galletti, D.E. De Rossi, and A.S. De Reggi, eds. *Medical Applications of Piezoelectric Polymers*. Gordon and Breach, New York, 1988.
13. J.B. Lando, H.G. Olf, and A. Peterlin. *J. Polym. Sci.* **A1**(4): 941 (1966).
14. M.K. Tamura, K. Ogasawara, N. Ono, and S. Hagiwara. *J. Appl. Phys.* **45**(9):3768 (1974).
15. R.G. Kepler and R.A. Anderson. *J. Appl. Phys.* **49**: 1232 (1978).
16. M. Tamura, S. Hagiwara, S. Matsumoto, and N. Ono. *J. Appl. Phys.* **48**: 513 (1977).
17. D. Naegele and D.Y. Yoon. *Appl. Phys. Lett.* **33**: 132 (1978).
18. S.C. Mathur, J.I. Scheinbeim, and B.A. Newman. *J. Appl. Phys.* **56**: 2419 (1984).
19. H. Von Berlepsch, W. Kunstler, A. Wedel, R. Danz, and D. Geiss. *IEEE Trans. Elect. Insul.* **24**: 357 (1989).
20. E. Fukada, S. Tasaka, and H.S. Nalwa. In H.S. Nalwa, ed., *Polyureas and Polythioureas, Ferroelectric Polymers*. Dekker, New York, chap. 9, 1995, pp. 353–392.
21. R.A. Ferren. *Application of Ferroelectric Polymers*. Blackie and Son, Glasgow, 1988, chap. 2.
22. T. Yagi and M. Tatamoto. *Polym. J.* **2**(6): 429 (1979).
23. K. Kimura and H. Ohigashi. *Jpn. J. Appl. Phys.* **25**: 383 (1986).
24. A.J. Lovinger. In D.C. Bassett, ed., *Developments in Crystalline Polymers*, Vol. 1. Applied Science Publishers, London, 1982, p. 195.
25. R. Hasegawa, Y. Takahashi, Y. Chatani, and H. Tadokoro. *Polym. J.* **3**: 600 (1972).
26. Y. Takahashi, Y. Matsubara, and H. Tadokoro. *Macromolecul.* **16**: 1588 (1983).
27. M. Bachmann, W.L. Gordon, S. Weinhold, and J.B. Lando. *J. Appl. Phys.* **51**: 5095 (1980).
28. A.J. Lovinger. *Macromolecul.* **14**: 322 (1981).
29. Y. Takahashi and H. Tadokoro. *Macromolecul.* **13**: 1317 (1980).
30. H. Okigashi and K. Koga. *Jpn. J. Appl. Phys.* **8**: L455 (1982).
31. *Lange's Handbook of Chemistry*, 13th ed., 1985, pp. 3–121.
32. K. Tashiro, K. Takano, M. Kobayashi, Y. Chatani, and H. Tadokoro. *Ferroelect.* **57**:297–326 (1984).
33. A.J. Lovinger, T. Furukawa, G.T. Davis, and M.G. Broadhurst. *Polym.* **24**:1225+ (1983).
34. B.L. Farmer, A.J. Hopfinger, and J.B. Lando. *J. Appl. Phys.* **43**:4293 (1972).
35. K. Tashiro and H. Tadokoro. *Macromolecul.* **16**:961 (1983).
36. N. Karasawa and W.A. Goddard. *Macromolecul.* **25**:7268 (1992).

37. G.J. Kavarnos and R. Holman. *Polym.* **35**: 5586 (1994).
38. G.J. Kavarnos, H.C. Robinson and R.W. Holman. *Ferroelectr.* **205**: 133 (1998).
39. N.C. Banik, F.P. Boyle, T.J. Sluckin, P.L. Taylor, S.K. Tripathy, and A.J. Hopfinger. *J. Chem. Phys.* **72**: 3191 (1980).
40. P.E. Bloomfield and M.A. Marcus. In T.T. Wang, J.M. Herbert, and A.M. Glass, eds., *Application of Ferroelectric Polymers*. Blackie and Son, Glasgow, 1988, chap. 3.
41. R.W. Holman and G.J. Kavarnos. *Polym.* **37**: 1697 (1996).
42. G.T. Davis, T. Furukawa, A.J. Lovinger, and M.G. Broadhurst. *Macromolecul.* **15**: 329 (1982).
43. A.J. Lovinger, G.T. Davis, T. Furukawa, and M.G. Broadhurst. *Macromolecul.* **15**: 323 (1982).
44. M.V. Fernandez, A. Suzuki, and A. Chiba. *Macromolecul.* **20**: 1806 (1987).
45. K. Tashiro and M. Kobayashi. *Rep. Progr. Polym. Phys. Jpn.* **29**: 169 (1986).
46. M.E. Lines and A.M. Glass. *Principles and Applications of Ferroelectrics and Related Materials*. Clarendon Press, Oxford, 1977.
47. F. Jona and G. Shirane. *Ferroelectric Crystals*. Dover, New York, 1993, p. 138.
48. A. Sharples. *Introduction to Polymer Crystallization*. St. Martin's, New York, 1966.
49. J. Scheinbeim, C. Nakafuku, B.A. Newman, and K.D. Pae. *J. Appl. Phys.* **50**: 4399 (1979).
50. S. Ducharme, V.M. Fridkin, and A.V. Bune. *Phys. Rev. Lett.* **84**(1): 178 (2000).
51. G.M. Stack and R.Y. Ting. *J. Polym. Sci.* **B26**: 55 (1988).
52. J.S. Green, B.I. Farmer, and J.F. Rabolt. *J. Appl. Phys.* **60**(8): 2690 (1986).
53. R. Tanaka, K. Tashiro, and M. Kobayashi. *Polym.* **40**: 3855 (1999).
54. Y. Oka, N. Koizumi, and Y. Murata. *J. Polym. Sci.* **B24**: 2059 (1986).
55. N. Koizumi, Y. Murata, and H. Tsunashima. *IEEE Trans. Electr. Insul.* **E1-21**: 543 (1986).
56. T. Furukawa. *Adv. Collo. Inter. Sci.* **71-72**: 183 (1997).
57. K. Kimura and H. Ohigashi. *Appl. Phys. Lett.* **43**: 834 (1983).
58. S. Palto, L. Blinov, A. Bune, E. Dubovik, V. Fridkin, N. Petukhova, K. Verkhovsakya, and S. Yudin. *Ferroelectr. Lett. Sect.* **19**: 65 (1995).
59. A.V. Bune, V.M. Fridkin, S. Ducharme, L.M. Blinov, S.P. Palto, A.V. Sorokin, S.G. Yudin, and A. Zlatkin. *Nature* **391**: 874 (1998).
60. V. Sundar and R.E. Newnham. *Ferroelectr.* **135**: 431 (1992).
61. S.C. Hwang and G. Arlt. *J. Appl. Phys.* **87**(2): 869 (2000).
62. T. Furukawa and N. Seo. *Jpn. J. Appl. Phys.* **29**(4): 675 (1990).
63. H. Dvey-Aharon, T.J. Sluckin, and P.L. Taylor. *Phys. Rev.* **B21**: 3700 (1980).
64. N. Takahashi and A. Odajima. *Ferroelectr.* **57**: 221 (1984).
65. Y. Takahashi, Y. Nakagawa, H. Miyaji, and K. Asai. *J. Polym. Sci.* **C25**: 153 (1987).
66. T. Takahashi, M. Dale, and E. Fukada. *Appl. Phys. Lett.* **37**(9): 791 (1980).
67. I.L. Guy and J. Unworth. *Appl. Phys. Lett.* **52**: 532 (1988).
68. J.F. Nye. *Physical Properties of Crystals*. Clarendon Press, Oxford, 1987.
69. IEEE Standard on Piezoelectricity. ANSI/IEEE Std 176-1987, IEEE, New York, 1988.
70. H. Wang, Q.M. Zhang, L.E. Cross, and A.O. Sykes. *J. Appl. Phys.* **74**: 3394 (1993).
71. H. Wang, Q.M. Zhang, and L.E. Cross. *Jpn. J. Appl. Phys.* **32**: L1281 (1993).
72. H. Dvey-Aharon and P.L. Taylor. *Ferroelectr.* **33**: 103 (1981).
73. R.G. Kepler and R.A. Anderson. *J. Appl. Phys.* **49**: 4490 (1978).
74. R.A. Anderson and R.G. Kepler. *Ferroelectr.* **32**: 13 (1981).
75. H. Schewe. *Ultrasonics Symp. Proc.*, Vol. 1, IEEE, New York, 1982, p. 519.
76. N.G. McCrum, B.E. Read, and G. Williams. *Anelastic and Dielectric Effects in Polymeric Solids*. Dover, New York, 1991, chap. 4.
77. R. Holland. *IEEE Trans. Sonics Ultrason.* **14**(1): 18 (1967).
78. A.F. Devonshire. *Philosophical Mag.* **3**(10): 86 (1954).
79. D.A. Berlincourt, D.R. Curran, and H. Jaffe. In W.P. Mason, ed., *Piezoelectric and Piezomagnetic Materials and Their Function in Transducers in Physical Acoustics*, Vol. 1. Academic Press, New York, 1964.
80. H. Ohigashi and T. Hattori. *Ferroelectro.* **171**: 11 (1995).
81. K. Omote, H. Ohigashi, and K. Koga. *J. Appl. Phys.* **81**(6): 2760 (1997).
82. Q.M. Zhang, J. Zhao, T. Shrout, N. Kim, L.E. Cross, A. Amin, and B.M. Kulwicki. *J. Appl. Phys.* **77**: 2549 (1995).
83. S. Ducharme, A.V. Bune, L.M. Blinov, V.M. Fridkin, S.P. Palto, A.V. Sorokin, and S.G. Yudin. *Phys. Rev.* **B57**: 25 (1999).
84. A. Lovinger. *Macromolecul.* **18**: 190 (1985).
85. F. Maachi, B. Daudin, and J.F. Legrand. *Nucl. Instr. Meth.* **B46**: 324 (1990).
86. A. Odajima, Y. Takasa, T. Ishibashi, and Y. Yuasa. *Jpn. J. Appl. Phys.* **24**: 881 (1985).
87. Q.M. Zhang, V. Bharti, and X. Zhao. *Science* **280**: 2101 (1998).
88. V. Bharti, X. Zhao, and Q.M. Zhang. *Mater. Res. Innovat.* **2**: 57 (1998).
89. V. Bharti, H.S. Xu, and Q.M. Zhang. *J. Appl. Phys.* **87**: 452 (2000).
90. Q.M. Zhang, Z.Y. Cheng, and V. Bharti. *Appl. Phys. A: Mater. Sci. Process* **70**: 307 (2000).
91. H. Vogel. *Z. Phys.* **22**: 645 (1921).
92. L.E. Cross. *Ferroelectro* **151**: 305 (1994).
93. D. Viehland, S.J. Jang, L.E. Cross, and M. Wuttig. *J. Appl. Phys.* **68**: 2916 (1990).
94. X. Zhao, V. Bharti, Q.M. Zhang, T. Ramotowski, F. Tito, and R. Ting. *Appl. Phys. Lett.* **73**: 2054 (1998).
95. Z.-Y. Cheng, T.-B. Xu, V. Bharti, S. Wang, and Q.M. Zhang. *Appl. Phys. Lett.* **74**: 1901-1903 (1999).
96. Z.-Y. Cheng, V. Bharti, T.B. Xu, S. Wang, Q.M. Zhang, T. Ramotowski, F. Tito, and R. Ting. *J. Appl. Phys.* **86**: 2208 (1999).
97. W. Kinase and T. Takahashi. *J. Phys. Soc. Jpn.* **10**: 942 (1955).
98. Y.M. Shkel and D.J. Klingenberg. *J. Appl. Phys.* **83**: 415 (1998).
99. C. Hom, S. Pilgrim, N. Shankar, K. Bridger, M. Masuda, and S. Winzer. *IEEE Trans. Ultrason. Ferro. Freq. Cntr.* **41**: 542-551 (1994).
100. V. Bharti, Z.-Y. Cheng, S. Gross, T.-B. Xu, and Q.M. Zhang. *Appl. Phys. Lett.* **75**: 2653 (1999).
101. S.J. Gross, Z.-Y. Cheng, V. Bharti, and Q.M. Zhang. *Proc. IEEE 1999 Int. Symp. Ultrasonics*, Lake Tahoe, NE, 1999, pp. 1019-1024.

POLYMER BLENDS, FUNCTIONALLY GRADED

YASUYUKI AGARI

Osaka Municipal Technical Research Institute

Joto-ku

Osaka, Japan

INTRODUCTION

Many reports have been published on functionally graded materials made of metals and ceramics (1). These graded materials have improved strength against thermal stress, electromagnetic, and optical properties. There have been particularly many reports on a functionally graded ceramic, which can be called a smart material. In this ceramic, the area of strong thermoelectric performance shifted with increasing temperature. Then, thermoelectric performance can be kept high across a wide temperature gradient.

There have been some reports on functionally graded polymeric materials (2–38). These functionally graded polymeric materials can be classified into four types based on the materials used, as shown in Table 1. Then, graded structures may be classified into six groups. However, reports on functionally graded polymer blends are few (4–9,14–25), although studies have been published on various types of blends. A functionally graded polymer blend has the structure shown in Fig. 1. The blend has two

different surfaces without an interface and can have both the advantages of a laminate and a homogenous blend.

Thus, we devised a new method, the dissolution–diffusion method for preparing functionally graded polymer blends (4–9,24,25). Here, graded polymer blends are classified into two types, and they were prepared by three methods except for our method, surface inclination in the melt state (14,15), surface inclination in solution (16,17), and diffusion in melt (19–23). The dissolution–diffusion method devised by us is only one method that can be used for preparing both types of graded polymer blends. Our method has the following advantages compared with other methods. The preparative time in our method is very short, 100 times shorter than the “diffusion in melt state” method. The optimum conditions can be easily determined because our method has many controllable factors. Further, chemical decomposition of molecules does not occur in preparation because the preparation is at a lower temperature. Therefore, our method is considered very useful.

In this report, I give a detailed description of the preparative mechanism for functionally graded polymer blends in the dissolution–diffusion method. Then, I explain how I determined the optimum conditions for the several types of functionally graded polymer blends, polyvinyl chloride (PVC)/(polymethyl methacrylate (PMMA), polyhexyl methacrylate (PHMA), or polycaprolactone (PCL), and bisphenol A type polycarbonate (PC)/polystyrene (PS), in characterizing graded structures of the blends by measuring FTIR spectra, Raman

Table 1. Various Types of Functionally Graded Polymeric Materials

Types of Materials Used	Structure	Preparative Method	Size of Dispersion Phase
Metal(or ceramic)/ Polymer	Composites	<ul style="list-style-type: none"> • Laminate method • Electric field method • Centrifugation method • Flame spraying method 	Big
Polymer/Polymer	Immiscible polymer blend	<ul style="list-style-type: none"> • Surface inclination in melt state method • Surface inclination in solution method • Dissolution–diffusion method 	
	Miscible polymer blend	<ul style="list-style-type: none"> • Diffusion in melt method • Dissolution–diffusion method 	Molecular order
Atom–atom (intramolecules)	Copolymer (random)	<ul style="list-style-type: none"> • Diffusion method of monomer in polymer gel 	Atom order
	Copolymer (tapered)	<ul style="list-style-type: none"> • Living anion or radical polymerization method 	
	Density of cross-linking	<ul style="list-style-type: none"> • Changing method of cross-linking conc. 	
High-order structure (same polymer)		<ul style="list-style-type: none"> • Changing method of cross-linking temp. 	Same atoms and molecules
	Crystal structure	<ul style="list-style-type: none"> • Injection molding method 	

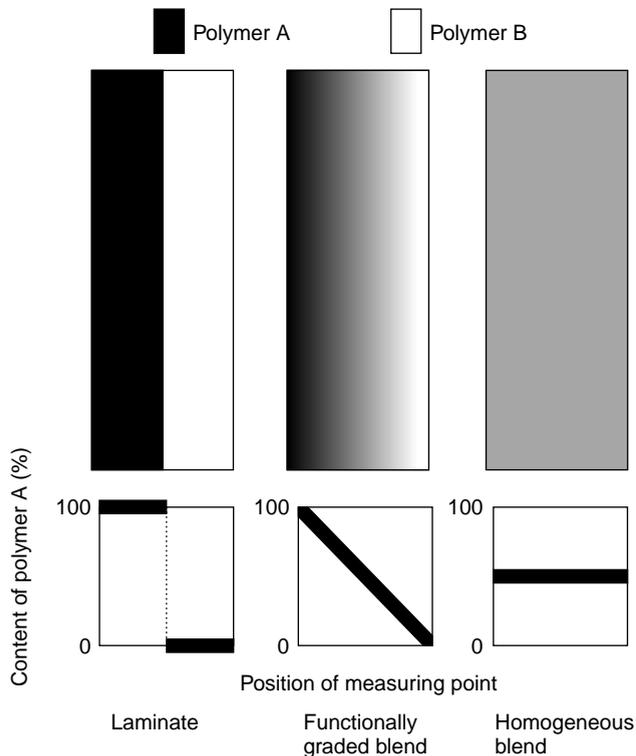


Figure 1. Schematic model of functionally graded blend.

microscopic spectra, thermal behaviors around the glass transition temperature (T_g) by the DSC method or by SEM-EDX (Scanning Electro Microscopy-Energy Dispersive X-ray Spectrometer) observation. Further, several types of functional properties, especially smart performance, are discussed, which result from the graded structure. Finally, the prospects of functionally graded polymer blends for applications are discussed.

MECHANISM OF DIFFUSION-DISSOLUTION METHOD

The mechanism of forming a graded structure is as follows. After a polymer B solution is poured on a polymer A film in a glass petri dish, polymer A begins to dissolve and diffuse in the solution to the air side (Fig. 2), but the diffusion is interrupted when all of the solvent evaporates. Thus, a blend film is produced that consists of a concentration gradient of polymer A/polymer B in the thickness direction.

Based on the steps of dissolution and diffusion of polymer A, the graded structures can be classified into three types (Fig. 3).

First Type. Polymer A begins to dissolve and then diffuses but does not yet reach the air side surface of the polymer B solution. The blend has three phases (polymer A, polymer B, and a thin graded structure).

Second Type. Just when all polymer A has finished dissolving, the diffusion frontier reaches to the air side surface of polymer B solution. The blend has one graded phase from

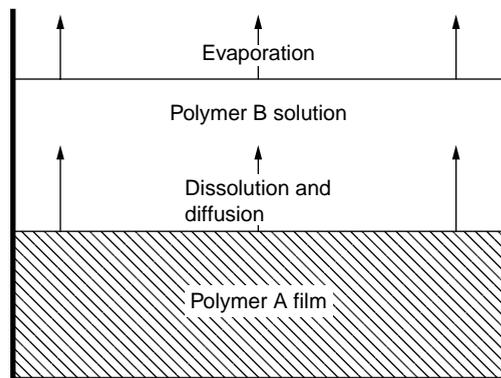


Figure 2. Schematic model of dissolution-diffusion method.

the surface to the other, and those surfaces are composed of polymer A only or polymer B only.

Third Type. After the dissolution and diffusion of polymer A reaches the air side surface of the polymer B solution, polymer A and polymer B molecules begin to mix with each other and become miscible. The concentration gradient begins to disappear.

The Formation of a concentration gradient depends on (1) the dissolution rate of polymer A in the polymer B solution, (2) the diffusion rate of polymer A in the polymer B solution, and (3) the interrupted time of the diffusion due to the completion of solvent evaporation. The factors that control these phenomena are (1) the type of solvent, (2) the casting temperature, (3) the molecular weight of polymer A, and (4) the amount of polymer B solution.

Until polymer A completely dissolves or reaches the surface of the polymer B solution in the formation of the first and second types of structure, the diffusion of polymer A in

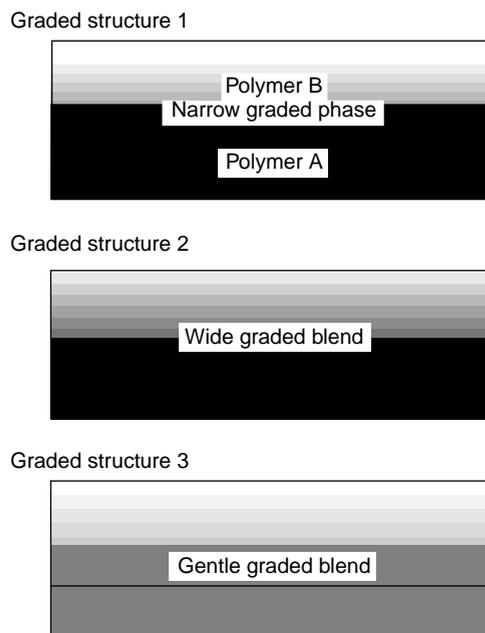


Figure 3. Schematic models of various types of graded structures.

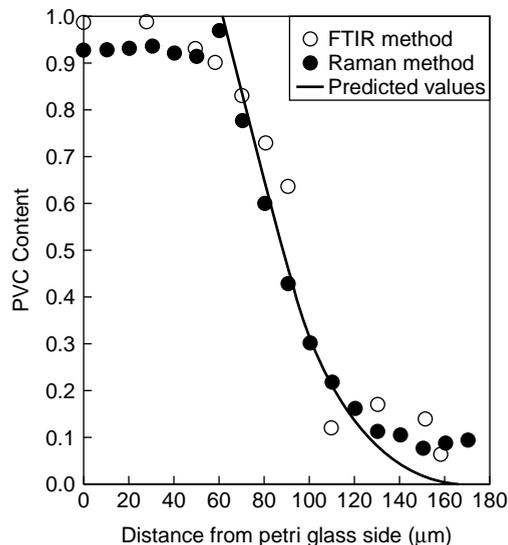


Figure 4. The change in PVC content in the thickness direction of PVC/PMMA graded blends.

the polymer B solution is considered to obey Fick's second law (Eq. 1) by assuming that the evaporation of the solvent from the polymer B solution can be neglected during diffusion:

$$\frac{\partial C_A}{\partial t} = D_{AB} \cdot \left(\frac{\partial^2 C_A}{\partial x^2} \right), \quad (1)$$

where C_A is the concentration of polymer A, t is time passed, x is the distance from the surface of the polymer A sheet, and D_{AB} is an apparent diffusion coefficient.

The point where C_A becomes one shifts to the petri glass side, as the dissolution of polymer A proceeds. Thus, by considering this effect and rearranging mathematically, Eq. (2), is obtained from Eq. (1):

$$C_A = \operatorname{erfc} \left(\frac{(x-b)}{2\sqrt{D_{AB}t}} \right), \quad (2)$$

$$\operatorname{erfc}(x) = \left(\frac{2}{\sqrt{\pi}} \right) \int_x^\infty \exp(-\xi^2) d\xi,$$

where b is the distance between the petri glass surface and the other side of the remainder of polymer A, which has not yet dissolved. Therefore, the gradient profile in the blend at t can be estimated from Eq. (2).

The fit of Eq. (2) to the experimental data was examined for the PVC/PMMA graded blend, and this is explained in detail in the next paragraph. The experimental data agreed approximately with the values predicted by Eq. (2), as shown in Fig. 4. D_{AB} and b were obtained as $6.38 \mu\text{m}^2/\text{s}$ and $57 \mu\text{m}$, respectively. The D_{AB} was much larger than the value in the "diffusion in melt state," and this means that this dissolution-diffusion method is very useful.

Further, a thicker and more excellently graded blend film can be prepared by the multiple step method, as illustrated in Fig. 5. Here, the graded blend was obtained by repeatedly changing the composition of the blend in the poured solution.

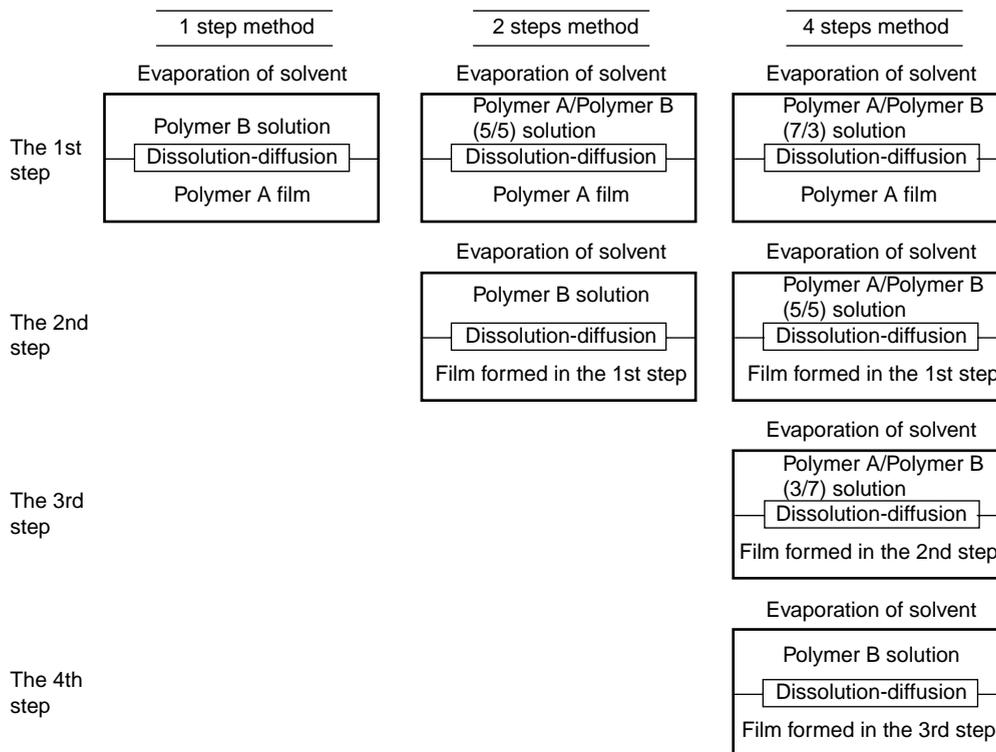


Figure 5. Schematic models of multiple step methods.

PREPARATION AND CHARACTERIZATION OF SEVERAL TYPES OF FUNCTIONALLY GRADED POLYMER BLENDS

Amorphous Polymer/Amorphous Polymer Miscible Blends

In the PVC/PMMA system (6,7), we prepared samples by changing the four controllable conditions: (1) the type of solvent, (2) the casting temperature, (3) the molecular weight of the PVC, and (4) the amount of the PMMA solution) and characterized the graded structures of the samples by FTIR-ATR, Raman microscopic spectroscopy, and DSC methods. Figure 6 shows the graded structure of the samples in the direction of thickness, measured by FTIR-ATR. In a similar blend that had graded structure 1, on a laminate, the PMMA content increased at 60% of the distance/thickness, and it was confirmed that it has a thin graded layer (about 10–20% of the distance/thickness). Then, in the blend of graded structure 3, the PMMA content was kept at about 50% in the entire range. However, in the blend of graded structure 2, the PMMA content gradually increased in the range from 0–100% of the distance/thickness. Thus, it was found that this blend had an excellently wide concentration gradient. Here, the PMMA content was estimated from the ratios of the absorption band intensities at 1728 cm^{-1} (stretching of the carbonyl group in PMMA) and 615 cm^{-1} (stretching of C–Cl bond in PVC). The change in PMMA content in the thickness direction of the blend film was estimated by measuring FTIR-ATR spectra on a sliced layer of the blend film.

The change in PVC content of the graded blend can be characterized by Raman microscopic spectroscopy method, similarly to the FTIR-ATR method, as shown in Fig. 4. Raman microscopic spectra were measured at the focused point, which was shifted by $10\text{ }\mu\text{m}$ from one surface area to the other. It was confirmed that the blend had a comparatively thick layer of a graded structure phase. This method is considered significantly useful because an easy and detailed estimate can be made for the graded profile of a blend.

Further, the graded structure was characterized by the DSC method. The DSC curve of the blend that has a widely graded structure (graded structure 2), shows more gradual steps around T_g than the others (Fig. 7). Similarly,

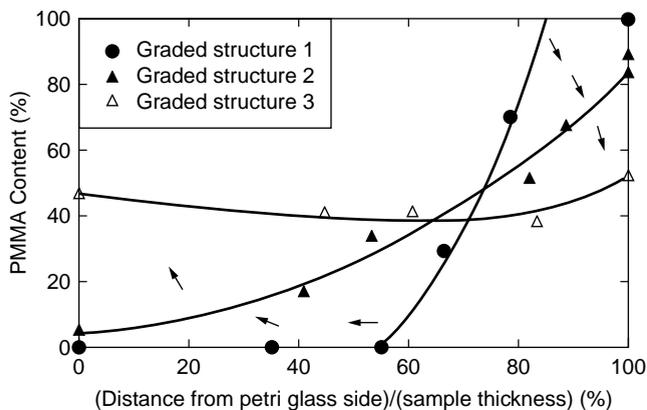


Figure 6. The change in PMMA content in the thickness direction of several types of PVC/PMMA graded blends.

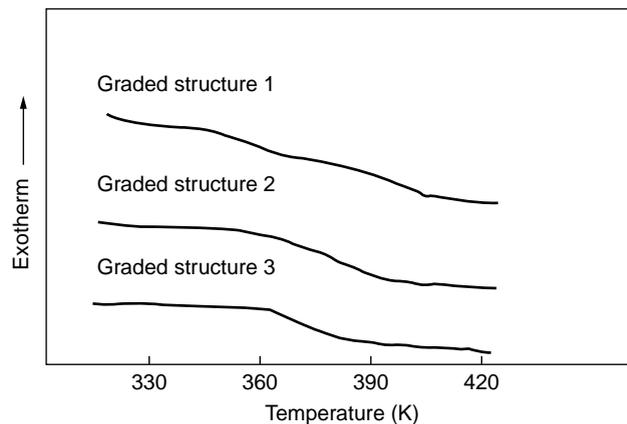


Figure 7. The DSC curves of several types of PVC/PMMA blends.

the structures of the samples, which were prepared under several types of conditions were investigated, and the optimum conditions (molecular weight of PVC: $M_n = 35600$, $MW = 60400$; type of solvent: THF/toluene(5/1); volume of solvent: 0.23 mL/cm^2 ; temperature: 333 K) were determined.

In the PVC/PHMA system (7), the graded structure of the sample could not be estimated by the FTIR-ATR and DSC methods, because PHMA was very soft at room temperature. Thus, the graded structure was measured by the SEM-EDX method (Fig. 8). The chlorine content in the sample increased gradually to the petri glass side, and then it was confirmed that it has a widely graded structure.

Further, the structures of the samples, which were prepared under several types of conditions were investigated, and the optimum conditions (molecular weight of PVC: $M_n = 35600$, $MW = 60400$; type of solvent: MEK; volume of solvent: 0.37 mL/cm^2 ; temperature: 313 K) were determined.

Amorphous Polymer/Crystalline Polymer Miscible Blends

In the PVC/PCL system (25), we obtained the optimum conditions for preparing a graded polymer blend that had a wider compositional gradient, similar to that of

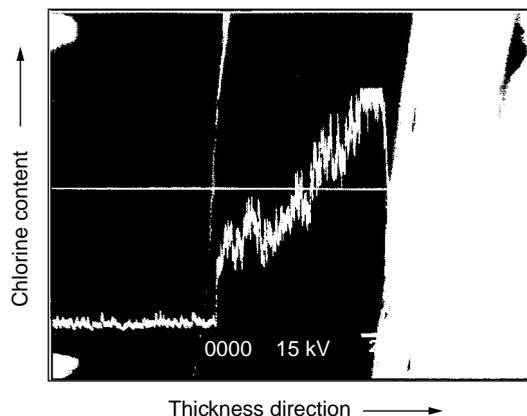


Figure 8. Chlorine content along the thickness of a PVC/PHMA graded blend (X 750, —; $20\text{ }\mu\text{m}$).

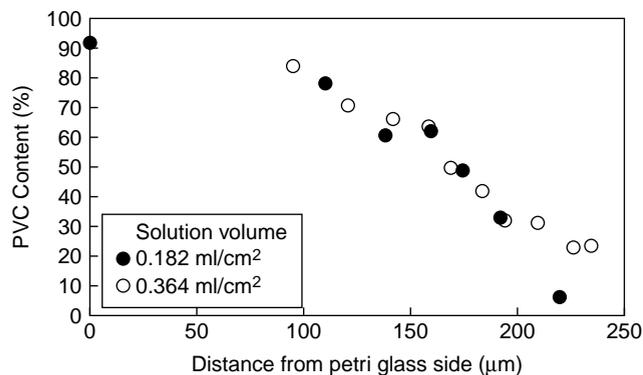


Figure 9. Graded structure of PVC/PCL graded blends.

the PVC/PMMA system. Figure 9 shows the PVC content of the samples in the direction of thickness, measured by the FTIR-ATR method. PVC began to decrease at about 70 μm from the petri glass side and decreased gradually until the surface of the air side, that is, about 240 μm away from the petri glass side in both solution volumes.

Then, the change of T_g in the thickness direction of the blend film was characterized by the DSC method (Fig. 10) for 0.364 mL/cm² of solution volume. T_g decreased at increasing distance from the petri glass side, similar to the PVC content. Thus, the graded structure in PVC content was confirmed by the graded profile in T_g .

Further, the change in PCL crystalline content was determined from the amount of heat diffusion of crystalline PCL, measured by the DSC method. The heat diffusion began to increase, after it was kept at zero until about 130 μm of the distance. Then, it increased immediately at about 180 μm . Thus, it was found that a graded structure in crystalline PCL was formed in the range from 130–240 μm of the distance. This means that the graded PVC/PCL blend obtained had both a gradient concentration of PVC and a gradient content of crystalline PCL, as shown in Fig. 11. The PCL content was about 30% at about 130 μm of the distance. This result indicates that crystalline PCL in the homogeneous PVC/PCL blend emerged at concentrations of more than 30% PCL (39). Then, it was concluded that the amorphous phase was made of a miscible amorphous

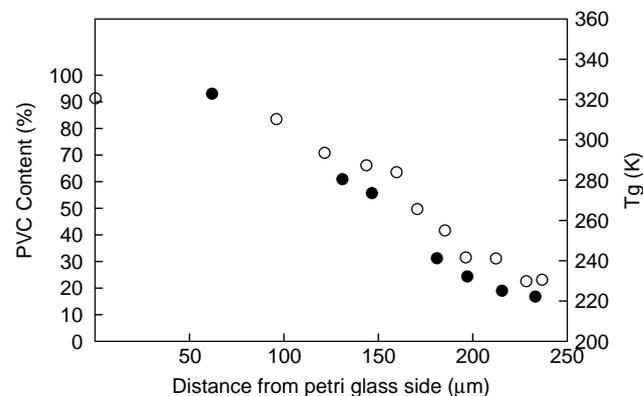


Figure 10. Graded structure of PVC/PCL graded blends (\bullet , T_g ; \circ , PVC content).

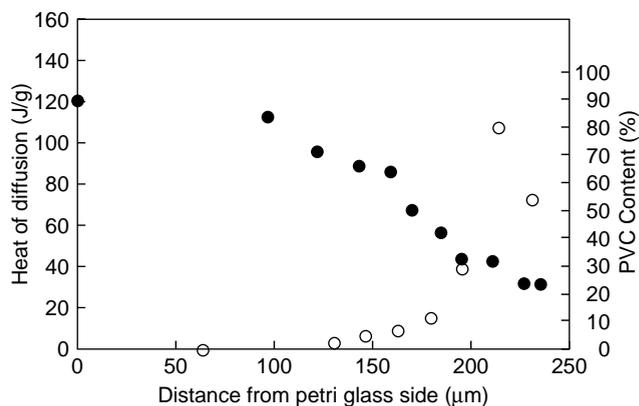


Figure 11. Graded structure of PVC/PCL graded blends (\bullet , PVC content; \circ , heat of diffusion).

PVC/amorphous PCL blend. Further, the PCL crystalline phase decreased again coming closer to the surface of the air side. It is thought that this phenomenon occurs because the formation of the amorphous phase is more thermodynamically stable than that of the crystalline phase. Therefore, it was believed that the graded structure of the PVC/PCL graded blend is as schematically illustrated in Fig. 12.

Amorphous Polymer/Amorphous Polymer Immiscible Blends

We attempted to prepare a graded PC/PS blend by the dissolution–diffusion method (24), similar to the PVC/PMMA system. In this case, PS solution was poured on PC film. However, we did not obtain a graded structure, but we did obtain a system of two homogeneous layers which were composed of about 50% and 0–10% PC, as shown in Fig. 13. Then, macrophase separation was observed in the former layer. It is believed that this results because of only three factors, the dissolution rate, diffusion rate, and evaporation time affect the process of forming a graded structure of a miscible blend. However, in forming a graded structure of an immiscible blend, three new factors, macrophase separation, surface inclination, and gravimetry, in addition to the former factors may significantly affect the process, as shown in Fig. 14. It is especially considered

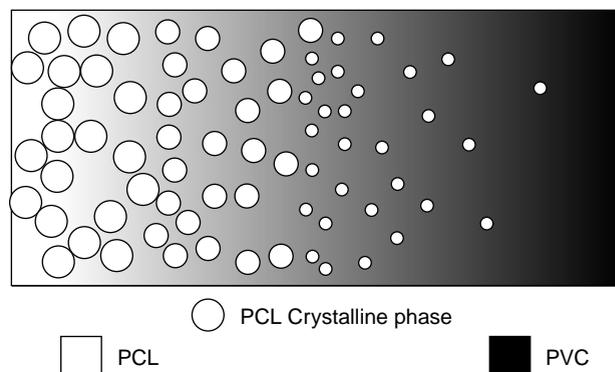


Figure 12. Schematic model of PC/PS graded blend.

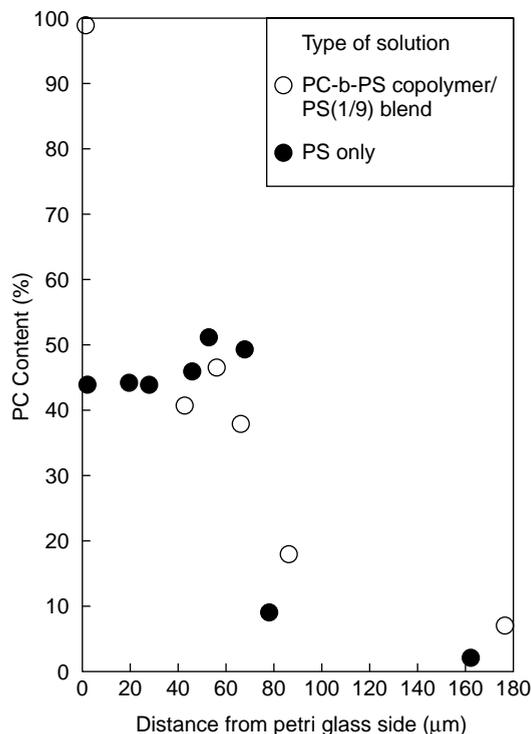


Figure 13. Graded structure of PC/PS graded blends with or without PS-b-PC block copolymer.

that macrophase separation may break a strongly graded structure on the way to formation, because it is concentrated by evaporation of solvent.

Thus, we attempted to protect the formation process of the graded structure from macrophase separation by adding a PS-b-PC copolymer (40) to the PS solution (PS-b-PC copolymer/PS = 1/9). Here, the PC segment content in the block copolymer was 46% (NMR measurement). It was found that the widely graded structure obtained in the

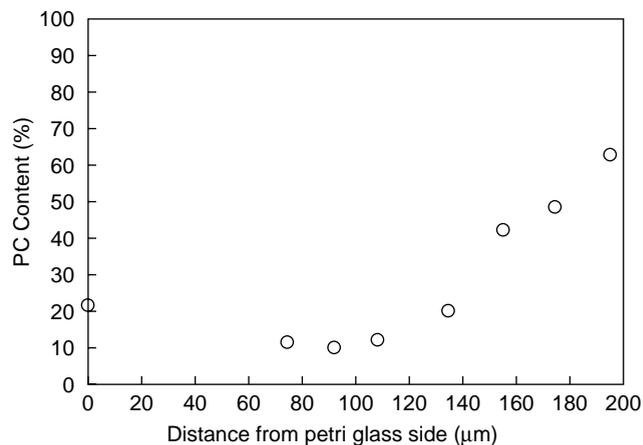


Figure 15. Graded structure of PC/PS graded blends when PC solution was poured on a PS film.

PC/PS blend was formed in the distance range of 0–100 μm from the petri glass side (Fig. 13).

Furthermore, we attempted to prepare a graded PC/PS blend by pouring a PC solution containing the block copolymer on the PS film. Figure 15 shows the change in PC content in the direction of the film thickness. The formation of a widely graded structure was confirmed at a long far distance from, and also, close to the petri glass side. This result was considered to mean that factor of surface inclination significantly influenced the formation of a graded structure.

Therefore, the graded immiscible PC/PS blend was obtained by adding a PC-b-PS copolymer. It is believed that the graded structure of the PC/PS graded blend is as schematically illustrated in Fig. 16.

FUNCTIONAL AND SMART PERFORMANCES AND THE PROSPECT FOR APPLICATION

Functional and Smart Performance

It was found in our study (6,7) that graded polymer blends had several types of functional properties, including smart performance. Thus, the functional properties of a PVC/PMMA blend that contains graded structure 2 (an extremely widely graded concentration) were explained by

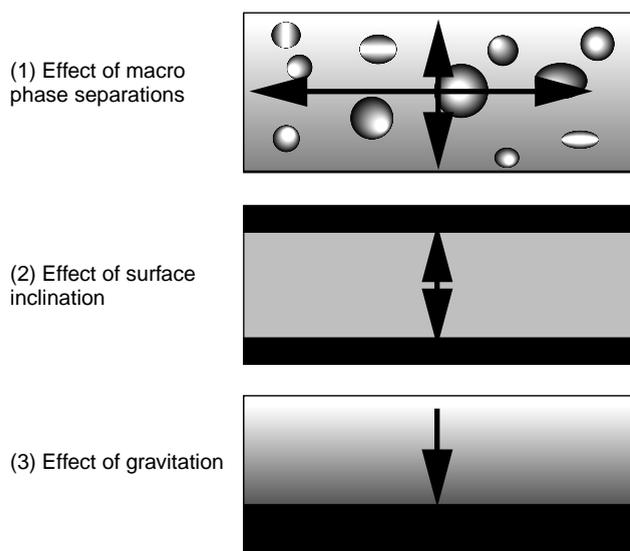


Figure 14. The other factors that affect the formation of a graded structure in an immiscible blend.

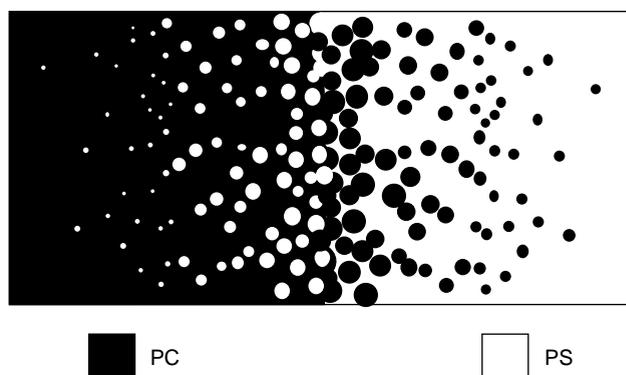


Figure 16. Schematic model of PC/PS graded blend.

comparison with those of a blend that contains graded structure 1 (similar to a laminate system), perfectly miscible blends (1/1), PVC only, and PMMA only.

Tensile Properties. The tensile properties of PVC, PMMA, a perfectly miscible blend (1/1), blends that have graded structures 1 and 2 (blend types 1 and 2), in the vertical direction of thickness are summarized in Table 2. The tensile strength of the homogeneously miscible blend is the highest, and the next is blend type 2, surpassing PVC, PMMA, and blend type 1. This phenomenon means that formation of a graded structure suppresses a break at the interface and also gives properties superior to the source materials. It is believed that this occurs because the blend phase that has the concentration gradient has a sufficiently high tensile strength. For elongation at break, blend type 2 appeared sufficiently good. The tensile modulus of blend type 2 is higher than that of blend type 1. It was found, thus, that the break in tensile stress could be suppressed by formation of a concentration gradient.

Thermal Shock Resistance. Thermal shock resistance was tested by moving the specimens from a box to another (kept at 253 K and 373 K) repeatedly (5 times) every 30 min. The specimens were then evaluated for thermal shock resistance by measuring a maximum angle of warp, as illustrated in Fig. 17, and adhesive strength in shear by tension loading.

The thermal shock resistance of blend type 2 that had a graded structure 2 was tested and those results (maximum value of warp angle and adhesive strength in shear by tension loading) were compared with those of blend type 1 that had graded structure 1, as shown in Table 2.

The film of blend type 1 was highly warped, whereas that of blend type 2 almost did not warp. The adhesive

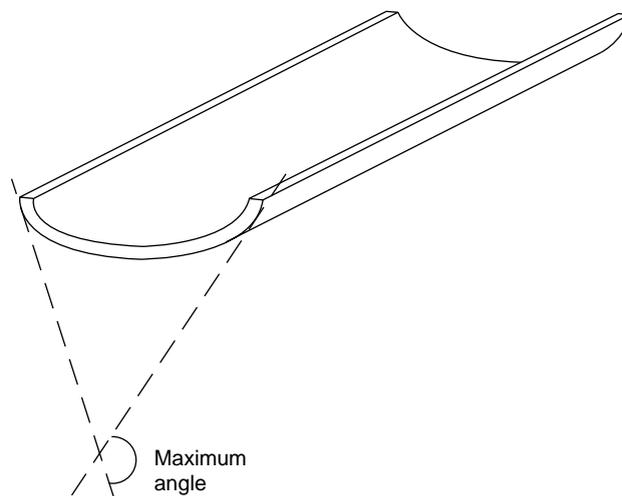


Figure 17. Method of measuring maximum angle of warp.

strength in shear by tension loading of blend type 2 was higher than that of blend type 1. It is believed that it occurs because the differences in the expansion of PVC (rubber state) and PMMA (glass state) at high temperature (395 K) concentrated the warp stress at the interface and decreased the strength of the interface. However, in blend type 2, the phase containing an excellently wide concentration gradient prevented the warp stress from concentrating. Thus, the thermal shock resistance of the blend (blend type 2) that has an excellently wide concentration gradient was superior to that of the similar blend (blend type 1) on a laminate film. It was found that the formation of an excellently wide concentration gradient improved the strength of the interface.

Table 2. Properties of PVC/PMMA Functionally Graded Blends

		PVC/PMMA Blend			PVC	PMMA
		Type 2 ^a	Type 1	P.M.T ^b		
<i>Tensile Properties</i>						
Tensile strength	(kgf/mm ²)	6.4	4.5 ^c	7.2	5.7	6.1
Elongation at break	(%)	4.5	2.8 ^c	5.2	3.9	3.1
Tensile modulus of elasticity	(kgf/mm ²)	200	190 ^c	220	230	230
<i>DMA Properties (Tensile Mode)</i>						
T_g width of storage modulus	(K)	20	8.6,11 ^c	11	—	—
Half temperature width T_g in $\tan \delta$	(K)	16	—	10	—	—
<i>Thermal Shock Resistance</i>						
Maximum warp angle	(°)	9	170 ^d	—	—	—
Adhesive strength in shear by tension loading	(kgf)	98	71 ^d	—	—	—

^aBlend containing graded structure 2.

^bPerfectly miscible blend.

^cPrepared by the hot press method.

^dBlend containing graded structure 1.

Smart Performance(DMA Properties). The change in tensile storage modulus and $\tan \delta$ of PVC/PMMA blend type 2 that has a wide concentration gradient around T_g was compared with the perfectly miscible blend (1/1) by a DMA measurement (rate of temperature increase: 1 K/min; frequency: 0.2 Hz). Then, the T_g width of storage modulus and half temperature of the T_g width of $\tan \delta$ were estimated, as shown in Table 2.

The half width of temperature of $\tan \delta$ for the former (16 K) was significantly larger than that of the latter (10 K). Thus, it was confirmed that blend type 2 has a continuous phase because of its wide range of T_g . Thus, $\tan \delta$ of the graded blends of PVC and several types of polyalkyl methacrylate(PMA) that contain graded structure 2 were measured, as shown in Fig. 18. $\tan \delta$ of the graded PVC/PHMA blend had the widest temperature range. Thus, it was confirmed that the wide temperature range is caused by the larger difference of the T_g in the polymer pairs of the graded PVC/PHMA blend.

Further, we investigated the optimum conditions for preparing a graded PVC /PHMA blend that had a wider temperature range of $\tan \delta$. Then, we obtained the PVC/PHMA blend that contained an excellently graded structure 2, which showed a peak of $\tan \delta$ in a much wider temperature range compared with those of a blend that contained graded structure 1 and a perfectly miscible blend (1/1), as shown in Fig. 19.

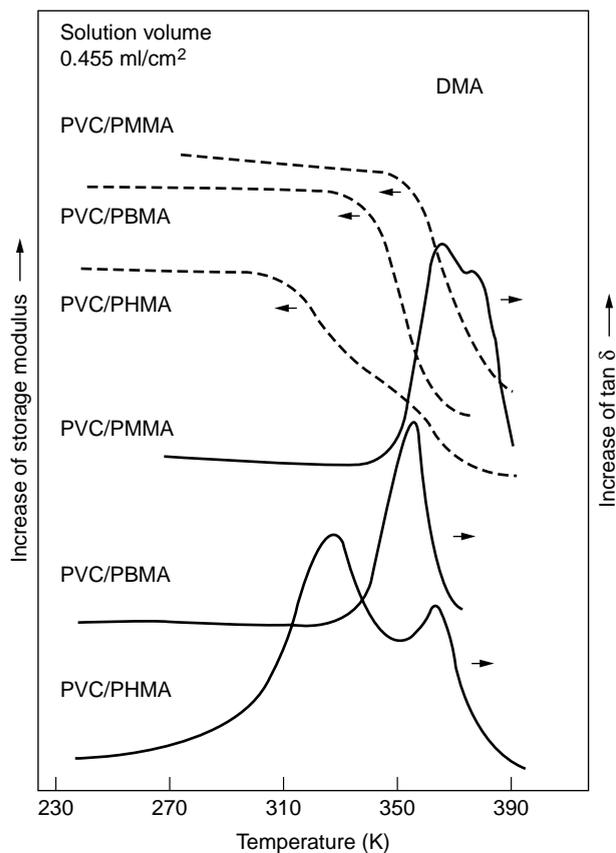


Figure 18. DMA data for PVC/PMA graded blend.

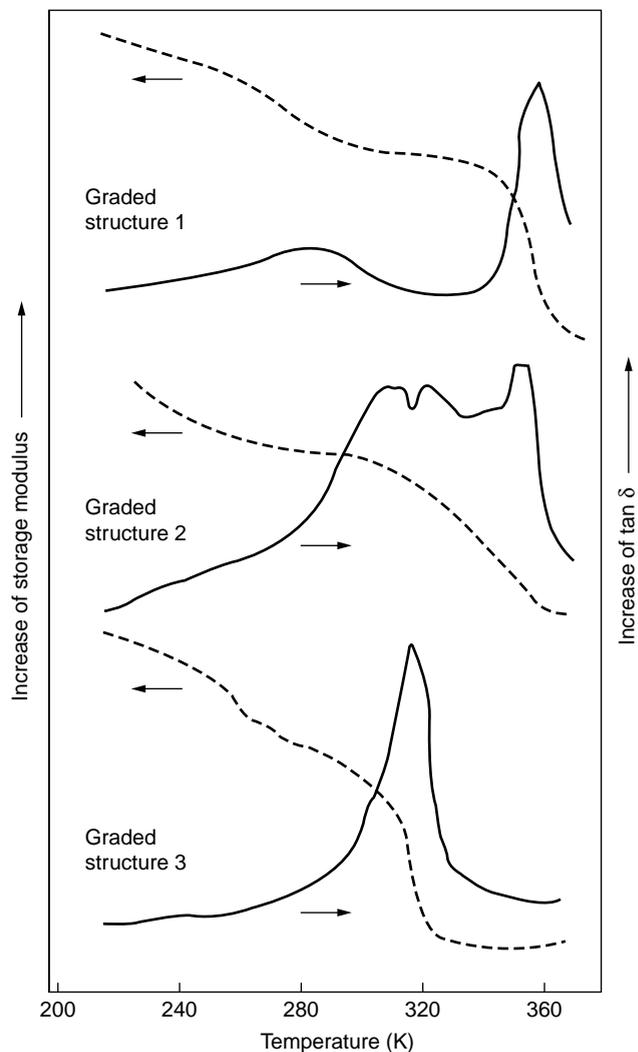


Figure 19. DMA data for PVC/PHMA graded blends.

Further, in both PVC/PMMA and PVC/PHMA blends, the tensile storage modulus of blend type 2 that contained an excellently graded structure 2 began to decrease at a lower temperature than that of a perfectly miscible blend (1/1) and did not have a terrace, whereas that of a similar blend that contained graded structure 1 on a laminate had some terraces.

Sandwiched steel beams combined by a polymer are used for damping materials (41), and it is known that the damping efficiency shows a maximum in the temperature range, at which the polymer used has a peak of $\tan \delta$. Then, it is expected that an excellently graded blend that has a peak of $\tan \delta$ in a much wider temperature range will be useful as a damping material in a wide temperature range. Graded polymer blends can be used as smart materials based on the following principle.

An excellently graded blend was used as the polymer that combined the steel plates shown at the right in Fig. 20. The T_g of the graded blend decreases with a shift from left to the right side of the figure. At the highest temperature, that is, the same temperature as the higher T_g of the polymer pairs in the blend, the area at the farthest left

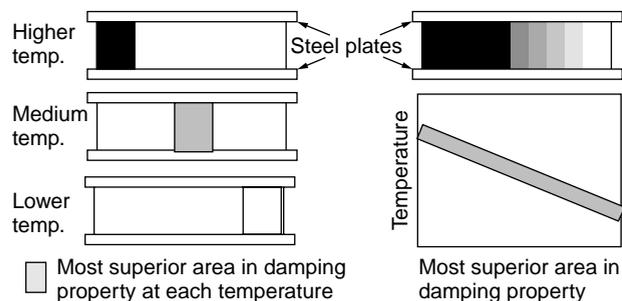


Figure 20. Schematic model of so-called smart performance in the damping property of a steel plates combined by a functionally graded blend.

side shows the best damping performance. And then, the area shifts to the right side as the temperature decreases. Finally, at the lowest temperature, that is, the same temperature as the lower T_g of the polymer pair in the blend, the area at the farthest right side shows the best damping performance. Therefore, the area that shows high damping performance shifts as the temperature changes. This performance is considered one of the so-called smart performances.

The Prospects for Application of Functionally Graded Blends

Functionally graded polymer blends are expected to be used in place of laminates because of their superior strength and thermal shock resistance. The superiority results from the lack of an interface that suppresses the break at the interface and thermal stress. Further, the excellently wide compositional gradient results in a graded structure that has several types of improved physical properties. Therefore, new functional performance is expected because of these physical property gradients that can be applied in the various fields shown in Table 3.

Table 3. Possibility of Applications of Functionally Graded Polymer Blends

Expected Functional Property	Application
Relaxation of thermal stress	<ul style="list-style-type: none"> Mechanical device for antiabrasion Sporting goods Construction materials
Prevention of vibration and sound	<ul style="list-style-type: none"> Vibration and sound proofing
Electromagnetic materials	<ul style="list-style-type: none"> Electromagnetic shield
Photo materials	<ul style="list-style-type: none"> Copy machine device Optic fiber Lens
Medical materials	<ul style="list-style-type: none"> Artificial internal organs Artificial blood vessels and organs
Packing materials	<ul style="list-style-type: none"> Waterproof adhesive
Chemical	<ul style="list-style-type: none"> Chemical resistance material

BIBLIOGRAPHY

- Society of Functionally Graded Material, ed., *Functionally Graded Materials*, Kogyo Chyosakai, Tokyo, 1993.
- T. Kitano, *Kogyo Zairyo* **43**(6), 112 (1994).
- M. Takayanagi, *23rd Colloq. Struct. Property Polym.*, Tokyo, 1993.
- Y. Agari, *Funct. Mater.* **16**(4), 32 (1996).
- Y. Agari, *Kobunshi Kako* **46**(6), 251 (1997).
- Y. Agari, M. Shimada, A. Ueda, and S. Nagai, *Macromol. Chem. Phys.*, **197**, 2017 (1996).
- Y. Agari, M. Shimada, A. Ueda, T. Anan, R. Nomura, and Y. Kawasaki, *Funct. Graded Mater.*, 1996, p. 761 (1997).
- Y. Agari, M. Shimada, M. Ueda, R. Nomura, and Y. Kawasaki, *Polym. Prepr.* **47**, 701 (1998).
- Y. Agari, M. Shimada, H. Shirakawa, R. Nomura, and Y. Kawasaki, *Polym. Prepr.* **48**, 698 (1999).
- J.Z. Yu, C. Lei, and F.K. Ko, *Soc. Plast. Eng. [Tech. Pap.]* **52**, 2352 (1994).
- M. Omori, A. Okubo, K. Gilhwan, and T. Hirai, *Funct. Graded Mater.*, 1996, p. 764 (1997).
- M. Funabashi and T. Kitano, *Seni-Gakkaishi* **50**(12), 573 (1994).
- C.M. Thai, T. Kato, and A. Yoshizumu, *J. Thermoset. Plas. Jpn.*, **16**(3), 126 (1995).
- Xu.M. Xie, M. Matsuoka, and K. Takemura, *Polymer* **33**(9), 1996 (1992).
- S. Kanayama and T. Umemura, *J. Jpn. Soc. Polym. Proc.*, **2**(4), 216 (1995).
- Y. Kano, S. Akiyama, H. Sano, and H. Yui, *J. Electron Microsc.* **44**(5), 344 (1995).
- S. Akiyama and Y. Kano, *Kagaku to Kogyo (Osaka)* **71**, 44 (1997).
- S. Murayama, S. Kuroda, and Z. Osawa, *Polymer* **34**(18), 3893 (1993).
- E. Jabbari and N.A. Peppas, *Macromolecules* **26**, 2175 (1993).
- P.F. Nealey, R.E. Cohen, and S. Argon, *Macromolecules* **27**, 4193 (1994).
- K.C. Farinas, L. Doh, S. Venkatraman, and R.O. Potts, *Macromolecules* **27**, 5220 (1994).
- T.E. Shearmur, A.S. Clough, D.W. Drew, M.G.D. van der Grinten, and R.A.L. Jones, *Macromolecules* **29**, 7269 (1996).
- M.A. Parker and D. Vesely, *J. Polym. Sci., Part B* **24**, 1869 (1986).
- Y. Agari, M. Shimada, A. Ueda, T. Koga, R. Nomura, and Y. Kawasaki, *Polym. Prepr., Jpn.* **45**, 2241 (1996).
- Y. Agari, M. Shimada, A. Ueda, T. Koga, R. Nomura, and Y. Kawasaki, *Polym. Prepr., Jpn.* **46**, 657 (1997).
- M. Kryszewski and G. Czeremuszkina, *Plaste Kautsch.* **11**, 605 (1980).
- P. Milczarek and M. Kryszewski, *Colloid Polym. Sci.* **265**, 481 (1987).
- Y. Koike, H. Hidaka, and Y. Ohtsuka, *Appl. Opt.* **22**, 413 (1983).
- Y. Koike, N. Tanio, E. Nihei, and Y. Ohtsuka, *Polym. Eng. Sci.* **29**(17), 1200 (1989).
- C.F. Jasso and E. Mendizabal, *Soc. Plast. Eng. [Tech. Pap.]* **50**, 2352 (1992).
- S. Ashai, *Polym. Prepr., Jpn.* **27**, 18 (1978).
- S. Ashai, *6th Symp. Funct. Graded Mater.*, p. 61 (1993).
- S. Ashai, *Kagaku to Kogyo (Osaka)* **71**, 50 (1997).

34. Y. Tsukahara, N. Nakamura, T. Hashimoto, and H. Kawai, *Polym. J.* **12**(12), 455 (1980).
35. D. Greszta, K. Matsuoka, and K. Matyaszewski, *Am. Chem. Soc., Polym. Repr. ACS* **37**, 569 (1996).
36. M. Furukawa, T. Okazaki, and T. Yokoyama, *Polym. Prepr., Jpn.* **45**, 2239 (1996).
37. G.B. Park, M. Hirata, Y. Kagari, T. Matsunaga, Jianping Gong, Y. Osada, and D.C. Lee, *Polym. Prepr., Jpn.* **45**, 1836 (1996).
38. Y. Ulcer, M. Cakmak, and C.M. Hsiung, *J. Appl. Polym. Sci.* **60**(1), 125 (1996).
39. Y. Agari and A. Ueda, *J. Polym. Sci., Part B* **32**, 59 (1994).
40. M. Shimada, Y. Agari, and Y. Makimura, *Polym. Prepr., Jpn.* **45**, 1958 (1996).
41. D.J. Mead, *J. Sound Vib.* **83**, 363 (1982).

POLYMERS, BIOTECHNOLOGY AND MEDICAL APPLICATIONS

I. YU. GALAEV
B. MATTIASSON
Lund University
Lund, Sweden

INTRODUCTION

Life is polymeric in its essence. The most important components of living cell, proteins, carbohydrates, and nucleic acids are polymers. Even lipids, which have lower molecular weights, can be regarded as methylene oligomers that have a polymerization degree around 20. Nature uses polymers as constructive elements and parts of complicated cell machinery. The salient feature of functional biopolymers is their all-or-nothing or at least highly nonlinear response to external stimuli. Small changes happen in response to varying parameters until the critical point is reached; then a transition occurs in the narrow range of the varied parameter, and after the transition is completed, there is no significant further response of the system. Such nonlinear response of biopolymers is warranted by highly cooperative interactions. Despite the weakness of each particular interaction in a separate monomer unit, these interactions, when summed through hundreds and thousands of monomer units, provide significant driving forces for the processes in such systems.

Not surprisingly, understanding the mechanism of cooperative interactions in biopolymers has opened the floodgates for attempts to mimic the cooperative behavior of biopolymers in synthetic systems. Recent decades witnessed the appearance of synthetic functional polymers, which respond in some desired way to a change in temperature, pH, electric, or magnetic fields or some other parameters. These polymers were nicknamed stimuli-responsive. The name "smart polymers" was coined due to the similarity of the stimuli-responsive polymers to biopolymers (1). We have a strong belief that nature has always striven for smart solutions in creating life. The goal of scientists is to mimic biological processes, and therefore understand them better, and also to create novel species and invent new processes.

The applications of smart polymer in biotechnology and medicine are discussed in this article. The highly nonlinear response of smart polymers to small changes in the external medium is of critical importance for the successful functioning of a system. Most applications of smart polymers in biotechnology and medicine include biorecognition and/or biocatalysis, which take place principally in aqueous solutions. Thus, only water-compatible smart polymers are considered; smart polymers in organic solvents or water/organic solvent mixtures are beyond the scope of the article. The systems discussed in the article are based on either soluble/insoluble transition of smart polymers in aqueous solution or on the conformational transition of macromolecules physically attached or chemically grafted to the surface. Systems that have covalently cross-linked networks of macromolecules, called smart hydrogels, are not considered.

One could define smart polymers used in biotechnology and medicine as *macromolecules that undergo fast and reversible changes from hydrophilic to hydrophobic microstructure triggered by small changes in their environments. These microscopic changes are apparent at the macroscopic level as precipitate formation in solutions of smart polymers or changes in the wettability of a surface to which a smart polymer is grafted. The changes are reversible, and the system returns to its initial state when the trigger is removed.*

SMART POLYMERS USED IN BIOTECHNOLOGY AND MEDICINE

The highly nonlinear transitions in smart polymers are driven by different factors, for example, neutralization of charged groups by either a pH shift (2) or the addition of an oppositely charged polymer (3), changes in the efficiency of hydrogen bonding and an increase in temperature or ionic strength (4), and critical phenomena in hydrogels and interpenetrating polymer networks (5). The polymer systems that have highly nonlinear response can be divided into three general groups: pH-sensitive smart polymers, thermosensitive smart polymers, and reversibly cross-linked networks.

pH-Sensitive Smart Polymers

The first group of smart polymers consists of polymers whose transition between the soluble and insoluble state is created by decreasing the net charge of the polymer molecule. The net charge can be decreased by changing the pH to neutralize the charges on the macromolecule and hence to reduce the hydrophilicity (increase the hydrophobicity) of the macromolecule. Copolymers of methylmethacrylate (hydrophobic part) and methacrylic acid (hydrophilic at high pH when carboxy groups are deprotonated but more hydrophobic when carboxy groups are protonated) precipitate from aqueous solutions by acidification to pH around 5, and copolymers of methyl methacrylate (hydrophobic part) with dimethylaminethyl methacrylate (hydrophilic at low pH when amino groups are protonated but more hydrophobic when amino groups are deprotonated) are soluble at low pH but precipitate in

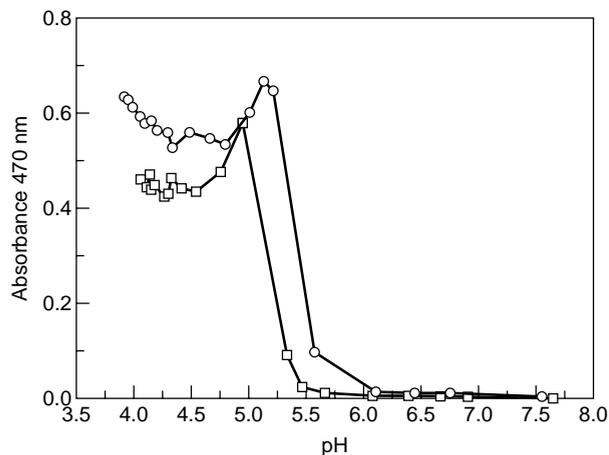


Figure 1. pH-induced precipitation of a random copolymer of methacrylic acid and methacrylate (commercialized as Eudragit S 100 by Röhm Pharma GMBH, Weiterstadt, Germany) (open squares) and *p*-amino-phenyl- α -D-glucopyranoside-modified copolymer (open circles) measured as turbidity at 470 nm. Some decrease in turbidity at lower pH values is caused by flocculation and sedimentation of the polymer precipitate [redrawn from (8)].

slightly alkaline conditions (6). Hydrophobically modified cellulose derivatives that have pending carboxy groups, for example, hydroxypropyl methyl cellulose acetate succinate are also soluble in basic conditions but precipitate in slightly acidic media (7).

The pH-induced precipitation of smart polymers is very sharp and usually requires a change in pH of not more than 0.2–0.3 units (Fig. 1). When some carboxy groups

are used to couple a biorecognition element, for example, noncharged sugar, the increased hydrophobicity of the copolymer results in precipitation at a higher pH (8). The copolymerization of *N*-acryloyl sulfametazine with *N,N*-dimethylacrylamide results in a pH-sensitive polymer whose reversible transition is in the physiological range of pH 7.0–7.5 (9).

The charges on the macromolecule can also be neutralized by adding an efficient counterion, for example, a low molecular weight counterion or a polymer molecule of opposite charges. The latter systems are combined under the name of polycomplexes. The cooperative nature of interaction between two polymers of opposite charges makes polycomplexes very sensitive to changes in pH or ionic strength (10). The complex formed by poly(methacrylic acid) (polyanion) and poly(*N*-ethyl-4-vinyl-pyridinium bromide) (polycation) undergoes reversible precipitation from aqueous solution at any desired pH value in the range 4.5–6.5 that depends on the ionic strength and polycation/polyanion ratio in the complex (Fig. 2) (11). Polyelectrolyte complexes formed by poly(ethylene imine) and poly(acrylic acid) undergo soluble–insoluble transition in an even broader pH range from pH 3–11 (12).

The pH of the transition of pH-sensitive polymers such as poly(methylmethacrylate-*co*-methacrylic acid) or poly(*N*-acryloyl sulfametazine-*co*-*N,N*-dimethylacrylamide) is strictly fixed for the given composition of comonomers. Thus, a new polymer should be synthesized for each desired pH value. The advantage of polyelectrolyte complexes is that by using only two different polymers and mixing them in different ratios, reversible precipitation can be achieved at any desired pH value in a rather broad pH-range.

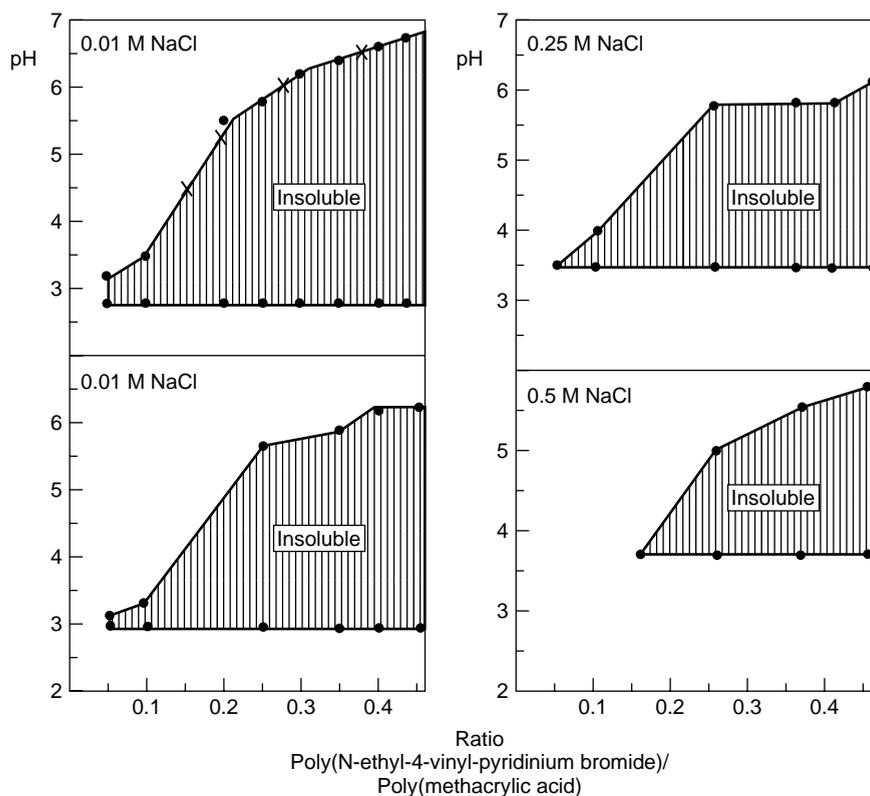


Figure 2. Phase diagram for the polyelectrolyte complex formed by poly(*N*-ethyl-4-vinyl-pyridinium bromide) (polymerization degree 530) and poly(methacrylic acid) (polymerization degree 1830). The dots (present pH values at which the turbidity of the polymer solutions was first observed at 470 nm. Ionic strength was 0.01 M NaCl (a), 0.1 M NaCl (b), 0.25 M NaCl (c) and 0.5 M NaCl (d). Dashed area represents pH/composition range where the complex is insoluble [reproduced from (11) with permission].

Thermosensitive Smart Polymers

The reversible solubility of thermosensitive smart polymers is caused by changes in the hydrophobic-hydrophilic balance of uncharged polymers induced by increasing temperature or ionic strength. Uncharged polymers are soluble in water due to hydrogen bonding with water molecules. The efficiency of hydrogen bonding lessens as temperature increases. The phase separation of a polymer occurs when the efficiency of hydrogen bonding becomes insufficient for the solubility of macro-molecule.

When the temperature of an aqueous solution of a smart polymer is raised above a certain critical temperature (which is often referred to as the transition temperature, lower critical solution temperature (LCST), or "cloud point"), phase separation takes place. An aqueous phase that contains practically no polymer and a polymer-enriched phase are formed. Both phases can be easily separated by decanting, centrifugation, or filtration. The temperature of the phase transitions depends on the polymer concentration and molecular weight (MW) (Fig. 3) (13,14). The phase separation is completely reversible, and the smart polymer dissolves in water when the temperature is reduced below the transition temperature.

Two groups of thermosensitive smart polymers are most widely studied and used:

- Poly(*N*-alkyl substituted acrylamides) and the most well-known of them, poly(*N*-isopropyl acrylamide) (poly(NIPAAM)), whose transition temperature is 32°C (14), and
- Poly(*N*-vinylalkylamides) such as poly(*N*-vinylisobutyramide) whose transition temperature is

39°C (15) or poly(*N*-vinyl caprolactam) whose transition temperature is 32–33°C (depending on the polymers molecular weight) (13)

A variety of polymers that have different transition temperatures from 4–5°C for poly(*N*-vinyl piperidine) to 100°C for poly(ethylene glycol) are available at present (16).

pH-sensitive smart polymers usually contain carboxy or amino groups that can be used for covalent coupling of biorecognition or biocatalytic elements (ligands). Thermosensitive polymers, on the contrary, do not have inherent reactive groups which could be used for ligand coupling. Thus, copolymers that contain reactive groups can be synthesized. *N*-Acryloylhydroxysuccinimide (17) or glycidyl methacrylate (18) have often been used as active comonomers in copolymerization with NIPAAM allowing further coupling of amino-group-containing ligands to the synthesized copolymers. The use of an initiator of polymerization (19) or chain transfer agent (20) that has an active group results in a polymer modified only at the end of the macromolecule. An alternative strategy is to incorporate a polymerizable double bond into the ligand, for example, by modification with acryoyl group, and then to copolymerize the modified ligand with NIPAAM (21,22).

An increase in the hydrophilicity of the polymer-accompanied incorporation of hydrophilic comonomers or coupling to hydrophilic ligands increases the transition temperature, whereas hydrophobic comonomers and ligands have the opposite effect (4). The pH-induced change in ligand hydrophobicity could have a dramatic effect on the thermoseparation of the ligand-polymer conjugate. A copolymer of NIPAAM and vinyl imidazole precipitates at about 35°C at pH 8.0 where imidazole moieties are non-charged and relatively hydrophobic, but no precipitation occurs even when heating the polymer solution to 80°C at pH 6 where imidazole groups are protonated and very hydrophilic (23).

Ligand-ligand interactions in a ligand-polymer conjugate also have a significant effect on the thermoseparation. The precipitation temperature for the previously mentioned copolymers of NIPAAM and vinyl imidazole increases as the imidazole content in the copolymer increases. On the contrary, the precipitation temperature decreases as the increase of imidazole content increases, when the polymer forms a Cu(II)-complex (23). Each Cu(II) ion interacts with two to three imidazole groups to cross-link the segments of the polymer molecule (24). The restricted mobility of the polymer segments results in a lower precipitation temperature.

Block copolymers that have a thermosensitive "smart" part that consists of poly(NIPAAM) form reversible gels on an increase in temperature, whereas random copolymers separate from aqueous solutions by forming a concentrated polymer phase (25). Thus, the properties of smart polymers that are important for biotechnological and medical applications could be controlled by the composition of comonomers and also by the polymer architecture.

The phase transition of thermosensitive polymers at increased temperature results from hydrophobic interactions between polymer molecules. Because hydrophobic interactions are promoted by high salt concentrations, the addition of salts shifts the cloud point to lower temperatures.

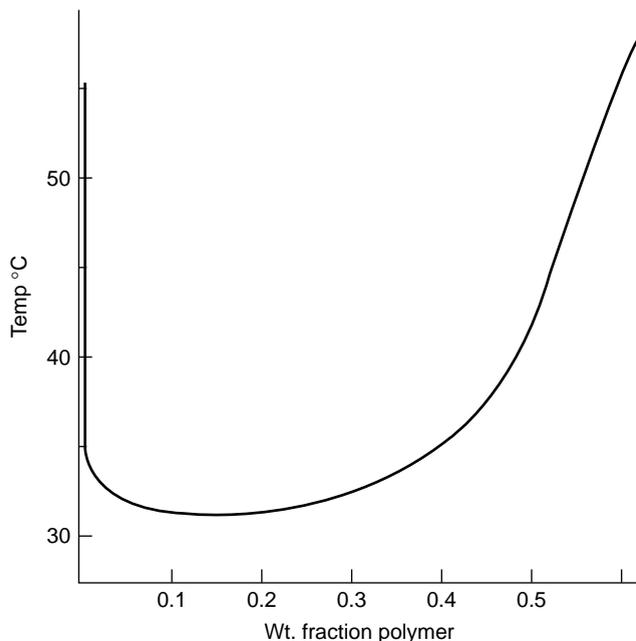


Figure 3. Phase diagram for poly(NIPAAM) in aqueous solution. The area under the binodal curve presents the range of temperatures/polymer concentrations for homogeneous solution. Separation into polymer-enriched and polymer-depleted phases takes place for any polymer concentration/temperature above the binodal curve [reproduced from (14) with permission].

When the transition temperature is below room temperature, polymer precipitation is achieved by just a salt addition without any heating. The addition of organic solvents, detergents, and chaotropic agents increases the transition temperature because these compounds deteriorate hydrophobic interactions.

Reversibly Cross-Linked Polymer Networks

Systems that have reversible noncovalent cross-linking of separate polymer molecules into a polymer network belong to the third group of smart polymers. When formed, reversibly cross-linked polymers either precipitate or form a physical gel. Polymers that have sugar ligands cross-linked by lectins with multibinding sites (26) and boronate-polyols (27–29) are the most widely used systems of this type. The reversible response in these systems is achieved by addition/removal of a low molecular weight analog of the polymer. For example, small sugars added at high concentrations compete with sugar-containing polymers for binding to lectin and destroy intrapolymer cross-links that result in disengagement of the network.

Heterogeneous Systems Using Smart Polymers

A solid surface acquires new properties when modified by adsorption or chemical grafting of smart polymers. Smart polymer that have terminal (only single-point attachment possible) or random (multipoint attachment possible) could be covalently coupled to the respective active groups on the surface (30). Single-point attachment could also be achieved by covalent modification of the surface using an initiator of polymerization and then carrying out polymerization of monomers in the solution that surrounds the support. The growth of polymer chains occurs only at the sites where initiator was coupled (31). Alternatively, the solid support is irradiated by light (32) or a plasma beam (33) when monomer is in the surrounding solution. Active radical sites on the surface, which appear as a result of irradiation, initiate the growth of polymer macromolecules. As a rule, irradiation methods give a higher density of grafted polymer, but polymerization is less controlled as in covalent coupling or using a covalently coupled initiator. Irradiation, especially at high monomer concentrations, could produce a cross-linked polymer gel attached to the solid support (34).

A separate group of smart polymers is represented by particulate systems. Liposomes that reversibly precipitate on salt addition and removal were prepared from a synthetic phospholipid that had a diacetylene moiety in the hydrophobic chain and an amino group in the hydrophilic head of the phospholipid, followed by polymerization of diacetylene bonds (35). Latices composed of thermosensitive polymers or a layer of thermosensitive polymer at the surface represent another example of insoluble but reversibly suspended particulate systems that respond to increasing/decreasing temperature (31).

APPLICATIONS

There are numerous potential applications for smart polymers in biotechnology and medicine. The main commercial

application of smart polymers is the production of “smart” pills where the shell of the smart polymer protects the pill from the harmful action of the stomach contents but allows the pill to dissolve in the intestine. There is not yet any other product on the market that applies smart polymers, but the interest in these applications is growing in both the academic community and industry. The following applications are considered in this article:

- smart pills that have an enteric coating
- smart polymers for affinity precipitation of proteins
- aqueous two-phase polymer systems formed by smart polymers and their application for protein purification
- smart surfaces for mild detachment of cultivated mammalian cells
- smart chromatographic matrices that respond to temperature
- smart polymers for controlled porosity of systems—“chemical valve”
- liposomes that trigger the release of their contents
- smart polymers for bioanalytical applications
- reversibly soluble biocatalysts

Smart Pills That Have an Enteric Coating

It is common knowledge that *peroral* introduction of medical preparations is the most convenient method compared to subcutaneous or intravenous injection and even to nasal sprays or eye droplets. The absorption of a swallowed pill takes place predominantly in the intestine and to reach the intestine the medicine must pass unharmed through the stomach that has a very low pH value of 1.4 and abundant hydrolytic enzymes that can degrade a broad variety of chemical structures. Many medicines are susceptible to damage in the stomach environment. The ideal condition for *peroral* introduction is to have a smart pill, which is insoluble in the stomach and hence passes through the stomach unaffected but easily dissolves at the higher pH in the intestine where the medicine is absorbed. Smart polymers provide the solution. Hydrophobic polymers such as poly(methylmethacrylate) or hydrophobically modified celluloses are insoluble in water *per se*, but the introduction of carboxy groups (either by partial hydrolysis of ester groups in methylmethacrylate or modification of cellulose HO groups by dicarboxylic acids such as succinic or phthalic acid) endows the polymers with pH-dependent solubility. The pill covered by a shell of such a polymer (enteric coating) is insoluble at low pH when the carboxy groups are protonated and uncharged, but easily soluble at a pH above 6 when carboxy groups are protonated and charged. Industrially produced polymers for enteric coating belong to two main groups, synthetic copolymers of methylmethacrylate and methacrylic acid and modified derivatives of cellulose, a natural polymer (Table 1). The first group of polymers is used mainly by European and U.S. manufacturers, and the second group is more popular in Japan.

Whenever the charge-bearing comonomer has an amino group instead of a carboxy group, the solubility of the polymer acquires opposite pH-dependence. The polymer is soluble at low pH values but insoluble in neutral

Table 1. Industrially Manufactured Smart Polymers for Producing Smart Pills

Polymer	Trade Name	Manufacturer	Country
Poly(methacrylic acid- <i>co</i> -methylmethacrylate) 1:1 monomer ratio, MW 135 000	Eudragit L	Röhm Pharma GmbH	U.S., Germany
Poly(methacrylic acid- <i>co</i> -methylmethacrylate) 1:2 monomer ratio, MW 135 000	Eudragit S	Röhm Pharma GmbH	U.S., Germany
Carboxymethylcellulose	CMEC	Freund Sangyo Co., Ltd.	Japan
Cellulose acetate phthalate	CAP	Wako Pure Chemicals Ltd.	Japan
Hydroxypropylmethylcellulose phthalate	HP-50, HP-55	Shin-Etsu Chemical Co., Ltd.	Japan
Hydroxypropylmethylcellulose acetate succinate	ASM, AS-H	Shin-Etsu Chemical Co., Ltd.	Japan
Poly(diethylaminoethyl methacrylate- <i>co</i> -methylmethacrylate) MW 150 000	Eudragit E	Röhm Pharma GmbH	U.S., Germany

and alkaline media. Poly(diethylaminoethylmethacrylate-*co*-methylmethacrylate) (commercialized as Eudragit E) is an example of such a polymer. The shell that is composed of this polymer protects the tablet against dissolution in the neutral saliva, and the mouth is not affected by the unpleasant taste of bitter medicine, but the polymer dissolves readily in the stomach.

Bioseparation—Affinity Precipitation

All bioseparation processes include three stages: preferential partitioning of target substance and impurities between two phases (liquid–liquid or liquid–solid), mechanical separation of the phases (e.g., separation of the stationary and mobile phases in a chromatographic column), and recovery of the target substance from the enriched phase. Because smart polymers can undergo phase transitions, they could facilitate the second and the third stages of bioseparation processes.

The ability of smart polymers to form *in situ* heterogeneous systems is exploited in affinity precipitation (Fig. 4). The technique is based on using a conjugate of a smart polymer that has a covalently coupled biorecognition moiety, that is, a ligand specific for a target protein. The conjugate forms a complex with the target protein but not with the other proteins in the crude extract. Phase separation of the complex is triggered by small changes in the environment resulting in transition of the polymer backbone into an insoluble state. The target protein specifically coprecipitates with the smart polymer, and the impurities in the crude remain in solution. Then, the target protein is either eluted from the insoluble macroligand–protein complex or the precipitate is dissolved. The protein is dissociated from the macroligand, and the ligand–polymer conjugate is precipitated again. Now without the protein that remains in the supernatant in purified form. A variety of different ligands such as triazine dyes, sugars, protease inhibitors, antibodies, nucleotides, double- or single-stranded DNA, and chelated metal ions were successfully used for affinity

precipitation (36). After elution of the target protein the ligand–polymer conjugate could be recovered and used in the next purification cycle (37).

Triazine dyes, robust affinity ligands for many nucleotide-dependent enzymes, were successfully used in conjugates with the pH-sensitive copolymer of methacrylic acid and methylmethacrylate which precipitates when pH decreases (Eudragit S 100) for purification of dehydrogenases from various sources by affinity precipitation (38,39). Sugar ligands constitute another attractive alternative and have been used in combination with Eudragit S 100 for bioseparation of lectins (40). Restriction endonuclease Hind III was successfully isolated using the thermosensitive conjugate of poly (NIPAAm) with phage λ DNA (21). Human IgG was specifically precipitated with a conjugate of protein A and galactomannan. Galactomannan polymer was reversibly precipitated by adding tetraborate (41).

The efficient precipitation of Cu(II)-loaded poly(*N*-vinylimidazole-*co*-NIPAAm) by high salt concentrations at mild temperature is very convenient for metal affinity precipitation of proteins that have inherent histidine residues at the surface or for recombinant proteins artificially provided with histidine tags (usually four to six residues). High salt concentration does not interfere with protein–metal ion–chelate interaction, and, on the other hand, it reduces the possibility of nonspecific binding of foreign proteins to the polymer both in solution and when precipitated (23). The flexibility of polymer chains in solution allows several imidazole ligands on a polymer molecule to come close enough to interact with the same Cu(II) ion and thus to provide sufficient strength of polymer–Cu(II) interactions to purify a variety of histidine-containing proteins (37).

Polyelectrolyte complexes that have pH-dependent solubility were successfully used in different bioseparation procedures. When an antigen, inactivated glyceraldehyde-3-phosphate dehydrogenase, from rabbit was covalently coupled to a polycation, the resulting complex was used to purify monoclonal antibodies specific toward

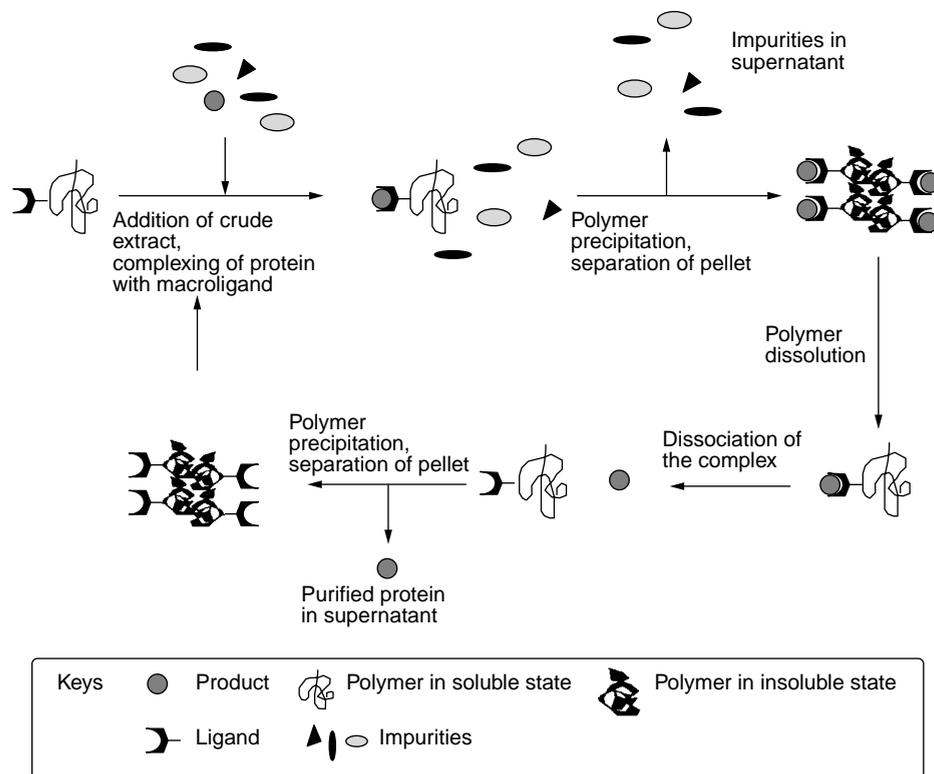


Figure 4. Schematic of affinity precipitation technique for protein purification.

inactivated glyceraldehyde-3-phosphate dehydrogenase (11). The successful affinity precipitation of antibodies using glyceraldehyde-3-phosphate dehydrogenase bound to a polyelectrolyte complex indicates that the ligand is exposed to the solution. This fact was used to develop a new method for producing monovalent Fab fragments of antibodies. Traditionally, Fab fragments are produced by proteolytic digestion of antibodies in solution followed by isolation of Fab fragments. In the case of monoclonal antibodies against inactivated subunits of glyceraldehyde-3-phosphate dehydrogenase, digestion with papain resulted in significant damage of binding sites of the Fab fragment. Proteolysis of monoclonal antibodies in the presence of the antigen–polycation conjugate followed by (1) precipitation induced by adding polyanion, poly(methacrylic) acid, and a pH shift from 7.3 to 6.5 and (2) elution at pH 3.0 that resulted in 90% immunologically competent Fab fragments. Moreover, the papain concentration required for proteolysis was 10 times less for antibodies bound to the antigen–polycation conjugate compared to that for free antibodies in solution (42). Active glyceraldehyde-3-phosphate dehydrogenase from rabbit muscle was separated from the inactivated enzyme by using monoclonal antibodies specific for the inactivated enzyme covalently coupled to the polyanion component of the polyelectrolyte complex. This system can be regarded as a simplified model of chaperone action in living cells that assist in separating active protein molecules from misfolded ones (43).

Apart from specific interactions between a target protein and a ligand–polymer conjugate, nonspecific interactions of protein impurities with the polymer backbone could take place. The nonspecific interactions limit the efficiency of the affinity precipitation technique, and

significant efforts were made to reduce these interactions. The advantage of polyelectrolyte complexes as carriers for affinity precipitation is low nonspecific coprecipitation of proteins when the polymer undergoes a soluble–insoluble transition (10).

Smart particles capable of reversible transition between aggregated and dispersed states were used for affinity precipitation of proteins. Thermosensitive (44) or pH-sensitive latices (45) or salt-sensitive liposomes that have polymerized membranes (35) are examples of such systems.

Two elements are required for successful affinity precipitation. The backbone of a smart polymer provides precipitation at the desired conditions (temperature, pH, ionic strength), and the biorecognition element is responsible for selective binding of the protein of interest. By proper choice of a smart polymer, precipitation could be achieved practically at any desired pH or temperature. For example, poly(*N*-acryloylpiperidine) terminally modified with maltose has an extremely low critical temperature (soluble below 4°C and completely insoluble above 8°C) and was used to purify thermolabile α -glucosidase (46).

Bioseparation—Partitioning in Aqueous Polymer Two-Phase Systems

Two aqueous polymer solutions become mutually incompatible when the threshold concentrations of polymers are exceeded. Both of the polymer phases formed contain about 90% water and hence present a very friendly environment for proteins and other biomolecules. Proteins partition selectively between two phases depending on their size, charge, hydrophobicity, nature, and the concentration of the phase-forming polymers. The partitioning could be also

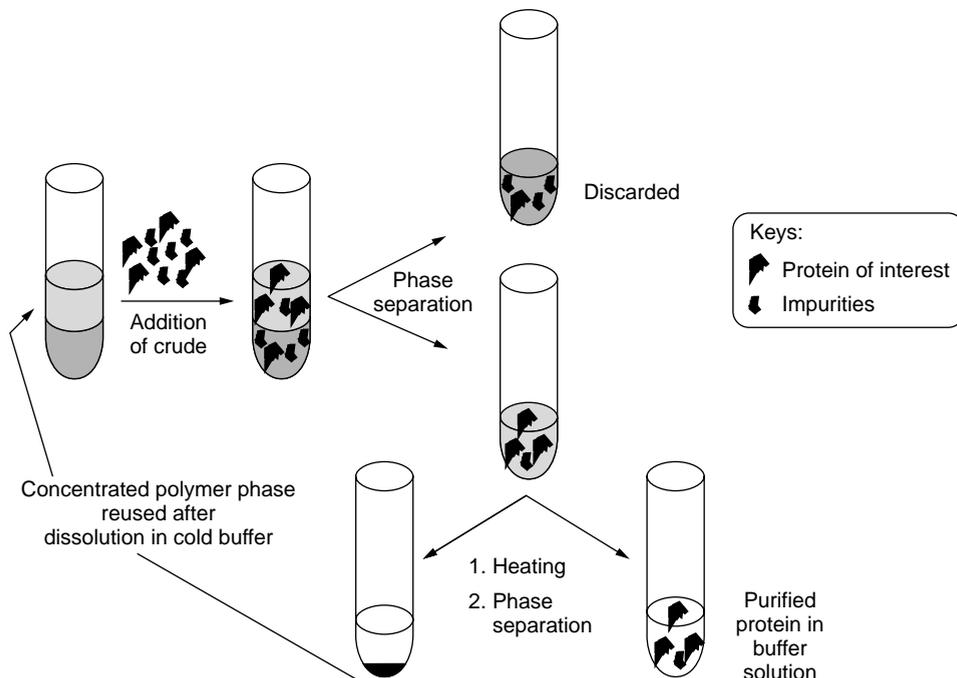


Figure 5. Schematic presentation of protein partitioning in aqueous two-phase polymer system formed by a smart (thermosensitive polymer) [reproduced from (140) with permission].

directed by adding some salt or coupling an affinity ligand specific for a given protein to one of the phase-forming polymers (47). The selective partitioning of proteins between the two phases formed has proven to be an efficient tool for purifying proteins and some low molecular weight substances. The main problem of the method—how to separate the target protein from the phase-forming polymer—has not yet been completely solved. Smart polymers provide an elegant solution to this problem—simple precipitation of the phase-forming polymer leaves protein in the supernatant (Fig. 5): (1) The crude protein extract is mixed with the aqueous two-phase polymer system, and the conditions are selected so that the protein of interest partitions into a phase formed by a smart polymer (for example by coupling affinity ligand to the smart polymer), and the impurities concentrate in the other phase. (2) The phases are separated mechanically and the phase formed by the smart polymer is subjected to conditions (pH or temperature) where the polymer undergoes phase separation; (3) two new phases are formed, a polymer-enriched phase of high polymer and low water concentration, which contains practically no protein, and a polymer-depleted aqueous phase that contains most of the purified protein and minute amounts of the polymer left after phase separation.

pH-sensitive acrylic copolymers (48) or thermoresponsive polymers, poly(ethylene oxide-co-propylene oxide) (49,50) or poly(*N*-vinyl caprolactam-co-vinyl imidazole) (51), form two-phase systems from relatively hydrophilic polymers such as dextran or modified starch and have been successfully used for protein purification. The pH- or thermoprecipitated polymer opposite dextran could be regenerated by dissolution at a lower temperature. Quite recently, an aqueous two-phase polymer system was developed where both phase-forming polymers, poly(*N*-isopropylacrylamide-co-vinyl imidazole) and poly(ethylene oxide-co-propylene oxide) end modified by

hydrophobic $C_{14}H_{29}$ groups, are thermoresponsive and could be recycled (52).

Smart Surfaces—Cell Detachment

The driving force behind phase separation of smart polymers is a sharp increase in hydrophobicity after a small change in environmental conditions. The hydrophobic “collapsed” polymer aggregates form a separate phase. When grafted to the surface, macromolecules of the smart polymer cannot aggregate, but the conformational transition from the hydrophilic to the hydrophobic state endows the surface with regulated hydrophobicity: the surface is hydrophilic when the smart polymer is in the expanded “soluble” conformation and hydrophobic when the polymer is in the collapsed “insoluble” conformation. The change of hydrophobicity of the surface by grafted poly(*N*-isopropylacrylamide) was demonstrated by contact angle measurements (53) and water absorbency (54).

The transition temperature for adsorbed (presumably via multipoint attachment) poly(NIPAAm) molecules is lower than that in bulk solution, and the properties of the layer of collapsed macromolecules formed above the transition temperature depend strongly on the speed by which the temperature increases. At a low speed of temperature increase, the “liquid-like” polymer layer is formed, whereas at high speeds, the polymer layer has more “solid-like” properties (55). When cooling, the collapsed polymer molecules return to the initial loopy adsorbed conformation via transitional extended conformation. The relaxation process for the extended-to-loopy adsorbed conformational transition occurs slowly and depends on the temperature observance of an Arrhenius law. Kinetic constraints, it is proposed, play an important role in this transition (56).

The change of surface properties from hydrophobic above the critical temperature of the polymer grafted

to hydrophilic below it has been successfully used for detaching mammalian cells. Mammalian cells are normally cultivated on a hydrophobic solid substrate and are detached from the substrate by protease treatment, which often damages the cells by hydrolyzing various membrane-associated protein molecules. The poly(NIPAAm)-grafted surface is hydrophobic at 37°C because this temperature is above the critical temperature for the grafted polymer and that cells that are growing well on it. A decrease in temperature results in transition of the surface to the hydrophilic state, where the cells can be easily detached from the solid substrate without any damage. Poly(NIPAAm) was grafted to polystyrene culture dishes using an electron beam. Bovine hepatocytes, cells that are highly sensitive to enzymatic treatment, were cultivated for 2 days at 37°C and detached by incubation at 4°C for 1 h. Nearly 100% of the hepatocytes was detached and recovered from the poly(NIPAAm)-grafted dishes by low-temperature treatment, whereas only about 8% of the cells was detached from the control dish (57). The technique has been extended to different cell types (58,59). It is noteworthy that hepatocytes recovered by cooling retained their native form had numerous bulges and dips, and attach well to the hydrophobic surface again, for example, when the temperature was increased above the conformational transition of poly(NIPAAm). On the contrary, enzyme-treated cells had a smooth outer surface and had lost their ability to attach to the surface. Thus, cells recovered by a temperature shift from poly(NIPAAm)-grafted surfaces have an intact structure and maintain normal cell functions (58).

The molecular machinery involved in cell-surface detachment was investigated using temperature-responsive surfaces (60). Poly(NIPAAm)-grafted and nongrafted surfaces showed no difference in attachment, spreading, growth, confluent cell density, or morphology of bovine aortic endothelial cells at 37°C. Stress fibers, peripheral bands, and focal contacts were established in similar ways. When the temperature was decreased to 20°C, the cells grown on poly(NIPAAm)-grafted support lost their flattened morphology and acquired a rounded appearance similar to that of cells immediately after plating. Mild agitation makes the cells float free from the surface without a trypsin treatment. Neither changes in cell morphology nor cell detachment occurred on ungrafted surfaces. Sodium azide, an ATP synthesis inhibitor, and genistein, a tyrosine kinase inhibitor, suppressed changes in cell morphology and cell detachment, whereas cycloheximide, a protein synthesis inhibitor, slightly enhanced cell detachment. Phalloidin, an actin filament stabilizer, and its depolymerizer, cytochalasin D, also inhibited cell detachment. These findings suggest that cell detachment from grafted surfaces is mediated by intracellular signal transduction and reorganization of the cytoskeleton, rather than by a simple changes in the "stickiness" of the cells to the surface when the hydrophobicity of the surface is changed.

One could imagine producing artificial organs using temperature-induced detachment of cells. Artificial skin could be produced as the cells are detached from the support not as a suspension (the usual result of protease-induced detachment) but preserving their intercellular contacts. Fibroblasts were cultured on the poly(NIPAAm)-collagen support until the cells completely covered the

surface at 37°C, followed by a decrease in temperature to about 15°C. The sheets of fibroblasts detached from the dish and within about 15 min floated in the culture medium (57). The detached cells could be transplanted to another culture surface without functional and structural changes (34). Grafting of poly(NIPAAm) onto a polystyrene surface by photolithographic technique creates a special pattern on the surface, and by decreasing temperature, cultured mouse fibroblast STO cells are detached only from the surface area on which poly(NIPAAm) was grafted (61). Lithographed films of smart polymer present supports for controlled interactions of cells with surfaces and can direct the attachment and spreading of cells (62). One could envisage producing artificial cell assemblies of complex architecture using this technique.

Smart Surfaces—Temperature Controlled Chromatography

Surfaces that have thermoresponsive hydrophobic/hydrophilic properties have been used in chromatography. HPLC columns with grafted poly(NIPAAm) have been used for separating steroids (63) and drugs (64). The chromatographic retention and resolution of the solutes was strongly dependent on temperature and increased as temperature increased from 5 to 50°C, whereas the reference column packed with nonmodified silica displayed much shorter retention times that decreased as temperature decreased. Hydrophobic interactions dominate in retaining solutes at higher temperature, and the preferential retention of hydrogen-bond acceptors was observed at low temperatures. The effect of temperature increase on the retention behavior of solutes separated on the poly(NIPAAm)-grafted silica chromatographic matrix was similar to the addition of methanol to the mobile phase at constant temperature (65).

The temperature response of the poly(NIPAAm)-silica matrices depends drastically on the architecture of the grafted polymer molecules. Surface wettability changes dramatically as temperature changes across the range 32–35°C (corresponding to the phase-transition temperature for NIPAAm in aqueous media) for surfaces where poly(NIPAAm) is terminally grafted either directly to the surface or to the looped chain copolymer of NIPAAm and *N*-acryloylhydroxysuccinimide which was initially coupled to the surface. The wettability changes for the loop-grafted surface itself were relatively large but had a slightly lower transition temperature (~27°C). The restricted conformational transitions for multipoint grafted macromolecules are probably the reason for the reduced transition temperature. The largest surface free energy changes among three surfaces was observed for the combination of both loops and terminally grafted chains (30).

Introduction of a hydrophobic comonomer, butylmethacrylate, in the polymer resulted in a decreased transition temperature of about 20°C. Retention of steroids in poly(NIPAAm-co-butylmethacrylate)-grafted columns increases as column temperature increases. The capacity factors for steroids on the copolymer-modified silica beads was much larger than that on poly(NIPAAm)-grafted columns. The effect of temperature on steroid retention on poly(NIPAAm-co-butylmethacrylate)-grafted stationary phases was more pronounced compared

to supports modified with poly(NIPAAm). Furthermore, retention times for steroids increased remarkably as the butylmethacrylate content increased in the copolymer. The temperature-responsive elution of steroids was strongly affected by the hydrophobicity of the grafted polymer chains on silica surfaces (63).

The mixture of polypeptides, consisting of 21–30 amino acid residues (insulin chain A, β -endorphin fragment 1–27 and insulin chain B) could not be separated at 5°C (below the transition temperature) on copolymer-grafted matrix. At this temperature, the copolymer is in an extended hydrophilic conformation that results in decreased interactions with peptides and hence short retention times insufficient to resolve them. The mixture has been easily separated at 30°C, when the copolymer is collapsed, hydrophobic interactions are more pronounced, and retention times sufficiently long for resolving polypeptides (66). Large protein molecules such as immunoglobulin G demonstrate less pronounced changes in adsorption above and below the transition temperature. Only about 20% of the protein adsorbed on poly(NIPAAm)-grafted silica at 37°C (above the LCST) were eluted after decreasing temperature to 24°C (below the transition temperature) (67). Quantitative elution of proteins adsorbed on the matrix via hydrophobic interactions has not yet been demonstrated, although protein adsorption on poly(NIPAAm)-grafted matrices could be somewhat controlled by a temperature shift. A successful strategy for temperature-controlled protein chromatography proved to be a combination of temperature-responsive polymeric grafts and biorecognition element, for example, affinity ligands.

The access of the protein molecules to the ligands on the surface of the matrix is affected by the transition of the polymer macromolecule grafted or attached to the chromatographic matrix. Triazine dyes, for example, Cibacron Blue, are often used as ligands for dye-affinity chromatography of various nucleotide-dependent enzymes (68). Poly(*N*-vinyl caprolactam), a thermoresponsive polymer whose critical temperature is about 35°C interacts efficiently with triazine dyes. Polymer molecules of 40000 MW are capable of binding up to seven to eight dye molecules hence, the polymer binds via multipoint interaction to the dye ligands available on the chromatographic matrix. At elevated temperature, polymer molecules are in a compact globule conformation that can bind only to a few ligands on the matrix. Lactate dehydrogenase, an enzyme from porcine muscle has good access to the ligands that are not occupied by the polymer and binds to the column. Poly(*N*-vinyl caprolactam) macromolecules undergo transition to a more expanded coil conformation as temperature decreases. Now, the polymer molecules interact with more ligands and begin to compete with the bound enzyme for the ligands. Finally, the bound enzyme is displaced by the expanded polymer chains. The temperature-induced elution was quantitative, and the first reported in the literature when temperature change was used as the only eluting factor without any changes in buffer composition (69). Small changes in temperature, as the only eluting factor, are quite promising because there is no need in this case to separate the target protein from an eluent, usually a competing nucleotide or high salt concentration in dye-affinity chromatography.

Smart Surfaces—Controlled Porosity, “Chemical Valve”

Environmentally controlled change in macromolecular size from a compact hydrophobic globule to an expanded hydrophilic coil is exploited when smart polymers are used in systems of environmentally controlled porosity, so called “chemical valves.” When a smart polymer is grafted to the surface of the pores in a porous membrane or chromatographic matrix, the transition in the macromolecule affects the total free volume of the pores available for the solvent and hence presents a means to regulate the porosity of the system.

Membranes of pH-sensitive permeability were constructed by grafting smart polymers such as poly(methacrylic acid) (70), poly(benzyl glutamate), poly(2-ethylacrylic acid) (71), poly(4-vinylpyridine) (72), which change their conformation in response to pH. Thermosensitive chemical valves have been developed by grafting poly(*N*-acryloylpyrrolidine), poly(*N*-*n*-propylacrylamide), or poly(acryloylpiperidine) (73), poly(NIPAAm) alone (33,74) or in copolymers with poly(methacrylic acid) (74) inside the pores. For example, grafted molecules of poly(benzyl glutamate) at high pH are charged and are in extended conformation. The efficient pore size is reduced, and the flow through the membrane is low (“off-state” of the membrane). As pH decreases, the macromolecules are protonated, lose their charge, and adopt a compact conformation. The efficient pore size and hence the flow through the membrane increases (“on-state” of the membrane) (71). The fluxes of bigger molecules (dextrans of molecular weights 4400–50600) across a temperature-sensitive, poly(NIPAAm)-grafted membrane were effectively controlled by temperature, environmental ionic strength, and degree of grafting of the membrane, while the flux of smaller molecules such as mannitol was not affected by temperature even at high degree of membrane grafting (75). The on-off permeability ratio for different molecules (water, Cl⁻ ion, choline, insulin, and albumin) ranged between 3 and 10 and increased as molecular weight increased (76). An even more abrupt change of the on-off permeability ratio was observed for a membrane that had narrow pores formed by heavy ion beams when poly(NIPAAm) or poly(acryloyl-L-proline methyl ester) were grafted (77).

Different stimuli could trigger the transition of the smart polymer making it possible to produce membranes whose permeabilities respond to these stimuli. When a copolymer of NIPAAm with triphenylmethane leucocyanide was grafted to the membrane, it acquires photosensitivity—UV irradiation increases permeation through the membrane (78). Fully reversible, pH-switchable permselectivity for both cationic and anionic redox-active probe molecules was achieved by depositing composite films formed from multilayers of amine-terminated dendrimers and poly(maleic anhydride-co-methylvinyl ether) on gold-coated silicon (79).

When the smart polymer is grafted inside the pores of the chromatographic matrix for gel permeation chromatography, the transition of grafted macromolecules regulates the pore size and as a result, the elution profile of substances of different molecular weights. As the temperature is raised, the substances are eluted progressively earlier indicating shrinking of the pores of the hydrogel

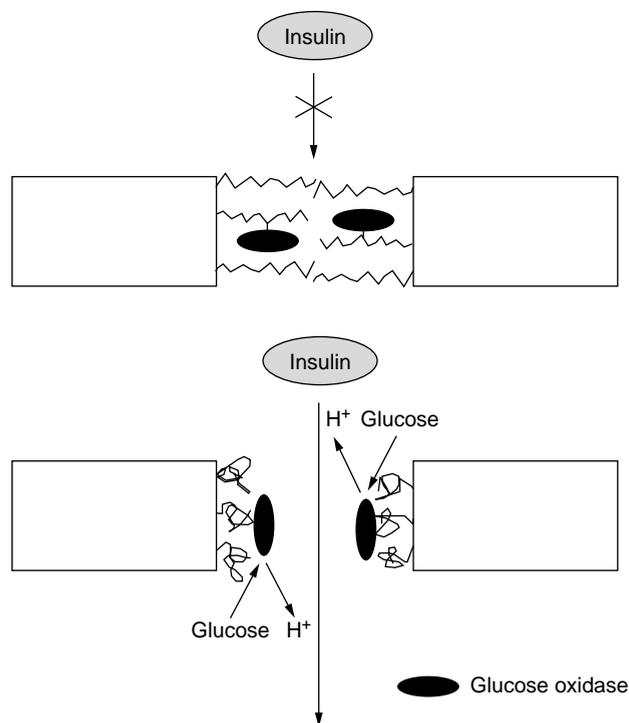


Figure 6. Schematic of a “chemical valve.” Glucose oxidase is immobilized on a pH-responsive polyacrylic acid grafted onto a porous polycarbonate membrane: (a) poly(acrylic acid) is in an expanded conformation that blocks insulin transport; (b) the oxidation of glucose is accompanied by a decrease in pH and the transition of poly(acrylic acid) into a compact conformation that results in opening of the pores and transport of insulin [redrawn from (82)].

beads composed of cross-linked poly(acrylamide-*co*-*N*-isopropylacrylamide) (80) or porous polymer beads with grafted poly(NIPAAm) (81).

When using a specific biorecognition element, which recognizes specific substances and translates the signal into a change of physicochemical properties, for example, pH, a smart membrane that changes its permeability in response to particular substances can be constructed. Specific insulin release in response to increasing glucose concentration, that is, an artificial pancreas, presents an everlasting challenge to bioengineers. One of the potential solutions is a “chemical valve” (Fig. 6). The enzyme, glucose oxidase, was used as a biorecognition element, capable of specific oxidation of glucose accompanied by a decrease in pH. The enzyme was immobilized on pH-responsive poly(acrylic acid) graft on a porous polycarbonate membrane. In neutral conditions, polymer chains are densely charged and have extended conformation that prevents insulin transport through the membrane by blocking the pores. Under exposure to glucose, the pH drops as the result of glucose oxidation by the immobilized enzyme, the polymer chains adopt a more compact conformation that diminishes the blockage of the pores, and insulin is transported through the membrane (82). Systems such as this could be used for efficient drug delivery that responds to the needs of the organism. A membrane that consists of poly(2-hydroxyethyl acrylate-*co*-*N,N*-diethylaminomethacrylate-*co*-4-trimethylsilylstyrene) undergoes a sharp transition

from a shrunken state at pH 6.3 to a swollen state at pH 6.15. The transition between the two states changes the membrane permeability to insulin 42-fold. Copolymer capsules that contain glucose oxidase and insulin increase insulin release five fold in response to 0.2 M glucose. After glucose removal, the rate of insulin release falls back to the initial value (83).

Alternatively, reversible cross-linking of polymer macromolecules could be used to control the porosity in a system. Two polymers, poly(*m*-acrylamidophenylboronic acid-*co*-vinylpyrrolidone) and poly(vinyl alcohol) form a gel because of strong interactions between boronate groups and the hydroxy groups of poly(vinyl alcohol). When a low molecular weight polyalcohol such as glucose is added to the gel, it competes with poly(vinyl alcohol) for boronate groups. The boronate–poly(vinyl alcohol) complex changes to a boronate–glucose complex that results in eventual dissolution of the gel (84). In addition to a glucose oxidase-based artificial pancreas, the boronate–poly(vinyl alcohol) system has been used for constructing glucose-sensitive systems for insulin delivery (29,85–87). The glucose-induced transition from a gel to a sol state drastically increases the release of insulin from the gel. The reversible response to glucose has also been designed using another glucose-sensitive biorecognition element, Concanavalin A, a protein that contains four sites that can bind glucose. Polymers that have glucose groups in the side chain such as poly(vinylpyrrolidone-*co*-allylglucose) (26) or poly(glucosyloxyethyl methacrylate) (88), are reversibly cross-linked by Concanavalin A and form a gel. The addition of glucose results in displacing the glucose-bearing polymer from the complex with Concanavalin A and dissolving the gel.

Reversible gel-formation of thermosensitive block copolymers in response to temperature could be utilized in different applications. Poly(NIPAAm) block copolymers with poly(ethylene oxide) which undergo a temperature-induced reversible gel–sol transition were patented as the basis for cosmetics such as depilatories and bleaching agents (89). The copolymer solution is liquid at room temperature and easily applied to the skin where it forms a gel within 1 min. Commercially available ethyl(hydroxyethyl)celluloses that have cloud points of 65–70°C have been used as redeposition agents in washing powders. Adsorption of the precipitated polymer on the laundry during the initial rinsing period counteracts read-sorption of dirt when the detergent is diluted (90).

Liposomes That Trigger Release of the Contents

When a smart polymer is attached somehow to a lipid membrane, the transition in the macromolecule affects the properties of the membrane and renders the system sensitive to environmental changes. To attach a smart polymer to a lipid membrane, a suitable “anchor” which could be incorporated in the membrane, should be introduced into the macromolecule. This could be achieved by copolymerizing poly(NIPAAm) with comonomers that have large hydrophobic tails such as *N,N*-didodecylacrylamide (91), using a lipophilic radical initiator (92) modifying copolymers (93), or polymers that have terminally active groups (94)

with a phospholipid. Alternatively, smart polymers have been covalently coupled to the active groups in the hydrophilic heads of the lipid-forming membrane (95).

Interesting and practically relevant materials for studying the behavior of smart polymers attached to lipid membranes, are liposomes, self assembled 50–200 nm vesicles that have one or more (phospho)lipid bilayers which encapsulate a fraction of the solvent. Liposomes are stable in aqueous suspension due to the repulsive forces that appear when two liposomes approach each other. Liposomes are widely used for drug delivery and in cosmetics (96).

The results of a temperature-induced conformational transition of a smart polymer on the liposomal surface depend significantly on the fluidity of the liposomal membrane. When the membrane is in a fluid state at temperatures both above and below the polymer transition temperature, the collapse of the polymer molecule forces anchor groups to move closer together by lateral diffusion within the membrane. The compact globules of collapsed polymer cover only a small part of the liposomal surface. Such liposomes have a low tendency to aggregate because the most of their surface is not covered by the polymer. Naked surfaces contribute to the repulsion between liposomes. On the other hand, when the liposomal membrane is in a solid state at temperatures both above and below the polymer transition temperature, the lateral diffusion of anchor groups is impossible, and the collapsed polymer cannot adopt a compact globule conformation but spreads over the most of the liposomal surface (97). Liposomes whose surfaces are covered to a large degree by a collapsed polymer repel each other less efficiently than intact liposomes. The stability of a liposomal suspension is thereby decreased, and aggregation and fusion of liposomes takes place, which is often accompanied by the release of the liposomal content into the surrounding medium (98).

When the liposomal membrane is perturbed by the conformational transition of the polymer, both the aggregation tendency and liposomal permeability for incorporated substances are affected. Poly(ethacrylic acid) undergoes a transition from an expanded to a compact conformation in the physiological pH range of 7.4–6.5 (99). The pH-induced transition of poly(ethacrylic acid) covalently coupled to the surface of liposomes formed from phosphatidylcholine results in liposomal reorganization into more compact micelles and concomitant release of the liposomal content into the external medium. The temperature-induced transition of poly(NIPAAM-*co*-*N,N*-didoceylacrylamide) (100) or poly(NIPAAM-*co*-octadecylacrylate) (101), incorporated into the liposomal membrane, enhanced the release of the fluorescent marker, calcein, encapsulated in copolymer-coated liposomes. Liposomes hardly release any marker at temperatures below 32°C (the polymer transition temperature), whereas the liposomal content is released completely within less than a minute at 40°C. To increase the speed of liposomal response to temperature change, the smart polymer was attached to the outer and inner sides of the lipid membrane. The polymer bound only to the outer surface if the liposomes were treated with the polymer after liposomal formation. When the liposomes were formed directly from the lipid-polymer mixture, the polymer was present on both sides of the liposomal membrane (91).

Changes of liposomal surface properties caused by polymer collapse affect liposomal interaction with cells. Liposomes modified by a pH-sensitive polymer, partially succinylated poly(glycidol), deliver calcein into cultured kidney cells of the African green monkey more efficiently compared to liposomes not treated with the polymer (102). Polymeric micelles formed by smart polymers and liposomes modified by smart polymers could be used for targeted drug delivery. Polymeric micelles have been prepared from amphiphilic block copolymers of styrene (forming a hydrophobic core) and NIPAAM (forming a thermosensitive outer shell). The polymeric micelles were very stable in aqueous media and had long blood circulation because of small diameter, unimodal size distribution (24 ± 4 nm), and, a low critical micellar concentration of around 10 $\mu\text{g/mL}$. At temperatures above the polymer transition temperature (32°C), the polymer chains that form an outer shell collapse, become more hydrophobic, and allow aggregation between micelles and favoring binding interactions with the surface of cell membranes. Thus, hydrophobic molecules incorporated into the micelles are delivered into the cell membranes. These micelles are capable of site-specific delivery of drugs to the sites as temperature changes, for example, to inflammation sites of increased temperature (103).

Smart Polymers in Bioanalytical Systems

Because smart polymers can recognize small changes in environmental properties and respond to them in a pronounced way, they could be used directly as sensors of these changes, for example, a series of polymer solutions that have different LCSTs could be used as a simple thermometer. As salts promote hydrophobic interactions and decrease the LCST, the polymer system could “sense” the salt concentration needed to decrease the LCST below room temperature. A poly(NIPAAM)-based system that can sense NaCl concentrations above 1.5% was patented (104). The response of the polymer is controlled by a balance of hydrophilic and hydrophobic interactions in the macromolecule. Using a recognition element that can sense external stimuli and translate the signal into the changes of the hydrophilic/hydrophobic balance of the smart polymer, the resulting system presents a sensor for the stimulus. If the conjugate of a smart polymer and a recognition element has a transition temperature T_1 in the absence and T_2 in the presence of stimuli, fixing the temperature T in the range $T_1 < T < T_2$ allows achieving the transition of a smart polymer isothermally by the external stimulus (105). An example of such a sensor was constructed using *trans-cis* isomerization of the azobenzene chromophore when irradiated by UV light. The transition is accomplished by an increase in the dipole moment of azobenzene from 0.5 D (for the *trans*-form) to 3.1 D (for the *cis*-form) and hence a significant decrease of hydrophobicity. Irradiation with UV light results in increasing the LCST from 19.4 to 26.0°C for the conjugate of the chromophore with poly(NIPAAM). The solution of the conjugate is turbid at $19.4^\circ\text{C} < T < 26.0^\circ\text{C}$, but when irradiated, the conjugate dissolves because the *cis*-form is below the LCST at this temperature. The system responds to UV light by transition from a turbid to

transparent solution. The termination of UV irradiation results in a slow return of the system to its initial turbid state (105). A few other light-sensitive systems were proposed that use different chromophores: triarylmethylcyanide (106) and leuconitriles (107).

The hydrophobicity of the recognition molecule was also changed by chemical signals. Poly(NIPAAm) containing 11.6 mol% of crown ether 9 has a LCST of 31.5°C in the absence of Na⁺ or K⁺ ions, 32°C in the presence of Na⁺, and 38.9°C in the presence of K⁺. Thus, the introduction of both Na⁺ and K⁺ ions leads to the dissolution of the insoluble polymer at that temperature. At 37°C, this effect is achieved only by K⁺ ions (108).

From better understanding of ligand–host interactions and development of new highly selective binding pairs (e.g., by using combinatorial libraries to find ligands of high affinity for particular biomolecules), one could expect that smart polymer systems will be used as “signal amplifiers” to visualize a physicochemical event, which takes place in a recognition element, by a pronounced change in the system—conversion of a transparent solution into a turbid one or vice versa.

Antibody–antigen interactions present nearly ideal analytical selectivity and sensitivity developed by nature. Not surprisingly, they are increasingly used for a broad variety of bioanalytical applications. Different analytical formats have been developed. The common feature of the most of them is the requirement for separating an antibody–antigen complex from a nonbound antibody or antigen. Traditionally, the separation is achieved by coupling one of the components of antibody–antigen pair to a solid support. The binding step is followed by washing nonbound material. Interactions of the soluble partner of the binding pair with the partner coupled to the support are often accompanied by undesired diffusional limitations, and hence, incubation times of several hours are required for analysis. Because smart polymers can undergo transition from the soluble to the insoluble state, they allow combining the advantages of homogeneous binding and, after the phase transition of the smart polymer has taken place, easy separation of the polymer precipitate from the supernatant. The essential features of an immunoassay that uses smart polymers (named PRECIPIA) are as follows. The covalent conjugate of poly(NIPAAm) with monoclonal antibodies to the κ -chain of human immunoglobulin G (IgG) are incubated for 1 h at room temperature (below the LCST of the conjugate), and the IgG solution is analyzed. Then plain poly(NIPAAm) (to facilitate thermoprecipitation of polymer–antibody conjugates) and fluorescently labeled monoclonal antibodies to the γ -chain of human IgG are added. The temperature is raised to 45°C, the precipitated polymer is separated by centrifugation, and fluorescence is measured in the supernatant (109). Immunoassay systems that use temperature-induced precipitation of poly(NIPAAm) conjugates with monoclonal antibodies are not inferior in sensitivity to the traditional heterogenous immunoassay methods, but because the antigen–antibody interaction takes place in solution, the incubation can be shortened to about 1 h (110,111). The limitations of PRECIPIA as an immunoassay technique are essentially the same as those of affinity precipitation, namely, nonspecific coprecipitation of analyzed

protein when poly(NIPAAm) precipitates. Polyelectrolyte complexes that have a low degree of nonspecific protein coprecipitation have also been successfully used as reversibly soluble carriers for PRECIPIA-type immunoassays (112). The conjugate of antibody and polyanion poly(methacrylic acid) binds to the antigen within a few minutes, and the polymer hardly exerts any effect on the rate of antigen–antibody binding. Subsequent addition of a polycation, poly(*N*-ethyl-4-vinyl-pyridinium bromide) in conditions where the polyelectrolyte is insoluble, results in quantitative precipitation of the antibody–polymer conjugate within 1 min. The total assay time is less than 15 min (10).

In principle, PRECIPIA-type immunoassays could be used for simultaneous assay of different analytes in one sample, provided that conjugates specific toward these analytes are coupled covalently to different smart polymers that have different precipitating conditions, for example, precipitation of one conjugate by adding a polymeric counterion followed by thermoprecipitation of the second conjugate by increasing temperature.

Reversibly Soluble Biocatalysts

The transition between the soluble and insoluble state of stimuli-responsive polymers has been used to develop reversibly soluble biocatalysts. A reversibly soluble biocatalyst catalyzes an enzymatic reaction in a soluble state and hence could be used in reactions of insoluble or poorly soluble substrates/products. As soon as the reaction is completed and the products are separated, the conditions (pH, temperature) are changed to promote precipitation of the biocatalyst. The precipitated biocatalyst is separated and can be used in the next cycle after dissolution. The reversibly soluble biocatalyst acquires the advantages of immobilized enzymes (ease of separation from the reaction mixture after the reaction is completed and the possibility for biocatalyst recovery and repeated use in many reaction cycles) but at the same time overcomes the disadvantages of enzymes immobilized onto solid matrices such as diffusional limitations and the impossibility of using them in reactions of insoluble substrates or products.

Biocatalysts that are reversibly soluble as a function of pH have been obtained by the covalent coupling of lysozyme to alginate (113); of trypsin to poly(acrolein-*co*-acrylic acid) (114); and of cellulase (115); amylase (115); α -chymotrypsin, and papain (116) to poly(methylmethacrylate-*co*-methacrylic acid). A reversibly soluble cofactor has been produced by the covalent binding of NAD to alginate (117). Reversibly soluble α -chymotrypsin, penicillin acylase, and alcohol dehydrogenase were produced by coupling to the polycation component of polyelectrolyte complexes formed by poly(methacrylic acid) and poly(*N*-ethyl-4-vinyl-pyridinium bromide) (118).

Biocatalysts that are reversibly soluble as a function of temperature have been obtained by the covalent coupling of α -chymotrypsin and penicillin acylase to a partially hydrolysed poly(*N*-vinylcaprolactam) (119); and of trypsin (120); alkaline phosphatase (121), α -chymotrypsin (122), and thermolysin (123,124) to NIPAAm copolymers that contain active groups suitable for covalent coupling of biomolecules. Lipase was coupled to a graft copolymer composed of NIPAAm grafts on a poly(acrylamide-*co*-acrylic

acid) copolymer (125). No significant differences in biocatalytic properties were found for α -amylase coupled to poly(NIPAAm) via single-point or multipoint mode. Both enzyme preparations demonstrated increased thermostability and the absence of diffusional limitation when hydrolyzing starch, a high molecular weight substrate (126). The temperature of a protein–ligand interaction was controlled by site-directed coupling of terminally modified poly(NIPAAm) to a specifically constructed site (close to a biotin binding site) on a genetically modified streptavidin (127).

Biocatalysts which are reversibly soluble as a function of Ca^{2+} concentration were produced by covalent coupling of phosphoglyceromutase, enolase, peroxidase, and pyruvate kinase to α_{s1} -casein. The enzyme casein conjugates are soluble at a Ca^{2+} concentration below 20 mM but precipitate completely at a Ca^{2+} concentration above 50 mM. The precipitate redissolves when EDTA, a strong Ca^{2+} -binding agent is added (128).

The reversible flocculation of latices has been used to produce thermosensitive reversibly soluble (more precisely reversibly dispersible) biocatalysts using trypsin (129), papain (130), and α -amylase (131). Latices sensitive to a magnetic field have been used to immobilize trypsin and β -galactosidase (132). Liposomes that have a polymerized membrane, that reversibly aggregates on changing salt concentration have been used to immobilize α -chymotrypsin (133).

The most attractive application of reversibly soluble biocatalysts is repeated use in a reaction which is difficult or even impossible to carry out using enzymes immobilized onto insoluble matrices, for example, hydrolysis of water-insoluble phlorizidin (134); hydrolysis of high molecular weight substrates such as casein (123,130) and starch (115); hydrolysis of insoluble substrates such as cellulose (135) and raw starch (corn flour) (7,134,136–138); production of insoluble products such as peptide, benzyloxycarbonyl-L-tyrosyl- N^{ω} -nitro-L-arginine (116) and phenylglycine (139).

The hydrolytic cleavage of corn flour to glucose is an example of successfully using a reversibly soluble biocatalyst, amylase coupled to poly(methylmethacrylate-co-methacrylic acid), in an industrially interesting process (136). The reaction product of the process, glucose, inhibits the hydrolysis. The use of a reversibly soluble biocatalyst improves the efficiency of the hydrolysis which is carried out at pH 5, at which the amylase–polymer conjugate is soluble. After each 24 h, the pH is reduced to 3.5, the unhydrolyzed solid residue and the precipitated conjugate are separated by centrifugation, the conjugate is resuspended in a fresh portion of the substrate at pH 5, and the hydrolysis is continued. The conversion achieved after 5 cycles is 67%, and the activity of the amylase after the fifth cycle was 96% of the initial value (136).

CONCLUSION

In the future, one looks forward to further developments and the commercial introduction of new smart polymers whose transition temperatures and pH are compatible with physiological conditions or conditions for maximal stability

of target biomolecules, such as temperatures of 4–15°C and pH values of 5–8. Additional prospects will stem from a better understanding of the mechanism of cooperative interactions in polymers and increasing knowledge of structure–property correlations to enable rational synthesis of smart polymers that have predefined properties. Due to the possibility of combining a variety of biorecognition or biocatalytic systems and the unique features of smart polymers, expectations are running high in this area. Only time and more experimentation will determine whether smart polymers will live up to their generous promises.

BIBLIOGRAPHY

1. R. Dagani, *Chem. Eng. News*, pp. 30–33 (1995).
2. H. Brønsted and J. Kopeček, in *Polyelectrolyte Gels. Properties, Preparation, and Applications*, R.S. Harland and R.K. Prud'homme, eds., American Chemical Society, Washington DC, 1992, pp. 285–304.
3. V.A. Kabanov, *Polym. Sci.* **36**: 143–156 (1994).
4. H.G. Schild, *Prog. Polym. Sci.* **17**: 163–249 (1992).
5. E.Y. Kramarenko and A.R. Khokhlov, *Polym. Gels Networks* **6**: 46–56 (1998).
6. Eudragit, Röhm Pharma GMBH Information Materials, 1993.
7. K. Hoshino, H. Yamasaki, C. Chida, S. Morohashi, M. Taniguchi, and M. Fujii, *J. Chem. Eng. Jpn.* **30**: 30–37 (1997).
8. E. Linné-Larsson and B. Mattiasson, *Biotechnol. Tech.* **8**: 51–56 (1994).
9. S.Y. Park and Y.H. Bae, *Macromol. Rapid Commun.* **20**: 269–273 (1999).
10. V.A. Izumrudov, I.Yu. Galaev, and B. Mattiasson, *Bioseparation* **7**: 207–220 (1999).
11. M.B. Dainiak, V.A. Izumrudov, V.I. Muronetz, I.Yu. Galaev, and B. Mattiasson, *Bioseparation* **7**: 231–240 (1999).
12. U. Dissing and B. Mattiasson, *J. Biotechnol.* **52**: 1–10 (1996).
13. A.A. Tager, A.P. Safronov, S.V. Sharina, and I.Yu. Galaev, *Colloid Polym. Sci.* **271** (1993).
14. M. Heskins and J.E. Guillet, *J. Macromol. Sci. Chem.* **A2**: 1441–1455 (1968).
15. K. Suwa, K. Morishita, A. Kishida, and M. Akashi, *J. Polym. Sci. Part A: Polym. Chem.* **35**: 3087–3094 (1997).
16. I.Yu. Galaev and B. Mattiasson, *Enzyme Microb. Technol.* **15**: 354–366 (1993).
17. F. Liu, F.H. Liu, R.X. Zhuo, Y. Peng, Y.Z. Deng, and Y. Zeng, *Biotechnol. Appl. Biochem.* **21**: 257–264 (1995).
18. S. Mori, Y. Nakata, and H. Endo, *Protein Expression Purif.* **5**: 151–156 (1994).
19. F.M. Winnik, A.D. Davidson, G.H. Hamer, and H. Kitano, *Macromolecules* **25**: 1876–1880 (1992).
20. G. Chen and A.S. Hoffman, *373*: 49–52 (1995).
21. M. Maeda, C. Nishimura, A. Inenaga, and M. Takagi, *Reactive Functional Polym.* **21**: 27–35 (1993).
22. D. Umeno, T. Mori, and M. Maeda, *Chem. Commun.* 1433–1434 (1998).
23. B. Mattiasson, A. Kumar, and I. Yu. Galaev, *J. Mol. Recognition* **11**: 211–216 (1998).
24. A. Kumar, I.Yu. Galaev, and B. Mattiasson, *Biotechnol. Bioeng.* **59**: 695–704 (1998).
25. A. Durand and D. Hourdet, *Polymer* **40**: 4941–4951 (1999).

26. S.J. Lee and K. Park, *J. Mol. Recognition* **9**: 549–557 (1997).
27. K.-Y.A. Wu and K.D. Wisecarver, *Biotechnol. Bioeng.* **39**: 447–449 (1992).
28. E. Kokufuta and S. Matsukawa, *Macromolecules* **28**: 3474–3475 (1995).
29. S. Kitano, I. Hisamitsu, Y. Koyama, K. Kataoka, T. Okano, M. Yokoyama, and Y. Sakurai, *Proc. 1st Int. Conf. Intelligent Mater.*, 1993, pp. 383–388.
30. T. Yakushiji, K. Sakai, A. Kikuchi, T. Aoyagi, Y. Sakurai, and T. Okano, *Langmuir* **14**: 4657–4662 (1998).
31. H. Kawaguchi, in *Biomedical Functions and Biotechnology of Natural and Artificial Polymers*, M. Yalpani, ed., ATL Press, 1996, pp. 157–168.
32. H. Kubota and N. Shiobara, *Reactive Functional Polym.* **37**: 219–224 (1998).
33. Y.M. Lee and J.K. Shim, *Polymer* **38**: 1227–1232 (1997).
34. H.A. VonRecum, S.W. Kim, A. Kikuchi, M. Okuhara, Y. Sakurai, and T. Okano, *J. Biomed. Mater. Res.* **40**: 631–639 (1998).
35. Y. Sun, K. Yu, X.H. Jin, and X.Z. Zhou, *Biotechnol. Bioeng.* **47**: 20–25 (1995).
36. I.Yu. Galaev, M.N. Gupta, and B. Mattiasson, *CHEMTECH* 19–25 (1996).
37. A. Kumar, I.Yu. Galaev, and B. Mattiasson, *Bioseparation* **7**: 185–194 (1999).
38. D. Guoqiang, M.A.N. Benhura, R. Kaul, and B. Mattiasson, *Biotechnol. Prog.* **11**: 187–193 (1995).
39. D. Guoqiang, R. Batru, R. Kaul, M.N. Gupta, and B. Mattiasson, *Bioseparation* **5**: 339–350 (1995).
40. E. Linné-Larsson, I.Yu. Galaev, and B. Mattiasson, *Bioseparation* **6**: 283–291 (1996).
41. A.P. Bradshaw and R.J. Sturgeon, *Biotechnol. Tech.* **4**: 67–71 (1990).
42. M.B. Dainiak, V.A. Izumrudov, V.I. Muronetz, I.Yu. Galaev, and B. Mattiasson, *Anal. Biochem.* (submitted).
43. M.B. Dainiak, V.A. Izumrudov, V.I. Muronetz, I.Yu. Galaev, and B. Mattiasson, *Biochim. Biophys. Acta* **1381**: 279–285 (1998).
44. A. Kondo, T. Kaneko, and K. Higashitani, *Biotechnol. Bioeng.* **44**: 1–6 (1994).
45. C.S. Chen, C.K. Lee, C.Y. Chen, and M.J. Yeh, *Colloids Surf. B: Biointerfaces* **6**: 37–49 (1996).
46. K. Hoshino, M. Taniguchi, T. Kitao, S. Morohashi, and T. Sasakura, *Biotechnol. Bioeng.* **60**: 568–579 (1998).
47. P.-Å. Albertson, *Partition of Cell Particles and Macromolecules*. Wiley, NY, 1986.
48. P. Hughes and C.R. Lowe, *Enzyme. Microb. Technol.* **10**: 115–122 (1988).
49. G. Johansson and F. Tjerneld, in *Highly Selective Separations in Biotechnology*, G. Street, ed., Blackie Academic & Professional, London, 1994, pp. 55–85.
50. M. Lu, P.-Å. Albertson, G. Johansson, and F. Tjerneld, *J. Chromatogr. B* **680**: 65–70 (1996).
51. T.T. Franko, I.Yu. Galaev, R. Hatti-Kaul, N. Holmberg, L. Bülow, and B. Mattiasson, *Biotechnol. Tech.* **11**: 231–235 (1997).
52. J. Persson, H.-O. Johansson, I.Yu. Galaev, B. Mattiasson, and F. Tjerneld, *Bioseparation*, **9**: 105–116 (2000).
53. Y.G. Takei, T. Aoki, K. Sanui, N. Ogata, Y. Sakurai, and T. Okano, *Macromolecules* **27**: 6163–6166 (1994).
54. T. Kondo, M. Koyama, H. Kubota, and R. Katakai, *J. Appl. Polym. Sci.* **67**: 2057–2064 (1998).
55. F.-J. Schmitt, C. Park, J. Simon, H. Ringsdorf, and J. Israelachvili, *Langmuir* **14**: 2838–2845 (1998).
56. P.W. Zhu and D.H. Napper, *J. Phys. Chem. B* **101**: 3155–3160 (1997).
57. N. Yamada, T. Okano, H. Sakai, F. Karikusa, Y. Sawasaki, and Y. Sakurai, *Makromol. Chem. Rapid Commun.* **11**: 571–576 (1990).
58. T. Okano, N. Yamada, M. Okuhara, H. Sakai, and Y. Sakurai, *Biomaterials* **16**: 297–303 (1995).
59. T. Okano, A. Kikuchi, Y. Sakurai, Y.G. Takei, T. Aoki, and N. Ogata, *J. Controlled Release* **36**: 125–133 (1995).
60. M. Yamato, M. Okuhara, F. Karikusa, A. Kikuchi, Y. Sakurai, and T. Okano, *J. Biomed. Mater. Res.* **44**: 44–52 (1999).
61. Y. Ito, G. Chen, Y. Guan, and Y. Imanishi, *Langmuir* **13**: 2756–2759 (1997).
62. T. Bohanon, G. Elender, W. Knoll, P. Koeberle, J. Lee, A. Offenhaeusser, H. Ringsdorf, E. Sackmann, and J. Simon, *J. Biomater. Sci. Polym. ed.*, **8**: 19–39 (1996).
63. H. Kanazawa, Y. Kashiwase, K. Yamamoto, Y. Matsushima, A. Kikuchi, Y. Sakurai, and T. Okano, *Anal. Chem.* **69**: 823–830 (1997).
64. K. Hosoya, K. Kimata, T. Araki, and N. Tanaka, *Anal. Chem.* **67**: 1907–1911 (1995).
65. H. Go, Y. Sudo, K. Hosoya, T. Ikegami, and N. Tanaka, *Anal. Chem.* **70**: 4086–4093 (1998).
66. H. Kanazawa, Y. Matsushima, and T. Okano, *Trends Anal. Chem.* **17**: 435–440 (1998).
67. H. Yoshioka, M. Mikami, T. Nakai, and Y. Mori, *Polym. Adv. Technol.* **6**: 418–420 (1995).
68. B. Mattiasson and I. Yu. Galaev, in *High Resolution Chromatography*, P.A. Millner, ed., IRL Press, 1996, pp. 153–190.
69. I.Yu. Galaev, C. Warrol, and B. Mattiasson, *J. Chromatogr. A* **684**: 37–43 (1994).
70. Y. Ito, Y.S. Park, and Y. Imanishi, *Macromol. Rapid Commun.* **18**: 221–224 (1997).
71. Y. Ito, Y. Ochiai, Y.S. Park, and Y. Imanishi, *J. Am. Chem. Soc.* **119**: 1619–1623 (1997).
72. Y. Okahata, K. Ozaki, and T. Seki, *J. Chem. Soc. Chem. Commun.* 519–521 (1984).
73. Y. Okahata, H. Noguchi, and T. Seki, *Macromolecules* **19**: 493–494 (1986).
74. T. Nonaka, K. Hashimoto, and S. Kurihara, *J. Appl. Polym. Sci.* **66**: 209–216 (1997).
75. S. Akerman, P. Viinikka, B. Svarfvar, K. Putkonen, K. Jarvinen, K. Kontturi, J. Nasman, A. Urtti, and P. Paronen, *Int. J. Pharm.* **164**: 29–36 (1998).
76. R. Spohr, N. Reber, A. Wolf, G.M. Alder, V. Ang, C.L. Bashford, C.A. Pasternak, O. Hideki, and M. Yoshida, *J. Controlled Release* **50**: 1–11 (1998).
77. H. Omichi, M. Yoshida, M. Asano, N. Nagaoka, H. Kubota, R. Katakai, R. Spohr, N. Reber, A. Wolf, G.M. Alder, V. Ang, C.L. Bashford, and C.A. Pasternak, *Nucl. Instrum. Methods Phys. Res. Sect. B* **131**: 350–356 (1997).
78. S. Kurihara, Y. Ueno, and T. Nonaka, *J. Appl. Polym. Sci.* **67**: 1931–1937 (1998).
79. Y. Liu, M. Zhao, D.E. Bergbreiter, and R.M. Crooks, *J. Am. Chem. Soc.* **119**: 8720–8721 (1997).
80. T.G. Park and A.S. Hoffman, *Biotechnol. Prog.* **10**: 82–86 (1994).
81. K. Hosoya, E. Sawada, K. Kimata, T. Araki, N. Tanaka, and J.M.J. Frechet, *Macromolecules* **27**: 3973–3976 (1994).

82. Y. Imanishi and Y. Ito, *Pure Appl. Chem.* **67**: 2015–2021 (1995).
83. K. Ishihara and K. Matsui, *J. Polym. Sci. Polym. Lett. ed.* **24**: 413–417 (1986).
84. S. Kitano, K. Kataoka, Y. Koyama, T. Okano, and Y. Sakurai, *Makromol. Chem., Rapid Commun.* **12**: 227–233 (1991).
85. K. Kataoka, H. Miyazaki, B. Miyazaki, T. Okano, and Y. Sakurai, *J. Am. Chem. Soc.* **120**: 12694–12695 (1998).
86. S. Kitano, Y. Koyama, K. Kataoka, T. Okano, and Y. Sakurai, *J. Controlled Release* **19**: 162–170 (1992).
87. D. Shiino, K. Kataoka, Y. Koyama, M. Yokoyama, T. Okano, and Y. Sakurai, *J. Controlled Release* **28**: 317–319 (1994).
88. K. Nakamae, T. Miyata, A. Jikihara, and A.S. Hoffman, *J. Biomater. Sci. Polymer Ed.* **6**: 79–90 (1994).
89. Cosmetic Bases and Cosmetics Containing Polymers with Sol-Gel Transition Temperature, Jpn. Kokai Tokkyo Koho JP 09227329 A2, Sep. 2, 1997, Y. Mori, M. Yoshioka, and T. Mukoyama.
90. K. Lindell, Doctoral Thesis, Lund University, Sweden, 1996.
91. H. Hayashi, K. Kono, and T. Takagishi, *Bioconjugate Chem.* **10**: 412–418 (1999).
92. H. Ringsdorf, J. Simon, and F.M. Winnik, in *Colloid-Polymer Intercations. Particulate, Amphiphilic, and Biological Surfaces*, P.L. Dubin and P. Tong, eds., American Chemical Society, Washington DC, 1993, pp. 216–240.
93. X.S. Wu, A.S. Hoffman, and P. Yager, *Polymer* **33**: 4659–4662 (1992).
94. K. Kono, R. Nakai, K. Morimoto, and T. Takagishi, *Biochim. Biophys. Acta* **1416**: 239–250 (1999).
95. D.A. Tirrell, in *Pulsed and Self-Regulated Drug Delivery*, J. Kost, ed., CRC, Boca Raton, 1990, pp. 110–116.
96. D.D. Lasic, *Trends Biotechnol.* **16**: 307–321 (1998).
97. H. Ringsdorf, J. Venzmer, and F.M. Winnik, *Angew. Chem. Int. Ed. Engl.* **30**: 315–318 (1991).
98. H. Hayashi, K. Kono, and T. Takagishi, *Bioconjugate Chem.* **9**: 382–389 (1998).
99. S. Sugaj, K. Nitta, N. Ohno, and H. Nakano, *Colloid Polym. Chem.* **261**: 159–165 (1983).
100. K. Kono, A. Henmi, H. Yamashita, H. Hayashi, and T. Takagishi, *J. Controlled Release* **59**: 63–75 (1999).
101. H. Hayashi, K. Kono, and T. Takagishi, *Biochim. Biophys. Acta* **1280**: 127–134 (1996).
102. K. Kono, T. Igawa, and T. Takagishi, *Biochim. Biophys. Acta* **1325**: 143–154 (1997).
103. S. Cammas, K. Suzuki, C. Sone, Y. Sakurai, K. Kataoka, and T. Okano, *J. Controlled Release* **48**: 157–164 (1997).
104. Sensors for Concentrations of Salts, Jpn. Kokai Tokkyo Koho JP 63206653 A2 Aug. 25, 1988, Y. Tomita.
105. M. Irie, in *Responsive Gels: Volume Transitions II.*, K. Dusek, ed., Springer-Verlag, Berlin, 1993, pp. 50–65.
106. R. Kröger, H. Menzel, and M.L. Hallensleben, *Macromol. Chem. Phys.* **195**: 2291–2298 (1994).
107. A. Mamada, T. Tanaka, D. Kungwachakun, and M. Irie, *Macromolecules* **23**: 157–1519 (1990).
108. M. Irie, Y. Misumi, and T. Tanaka, *Polymer* **34**: 4531–4535 (1993).
109. N. Monji and A.S. Hoffman, *Appl. Biochem. Biotechnol.* **14**: 107–120 (1987).
110. Q.-Z. Zhu, F. Liu, J.-G. Xu, W. Su, and J. Huang, *Anal. Chim. Acta* **375**: 177–185 (1998).
111. Y.G. Takei, M. Matsukata, T. Aoki, K. Sanui, N. Ogata, A. Kikuchi, Y. Sakurai, and T. Okano, *Bioconjugate Chem.* **5**: 577–5582 (1994).
112. B.B. Dzantiev, A.N. Blintsov, A.F. Bobkova, V.A. Izumrudov, and A.B. Zezin, *Dokl. Biochem.* (in Russian) **342**: 549–552 (1995).
113. M. Charles, R.W. Coughlin, and F.X. Hasselberger, *Biotechnol. Bioeng.* **16**: 1553–1556 (1974).
114. E. VanLeemputten and M. Horisberger, *Biotechnol. Bioeng.* **18**: 587–590 (1976).
115. M. Taniguchi, K. Hoshino, Y. Netsu, and M. Fujii, *J. Chem. Eng. Jpn.* **22**: 313–314 (1989).
116. M. Fujimura, T. Mori, and T. Tosa, *Biotechnol. Bioeng.* **29**: 747–752 (1987).
117. R.W. Coughlin, M. Aizawa, and M. Charles, *Biotechnol. Bioeng.* **18**: 199–208 (1976).
118. A.L. Margolin, V.A. Izumrudov, V.K. Svedas, and A.B. Zezin, *Biotechnol. Bioeng.* **24**: 237–240 (1982).
119. Y.E. Kirsh, I. Yu. Galaev, T.M. Karaputadze, A.L. Margolin, and V.K. Svedas, *Biotechnology* (Russian), 184–189 (1987).
120. H. Lee and T.G. Park, *Biotechnol. Prog.* **14**(3): 508–516 (1999).
121. A.L. Nguyen and J.H.T. Luong, *Biotechnol. Bioeng.* **34**: 1186–1190 (1989).
122. J.-P. Chen, *J. Chem. Technol. Biotechnol.* **73**: 137–143 (1995).
123. F. Liu, G. Tao, and R. Zhuo, *Polymer J.* **25**: 561–567 (1993).
124. K. Hoshino, M. Taniguchi, H. Kawaberi, Y. Takeda, S. Morohashi, and T. Sasakura, *J. Ferment. Bioeng.* **83**: 246–252 (1997).
125. S. Takeuchi, I. Omodaka, K. Hasegawa, Y. Maeda, and H. Kitano, *Makromol. Chem.* **194**: 1991–1999 (1993).
126. J.-P. Chen, D.-H. Chu, and Y.-M. Sun, *J. Chem. Technol. Biotechnol.* **69**: 421–428 (1997).
127. P.S. Stayton, T. Shimoboji, C. Long, A. Chilkoti, G. Chen, J.M. Harris, and A.S. Hoffman, *Nature* **378**: 472–474 (1995).
128. K. Okumura, K. Ikura, M. Yoshikawa, R. Sasaki, and H. Chiba, *Agric. Biol. Chem.* **48**: 2435–2440 (1984).
129. H. Kitano, C. Yan, and K. Nakamura, *Makromol. Chem.* **192**: 2915–2923 (1991).
130. A. Kondo, K. Imura, K. Nakama, and K. Higashitani, *J. Ferment. Bioeng.* **78**: 241–245 (1994).
131. T. Kondo, T. Urabe, and K. Yoshinaga, *Colloids Surf. A: Physicochem. Eng. Aspects* **109**: 129–136 (1996).
132. A. Kondo and H. Fukuda, *J. Ferment. Bioeng.* **84**: 337–341 (1997).
133. Y. Sun, X.H. Jin, X.-Y. Dong, K. Yu, and X.Z. Zhou, *Appl. Biochem. Biotechnol.* **56**: 331–339 (1996).
134. K. Hoshino, M. Taniguchi, H. Ueoka, M. Ohkuwa, C. Chida, S. Morohashi, and T. Sasakura, *J. Ferment. Bioeng.* **82**: 253–258 (1996).
135. M. Taniguchi, M. Kobayashi, and M. Fujii, *Biotechnol. Bioeng.* **34**: 1092–1097 (1989).
136. K. Hoshino, M. Taniguchi, Y. Netsu, and M. Fujii, *J. Chem. Eng. Jpn.* **22**: 54–59 (1989).
137. K. Hoshino, M. Taniguchi, H. Marumoto, and M. Fujii, *Agric. Biol. Chem.* **53**: 1961–1967 (1989).
138. K. Hoshino, M. Taniguchi, H. Marumoto, and M. Fujii, *J. Ferment. Bioeng.* **69**: 228–233 (1990).
139. Yu.V. Galaev, *Prikl. Bioxim. Mikrobiol.* **30**: 167–170 (1994).
140. I.Yu. Galaev and B. Mattiasson, *Trends Biotechnol.* **17**: 335–340 (1999).

POLYMERS, FERROELECTRIC LIQUID CRYSTALLINE ELASTOMERS

RUDOLF ZENTEL
University of Mainz
Mainz, Germany

INTRODUCTION

Ferroelectric materials are a subclass of pyro- and piezoelectric materials (Fig. 1). They are very rarely found in crystalline organic or polymeric materials because ferroelectric hysteresis requires enough molecular mobility to reorient molecular dipoles in space. So semicrystalline polyvinylidene fluoride (PVDF) is nearly the only known compound (1). On the contrary, ferroelectric behavior is very often observed in chiral liquid crystalline materials, both low molar mass and polymeric. For an overview of ferroelectric liquid crystals, see (2). Tilted smectic liquid crystals that are made from chiral molecules lack the symmetry plane perpendicular to the smectic layer structure (Fig. 2). Therefore, they develop a spontaneous electric polarization, which is oriented perpendicular to the layer normal and perpendicular to the tilt direction. Due to the liquid-like structure inside the smectic layers, the direction of the tilt and thus the polar axis can be easily switched in external electric fields (see Figs. 2 and 4).

Here, we discuss materials (LC-elastomers) that combine a liquid crystalline phase and ferroelectric properties (preferable the chiral smectic C^* phase) in a polymer network structure (see Fig. 3). The coupling of the liquid crystalline director to the network or the softness of the network is chosen so that reorientation of the polar axis is still possible. Thus densely cross-linked systems, that possess a polar axis but cannot be switched (3) will be excluded.

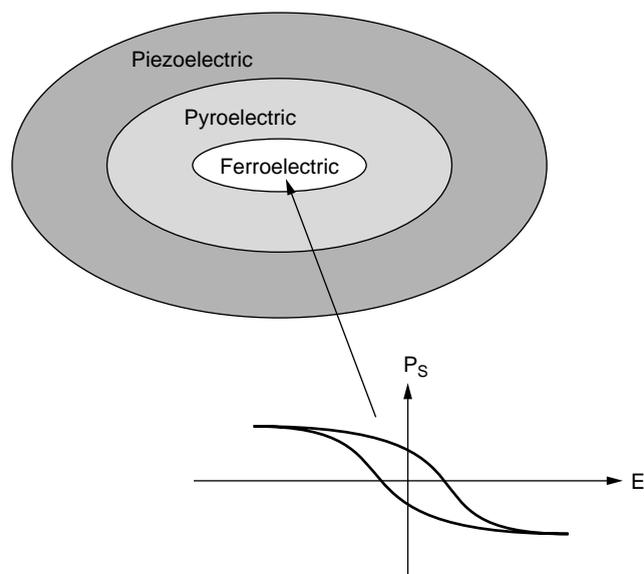


Figure 1. Ferroelectric hysteresis that shows the spontaneous polarization P_s of a ferroelectric material reversed by an applied electric field E .

It is the role of the network (1) to form a rubbery matrix for the liquid crystalline phase and (2) to stabilize a director configuration. LC-materials that have these properties can be made either (see Fig. 3) by covalently linking the mesogenic groups to a slightly cross-linked rubbery polymer network structure (see Fig. 3a) (4–6) or by dispersing a phase-separated polymer network structure within a low molar mass liquid crystal (see Fig. 3b) (8,9). Both systems possess locally a very different structure. They may show, however, macroscopically similar properties.

LC-elastomers (see Fig. 3a) have been investigated in detail (4–7). Although the liquid crystalline phase transitions are nearly unaffected by the network, the network retains the memory of the phase and director pattern during cross-linking (7). In addition, it freezes fluctuations of the smectic layers and leads to a real long range order in one dimension (11). An attempt to change the director pattern by electric or magnetic fields in LC-elastomers leads to a deformation of the network and to an elastic response (see Fig. 4). As a consequence of this, nematic LC-elastomers could never be switched in electric fields, if the shape of the elastomer was kept fixed. For freely suspended pieces of nematic LC-elastomers, shape variations in electric fields have been observed sometimes (12,13). In ferroelectric liquid crystals, the interaction with the electric field is, however, much larger. Thus, it has been possible to prepare real ferroelectric LC-elastomers (see Fig. 4) (14,15). In these systems, the polymer network stabilizes one switching state like a soft spring. It is, however, soft enough to allow ferroelectric switching. Therefore the ferroelectric hysteresis can therefore be measured in these systems. It is, however, shifted away from zero voltage (see Fig. 4).

SYNTHESIS OF FERROELECTRIC LC-ELASTOMERS

The ferroelectric LC-elastomers described so far (14–17, 44–46) are mostly prepared from cross-linkable ferroelectric polysiloxanes (see Fig. 5), which are prepared by hydrosilylation of precursor polysiloxanes (18). The cross-linking is finally initiated by irradiating a photoradiation generator, which leads to oligomerization of acrylamide or acrylate groups (see Fig. 5). The functionality of the net points is thus high (equal to the degree of polymerization) and varies with the cross-linking conditions.

The advantage of this photochemical-initiated cross-linking is that the crosslinking can be started—at will after the liquid crystalline polymer is oriented and sufficiently characterized in the uncross-linked state (see Fig. 6). The advantage of using polymerizable groups (acrylates) for cross-linking is that small amounts of these groups are sufficient to transform a soluble polymer into a polymer gel and that the chemical reactions happens far away from the mesogen. Cinnamoyl moieties, on the other hand (19), require a high concentration of these groups for cross-linking. The dimers thus formed are, in addition, nonmesogenic. Figure 7 summarizes the ferroelectric LC-elastomers discussed in this article. Two different positions of cross-linkable groups are used. In polymer P1, the cross-linking group is close to the siloxane chains, which are known to

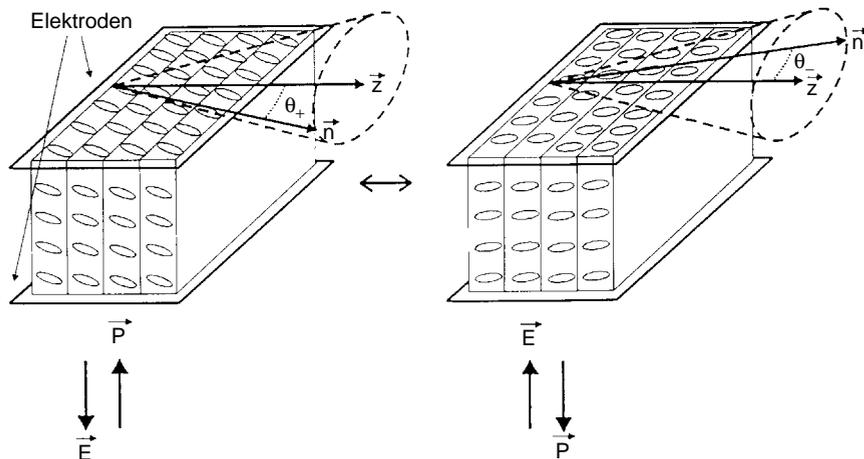


Figure 2. Schematic of the bistable switching of a ferroelectric liquid crystal in the “surface stabilized FLC” configuration.

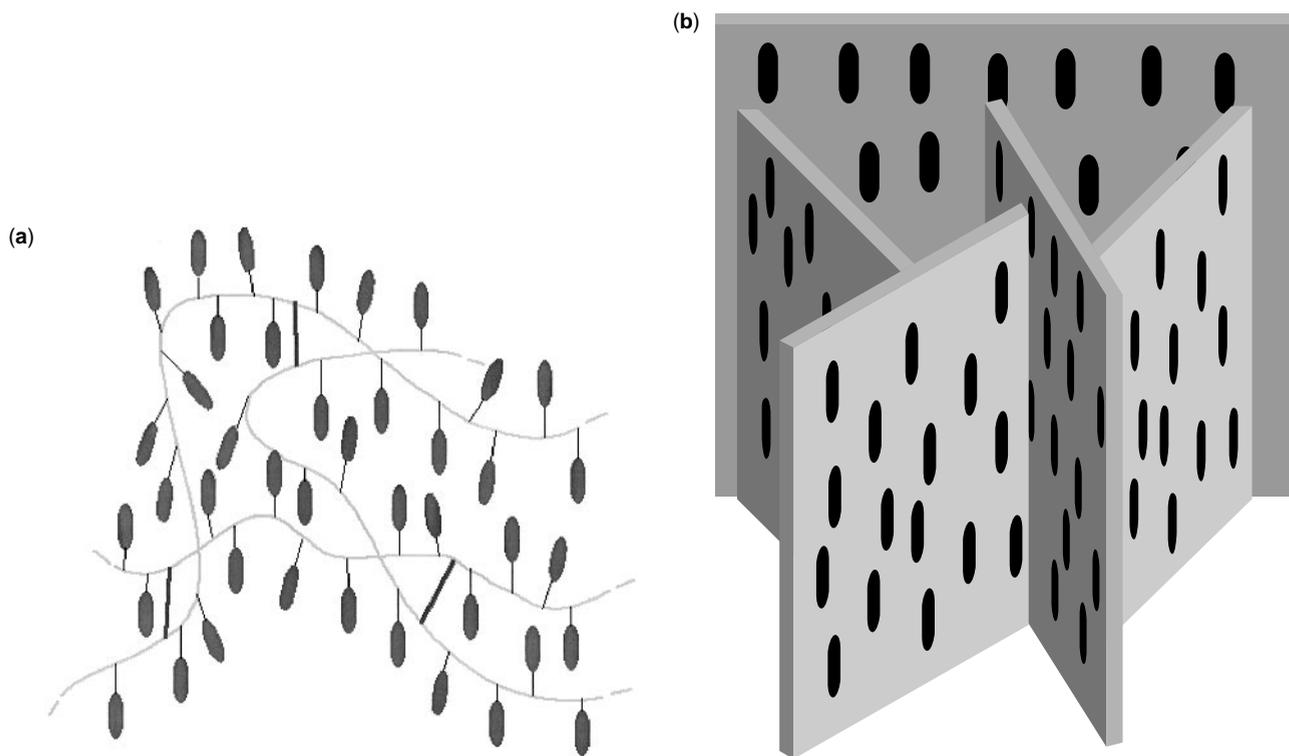


Figure 3. Network: soft, can be transformed like rubber band, but retains its shape and couples to director orientation because (a) director is preferably parallel (or perpendicular) to polymer chains (LC-elastomer) (4–8). (b) Director aligns (parallel) to chains in oriented phase-separated polymer network structure (low molar mass LC in LC-thermoset) (8,37).

microphase separate from the mesogenic groups (18,19). Therefore, the crosslinking should proceed mostly within the siloxane sublayers. In polymers P2 and P3, the crosslinking group is located at the end of mesogens. Therefore, the crosslinking should proceed mostly between different siloxane layers (see Fig. 7). A comparison of these elastomers allows evaluating structure–property relationships (17,23,26).

Properties and Characterization

Ferroelectric Characterization (Uncross-linked Systems). Before cross-linking, polarization, tilt angle, and switching times can be determined in the usual way (17,18,21). Figure 8 shows the temperature dependence of the spontaneous polarization for polymers P1–P3. For the homopolymer related to polymer P2, all relevant parameters

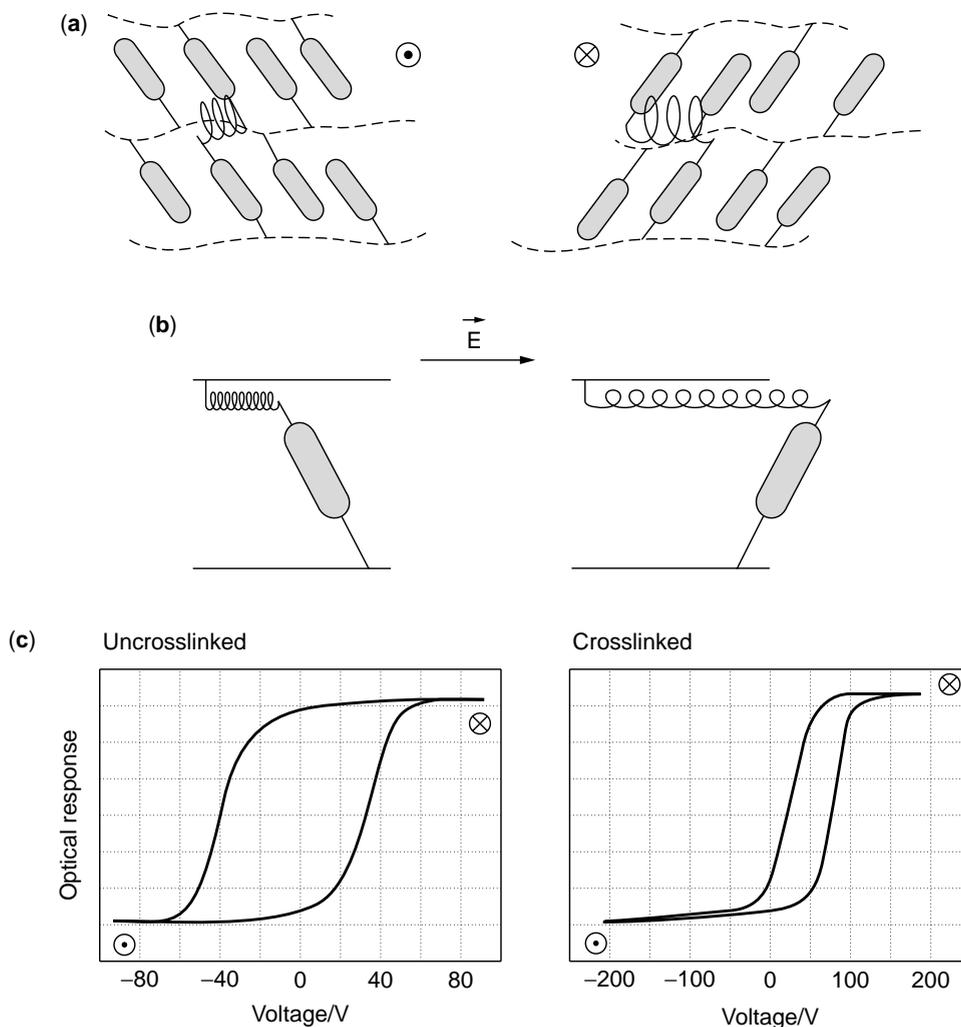


Figure 4. Schematic of the ferroelectric LC-elastomer and its two switching states (14): (a) A polymer chain acts as cross-linking point by connecting different mesogenic groups attached to the main polymer chains. A ferroelectric switching in this elastomer extends polymer chains. (b) The entropy elasticity arising from this acts like a spring that stabilizes one state. (c) For the uncrosslinked system (left) the hysteresis is symmetrical to zero voltage and both states are equal. After cross-linking in one polar state (right), only that state is stable with no electric field, and the hysteresis is no longer symmetrical to zero voltage.

were determined in a careful study by Kocot et al. (22). It seems that the electroclinic effect is especially strong in these polysiloxanes (15). This has implications for the freezing of a memory of the tilt angle present during cross-linking. Therefore, ferroelectric elastomers, which have been crosslinked in the smectic A phase while applying an electric field, produce a stable macroscopic polarization (tilt) after cooling into the smectic C* phase (17).

Properties of Ferroelectric LC-Elastomers. The crosslinking reactions of a series of copolymers analogous to polymer P2, but differing in the amount of cross-linkable groups were studied by FTIR spectroscopy (16). These measurements show a decrease of the acrylamide double bond on irradiation. Conversions between 60 to 84% were observed. The uncertainty of the conversion, however, is high because only very few double bonds are present in polymer

P2 and they are visible in the infrared spectrum at rather low intensity.

Mechanical measurements, which show how this photochemical crosslinking (conversion of double bonds) leads to an elastic response of the network are, however, still at the beginning because photo-cross-linking can be performed only in thin layers of some microns. It is best performed between two glass slides to exclude oxygen.

AFM measurements of photo-cross-linked free standing films show changes in topology during stretching (see Properties of Ferroelectric LC-Elastomers—AFM Imaging of Thin Films) (23). They do, however, not allow measuring elastic moduli.

The most promising approach to obtaining elastic data for these ferroelectric elastomers is investigation of LC-elastomer balloons (25,26). For this purpose, an experimental setup was developed on the basis of an apparatus

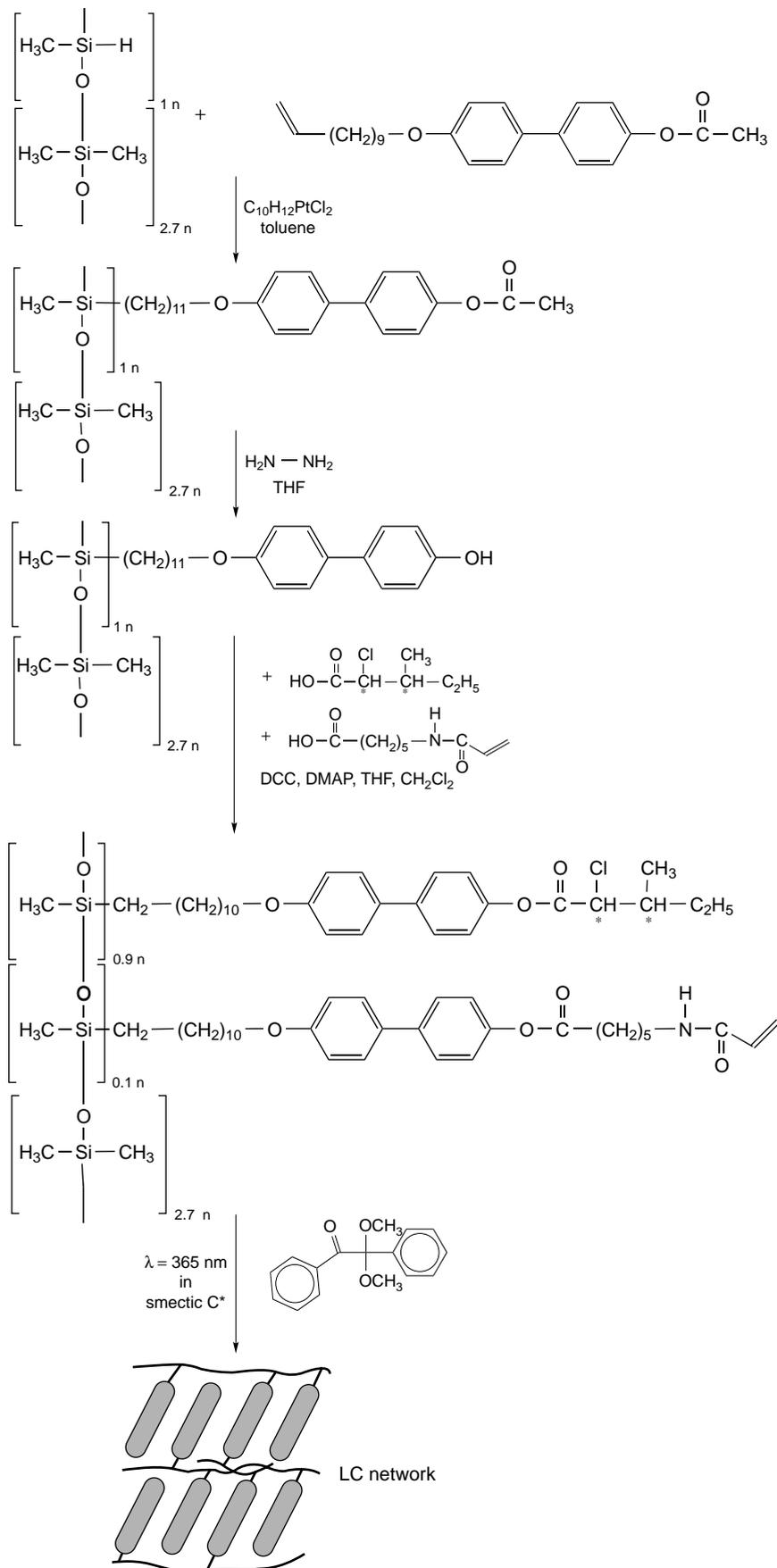


Figure 5. Synthetic route to the cross-linkable polysiloxane P2 and the following preparation of the oriented smectic C* network using UV light in the presence of a photoinitiator (1,1-dimethoxy-1-phenylacetophenone) (14).

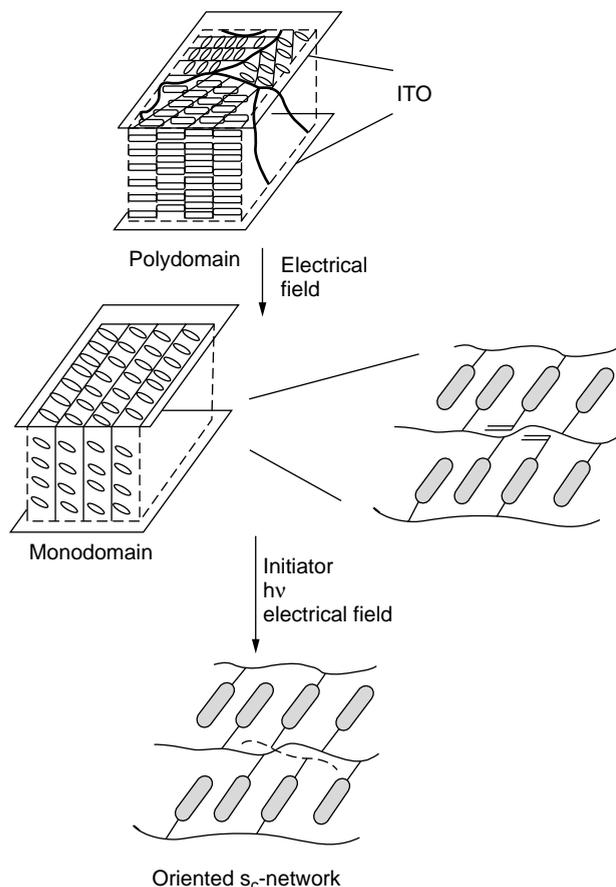


Figure 6. Preparation of polar smectic C* monodomains (14,15) (ITO: indium tin oxide).

designed to study smectic bubbles (25). Freely suspended films of the uncross-linked material behave like ordinary smectic films. They can be inflated to spherical bubbles several mm in diameter (the thickness of a smectic-layer skin is about 50 nm). These bubbles are stabilized by the smectic-layer structure and their inner pressure p is related to the surface tension and the bubble radius R by the Laplace–Young equation, $p \propto 1/R$. After exposure to UV light, the material is cross-linked, and an anisotropic elastomer is formed. When the cross-linked bubbles are inflated/ deflated, the radius–pressure curve reverses its slope and gives direct access to the elastic moduli of the material (26). Because the deformation during inflation of the balloon is isotropic in the layer plane, the material should contract in the direction of the layer normal.

Mechanical measurements of chemically crosslinked LC-elastomers have been made extensively (4,5,27,28, 36,41–43). For these systems, it can be shown that stretching allows orientation of the liquid crystalline phase. In ideal situations, it is thus possible to prepare a ferroelectric monodomain by stretching (28,30,36). This result can be rationalized as a two-stage deformation process (see Fig. 9) (36). This possibility of orienting or reorienting the polar axis mechanically is the basis for the piezoelectric properties to be discussed later. Ferroelectric switching could not be observed for any of the chemically crosslinked systems.

This may occur because chemically cross-linked films are too thick (several 100 μm compared to about 10 μm for photochemically cross-linked systems) and the electric field applied is therefore too small. In addition, the cross-linking density in chemically cross-linked systems is presumably higher.

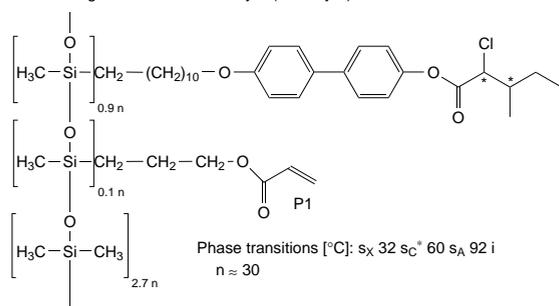
Ferroelectric Properties LC-Elastomers. The ferroelectric properties of the photochemically crosslinked elastomers E1 to E3 differ significantly and depend on the topology of the network formed. For the systems that have interlayer cross-linking (see Fig. 7, E2 and E3), the switching time is increased greatly. Therefore, spontaneous polarization can no longer be determined by the triangular wave method. Slow switching is, however, still possible and therefore ferroelectric hysteresis can be measured optically (see Figs. 4c and 10) (14,15). After photochemical cross-linking in a ferroelectric monodomain, the ferroelectric hysteresis shows stabilization of the orientation present during cross-linking. At zero external voltage, only this state is stable. The second switching state can, however, be reached. Therefore, the network acts like a spring that stabilized one state because switching to the other state leads to a deviation from the most probable conformation of the polymer chain (32) (see Fig. 10). Then, the shift of the center of the hysteretic loop away from zero voltage gives the magnitude of the electric field necessary to balance the mechanical field of the network. The asymmetry of the hysteresis increases with the cross-linking density (17). For high cross-linking densities, switching remains possible only if the spontaneous polarization is rather high (17). Otherwise, the network prohibits switching. The asymmetry of ferroelectric switching could also be proven by polarized FT-IR spectroscopy (33). Increasing the temperature of this ferroelectric elastomer leads to narrowing of the hysteretic loop, which is lost at the transition to the smectic A phase (see Fig. 10).

This behavior is best interpreted by plotting the liquid crystalline potential, the elastic potential of the network, and their superposition in one graph (15) (see Fig. 11). As the network is formed in the smectic C* phase, an internal elastic field is created, which has its minimum value for the tilt angle and tilt direction during cross-linking. Other tilt angles are destabilized.

The elastomer that has preferable intralayer cross-linking (E1, see Fig. 7) shows completely different behavior (see Fig. 12) (17,34). In this case, the switching time increases by less than a factor of 2, the polarization can still be determined, and measurement of the ferroelectric hysteresis shows no stabilization of the switching state present during cross-linking. Then, the coupling between the orientation of the mesogens and the network conformation is obviously very weak. The network stabilizes the smectic layer structure (see Properties of Ferroelectric LC-Elastomers—AFM Imaging of Thin Films), but it does not stabilize the tilt direction. Therefore, the polar axis can be switched easily. This is the result of the network topology (see Fig. 7) in which interlayer cross-linking is rare.

Properties of Ferroelectric LC-Elastomers—AFM Imaging of Thin Films. Freestanding films can be prepared from

Crosslinking within siloxane sublayer (intra layer)



Crosslinking between siloxane sublayers (inter layer)

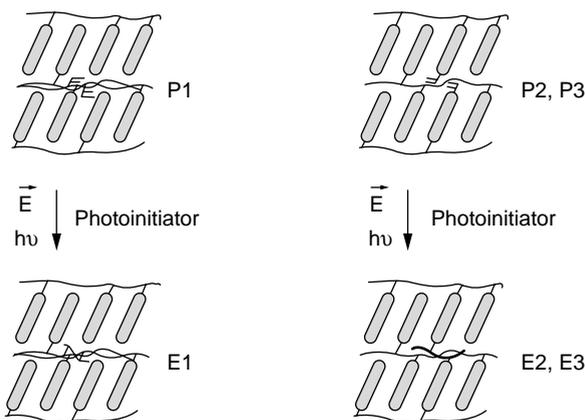
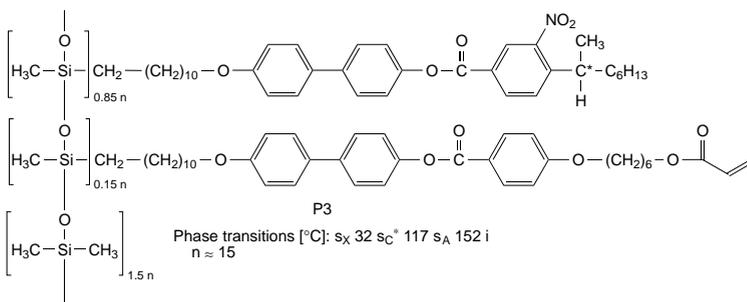
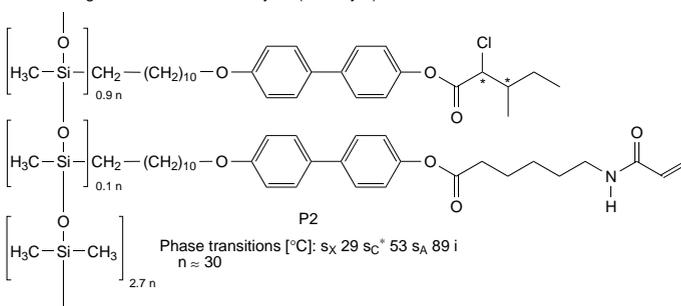


Figure 7. Chemical structure and phase transition temperatures of polymers P1–3 (17). (a) P1 is designed to favor intralayer cross-linking. (b) P2–3 forming an interlayer network.

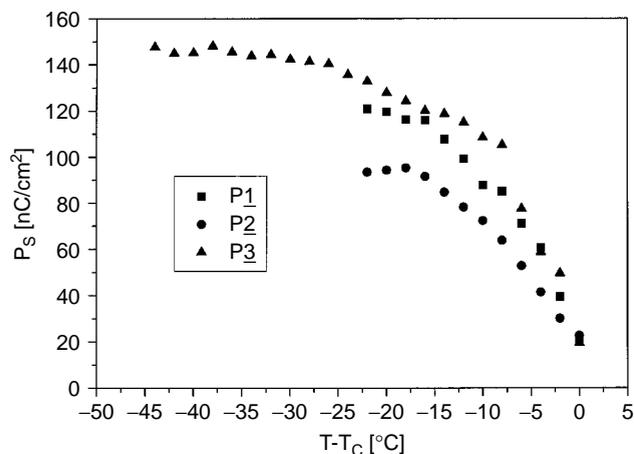


Figure 8. Temperature dependence of the spontaneous polarization P_S for the polymers P1–3 measured by the triangular wave method (17).

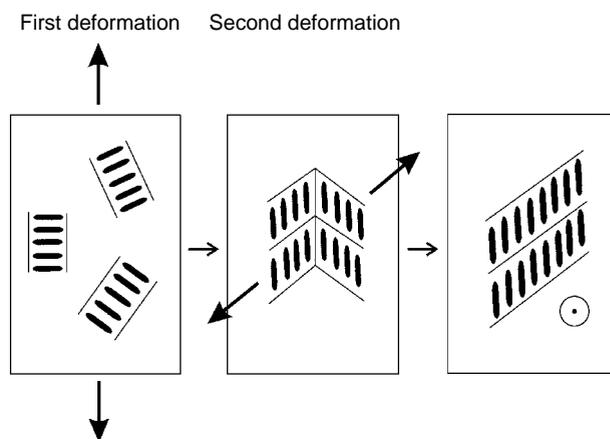


Figure 9. Two-step deformation process of a chiral smectic C^* elastomer that displays macroscopic polarization at the end (36).

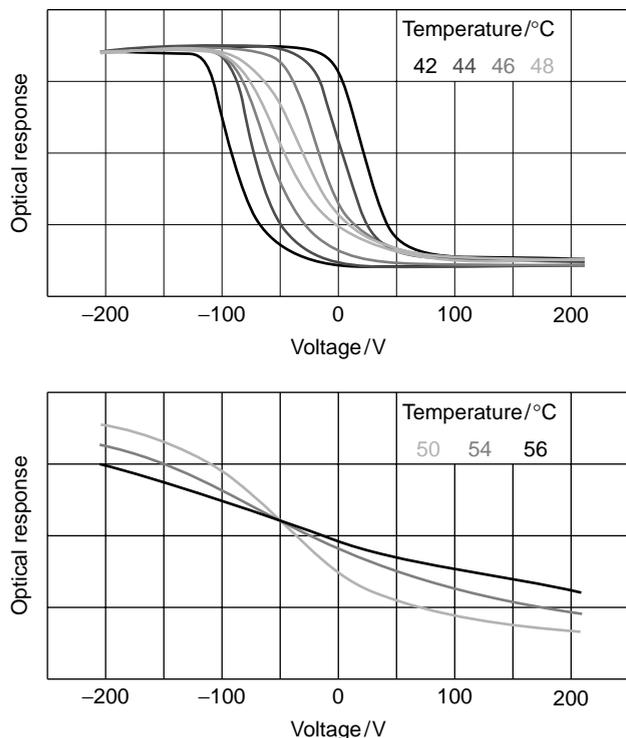


Figure 10. Temperature dependence of the optical hysteresis of elastomer E2 (S_C^* 49°C S_A) (15). (a) Ferroelectric behavior of the S_C^* phase (42, 44, 46, and 48°C, respectively). (b) Electroclinic behavior of the S_A phase (50, 54, and 56°C, respectively).

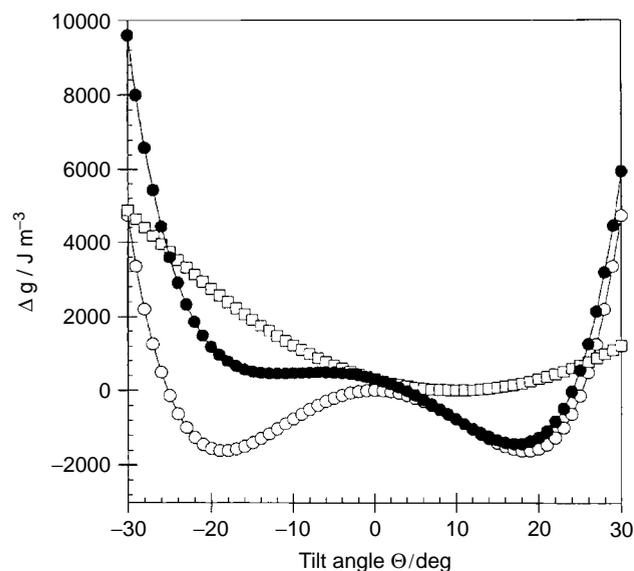


Figure 11. Effect of network force on the free energy density (15) 2 K below the phase transition, S_C^* phase: (○) calculated potential of the S_C^* phase, (◇) force due to the network (●) superposition of both.

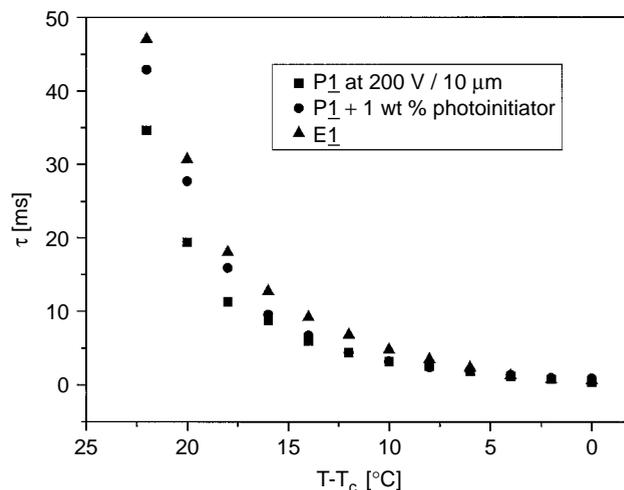


Figure 12. Temperature dependence of the switching time τ (defined as 0–100% change in transmission) for P1 and E1 [see (17) for comparison].

uncross-linked polymers (see also the smectic balloons in this context). They can be photo-cross-linked and transferred to a solid substrate. Thereafter, the topology of the films can be imaged by AFM, which gives direct visualization of the smectic layer structure at low temperatures. The uncross-linked polymers can only be imaged at low temperatures, deep inside the smectic phase and in the tapping mode, which does not induce strong lateral forces. At higher temperatures, the sample is too soft and mobile to allow imaging. Cross-linked elastomers, on the other hand, are mechanically stable, and films sustain the tapping mode and also the contact mode of the atomic force microscope (35). This holds both for intra- and interlayer cross-linked systems. Because measurements can be done in all phases, it is also possible to determine the change of the smectic layer thickness at the phase transitions directly. For elastomer E1, for example, the smectic layer thickness is 4.2 nm in the smectic C^* phase (36°C, tilt angle about 30°). It increases to 4.4 nm at 50°C in the smectic A phase (35). This corresponds to X-ray measurements.

To analyze the impact of the molecular structure on network properties, elastomers are compared, which are identical except for the molecular position of the cross-linkable group: (1) elastomer E1 that has cross-linkable groups attached to the backbone via a short spacer (intralayer cross-linking) and (2) elastomer E2 where the cross-linkable group is in the terminal position of a mesogenic side group (intralayer cross-linking) (23,24). When mechanical stress (stretching) is imposed on thin films in homeotropic orientation, the two elastomers react differently to the deformation (23,24), as seen by AFM imaging of the surface topology (see Fig. 13). For elastomer E1, “intralayer” cross-linking results in two-dimensional networks in the backbone layers, separated by liquid-like FLC side-group layers. Because there are practically no vertical connections in this intralayer network, no vertical distortions occur. Therefore, this elastomer can be stretched up to 100%, the surface remains smooth, and the layers deform

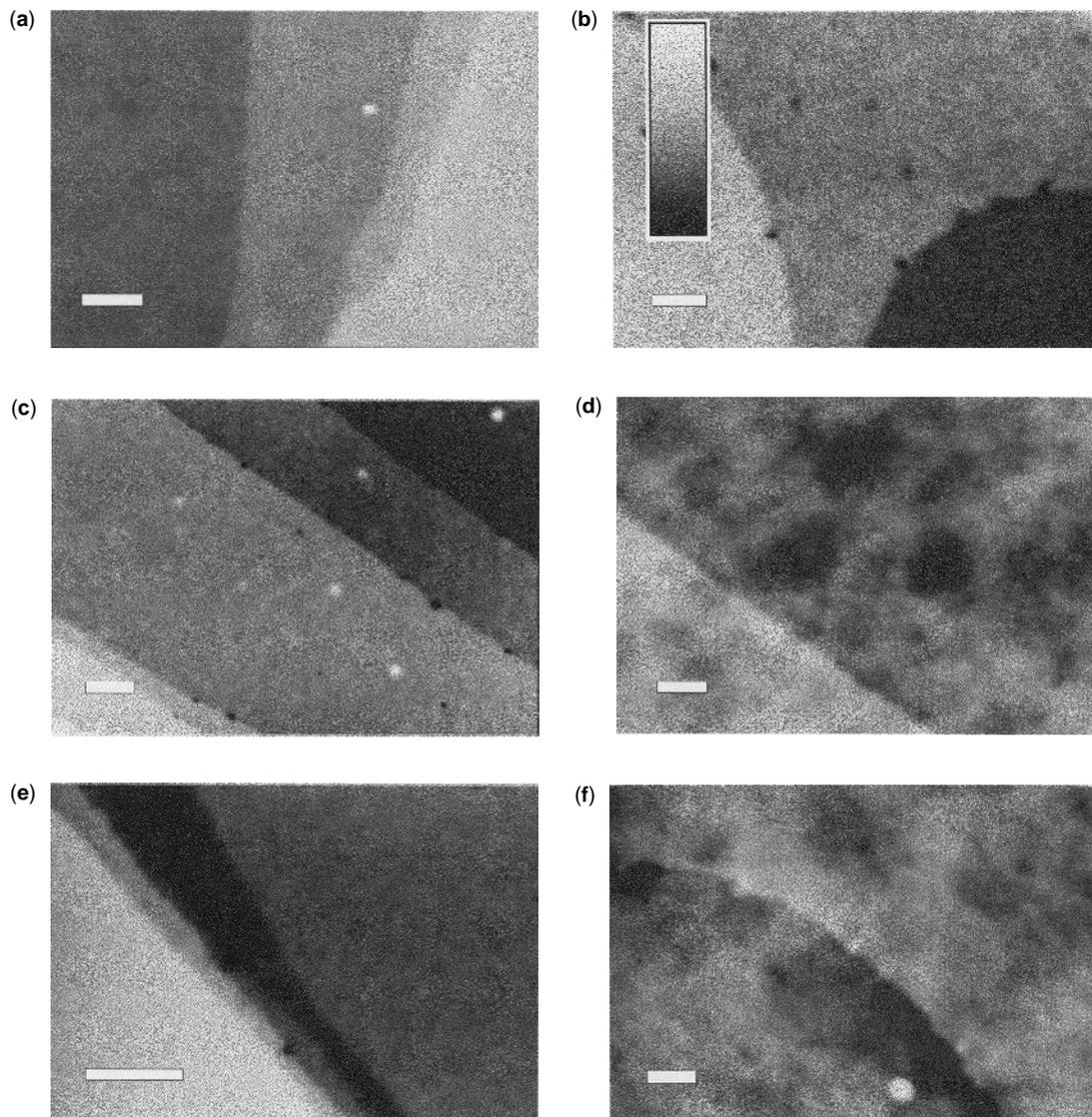


Figure 13. Surface topography of as prepared and stretched transferred films of elastomer $E1_{x=0.1}$ (a: $\lambda = 0$, b: $\lambda = 10\%$), $E2_{x=0.07}$ (c: $\lambda = 0$, d: $\lambda = 12\%$), and $E2_{x=0.25}$ (e: $\lambda = 0$, f: $\lambda = 2\%$). Scale bars $1\ \mu\text{m}$, height scale $25\ \text{nm}$ valid for all images. The surface of all polymers show plateau patterns. In the E2 samples, the lateral strain leads to surface deformation (24).

affinely. In elastomer E2, a three-dimensional, “interlayer” network is formed; the system reacts by distorting the smectic layering. Therefore, only smaller stretching ratios are accessible and the surface roughens and buckles during stretching. The distortion strength increases with a higher cross-linking density.

Piezoelectric Properties of Ferroelectric LC-Elastomers.

Because a ferroelectric material has to be piezoelectric (see Fig. 1), observation of a piezoresponse is natural. It has been observed for ferroelectric LC-elastomers (14,15,44–46) and also for more densely cross-linked systems (36–39) for which no ferroelectric switching could be observed. For the elastomers described here it is, however, possible to change the piezoresponse (14) by reorienting the polar axis in an external field (see E2 in Fig. 14). For this

experiment, the polar axis was kept in one orientation during cross-linking. This resulted in a positive piezoresponse (see Fig. 14). Thereafter, the direction of the polar axis was inverted by applying an external field of opposite direction. Then, the external field was removed and the piezocoefficient was measured. At first, a piezoresponse of opposite sign (negative) but identical value is determined (see Fig. 14). In the field-free state, this piezoresponse continuously decreases, it goes through zero, increases again, and finally reaches the original positive value. This experiment is comparable to the hysteresis measurements of Figs. 4 and 10 because it shows that two polar states are accessible, but the one present during cross-linking is stabilized.

The shape variation under application of an external electric field was most intensively studied for microtomed

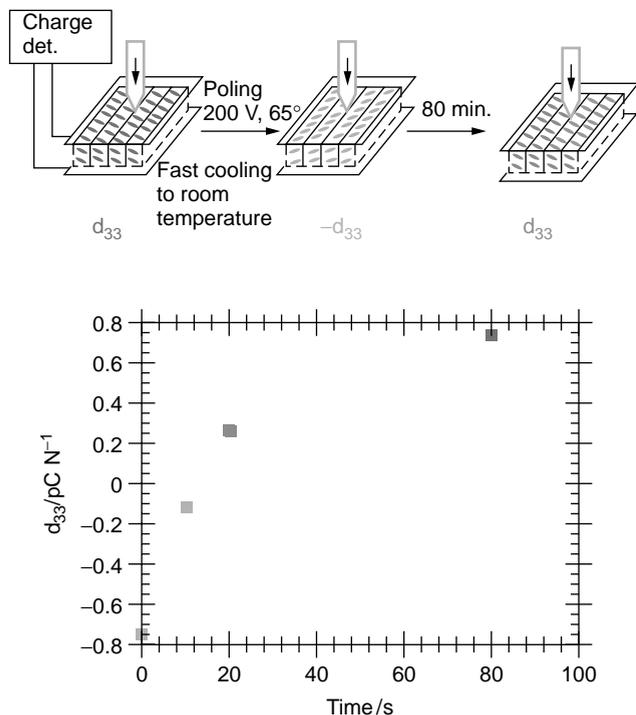


Figure 14. Relaxation of the piezoelectric coefficient d_{33} of elastomer E2 at room temperature after reversal poling at 65°C (S_A phase) (14).

piezes of ferroelectric elastomers, which had been oriented by drawing (44–46). These experiments show only a small shape variation if the field is applied parallel to the polar axis of the monodomain. The effects become, on the other hand, rather large if the smectic layer structure (chevron

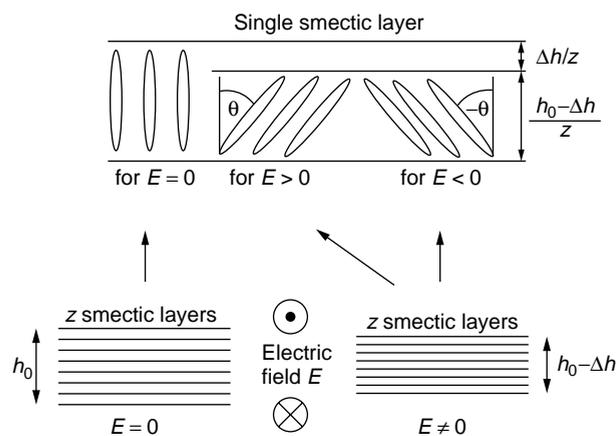


Figure 15. In the S_A^* phase, the mesogenic parts (depicted as ellipsoids) of the elastomeric macromolecule stand upright ($\theta = 0^\circ$) inside the single smectic layers. By applying a lateral electric field (perpendicular to the plane of the paper), a tilt angle θ that is proportional to the electric field E can be induced (electroclinic effect). The sign of the tilt depends on the sign of the electric field E . Hence, each smectic layer shrinks by $\Delta h/z$ twice during one period of the electric field. The shrinkage Δh of the whole film is measured by the interferometer as an optical phase shift between the sample beam and the reference beam.

Table 1.

Material	Electrostrictive strain ε (at electric field strength E)	Ref.
Free-standing FLCE film	4% lateral strain (1.5 MV/m)	40
Spin-coated FLCE	~1% lateral layer thickness shrinkage (2.0 MV/m)	40
PMN _{0.7} PT _{0.3}	0.15% (1 MV/m)	47
Lead magnesium niobate— Lead titanate ceramics		
P(VDF-TrFE)	4% (150 MV/m)	47
electron irradiated poly(vinylidene fluoride trifluoroethylene) copolymer		
PBLG monomolecular, grafted layer of poly- γ -benzyl-L-glutamate	0.005% (300 MV/m)	49

texture) rearranges (44,46). To get a large electrostrictive response, which can be understood on a molecular level, the geometry presented in Fig. 15 was chosen (40). The application of an electric field parallel to a smectic layer leads to a tilt of the mesogens (electroclinic effect). Thereby, the thickness of the layer decreases. For a stack of layers, the effect sums up over all layers. As a result, the thickness perpendicular to the smectic layers decreases if a field is applied parallel to the layers (see Fig. 15). Because the electroclinic effect is relatively large in these polymers (15,33), a large variation in thickness is expected.

X-ray diffraction measurements prove the electrically induced shrinkage of single smectic layers. A freestanding ultrathin (75 nm) film of a ferroelectric liquid crystalline elastomer (similar to E2) was used to measure the shape variation (electrostrictive response) associated with this (40). It was measured by a high precision (± 3 pm at 133 Hz) Michelson interferometer. The measurements (see Fig. 16) exhibit extremely high electrostrictive strains of 4% in an electric field of only 1.5 MV/m, which is, to our knowledge, a new world record for the corresponding electrostrictive coefficient a . The effect exhibits typical electroclinic behavior, which means that it is caused by an electrically induced tilt of the chiral LC molecules. As a consequence of chirality, the primary strain is perpendicular to the applied field. Hence, a new material that has a giant electrostriction effect is introduced, where the effect can be fully understood on a molecular level. The characterization of these materials is summarized in Table 1.

CONCLUSION

Ferroelectric LC-elastomers represent an interesting class of material because they combine the ordering of liquid crystalline ferroelectric phases and the rubber elasticity of polymer networks. Switching of the electric polarization leads to deformation of the polymer network, equivalent to

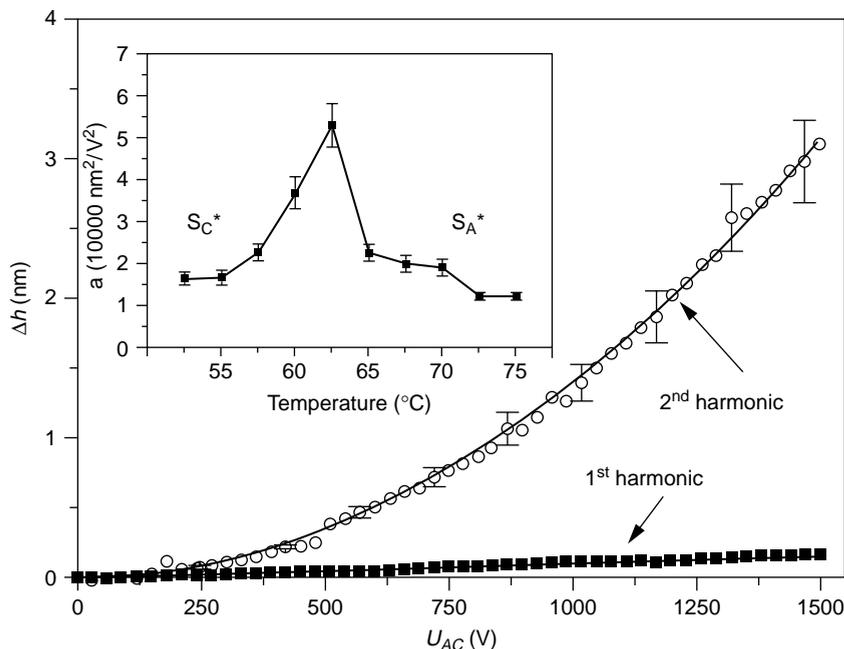


Figure 16. Electrostriction of a ferroelectric LC-elastomer (40). Big diagram: Thickness variation Δh as a function of the applied ac voltage U_{ac} . Interferometric data were obtained at the fundamental frequency of the electric field (piezoelectricity, first harmonic: +) and at twice the frequency (electrostriction, second harmonic: o). Sample temperature: 60°C. Inset: Electrostrictive coefficient a (+) versus temperature. At the temperature where the non-cross-linked polymer would have its phase transition $S_C^* - S_A^*$ (about 62.5°C), the tilt angle of 0° is unstable. That is why the electroclinic effect is most effective at this temperature. An electric field of only 1.5 MV/m is sufficient to induce lateral strains of more than 4%.

stretching a spring, and creates a stress in the network of polymer chains.

The interaction of mesogens and the network can be varied by using different topologies of net points: Crosslinking is carried out either within the siloxane sublayers (leading to fast switching elastomers) or between the siloxane sublayers (resulting in an elastomer that favors the ferroelectric switching state in which the cross-linking reaction took place).

Because the orientation of the smectic phase couples to the polymer network, electromechanical measurements show a piezoelectric effect. Mechanical deformation leads to polarization or an external electric field to deformation of the sample. Applying the electric field parallel to the smectic layer structure leads to an extremely high electrostrictive strain of 4% in an electric field of only 1.5 MV/m.

ACKNOWLEDGMENT

The work in this summary was possible only by close cooperation with several groups from physics and physical chemistry. We give special thanks to the groups of Profs. F. Kremer and R. Stannarius (Leipzig). A. Helfer is thanked for setting up the manuscript.

BIBLIOGRAPHY

1. M.G. Broadhurst and G.T. Davis. in *Medical Applications of Piezoelectric Polymers*. P.M. Galetti, D.E. De Rossi, and A.S. De Reggi, eds., Gordon and Breach, Amsterdam, 1988, pp. 3–14.
2. J.W. Goodby, R. Bline, N.A. Clark, S.T. Lagerwall, M.A. Osipov, S.A. Pikin, T. Sakurai, K. Yoshino, and B. Zeks. *Ferroelectric Liquid Crystals, Principles, Properties and Applications*,

Ferroelectricity and Related Phenomena. Gordon Breach, Philadelphia, 1991.

3. A. Hult, F. Sahlén, Trollsås, S. T. Lagerwall, D. Hermann, L. Komitov, P. Rudquist, and B. Stebler. *Liq. Cryst.* **20**: 23–28 (1996).
4. R. Zentel. *Angew. Chem.* **101**: 1437–1445 (1989).
5. M. Brehmer and R. Zentel. *Mol. Cryst. Liq. Cryst.* **243**: 353–376 (1994).
6. F.J. Davis. *J. Mater. Chem.* **3**: 551–562 (1993).
7. W. Gleim, H. Finkelmann. in *Side Chain Liquid Crystalline Polymers*. C.B. McArdle, ed., Blackie and Son, Glasgow, 1989; H. Finkelmann. in *Liquid Crystallinity in Polymers*, A. Ciferri, ed., VCH, Weinheim, 1991.
8. R.A.M. Hikmet, H.M. Boots, and M. Michielsen. *Liq. Cryst.* **19**: 65–76 (1995).
9. R.A.M. Hikmet and H. Kempermann. *Nature*. **392**: 476–478 (1998).
10. G.R. Mitchell, P.M.S. Roberts, K. Ahn, F.J. Davis, C. Hasson, H. Hirschmann, and J.A. Pople, *Macromol. Symp.* **117**: 21–31 (1997).
11. G.C.L. Wong, W.H. de Jeu, H. Shao, K.S. Liang, and R. Zentel. *Nature*. **389**: 576–579 (1997).
12. R. Zentel, *Field. Liq. Cryst.* **1**: 589 (1986).
13. N.R. Barnes, F.J. Davis, and G.R. Mitchell, *Mol. Cryst. Liq. Cryst.* **13**: 168 (1989).
14. M. Brehmer, R. Zentel, G. Wagenblast, and K. Siemensmeyer, *Macromol. Chem. Phys.* **195**: 1891–1904 (1994).
15. M. Brehmer, R. Zentel, F. Giesselmann, R. Germer, and P. Zugenmaier, *Liq. Cryst.* **21**: 589–596 (1996).
16. E. Gebhard, M. Brehmer, R. Zentel, J. Reibel, G. Decher, H.M. Brodowsky, and F. Kremer. *The Wiley Polymer Networks Group Review Series, Vol. 1, Chemical and Physical Networks*, K. te Nijenhuis and W.J. Mijs, eds., Wiley, Chichester, 1998, pp. 387–397.
17. E. Gebhard and R. Zentel, *Macromol. Chem. Phys.* **201**: 902–910; 911–922 (2000).

18. H. Poths and R. Zentel, *Liq. Cryst.* **16**: 749–767 (1994).
19. L.-C. Chien and L.G. Cada, *Macromolecules.* **27**: 3721 (1994).
20. S. Diele, S. Oelsner, F. Kuschel, B. Hisgen, H. Ringsdorf, and R. Zentel, *Makromol. Chem.* **188**: 1993 (1987).
21. H. Poths, E. Wischerhoff, R. Zentel, A. Schönfeld, G. Henn, and F. Kremer, *Liq. Cryst.* **18**: 811–818 (1995).
22. A. Kocot, R. Wrzalik, J.K. Vij, M. Brehmer, and R. Zentel, *Phys. Rev. B* **50**: 16346–16357 (1994).
23. H.M. Brodowsky, U.-C. Boehnke, F. Kremer, E. Gebhard, and R. Zentel, *Langmuir.* **15**: 274–278 (1999).
24. H.M. Brodowsky, F. Kremer, E. Gebhard, and R. Zentel, *Ferroelectrics.* **243**: 115–123 (2000).
25. R. Stannarius and C. Cramer, *Europhys. Lett.* **42**: 43 (1998).
26. H. Schüring, R. Stannarius, C. Tolksdorf, and R. Zentel, *Macromolecules.* **34**: 3962–3972 (2001).
27. J. Küpfer and H. Finkelmann, *Macromol. Chem. Phys.* **195**: 1353–1367 (1994).
28. K.-H. Hanus, H. Kapitza, F. Kremer, G. Reckert, S. Vallerien, and R. Zentel, *Polym. Prepr.* **30**: 493 (1989).
29. R. Zentel, *Liq. Cryst.* **3**: 531 (1988).
30. R. Zentel, G. Reckert, S. Bualek, and H. Kapitza, *Makromol. Chem.* **190**: 2869 (1989).
31. J. Küpfer and H. Finkelmann, *Makromol. Chem. Rapid Commun.* **12**: 717 (1991).
32. L.R.G. Treloar, *The Physics of Rubber Elasticity*, Clarendon, Oxford, 1975.
33. S. Shilov, E. Gebhard, H. Skupin, R. Zentel, and F. Kremer, *Macromolecules.* **32**: 1570–1575 (1999).
34. M. Brehmer and R. Zentel, *Macromol. Chem. Rapid Commun.* **16**: 659–662 (1995).
35. H.M. Brodowsky, U.-C. Boehnke, F. Kremer, E. Gebhard, and R. Zentel, *Langmuir* **13**: 5378–5382 (1997).
36. I. Benné, K. Semmler, and H. Finkelmann, *Macromolecules.* **28**: 1854–1858 (1995).
37. R.A.M. Hikmet and J. Lub, *Prog. Polym. Sci.* **21**: 1165–1209 (1996).
38. M. Mauzac, H.T. Nguyen, F.G. Tournilhac, and S.V. Yablonsky, *Chem. Phys. Lett.* **240**: 461 (1995).
39. S.U. Vallerien, F. Kremer, E.W. Fischer, H. Kapitza, R. Zentel, and H. Poths, *Makromol. Chem., Rapid Commun.* **11**: 593 (1990).
40. W. Lehmann, H. Skupin, C. Tolksdorf, E. Gebhard R. Zentel, P. Krüger, M. Lösche, and F. Kremer, *Nature* **410**: 447–450 (2001).
41. E. Nishikawa, H. Finkelmann, and H.-R. Brand, *Macromol. Rapid Commun.* **18**: 65 (1997).
42. J. Kundler and H. Finkelmann, *Macromol. Chem. Phys.* **199**: 677–686 (1998).
43. E. Nishikawa and H. Finkelmann, *Macromol. Chem. Phys.* **200**: 312–322 (1999).
44. F. Kremer, W. Lehmann, H. Skupin, L. Hartmann, P. Stein, and H. Finkelmann, *Polym. Adv. Technol.* **9**: 672–676 (1998).
45. W. Lehmann, P. Gattinger, M. Keck, F. Kremer, P. Stein, T. Eckert, and H. Finkelmann, *Ferroelectrics* **208**(1–4): 373–383 (1998).
46. W. Lehmann, L. Hartmann, F. Kremer, P. Stein, H. Finkelmann, H. Kruth, and S. Diele, *J. Appl. Phys.* **86**: 1647–1652 (1999).
47. J. Zhao, Q.M. Zhang, N. Kim, and T. Shrout. *Jpn. J. Appl. Phys.* **34**: 5658 (1995).
48. Q.M. Zhang, V. Bharti, and X. Zhao, *Science* **280** (5372): 2101–2104 (1998).
49. T. Jaworek, D. Neher, G. Wegner, R.H. Wieringa, and A.J. Schouten. *Science* **279**: 57 (1998).

POLYMERS, PIEZOELECTRIC

J. S. HARRISON

NASA Langley Research Center
Hampton, VA

Z. OUNAIES

ICASE/NASA Langley Research Center
Hampton, VA

INTRODUCTION

Piezoelectric polymers have been known for more than forty years, but in recent years they have gained repute as a valuable class of “smart materials.” There is no standard definition for smart materials, and terms such as intelligent materials, smart materials, adaptive materials, active devices, and smart systems are often used interchangeably. The term “smart material” generally designates a material that changes one or more of its properties in response to an external stimulus.

The most popular smart materials are piezoelectric materials, magnetostrictive materials, shape-memory alloys, electrorheological fluids, electrostrictive materials, and optical fibers. Magnetostrictives, electrostrictives, shape-memory alloys, and electrorheological fluids are used as actuators; optical fibers are used primarily as sensors.

Among these active materials, piezoelectric materials are most widely used because of their wide bandwidth, fast electromechanical response, relatively low power requirements, and high generative forces. A classical definition of piezoelectricity, a Greek term for “pressure electricity,” is the generation of electrical polarization in a material in response to mechanical stress. This phenomenon is known as the direct effect. Piezoelectric materials also display the converse effect: mechanical deformation upon application of an electrical charge or signal. Piezoelectricity is a property of many noncentrosymmetric ceramics, polymers, and other biological systems. Pyroelectricity is a subset of piezoelectricity, whereby the polarization is a function of temperature. Some pyroelectric materials are ferroelectric, although not all ferroelectrics are pyroelectric. Ferroelectricity is a property of certain dielectrics that exhibit spontaneous electric polarization (separation of the center of positive and negative electric charge that makes one side of the crystal positive and the opposite side negative) that can be reversed in direction by applying an appropriate electric field. Ferroelectricity is named by analogy with ferromagnetism, which occurs in materials such as iron. Traditionally, ferroelectricity is defined for crystalline materials, or at least in the crystalline region of semicrystalline materials. In the last couple of years, however, a number of researchers have explored the possibility of ferroelectricity in amorphous polymers, that is, ferroelectricity without a crystal lattice structure (1).

Table 1. Comparison of Properties of Standard Piezoelectric Polymer and Ceramic Materials

	d_{31}^a (pm/V)	g_{31}^a (mV-m/N)	k_{31}	Salient Features
Polyvinylidene fluoride (PVDF)	28	240	0.12	Flexible, lightweight, low acoustic and mechanical impedance
Lead zirconium titanate (PZT)	175	11	0.34	Brittle, heavy, toxic

^aValues shown are absolute values of constants.

Characteristics of Piezoelectric Polymers

The properties of polymers are very different from those of inorganics (Table 1), and they are uniquely qualified to fill niche areas where single crystals and ceramics cannot perform as effectively. As noted in Table 1, the piezoelectric strain constant (d_{31}) for the polymer is lower than that of the ceramic. However, piezoelectric polymers have much higher piezoelectric stress constants (g_{31}) which indicates that they are much better sensors than ceramics. Piezoelectric polymeric sensors and actuators offer the advantage of processing flexibility because they are lightweight, tough, readily manufactured in large areas, and can be cut and formed into complex shapes. Polymers also exhibit high strength and high impact resistance (2). Other notable features of polymers are low dielectric constant, low elastic stiffness, and low density, which result in high voltage sensitivity (excellent sensor characteristic) and low acoustic and mechanical impedance (crucial for medical and underwater applications). Polymers also typically possess high dielectric breakdown and high operating field strength, which means that they can withstand much higher driving fields than ceramics. Polymers offer the ability to pattern electrodes on the film surface and pole only selected regions. Based on these features, piezoelectric polymers possess their own established area for technical applications and useful device configurations.

Structural Requirements for Piezoelectric Polymers

The piezoelectric mechanisms for semicrystalline and amorphous polymers differ. Although the differences are distinct, particularly with respect to polarization stability, in the simplest terms, four critical elements exist for all piezoelectric polymers, regardless of morphology. These essential elements are: (1) the presence of permanent molecular dipoles, (2) the ability to orient or align the molecular dipoles, (3) the ability to sustain this dipole alignment once it is achieved, and (4) the ability of the material to undergo large strains when mechanically stressed (3).

SEMICRYSTALLINE POLYMERS

Mechanism of Piezoelectricity in Semicrystalline Polymers

Semicrystalline polymers must have a polar crystalline phase to render them piezoelectric. The morphology of such

polymers consists of crystallites dispersed within amorphous regions, as shown in Fig. 1a. The amorphous region has a glass transition temperature that dictates the mechanical properties of the polymer, and the melting temperature of the crystallites dictates the upper limit of the use

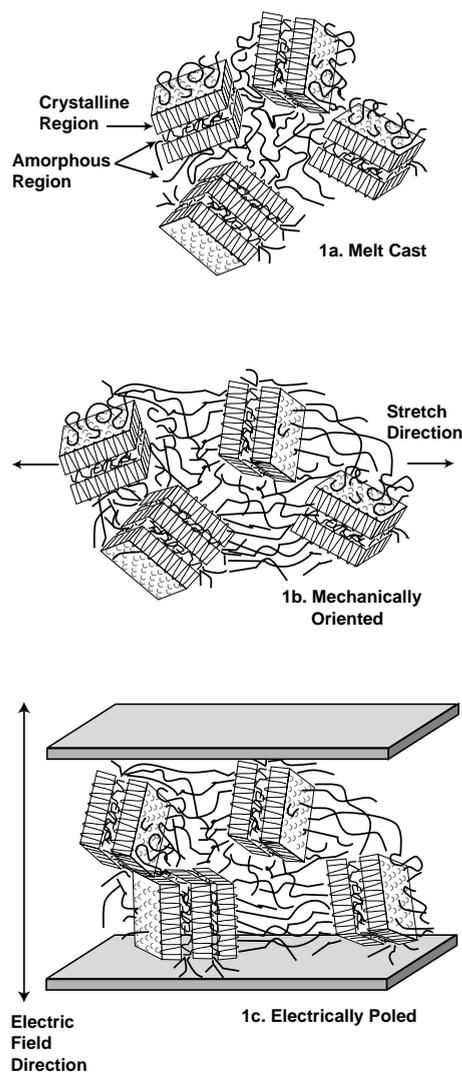


Figure 1. Schematic illustration of random stacks of amorphous and crystal lamellae in the PVDF polymer: (a) the morphology after the film is melt cast; (b) after orientation of the film by mechanically stretching several times its original length; and (c) after depositing metal electrodes and poling through the film thickness.

temperature. The degree of crystallinity in such polymers depends on the method of preparation and thermal history. Most semicrystalline polymers have several polymorphic phases, some of which may be polar. Mechanical orientation, thermal annealing, and high-voltage treatment are all it has been shown, effective in inducing crystalline phase transformations. Stretching the polymer aligns the amorphous strands in the film plane, as shown in Fig. 1b, and facilitates uniform rotation of the crystallites by an electric field. Depending on whether stretching is uniaxial or biaxial, the electrical and mechanical properties (and therefore the transduction response) are either highly anisotropic or isotropic in the plane of the polymer sheet. Electrical poling is accomplished by applying an electric field across the thickness of the polymer, as depicted in Fig. 1c. An electric field of the order of 50 MV/m is typically sufficient to effect crystalline orientation. Polymers can be poled by using a direct contact method or corona discharge. The latter is advantageous because contacting electrodes is not required and samples of large area can be poled continuously. This method is used to manufacture commercial poly(vinylidene fluoride) (PVDF) film. Some researchers have also successfully poled large areas of polymer films by sandwiching them between polished metal plates under a vacuum. This method eliminates electrical arcing of samples and the need for depositing metal electrodes on the film surface. The amorphous phase of semicrystalline polymers supports the crystal orientation, and polarization is stable up to the Curie temperature. This polarization can remain constant for many years if it is not degraded by moisture uptake or elevated temperatures.

Piezoelectric Constitutive Relationships

The constitutive relationships that describe piezoelectric behavior in materials can be derived from thermodynamic

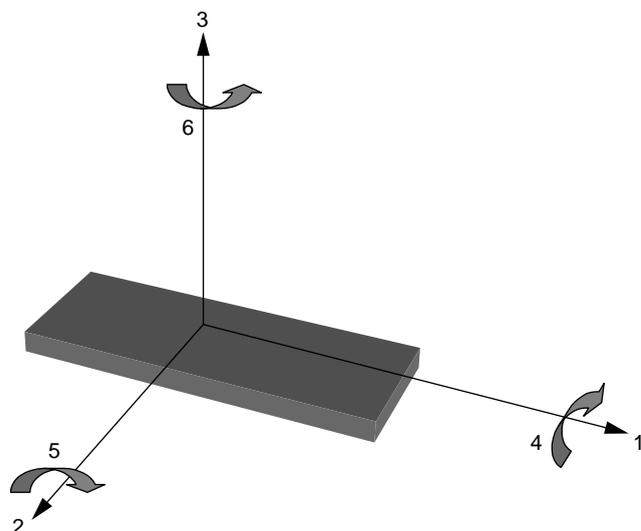


Figure 2. Tensor directions for defining the constitutive relationships.

principles (4). A tensor notation is adopted to identify the coupling between the various entities through mechanical and electrical coefficients. The common practice is to label directions as depicted in Fig. 2. The stretch direction is denoted as “1.” The “2” axis is orthogonal to the stretch direction in the plane of the film. The polarization axis (perpendicular to the surface of the film) is denoted “3.” The shear planes, indicated by the subscripts “4,” “5,” and “6,” are perpendicular to the directions “1,” “2,” and “3,” respectively. By reducing the tensor elements and using standard notations (5), the resulting equations can be displayed in matrix form as follows:

$$\begin{pmatrix} S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \end{pmatrix} = \begin{pmatrix} d_{11} & d_{12} & d_{13} \\ d_{21} & d_{22} & d_{23} \\ d_{31} & d_{32} & d_{33} \\ d_{41} & d_{42} & d_{43} \\ d_{51} & d_{52} & d_{53} \\ d_{61} & d_{62} & d_{63} \end{pmatrix} \begin{pmatrix} E_1 \\ E_2 \\ E_3 \end{pmatrix} + \begin{pmatrix} s_{11}^E & s_{12}^E & s_{13}^E & s_{14}^E & s_{15}^E & s_{16}^E \\ s_{21}^E & s_{22}^E & s_{23}^E & s_{24}^E & s_{25}^E & s_{26}^E \\ s_{31}^E & s_{32}^E & s_{33}^E & s_{34}^E & s_{35}^E & s_{36}^E \\ s_{41}^E & s_{42}^E & s_{43}^E & s_{44}^E & s_{45}^E & s_{46}^E \\ s_{51}^E & s_{52}^E & s_{53}^E & s_{54}^E & s_{55}^E & s_{56}^E \\ s_{61}^E & s_{62}^E & s_{63}^E & s_{64}^E & s_{65}^E & s_{66}^E \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \end{pmatrix}, \quad (1)$$

$$\begin{pmatrix} D_1 \\ D_2 \\ D_3 \end{pmatrix} = \begin{pmatrix} \varepsilon_{11}^T & \varepsilon_{12}^T & \varepsilon_{13}^T \\ \varepsilon_{21}^T & \varepsilon_{22}^T & \varepsilon_{23}^T \\ \varepsilon_{31}^T & \varepsilon_{32}^T & \varepsilon_{33}^T \end{pmatrix} \begin{pmatrix} E_1 \\ E_2 \\ E_3 \end{pmatrix} + \begin{pmatrix} d_{11} & d_{12} & d_{13} & d_{14} & d_{15} & d_{16} \\ d_{21} & d_{22} & d_{23} & d_{24} & d_{25} & d_{26} \\ d_{31} & d_{32} & d_{33} & d_{34} & d_{35} & d_{36} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \end{pmatrix}. \quad (2)$$

Piezoelectricity is a cross coupling among the elastic variables, stress X , and strain S , and the dielectric variables, electric charge density D and electric field E . Note that D is named in analogy to the B field in ferromagnetism, although some authors also refer to it as dielectric or electric displacement. There does not seem to be a standard nomenclature; however, it is the opinion of the authors of this article that electric charge density is a better description of this property. The combinations of these variables define the piezoelectric strain constant d , the material compliance s , and the permittivity ε . Other piezoelectric properties are the piezoelectric voltage constant g , stress constant e , and strain constant h given by the equations in Table 2. For a given constant, the first definition in the table refers to the direct effect, and the second refers to the converse effect. The piezoelectric constants are interrelated through the electrical and mechanical properties

Table 2. Definitions of Piezoelectric Constants

Equations	Units
$d = (dD/dX)_E = (dS/dE)_X$	(C/N or m/V)
$e = (dD/dS)_E = -(dX/dE)_S$	(C/m or N/Vm)
$g = (dE/dX)_D = (dS/dD)_X$	(Vm/N or m ² /C)
$h = (dE/dS)_D = -(dX/dD)_S$	(V/m or N/C)

of the material. Electric field strength and displacement charge density are related through the dielectric constant, $\epsilon\epsilon_0$ (where ϵ_0 is the permittivity of free space), and stress and strain are related through the compliance according to

$$d_{ij} = \epsilon_0\epsilon_i g_{ij}, \quad (3)$$

$$e_{ij} = s_{ij}d_{ij}. \quad (4)$$

The polarization P is a measure of the degree of piezoelectricity in a given material. In a piezoelectric material, a change in polarization ΔP results from an applied stress X or strain S under conditions of constant temperature and zero electric field. A linear relationship exists between ΔP and the piezoelectric constants. Due to material anisotropy, P is a vector that has three orthogonal components in the 1, 2, and 3 directions. Alternatively, the piezoelectric constants can be defined as

$$\Delta P_i = d_{ij}X_j, \quad (5)$$

$$\Delta P_i = g_{ij}S_j. \quad (6)$$

The electrical response of a piezoelectric material is a function of the electrode configuration relative to the direction of the applied mechanical stress. For a coefficient d_{ij} , the first subscript is the direction of the electric field or charge displacement, and the second subscript gives the direction of the mechanical deformation or stress. The C_{2v} crystallographic symmetry typical of synthetic oriented, poled polymer film leads to cancellation of all but five of the d_{ij} components (d_{31} , d_{32} , d_{33} , d_{15} , and d_{24}). If the film is poled and biaxially oriented or unoriented, $d_{31} = d_{32}$, and $d_{15} = d_{24}$. Most natural biopolymers possess D_∞ symmetry which yields a matrix that possesses only the shear piezoelectricity components d_{13} and d_{25} . Because the d_{33} constant is difficult to measure without constraining the lateral dimension of the sample, it is typically determined from Eq. (7) which relates the constants to the hydrostatic piezoelectric constant, d_{3h} .

$$d_{3h} = d_{31} + d_{32} + d_{33}. \quad (7)$$

The electromechanical coupling coefficient k_{ij} represents the conversion of electrical energy into mechanical energy and vice versa. The electromechanical coupling can be considered a measure of transduction efficiency and is always less than unity as shown here:

$$k^2 = \frac{\text{electrical energy converted to mechanical energy}}{\text{input electrical energy}}, \quad (8a)$$

$$k^2 = \frac{\text{mechanical energy converted to electrical energy}}{\text{input mechanical energy}}. \quad (8b)$$

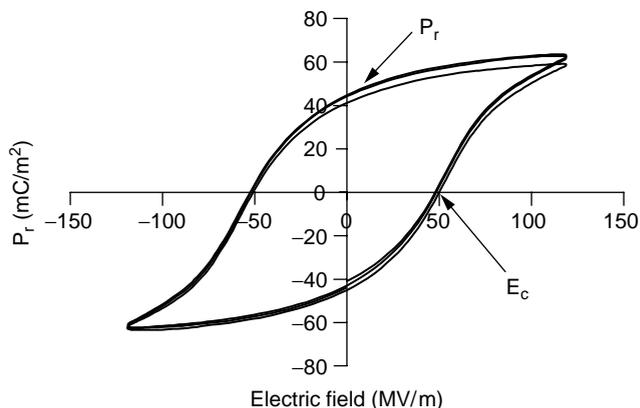


Figure 3. Typical ferroelectric hysteresis behavior for PVDF.

Some k coefficients can be obtained from a measured d constant as follows:

$$k_{31} = \frac{d_{31}}{\sqrt{s_{11}^E \epsilon_3^T}}. \quad (9)$$

Ferroelectricity in Semicrystalline Polymers

At high electric fields, the polarization in semicrystalline polymers such as PVDF is nonlinear with the applied electric field. This nonlinearity in polarization is defined as hysteresis. The existence of spontaneous polarization together with polarization reversal (as illustrated by a hysteresis loop) is generally accepted as proof of ferroelectricity. Figure 3 is an example of the typical hysteresis behavior of PVDF. Two other key properties typically reported for ferroelectric materials are the coercive field and the remanent polarization. The coercive field E_c , which marks the point where the hysteresis intersects the horizontal axis, is about 50 MV/m at room temperature for many ferroelectric polymers. The remanent polarization P_r corresponds to the point where the loop intersects the vertical axis. The values of E_c and P_r depend on temperature and frequency. The Curie temperature T_c , is generally lower than but close to the melting temperature of the polymer. Below T_c , the polymer is ferroelectric and above T_c , the polymer loses its noncentrosymmetric nature.

Although ferroelectric phenomenon has been well documented in ceramic crystals, the question of whether polymer crystallites could exhibit dipole switching was debated for about a decade after the discovery of piezoelectricity in PVDF. Inhomogeneous polarization through the film thickness that yielded higher polarization on the positive electrode side of the polymer led to speculations that PVDF was simply a trapped charge electret. These speculations were dispelled when X-ray studies (6) demonstrated that polarization anisotropy vanishes due to high poling field strengths and that true ferroelectric dipole reorientation occurs in PVDF. Luongo used infrared to attribute the polarization reversal in PVDF to 180° dipole rotation (7). Scheinbeim documented the same via X-ray

pole analysis and infrared techniques for odd-numbered nylons (8).

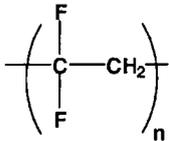
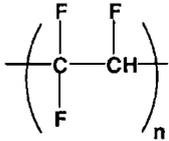
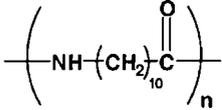
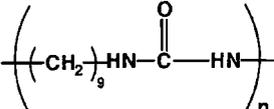
State of the Art

Pioneering work in the area of piezoelectric polymers (9) led to the development of strong piezoelectric activity in polyvinylidene fluoride (PVDF) and its copolymers with trifluoroethylene (TrFE) and tetrafluoroethylene (TFE). These semicrystalline fluoropolymers represent the state of the art in piezoelectric polymers and are currently the only commercial piezoelectric polymers. Odd-numbered nylons, the next most widely investigated semicrystalline piezoelectric polymers, have excellent piezoelectric properties at elevated temperatures but have not yet been embraced in practical application. Other semicrystalline polymers including polyureas, liquid crystalline polymers, biopolymers and an array of blends have been studied for their piezoelectric potential and are summarized in the following section. The chemical repeat unit and piezoelectric constants of several semicrystalline polymers are listed in Table 3.

Polyvinylidene Fluoride (PVDF). Interest in the electrical properties of PVDF began in 1969 when it was shown (9) that poled thin films exhibit a very large piezoelectric coefficient, 6–7 pC/N⁻¹, a value about ten times larger than had been observed in any other polymer. As seen in Table 3, PVDF is inherently polar. The spatially symmetrical disposition of the hydrogen and fluorine atoms along the polymer chain gives rise to unique polar effects that influence the electromechanical response, solubility, dielectric properties, and crystal morphology and yield an unusually high dielectric constant. The dielectric constant of PVDF is about

12, which is four times greater than that of most polymers, and makes PVDF attractive for integration into devices because the signal-to-noise ratio is smaller for higher dielectric materials. The amorphous phase in PVDF has a glass transition that is well below room temperature (–35°C); hence, the material is quite flexible and readily strained at room temperature. PVDF is typically 50–60% crystalline, depending on thermal and processing history, and has at least four crystal phases (α , β , γ , and δ); at least three are polar. The most stable, nonpolar α phase results upon casting PVDF from a melt and can be transformed into the polar β phase by mechanical stretching at elevated temperatures or into the polar δ phase by rotating the molecular chain axis in a high electric field (~130 MV/m) (10). The β phase is most important for piezoelectricity and has a dipole moment perpendicular to the chain axis of 2.1 D that corresponds to a dipole concentration of 7×10^{-30} Cm. After poling PVDF, the room temperature polarization stability is excellent; however, polarization and piezoelectricity degrade as temperature increases and are erased at its Curie temperature. Previously, it was believed that polarization stability was defined only by the melting temperature of the PVDF crystals. Recently, however, some researchers suggest that the polarization stability of PVDF and its copolymers is associated with coulombic interactions between injected, trapped charges and oriented dipoles in the crystals (11). They hypothesize that the thermal decay of the polarization is caused by the thermally activated removal of the trapped charges from the traps at the surface of the crystals. The role of trapped charges in stabilizing orientation in both semicrystalline and amorphous polymers is still a subject that needs further study. The electromechanical properties of PVDF have been widely investigated. For more details, the

Table 3. Comparison of Piezoelectric Properties of Some Semicrystalline Polymeric Materials

Polymer	Structure	T_g (°C)	T_m (°C)	Max Use Temp (°C)	d_{31} (pC/N)	Ref.
PVDF		–35	175	80	20–28	2
PTrFE		32	150	90–100	12	2
Nylon-11		68	195	185	3 at 25°C 14 at 107°C	22
Polyurea-9		50	180	—	—	28

reader is referred to the wealth of literature that exists on the subjects of piezoelectric, pyroelectric, and ferroelectric properties (2,6,12,13), and the morphology (14–16) of this polymer.

Poly(Vinylidene Fluoride–Trifluoroethylene and Tetrafluoroethylene) Copolymers. Copolymers of polyvinylidene fluoride with trifluoroethylene (TrFE) and tetrafluoroethylene (TFE) also exhibit strong piezoelectric, pyroelectric, and ferroelectric effects. These polymers are discussed together here because they behave similarly when copolymerized with PVDF. An attractive morphological feature of the comonomers is that they force the polymer into an all-trans conformation that has a polar crystalline phase, which eliminates the need for mechanical stretching to yield a polar phase. P(VDF–TrFE) crystallizes to a much greater extent than PVDF (up to 90% crystalline) and yields a higher remanent polarization, a lower coercive field, and much sharper hysteretic loops. TrFE also extends the use temperature by about 20°C to close to 100°C. Conversely, copolymers with TFE exhibit a lower degree of crystallinity and a suppressed melting temperature, compared to the PVDF homopolymer. Although the piezoelectric constants for the copolymers are not as large as those of the homopolymer, the advantages of P(VDF–TrFE) in processibility, enhanced crystallinity, and higher use temperature make it favorable for applications.

Researchers have recently reported that highly ordered, lamellar crystals of P(VDF–TrFE) can be made by annealing the material at temperatures between the Curie temperature and the melting point. They refer to this material as a “single crystalline film.” A relatively large single crystal of P(VDF–TrFE) 75/25 mol% copolymer was grown that exhibits a room temperature $d_{33} = -38$ pm/V and a coupling factor $k_{33} = 0.33$ (17).

The result of introducing defects into the crystalline structure of P(VDF–TrFE) copolymer on electroactive actuation has been studied using high electron irradiation (18). Extensive structural investigations indicate that electron irradiation disrupts the coherence of polarization domains (all-trans chains) and forms localized polar regions (nanometer-sized, all-trans chains interrupted by trans and gauche bonds). After irradiation, the material exhibits behavior analogous to that of relaxor ferroelectric systems in inorganic materials. The resulting material is no longer piezoelectric but rather exhibits a large electric field-induced strain (5% strain) due to electrostriction. The basis for such large electrostriction is the large change in the lattice strain as the polymer traverses the ferroelectric to paraelectric phase transition and the expansion and contraction of the polar regions. Piezoelectricity can be measured in these and other electrostrictives when a dc bias field is applied. Irradiation is typically accomplished in a nitrogen atmosphere at elevated temperatures using irradiation dosages up to 120 Mrad.

Polyamides. A low level of piezoelectricity was first reported in polyamides (also known as nylons) in 1970 (19).

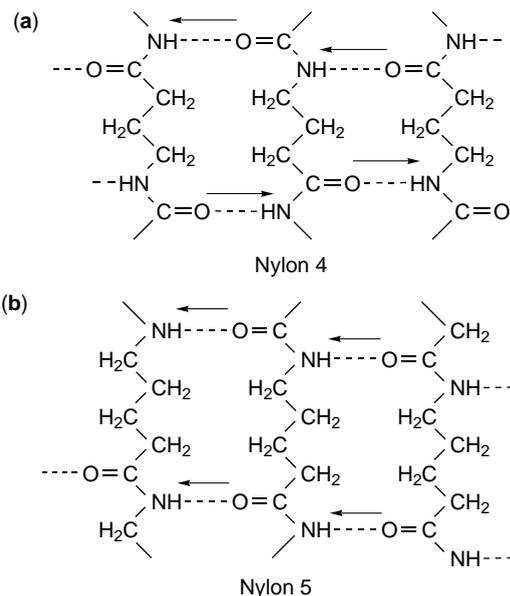


Figure 4. Schematic depiction of hydrogen-bonded sheets showing dipole directions in the crystal lattices of (a) even (nylon 4) and (b) odd polyamides (nylon 5).

A systematic study of odd-numbered nylons, however, initiated in 1980 (20), served as the impetus for more than 20 years of subsequent investigations of piezoelectric and ferroelectric activity in these polymers. The monomer unit of odd nylons consists of even numbers of methylene groups and one amide group whose dipole moment is 3.7 D. Polyamides crystallize in all-trans conformations and are packed to maximize hydrogen bonding between adjacent amine and carbonyl groups, as seen in Fig. 4 for an even-numbered and an odd-numbered polyamide. The amide dipoles align synergistically in the odd-numbered monomer, resulting in a net dipole moment. The amide dipole cancels in an even-numbered nylon, although remanent polarizations have been measured for some even-numbered nylons, as discussed later in this article. The unit dipole density depends on the number of methylene groups present, and polarization increases from 58 mC/m² for nylon-11 to 125 mC/m² for nylon-5 as the number of methylene groups decreases (8).

Polyamides are known hydrophilics. Because water absorption is associated with hydrogen bonding to the polar amide groups, the hydrophilicity increases as the density of amide groups increases. Water absorption in nylon-11 and nylon-7 has been shown to be as high as 4.5% (by weight) and more than 12% for nylon-5 (21), whereas it is less than 0.02% for PVDF and its copolymers. Studies have shown that water absorption can have a dramatic effect on the dielectric and piezoelectric properties of nylons; however, water does not affect the crystallinity or orientation in thermally annealed films (21). Thus, films can be dried to restore their original properties.

At room temperature, odd-numbered nylons have lower piezoelectric constants than PVDF; however, when

examined above their glass transition temperatures, they exhibit comparable ferroelectric and piezoelectric properties and much higher thermal stability. The piezoelectric d and e constants increase rapidly as temperature increases. Maximum stable d_{31} values of 17 pC/N and 14 pC/N are reported for nylon-7 and nylon-11, respectively. Corresponding values of the electromechanical coupling constant k_{31} are 0.054 and 0.049. Studies have also shown that annealing nylon films enhances their polarization stability because it promotes denser packing of the hydrogen-bonded sheet structure in the crystalline regions and hinders dipole switching due to lowered free volume for rotation (22).

Though widely studied, piezoelectric polyamides have not been widely used in applications due in part to their low room temperature piezoelectric response and the problem of moisture uptake.

Liquid Crystalline Polymers. Liquid crystals consist of highly ordered rodlike or disklike molecules. At their melting points they partially lose crystalline order and generate a fluid but ordered state. They can form layered structures called smectic phases or nematic phases that have an approximately parallel orientation of the molecular long axis. It was first predicted in 1975 that spontaneous polarization could be achieved in liquid crystals based on symmetry arguments (23). Subsequently, it has been shown that liquid crystalline molecules whose chiral carbon atoms link a mesogenic group and end alkyl chains may exhibit ferroelectric behavior in the smectic C phase (SmC^*) (24). In this phase, the molecular axis tilts from the normal to the layer plane, and the molecular dipoles align in the same direction, yielding a net polarization. If such liquid crystalline molecules are introduced into the backbone or as a side group on a polymer, a ferroelectric liquid crystalline polymer can be obtained. There are three requirements for spontaneous polarization in a liquid crystal: a center of chirality, a dipole moment positioned at the chiral center that acts transverse to the molecular long axis, and a tilted smectic phase (25).

Polyureas. Polyureas are thermosets, long used as insulators in a number of applications. Until a few years ago, ureas were available mostly as insoluble powders or highly cross-linked resins. In 1987, a vapor deposition polymerization method was successfully developed that was later applied to synthesizing polyureas (26a,27). Typically, a vapor deposition technique is used by evaporating $\text{OCN-R}_1\text{-NH}_2$ and $\text{H}_2\text{N-R}_2\text{-NH}_2$ monomers simultaneously on a substrate (where R_1 and R_2 are various aliphatic or aromatic groups). This prevents cross-linking and allows processing thicknesses in the hundreds of nanometers to tens of micrometers.

An exploration (27) of the dielectric and pyroelectric properties of polyureas films led to the discovery of their piezoelectricity. From the early 1990s to the present, various aromatic and aliphatic polyureas were synthesized, and it was shown that they are piezoelectric (28,29). Aromatic polyureas were the first polyurea structures identified as piezoelectric. They exhibit a piezoelectric

e constant of 15 mC/m^2 and have high temperature stability which remains independent of temperature up to 200°C . Their pyroelectric coefficient is high due to their low dielectric loss compared to other polymers. The d constant is about 5 pC/N at room temperature and increases as temperature increases (28).

Owing to their structures, aliphatic polyureas possess higher flexibility in their molecular chains. Similarly to polyamides, hydrogen bonds play a large role in stabilizing the orientation polarization that is imparted. Polyureas that have an odd number of methyl groups exhibit overall nonzero polarization. Polyurea-9 was synthesized and processed first, and an e constant of 5 mC/m^2 was reported (28). Then, polyureas that have a smaller number of carbons were attempted because it was surmised that they should lead to a higher density of urea bond dipoles. Toward that end, polyurea-5 was synthesized, and it was found that the e and d constants are twice those of polyurea-9. Aliphatic polyureas exhibit ferroelectric hysteresis and in addition, are piezoelectric when they have odd numbers of methyl groups. Their thermal stability and piezoelectric coefficients depend highly on the poling temperature (typically $70\text{--}150^\circ\text{C}$) but are lower than those of aromatic polyureas.

Biopolymers. Piezoelectricity of biopolymers was first reported in keratin in 1941 (30). When a bundle of hair was immersed in liquid air, an electric voltage of a few volts was generated between the tip and the root. When pressure was applied on the cross section of the bundle, an electric voltage was generated. Subsequently piezoelectricity has been observed in a wide range of other biopolymers including collagen (31,32), polypeptides like poly- γ -methylglutamate and poly- γ -benzyl-L-glutamate (33,34), oriented films of DNA (35), poly-lactic acid (36), and chitin (37). Most natural biopolymers possess D_∞ symmetry, so they exhibit shear piezoelectricity. A shear stress in the plane of polarization produces an electric displacement perpendicular to the plane of the applied stress and results in a $-d_{14} = d_{25}$ piezoelectric constant. The piezoelectric constants of biopolymers are small relative to synthetic polymers; they range in value from 0.01 pC/N for DNA to 2.5 pC/N for collagen. The electromechanical effect in such polymers is attributed to the internal rotation of polar atomic groups linked to asymmetrical carbon atoms. Keratin and some polypeptide molecules assume an α -helical or a β -crystalline structure in which the CONH dipoles align synergistically in the axial direction.

Currently, the physiological significance of piezoelectricity in many biopolymers is not well understood, but it is believed that such electromechanical phenomena may have a distinct role in biochemical processes. For example, it is known that electric polarization in bone influences bone growth (38). In one study, a piezoelectric PVDF film was wrapped around the femur of a monkey. Within weeks, a remarkable formation of new bone was observed. The motion of the animal caused deformation of the film that produced a neutralizing ionic current in the surrounding tissue. This minute fluctuating current appears to stimulate the metabolic activity of bone cells and leads to the proliferation of bone.

AMORPHOUS POLYMERS

The purpose of this section is to explain the mechanism and key components required for developing piezoelectricity in amorphous polymers and to present a summary of the polarization and electromechanical properties of the amorphous polymers currently under investigation.

Mechanism of Piezoelectricity

Dielectric Theory. The piezoelectricity in amorphous polymers differs from that in semicrystalline polymers and inorganic crystals in that the polarization is not in a state of thermal equilibrium, but rather a quasi-stable state due to the freezing-in of molecular dipoles. The result is a piezoelectric-like effect. A theoretical model for polymers that have frozen-in dipolar orientation was presented to explain piezoelectricity and pyroelectricity in amorphous polymers such as polyvinyl chloride (39).

One of the most important properties of an amorphous piezoelectric polymer is its glass transition temperature (the temperature below which the material exhibits glass-like characteristics, and above which it has rubber-like properties) because it dictates use temperature and defines the poling process conditions. Orientation polarization of molecular dipoles is responsible for piezoelectricity in amorphous polymers. It is induced, as shown in Fig. 5, by applying an electric field E_p at an elevated temperature ($T_p \geq T_g$) where the molecular chains are sufficiently mobile and allow dipole alignment with the electric field. Partial retention of this orientation is achieved by lowering the temperature below T_g in the presence of E_p , resulting in a piezoelectric-like effect. The remanent polarization P_r is

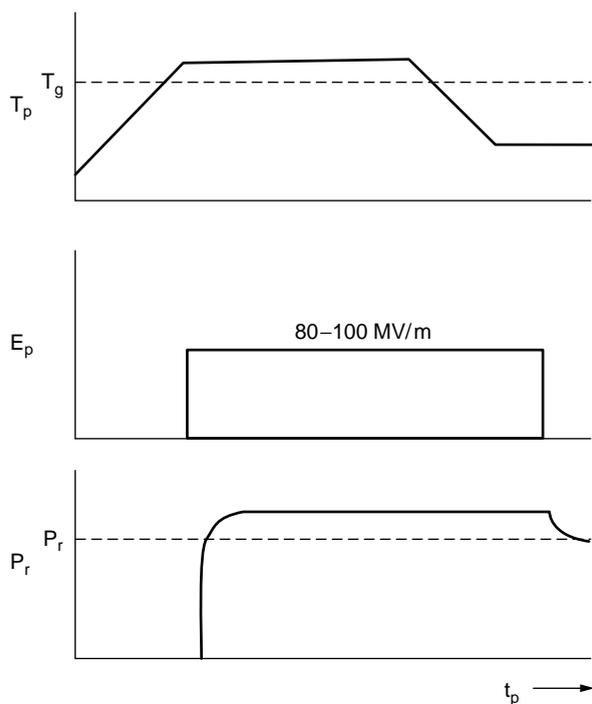


Figure 5. Poling profile of an amorphous polymer.

directly proportional to E_p and the piezoelectric response. The procedure used to prepare a piezoelectric amorphous polymer clearly results in both oriented dipoles and space or real charge injection. The real charges are usually concentrated near the surface of the polymer, and they are introduced due to the presence of the electrodes. Interestingly, some researchers (40,41) have shown that the presence of space charges does not significantly affect piezoelectric behavior. The reason is twofold. The magnitude of space charges is usually not significant with respect to polarization charges. Secondly, space charges are essentially symmetrical with respect to the thickness of the polymer; therefore, when the material is strained uniformly, the contribution to the piezoelectric effect is negligible.

In what follows, the origins of the dielectric contribution to the piezoelectric response of amorphous polymers are addressed. The potential energy U of a dipole μ at an angle θ with the applied electric field is $U = \mu \mathbf{E} \cos \theta$. Using statistical mechanics and assuming Boltzmann's distribution of dipole energies, the mean projection of the dipole moment $\langle \mu_E \rangle$ in the direction of the applied electric field is obtained:

$$\frac{\langle \mu_E \rangle}{\mu} = \coth \frac{\mu E_p}{kT} - \frac{kT}{\mu E}. \quad (8)$$

This is the Langevin equation that describes the degree of polarization in a sample when an electric field E is applied at temperature T . Experimentally, a poling temperature in the vicinity of T_g is used to maximize dipole motion. The maximum electric field that may be applied, typically 100 MV/m, is determined by the dielectric breakdown strength of the polymer. For amorphous polymers, $\mu E/kT$ is much less than one; this places these systems well within the linear region of the Langevin function. The remanent polarization P_r is simply the polarization during poling minus the electronic and atomic polarizations that relax at room temperature, once the field E_p is removed. The following linear equation for remanent polarization results when the Clausius–Mossotti equation is used to relate the dielectric constant to the dipole moment (42):

$$P_r = \Delta \epsilon \epsilon_0 E_p. \quad (9)$$

It can be concluded that remanent polarization and hence the piezoelectric response of a material are determined by $\Delta \epsilon$; this makes it a practical criterion to use when designing piezoelectric amorphous polymers. The dielectric relaxation strength $\Delta \epsilon$ may be the result of either free or cooperative dipole motion. Dielectric theory yields a mathematical approach for examining the dielectric relaxation $\Delta \epsilon$ due to free rotation of the dipoles. The equation incorporates Debye's work based on statistical mechanics, the Clausius–Mossotti equation, and the Onsager local field and neglects short range interactions (43):

$$\Delta \epsilon_{\text{calculated}} = \frac{N \mu^2}{3kT \epsilon_0} \left(\frac{n^2 + 2}{3} \right)^2 \left[\frac{3\epsilon(0)}{2\epsilon(0) + n^2} \right]^2, \quad (10)$$

where N is the number of dipoles per unit volume, k is the Boltzmann constant, $\epsilon(0)$ is the static dielectric constant,

and n is the refractive index. One way to measure P_r in amorphous polymers is the thermally stimulated current (TSC) method (refer to section on characterization). P_r can be calculated from the liberated charge during TSC, and by reconciling that with the Onsager relationship, the dipole density can be calculated:

$$P_r = \frac{N\mu^2 E_p}{3kT_p} \left(\frac{\varepsilon_\infty + 2}{3} \right)^2 \left[\frac{3\varepsilon(0)}{2\varepsilon(0) + \varepsilon_\infty} \right]. \quad (11)$$

The piezoelectric constants are related to the polarization. From basic thermodynamics,

$$d_{3i} = \left(\frac{\partial P}{\partial \sigma_i} \right)_{\gamma, T}. \quad (12)$$

A molecular theory of the direct piezoelectric effect in poled amorphous piezoelectric polymers has been developed. An expression for the hydrostatic coefficient appears in the original paper (41). Later, this theory was extended, and an equation for d_{31} was obtained (44,45). By differentiating Eq. (11) and modifying it to account for dimensional effects such as stretching (44,46),

$$d_{31} = P_r(1 - \gamma)S_{11} + \frac{P_r(1 - \gamma)}{3}(\varepsilon_\infty - 1)S_{11}, \quad (13)$$

where γ is Poisson's ratio, ε_∞ is the permittivity at high frequencies, and S_{11} is the compliance of the polymer. The first term accounts for dimensional effects, and the second term gives the contribution of the local field effect.

Polarizability and Poling Conditions. Designing an amorphous polymer that has a large dielectric relaxation strength and hence piezoelectric response requires the ability to incorporate highly polar groups at high concentrations and cooperative dipole motion. A study of the relationship between relaxation times, poling temperatures, and poling fields is crucial to achieving optimal dipole alignment. Theoretically, the higher the electric field, the better the dipole alignment. The value of the electric field is limited, however, by the dielectric breakdown of the polymer. In practice, 100 MV/m is the maximum field that can be applied to these materials. Poling times need to be of the order of the relaxation time of the polymer at the poling temperature.

During poling, the temperature is lowered to room temperature, while the field is still on, to freeze in the dipole alignment. In a semicrystalline material, however, locking in the polarization is supported by the crystalline structure of the polymer, and it is therefore stable above the glass transition temperature of the polymer. Because the remanent polarization in amorphous polymers is lost in the vicinity of T_g , their use is limited to temperatures well below T_g . This means that the polymers are used in their glassy state, where they are quite stiff and thus limit the ability of the polymer to strain as stress is applied. A piezoelectric amorphous polymer may be used at temperatures near its T_g to optimize its mechanical properties, but not too close so as to maintain the remanent polarization.

Although there are few data that address the stability of piezoelectric activity in amorphous polymers, it is clear that time, pressure, and temperature can contribute to dipole relaxation in these polymers. For a given application and use temperature, the effect of these parameters on the stability of the frozen-in dipole alignment should be determined.

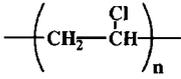
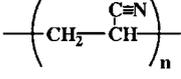
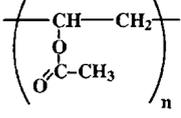
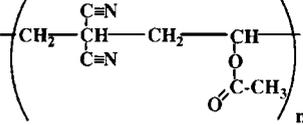
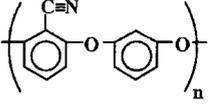
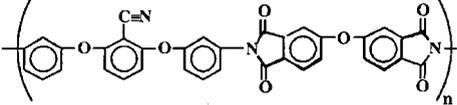
Examples of Amorphous Piezoelectric Polymers

The literature on amorphous piezoelectric polymers is much more limited than that for semicrystalline systems this is in part because no amorphous piezoelectric polymers; have responses high enough to attract commercial interest. Much of the previous work on amorphous piezoelectric polymers was in nitrile-substituted polymers, including polyacrylonitrile (PAN) (47–49), poly(vinylidene cyanide vinyl acetate) (PVDCN/VAc) (50–54), polyphenylethynitrile (PPEN) (55,56), and poly(1-bicyclobutanecarbonitrile) (57). Weak piezoelectric activity in polyvinyl chloride (PVC) and polyvinyl acetate (PVAc) has also been found (11,41,58,59). The most promising of these materials are vinylidene cyanide copolymers that exhibit large dielectric relaxation strengths and strong piezoelectricity. Table 4 shows the molecular structures of the most commonly encountered amorphous piezoelectric polymers.

Polyvinylidene Chloride. The carbon–chlorine dipole in polyvinylidene chloride (PVC) has been oriented to produce a low level of piezoelectricity. The piezoelectric and pyroelectric activities generated in PVC are stable and reproducible. PVC has been used as a basis for understanding and studying piezoelectricity in amorphous polymers (39). The piezoelectric coefficients d_{31} of PVC are reportedly in the range of 0.5–1.3 pC/N. This response was improved by simultaneous stretching and corona poling of film (44). The enhanced piezoelectric coefficient d_{31} ranged from 1.5–5.0 pC/N.

PVDCN Copolymers. In 1980, exceptionally strong piezoelectric activity was found (50) in the amorphous copolymer of VDCN and VAc. The copolymer was poled at 150°C (20°C below its T_g) and cooled to room temperature in the electric field. A $P_r = 55$ mC/m² was obtained in a poling field of 50 MV/m. That is comparable to the P_r of PVDF. When local ordering, or paracrystallinity, is inherent in the polymer or is induced by mechanical stretching, an increase in the value of the remanent polarization is observed. For example, some researchers (51) assert that the large discrepancy between the measured and calculated $\Delta\varepsilon$ for PVDCN-VAc may be attributed to locally ordered regions in the polymer. $\Delta\varepsilon_{\text{calculated}} = 30$ for the copolymer PVDCN/VAc, and $\Delta\varepsilon_{\text{measured}} = 125$ (51). This large discrepancy in the values of $\Delta\varepsilon$ is indicative of cooperative motion of several nitrile dipoles within the locally ordered regions of the polymers. Cooperativity means that multiple nitrile dipoles respond to the applied electric field in a unified manner, instead of each dipole acting independently. Although the existence of cooperative dipole motion clearly increases the piezoelectric response of amorphous polymers, the mechanisms by which cooperativity can be

Table 4. Structure, Polarization, and T_g of Piezoelectric Amorphous Polymers

Polymer	Structure	T_g (°C)	d_{31} (pC/N)	P_r (mCm ²)	Ref.
PVC		80	5	16	44
PAN		90	2	25	49
PVAc		30	—	5	59
P(VDCN-VAc)		170	10	50	51
PPEN		145	—	12	55
(β -CN) APB/ODPA		220	5 at 150°C	20	59

systematically incorporated into the polymer structure remain unclear at this time (59).

The large relaxation strength exhibited by PVDCN/VAc gives it the largest value of P_r and hence d_{31} of all of the amorphous polymers. A number of authors have suggested that PVDCN-VAc also exhibits ferroelectric-like behavior (51–53) due to switching of the nitrile dipoles in an ac field. The switching time is long compared to that of a normal ferroelectric polymer.

Other VDCN Polymers. The homopolymer of vinylidene cyanide is thermally unstable (60) and highly sensitive to moisture, but VDCN can be polymerized with a variety of monomers in addition to VAc, such as vinyl benzoate (VBz), methyl methacrylate (MMA) and others that form highly alternating chains. All of these copolymers show some degree of piezoelectricity although lower than PVDCN-VAc, which is explained by different activation energies for dipole orientation in the glassy state and different chain mobility that depends on the side group.

Polyacrylonitrile. Polyacrylonitrile (PAN) is one of the most widely used polymers. Shortly after it was shown that the PVDCN-VAc system is piezoelectric, researchers turned their attention to PAN due to its similarity to the aforementioned polymers. The presence of the large nitrile dipole in PAN indicated that it can be oriented by an applied electric field. PAN presented some challenges not encountered in other nitrile-substituted polymers, however. Although theoretical calculations predicted strong piezoelectric behavior, it was difficult to pole. Several investigators (47–49) proposed that the difficulty of poling PAN in

the unstretched state is related to the strong dipole–dipole interaction of nitrile groups of the same molecule that repel each other, and thus prevent normal polarization. Upon stretching, the intermolecular dipole interactions facilitate packing of the individual chains and give rise to ordered zones. The remanent polarization of both unstretched and stretched PAN has been measured using the thermally stimulated current method (TSC), and a twofold increase in remanent polarization (TSC peak at 90°C) was observed for PAN that was stretched to four times its original length (47). Another approach is the copolymerization of PAN with another monomer. Researchers have reported reduction of the hindering effect of the dipole–dipole interactions and enhancement of the internal mobility of the polymer segments when PAN is copolymerized with polystyrene or methyl methacrylate. Ferroelectric behavior has been observed in P(AN-MMA), where, for given temperature and field conditions, a characteristic hysteretic loop is obtained (49). It was concluded that it may be one rare example where both ferroelectric and frozen-in dipole orientations are superimposed.

Nitrile-Substituted Polyimide. Amorphous polyimides that contain polar functional groups were synthesized (61–63) and investigated for use as high temperature piezoelectric sensors. (β -CN) APB/ODPA polyimide is one such system. The (β -CN) APB/ODPA polyimide possesses the three dipole functionalities shown in Table 5. Typically, the functional groups in amorphous polymers are pendent to the main chain. The dipoles, however, may also reside within the main chain of the polymer, such as the anhydride units in the (β -CN) APB/ODPA polyimide. The nitrile

Table 5. Values of Dipole Moments Within Nitrile-Substituted Polyimide

Dipoles	Dipole Identity	Dipole Moment (Debye)
	Pendent nitrile group	4.18
	Main chain dianhydride group	2.34
	Main chain diphenylether group	1.30

to dipole is pendent a phenyl ring ($\mu = 4.2$ D), and the two anhydride dipoles ($\mu = 2.34$ D) are within the chain, resulting in a total dipole moment of 8.8 D per repeat unit.

The remanent polarization P_r of the (β -CN)APB/ODPA polymer found by the thermally stimulated current method (TSC) was approximately 20 mC/m² when poled at 80 MV/m above the T_g for 1 hour (64). Excellent thermal stability was observed up to 100°C, and no loss of the piezoelectric response was seen after aging at 50°C and 100°C for as long as 500 hrs.

Partially cured films of the (β -CN)APB/ODPA system were simultaneously corona poled and cured to enhance dipolar orientation and minimize localized arcing during poling. The aligned polar groups should be immobilized by additional imidization and subsequent cooling in an electric field. Both the T_g and the degree of imidization increase almost linearly as the final cure temperature is increased (64). The value of P_r was higher for films cured at lower temperatures. The mobility of the molecules in a partially imidized state should be higher than that in the fully cured state and therefore produce a higher degree of dipole orientation.

The importance of dipole concentration in ultimate polarization is evident from a comparison of polyacrylonitrile (PAN) and the polyimide (β -CN) APB/ODPA. PAN has a single nitrile dipole per repeat unit ($\mu = 3.5$ D) resulting in a dipole concentration of 1.34×10^{28} m⁻³. This translates into an ultimate polarization of 152 mC/m² (20). The (β -CN) APB/ODPA polyimide, on the other hand, has a total dipole moment of 8.8 D per monomer. The dipole concentration of (β -CN) APB/ODPA, however, is only 0.136×10^{28} m⁻³, resulting in an ultimate polarization of 40 mC/m², which is less than a fourth of that of PAN. As a result, similar polyimides that have increased nitrile concentrations were synthesized and characterized. Studies of these polymers show that polarization is significantly increased by increasing dipole concentration. Structure-property investigations designed to assess the effects of these dipoles on T_g , thermal stability, and overall polarization behavior are currently being pursued.

Even-Numbered Nylons. Nylon 6I and 6I/6T exhibit a D - E hysteretic loop across a temperature range of 30–65°C at a fixed maximum field of 168 MV/m (65). The remanent polarization increases as the temperature increases. Note that nylon 6I and 6I/6T are completely amorphous. The P_r is about 30 mC/m².

Aliphatic Polyurethane. Some researchers (1) have suggested that aliphatic polyurethane systems exhibit ferroelectricity that stems from the amorphous part at temperatures above the glass transition temperature. This “liquid state” ferroelectricity is very peculiar, seems to exist, and is supported by the hydrogen bonds present.

CHARACTERIZATION AND MODELING

Most piezoelectric characterization methods were developed for crystalline ceramics and had to be adapted for piezoelectric polymers. Methods based on resonance analysis and equivalent circuits that can be used to characterize semicrystalline PVDF and its copolymers are outlined in IEEE standards (66). Details for applying resonance analysis to piezoelectric polymers have recently been explored (67). Due to the lossy nature of some polymers, the IEEE standards are not adequate, and other techniques are needed to describe piezoelectric properties more accurately.

Quasi-static direct methods are both versatile and well suited to investigating fully the piezoelectric response of polymers. Direct methods of this type are especially appropriate for amorphous polymers. Thermally stimulated current measurements (TSC) (68) are used to measure the remanent polarization imparted to a polymer, and direct strain or charge measurements are used to investigate the piezoelectric coefficients with respect to the electric field, frequency, and stress.

TSC is a valuable tool for characterizing piezoelectric polymers. After poling a polymer, a measure of the current dissipation and the remanent polarization as a function of temperature can be obtained by TSC. As the sample is heated through its glass transition temperature (or Curie temperature for a semicrystalline polymer) at a slow rate (typically 1–4°C/min), the depolarization current is measured by an electrometer. The remanent polarization is equal to the charge per unit area and is obtained from the data by integrating the current with respect to time and plotting it as a function of temperature:

$$P_r = \frac{Q}{A} = \frac{1}{A} \int I(t) dt. \quad (14)$$

Figure 6 illustrates a typical TSC result. Because permanent dipoles are immobile at temperatures well below T_g , the current discharge remains low in this temperature range. As temperature increases to and beyond the T_g , however, the onset of dipole mobility contributes to a significant increase in the current peak. The peak in the current and the subsequent polarization maximum usually occurs in the vicinity of the T_g .

Direct methods for measuring the strain that results from applying a field or vice versa, applying a strain, and measuring the accumulated charge, are abundant. Interferometers, dilatometers, fiber-optic sensors, optical levers, linear variable displacement transducers, and optical methods are employed to evaluate the piezoelectric strain (converse effect) (69–72). The “out-of-plane” or thickness piezoelectric coefficient d_{33} can be ascertained as a function of the driving field and frequency. The coefficient

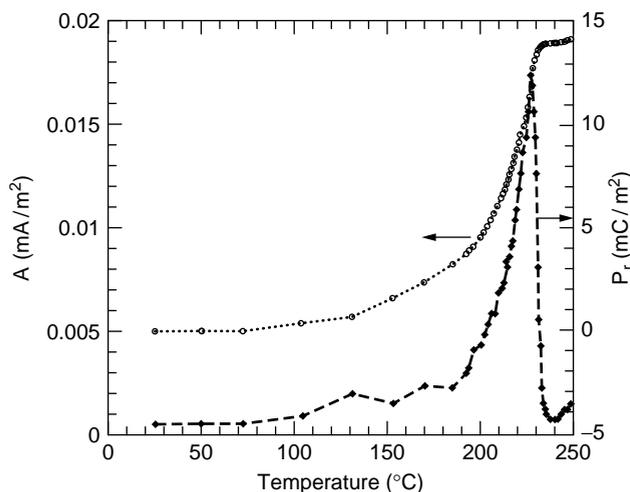


Figure 6. Plot of thermally stimulated current for a typical amorphous poled polymer.

is measured based on the equation,

$$S_{33} = d_{33}E_3, \quad (15)$$

where S_{33} is the strain and E_3 is the applied electric field.

A modified Rheovibron, or similar techniques, have been used to measure the direct piezoelectric effect, where charges accumulated on the surfaces of the polymer are measured (59). The piezoelectric coefficient d_{31} can be obtained by straining the polymer in the direction of applied stress using a force F . A charge Q is generated on the surface of the electrodes. A geometric factor is used to produce a geometrically independent parameter, surface charge density per unit applied stress:

$$d_{31} = \frac{Q/(WL)}{F/(Wt)}, \quad (16)$$

which has units of pC/N. W , L , and t are the width, length, and thickness of the sample, respectively.

Modeling

The methodology for modeling piezoelectric behavior in polymers varies, depending on the targeted properties. Approaches cover the range from macroscale to micro and atomistic scales. A detailed review of computational methods applied to electroactive polymers has been published (73).

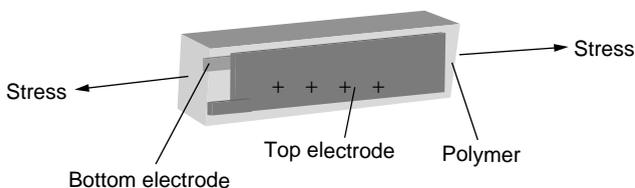


Figure 7. Direct effect in polymers.

In some cases, modeling can predict behavior where experiments cannot. Using molecular dynamics, the orientation polarization of the (β -CN) APB/ODPA polymer was assessed by monitoring the angle θ that the dipoles make with an applied electric field (74). The bulk P_r was calculated, and the results agreed extremely well with experimental results (61). Computational modeling, however, gave insight into the contributions of the various dipoles present in a way experimental results could not. The model predicted that 40% of the orientation polarization was due to the dianhydride within the backbone of the ODPA monomer and demonstrated the importance of the flexible ether linkage (oxygen atom) in facilitating dipole alignment. Modeling insight of this kind is invaluable in guiding the synthesis of new materials.

Modeling of PVDCN-VAc can also play a role in understanding the cooperative motion responsible for the high dielectric relaxation strength of this class of polymers, not possible experimentally (75). Recently, mesoscale simulation was used to describe polarization reversal in PVDF films (76).

Applications and Future Considerations

The potential for applying piezoelectric and other electroactive polymers is immense. To date, ferroelectric polymers have been incorporated into numerous sensing and actuating devices for a wide array of applications. Typical applications include devices in medical instrumentation, robotics, optics, computers, and ultrasonic, underwater, and electroacoustic transducers. One important emerging application area for electroactive polymers is the biomedical field where polymers are being explored as artificial muscle actuators, as invasive medical robots for diagnostics and microsurgery, as actuator implants to stimulate tissue and bone growth, and as sensors to monitor vascular grafts and to prevent blockages (77,78). Such applications are ideal for polymers because they can be made biocompatible, and they have excellent conformability and impedance that match body fluids and human tissue. The intent of this article is not to detail specific applications; the interested reader may consult excellent sources on the applications of piezoelectric and ferroelectric polymers (79–81).

In the future, we believe that fertile research areas for piezoelectric polymers will include work to enhance their properties, to improve their processibility for incorporation in devices, and to develop materials that have a broader use temperature range. Fundamental structure–property understanding has enabled the development of numerous semicrystalline and amorphous polymers. Based on this knowledge, future research that focuses on property enhancement via new chemistries that have higher dipole concentrations and incorporate dipole cooperativity may yield improved materials. Property enhancements may also be gained from processing studies to alter polymer morphology such as those used to make “single crystalline” fluoropolymers. Development of materials that can operate in extreme environments (high temperature and subambient temperature) is also important for expanding the use of piezoelectric polymers. Piezoelectric and pyroelectric constants of polymers are considerably lower than those of

ferroelectric inorganic ceramics. Improvements in properties by incorporating polymers in composites with inorganics to obtain higher electromechanical properties and better mechanical properties are also valuable. To date, piezoelectric polymer-ceramic composites have been made wherein the polymer serves only as an inactive matrix for the active ceramic phase. This is due to the mismatch in permittivity between the polymer and ceramic which makes it difficult to pole both phases. Research that results in active polymer and ceramic phases could yield interesting electromechanical properties.

ACKNOWLEDGMENTS

The authors express sincere appreciation to Dr. J.A. Young (Lawrence Livermore) for her technical insight in the area of amorphous piezoelectric polymers. The authors thank Suzanne Waltz (NASA Langley) for graphics assistance.

BIBLIOGRAPHY

1. T. Yuki, K. Shida, T. Koda, and S. Ikeda, *10th Int. Symp. Electrets*, 1999, pp. 675–677.
2. G.T. Davis, in *Polymers for Electronic and Photonic Applications*, C.P. Wong, ed., Academic Press, Boston, MA, 1993, p. 435.
3. M.G. Broadhurst, and G.T. Davis, in *Electrets*, G.M. Sessler, ed., Springer-Verlag, NY, 1980, Vol. 33, p. 283.
4. W.P. Mason, *Physical Acoustics and the Properties of Solids*. Van Nostrand, Princeton, NJ, 1958.
5. J.F. Nye, *Physical Properties of Crystals*. Clarendon, Oxford, 1957.
6. R.G. Kepler and R.A. Anderson, *J. Appl. Phys.* **49** (3): 1232–1235 (1978).
7. J.P. Luongo, *J. Polym. Sci. A-2* **10**: 1119–1123 (1972).
8. B.Z. Mei, J.I. Schienbeim, and B.A. Newman, *Ferroelectrics* **144**: 51–60 (1993).
9. H. Kawai, *Jpn. J. Appl. Phys.* **8**: 975 (1969).
10. K. Tashiro, H. Tadokoro, and M. Kobayashi, *Ferroelectrics* **32**: 167 (1981).
11. T. Furukawa, *IEEE Trans. Electr. Insul.* **24**: 375–393 (1989).
12. G.M. Sessler, *J. Acoust. Soc. Am.* **70** (6): 1596–1608 (1981).
13. A.J. Lovinger, *Science* **220** (4602): 1115–1121 (1983).
14. J.B. Lando and W.W. Doll, *J. Macromol. Sci.-Phys. B2*, 205 (1968).
15. A.J. Lovinger, in *Developments in Crystalline Polymers*, D.C. Bassett, ed., Applied Science, London, UK, 1982, Vol. 1, p. 242.
16. G.T. Davis, J.E. McKinney, M.G. Broadhurst, and S.C. Roth, *J. Appl. Phys.* **49**: 4998 (1978).
17. K. Omote, H. Ohigashi, and K. Koga, *J. Appl. Phys.* **81**(6): 2760 (1997).
18. Q.M. Zhang, V. Bharti, and X. Zhao, *Science* **280**: 2101–2104 (1998).
19. H. Kawai and I. Henji, *Oyo Buturi* **39**: 413 (1970).
20. B.A. Newman, P. Chen, K.D. Pae, and J.I. Scheinbeim, *J. Appl. Phys.* **51**(10): 5161–5164 (1980).
21. B.A. Newman, K.G. Kim, and J.I. Scheinbeim, *J. Mater. Sci.* **25**: 1779–1783 (1990).
22. Y. Takase, J.W. Lee, J.I. Scheinbeim, and B.A. Newman, *Macromolecules* **24**: 6644–6652 (1991).
23. R.B. Meyer, L. Strzelecki, L. Liebert, and P. Keller, *J. Phys. Lett. (Paris)* **36**: 69 (1975).
24. P. Le Barny and J.C. Dubois, in *Side Chain Liquid Crystal Polymers*, C.B. McArdle, ed., Blackie, Chapman and Hall, NY, 1989, p. 130.
25. G. Scherowsky, in *Ferroelectric Polymers, Chemistry, Physics and Application*, H.S. Nalwa, ed., Marcel Dekker, NY, 1995, p. 435.
26. Y. Takahashi, M. Iijima, K. Inagawa, and A. Itoh, *J. Vac. Sci. Technol. A* **5**: 2253 (1987).
27. Y. Takahashi, M. Iijima, and E. Fukada, *Jpn. J. Appl. Phys.* **28**: L408–L410 (1989).
28. T. Hattori, Y. Takahashi, M. Iijima, and E. Fukada, *J. Appl. Phys.* **79**: 1713–1721 (1996).
29. E. Fukada, *IEEE Trans. Ultrasonics Ferroelectrics Frequency Control* **47**: 1277–1290 (2000).
30. A.J.P. Martin, *Proc. Phys. Soc.* **53**: 183 (1941).
31. E. Fukada, in *Ferroelectric Polymers, Chemistry, Physics and Application*, Marcel Dekker, H. S. Nalwa, ed., NY, 1995, p. 393.
32. H. Maeda and E. Fukada, *Biopolymers* **21**: 2055 (1982).
33. M. Date, S. Takashita, and E. Fukada, *J. Polym. Sci. A-2* (8): 61 (1970).
34. T. Furukawa and E. Fukada, *J. Polym. Sci. Polym. Phys.* **14**: 1979 (1976).
35. Y. Ando and E. Fukada, *J. Polym. Sci. Polym. Phys.* **14**: 63 (1976).
36. E. Fukada, *Rep. Prog. Polym. Phys. Jpn.* **34**: 269 (1991).
37. Y. Ando, E. Fukada, and M.J. Glimcher, *Biorheology* **14**: 175 (1977).
38. E. Fukada, in *Mechanisms of Growth Control*, R.O. Becker, ed., Charles C. Thomas, Springfield, IL, 1981, p. 192.
39. M.G. Broadhurst, C.G. Malmberg, F.I. Mopsik, and W.P. Harris, in *Electrets: Charge Storage and Transport in Dielectrics*, M.M. Perlman, ed., The Electrochemical Society, Princeton, NJ, 1973, pp. 492–504.
40. M.G. Broadhurst, W.P. Harris, F.I. Mopsik, and C.G. Malmberg, *Polym. Prepr.* **14**: 820 (1973).
41. F.I. Mopsik and M.G. Broadhurst, *J. Appl. Phys.* **46**: 4204–4208 (1975).
42. B. Hilczler and J. Malecki, in *Electrets: Studies in Electrical and Electronic Engineering 14*. Elsevier, NY, 1986, p. 19.
43. H. Frohlick, *Theory of Dielectrics: Monographs on the Physics and Chemistry of Materials 42*. Oxford, U.K., 1958, p. 15.
44. V. Bharti, T. Kaura, and R. Nath, *IEEE Trans. Dielectr. Electr. Insul.* **2** (6): 1995.
45. H. Stefanou, *J. Appl. Phys.* **50**: 1486–1490 (1979).
46. R. Meyrueix and O. Lemonnier, *Phys. D Appl. Phys.* **27**: 379–386 (1994).
47. R.J. Comstock, S.I. Stupp, and S.H. Carr, *J. Macromol. Sci. Phys.* **B13** (1): 101–115 (1977).
48. H. Ueda and S.H. Carr, *Polym. J.* **16**: 661–667 (1984).
49. H. von Berlepsch, W. Kunstler, A. Wedel, R. Danz, and D. Geib, *IEEE Trans. Electr. Insul.* **24**: 357–362 (1989).
50. S. Miyata, M. Yoshikawa, S. Tasaka, and M. Ko, *Polym. J.* **12**: 857–860 (1980).

51. T. Furukawa, M. Date, K. Nakajima, T. Kosaka, and I. Seo, *Jpn. J. Appl. Phys.* **25**: 1178–1182 (1986).
52. I. Seo, *Ferroelectrics* **171**: 45–55 (1995).
53. P.A. Mirau and S.A. Heffner, *Polymer* **33**: 1156–1161 (1992).
54. M. Sakurai, Y. Ohta, Y. Inoue, and R. Chujo, *Polym. Commun.* **32**: 397–399 (1991).
55. S. Tasaka, T. Toyama, and N. Inagaki, *Jpn. J. Appl. Phys.* **33**: 5838–5844, (1994).
56. T. Takahashi, H. Kato, S.P. Ma, T. Sasaki, and K. Sakurai, *Polymer* **36**: 3803–3808 (1995).
57. H.K. Hall, R.J.H. Chan, J. Oku, O.R. Hughes, J. Scheinbeim, and B. Newman, *Polym. Bull.* **17**: 135–136 (1987).
58. V. Bharti and R. Nath, *J. Appl. Phys.* **82**: 3488–3492 (1997).
59. Z. Ounaies, J.A. Young, and J.S. Harrison, in *An Overview of the Piezoelectric Phenomenon in Amorphous Polymers*, I.M. Khan and J.S. Harrison, eds., American Chemical Society, Washington, DC, 1999.
60. S. Tasaka, N. Inagaki, T. Okutani, and S. Miyata, *Polymer* **30**: 1639–1642 (1989).
61. Z. Ounaies, J.A. Young, J.O. Simpson, and B.L. Farmer, *Mater. Res. Soc. Proc.: Mater. Smart Syst. II*, Materials Research Society, Pittsburgh, PA, 1997, Vol. 459, p. 59.
62. J.O. Simpson, Z. Ounaies, and C. Fay, *Mater. Res. Soc. Proc.: Mater. Smart Syst. II*, Materials Research Society, Pittsburgh, PA, 1997, Vol. 459, p. 53.
63. Z. Ounaies, C. Park, J.S. Harrison, J.G. Smith, and J. Hinkley, *SPIE Proc., Electroactive Polym. Actuators Devices*, Newport Beach, CA, 1999, Vol. 3669, p. 171.
64. C. Park, Z. Ounaies, J. Su, J.G. Smith, Jr., and J.S. Harrison, *Mater. Res. Soc. Proc.: Electroactive Polym.*, 1999, Vol. 600.
65. Y. Murata, K. Tsunashima, and N. Koizumi, *8th Int. Symp. Electrets (ISE 8)*, 1994, pp. 709–714.
66. IEEE Standard on Piezoelectricity, (IEEE Standard 176-1987); Institute of Electrical and Electronic Engineers, NY, 1989.
67. S. Sherritt and Y. Bar-Cohen, in *Electroactive Polymer (EAP) Actuators as Artificial Muscles: Reality, Potential and Challenges*, Y. Bar-Cohen, ed., SPIE Press, Bellingham, WA, 2001, p. 405.
68. J. van Turnhout, *Polym. J.* **2**: 173–191 (1971).
69. J.T. Dawley, G. Teowee, B.J.J. Zelinski, and D.R. Uhlmann, *Piezoelectric Characterization of Bulk and Thin Film Ferroelectric Materials Using Fiber Optics*. MTI Instruments Inc., Application Note.
70. W.Y. Pan, H. Wang, and L.E. Cross, *Jpn. J. Appl. Phys.* **29**: 1570 (1990).
71. Q.M. Zhang, W.Y. Pan, and L.E. Cross, *J. Appl. Phys.* **63**: 2429 (1988).
72. T.L. Jordan, Z. Ounaies, and T.L. Turner, *Mater. Res. Soc. Symp. Proc.* **459**: 231 (1997).
73. K.E. Wise, in *Electroactive Polymer (EAP) Actuators as Artificial Muscles: Reality, Potential and Challenges*, Y. Bar-Cohen, ed., SPIE Press, Bellingham, WA, 2001, pp. 267–284.
74. J.A. Young, B.L. Farmer, J.A. Hinkley, *Polymer* **40**: 2787 (1999).
75. K.E. Wise, *3rd SIAM Conf. Math. Aspects Mater. Sci.*, Philadelphia, PA, May 21–24, 2000.
76. T. Koda, K. Shibasaki, and S. Ikeda, *Comp. Theor. Polym. Sci.* **10**: 335–343 (2000).
77. Y. Bar-Cohen, in *Electroactive Polymer (EAP) Actuators as Artificial Muscles: Reality, Potential and Challenges*,

Y. Bar-Cohen, ed., SPIE Press, Bellingham, WA, 2001, p. 615.

78. D. DeRossi and P. Dario, in *Medical Applications of Piezoelectric Polymers*, P.M. Galletti, D.E. DeRossi, and A.S. DeReggi, eds., Gordon and Breach, 1988, p. 83.
79. G.M. Garner, in *The Applications of Ferroelectric Polymers*, T. Wang, J. Herbert, and A. Glass, eds., Blackie, London, UK, 1988, p. 190.
80. T.R. Meeker, in *The Applications of Ferroelectric Polymers*, T. Wang, J. Herbert, and A. Glass, eds., Blackie, London, UK, 1988, p. 305.
81. E. Yamaka, in *The Applications of Ferroelectric Polymers*, T. Wang, J. Herbert, and A. Glass, eds., Blackie, London, UK, 1988, p. 329.

POWER INDUSTRY APPLICATIONS

J. STRINGER

EPRI
Palo Alto, CA

R. H. RICHMAN
Daedalus Associates
Mountain View, CA

INTRODUCTION

Many industrial and commercial sectors are made up of separate, relatively small components: aerospace and ground transport, for example. In such applications, truly smart materials systems (SMS) are defined as those that respond autonomously to changes in their operating conditions. That is, they detect the onset of “illness” and take steps to effect a cure. In contrast, the electric power industry is a vast, interconnected enterprise, not a collection of individual entities. It is a far-flung grid, fed and drained continuously at variable rates. SMS must be consistent with this character: owing to the extent of the grid and the way it operates, just detecting illness *and determining its location* constitutes smartness. Moreover, users of electricity expect reliability everywhere the grid reaches, so avoiding failures (outages) is paramount. Thus, *distributed* SMS are often required to achieve reliability. Some other sectors (e.g., highways) also will rely on distributed SMS, but in electricity systems the SMS must function under more hostile conditions of temperature, pressure, aggressive chemicals, and especially, intense electric and/or magnetic fields. Early detection of deviations or trouble is vital. In a sense, successful performances of SMS in the power industry equates to buying time for a rational response. Whatever else they might do by way of autonomous responses (derating a unit, eliminating an ailing component from a redundant set, etc.), the crucial action for SMS in electricity systems is to notify a central authority about an (impending) illness and where that illness is.

Although investment in SMS research by the power industry has been substantial, utilization of SMS within the industry is still in the early stages. There is little doubt, however, that the influence of SMS will be profound in the future.

Table 1. Electric Power Industry Segments

Segment	Number of Entities	% of U.S. Consumers Served	Median Number of Customers per Organization
Investor-owned utilities	264	75%	~210,000
Rural electric cooperatives	947	10%	~8000
	888 (distribution only)		
	59 (generation and transmission)		
Public power	~2,000	15%	~1750
Federal systems	10	<0.1%	n/a
		(supplier to other segments)	

OVERVIEW OF THE ELECTRIC POWER INDUSTRY

Size, Composition, and Economic Pressures

There are four sectors in the electric power industry: investor-owned utilities, rural electric cooperatives, public power systems, and federal power agencies. The number of organizations in each category, percentage of U.S. customers served by each segment, and the median number of customers served by each are summarized in Table 1. Total installed capacity is on the order of 750 GW (1), with approximately 627,400 miles of transmission lines of 11 kV or more (2), and 3,600,000 miles of distribution lines (3). Annual expenditures for generation and for transmission and distribution are approximately equal at about \$12 billion each.

In the past, utilities have been highly regulated, with specified rates of return, protected service territories, and the obligation to provide power to any customer within a defined service territory. Over the last few years, protected markets and assured customers have come to an end, and nonutility generators (NUGs) have appeared on the scene. It is estimated that there are now over 3000 NUGs in the United States and that 75% of all new capacity since 1980 has been supplied by NUGs. Competitive pressure is pervasive; lowering costs, improving and expanding customer services, diversifying into other businesses, and striving for better utilization of corporate assets have become the primary drivers.

Problem Costs and Solution Benefits

A principal role for smart materials in the power industry would be to lower costs and improve system reliability. An indication of the scale of the challenge is provided by the three rules of thumb that follow.

1. A typical coal-fired power plant is approximately 35% efficient. For a 500 MW generating unit, loaded at 70% plant factor, a 0.1% efficiency improvement would reduce annual costs by about \$150,000 at a typical busbar electricity cost of \$0.017/kWh. For a six-unit plant, annual savings would therefore be about \$900,000.
2. Most electricity is generated by large units and there is a high premium for keeping those units on line.

When high-merit units are out of service, replacement power must often be provided by less-efficient (more costly) sources. Unscheduled outages caused by equipment failure or significant unit degradation are particularly costly. Outage lengths depend on availability of spares, difficulty of access, degree of damage, and so on, and can range from hours to months. Replacement power cost for a one-day outage of a 500 MW unit is on the order of \$240,000 (at \$0.02/kWh). Emergency (unscheduled) power purchases can cost even more; \$0.03–0.04/kWh is not unusual.

3. Most generation and power delivery equipment is operated well below theoretical limits so as to provide reasonable factors of safety. Furthermore, when deterioration of structural or functional integrity is suspected, the afflicted equipment is derated, that is, operated at a lower level of performance. Derating a 500 MW unit by 5% (25 MW) is equivalent to \$12,000 per day or \$4,380,000 per year of lost revenue (at \$0.02/kWh). A transmission line that must be operated below its normal thermal limit can reduce revenues by \$1,750,000 per year.

The Power Industry Environment

Although combustion turbines, often combined with heat-recovery boilers, are increasingly popular for new installations, the generation of power from thermal sources is still preponderantly by steam turbines. Hence, a primary activity for many utilities is the raising of steam. Fossil-fueled boilers in large central stations produce as much as 1260 kg of steam per second (10,000,000 lb of steam per hour) at pressures ranging from 13.8 to 24.8 MPa (2000 to 3600 psig) and temperatures up to 580°C (1100°F). Nuclear-fueled, pressurized-water reactors operate at coolant pressures of 15.2 to 15.9 MPa (2200–2300 psig) and a temperature of about 316°C (600°F) to produce steam at 6.2 to 7.3 MPa (900–1050 psig). In other words, steam plants are characterized by hostile environments: high temperatures, high pressures, and aggressive fluids. The damage mechanisms that ensue when plant components are exposed to these conditions are summarized in Table 2.

Steam plant environments are demanding, but most of them are not unique; high temperatures and pressures,

Table 2. Damage Mechanisms in Large Components of Fossil-Fueled Generating Plants

Plant Component or System	Environmentally- Assisted Fatigue					Creep- Fatigue	Erosion	Wear (Note 2)	Embrittle- ment (Note 3)
	Fatigue	(Note 1)	Corrosion	Creep					
Boiler tubes	×	×	×	×	×	×	×	×	×
High-temperature headers	×			×	×				×
Low-temperature headers, steam, and lower Drums	×	×	×				×		
Main steam and hot reheat piping	×	×		×	×	×			
Steam turbine casings, steam chests, valve casings	×			×	×	×			×
High-pressure (HP) and intermediate-pressure (IP) turbine rotors	×			×	×				×
Low-pressure (LP) turbine rotors	×	×							
Steam turbine blades, rotating and stationary	×	×	×				×		
Condensers and heat exchangers		×	×				×		
Feedwater heaters	×		×				×		
Coal mills	×							×	

Notes: 1. Environmentally assisted fatigue includes mechanisms such as corrosion fatigue and stress corrosion cracking.
2. Wear includes abrasion, galling, rolling contact, fretting, etc.
3. Embrittlement includes temper embrittlement, graphitization, and hydrogen embrittlement.

and reactive process streams, are also found, for instance, in the chemical processing and metal production industries. In contrast, the environments associated with electricity generation and delivery are special in that they involve strong electromagnetic fields and large electric currents. As a consequence, the health of insulation materials, with their extreme sensitivity to relatively low temperatures (100–180°C, or 212–356°F) and mild mechanical forces, dominates concerns about reliability and failure avoidance. Of course, there are also the usual threats to integrity, such as cyclic stresses, corrosion, and wear. Damage mechanisms operative in this sector of the power industry are given in Table 3.

A final point about the power industry environment: electricity suppliers keep their primary equipment (boilers, turbines, transformers, transmission and distribution lines) in service for longer times than is common for most other industries. Plant and equipment designs aim for a 40-year life, and nowadays the utilities often strive for life extension to even longer times. Inevitably, even normal wear and tear take their toll and apparatus becomes increasingly fragile with age. However, the reliability requirement is not relaxed for these older components, so that early warning of impending dysfunction becomes even more critical. In addition, there are implications about longevity and redundancy requirements for SMS,

Table 3. Damage Mechanisms in Electrical and Hydroelectric Apparatus

Component or System	Degradation of Insulation					Mechanical or Environ- mentally Assisted Fatigue	Corro- sion	Mechanical Over- load	Mass Trans- fer by Arcing	Cavi- tation Erosion	Wear
	Thermal	Mechanical	Chemical	Electrical	Aging						
Generators											
Rotating fields	×	×	×	×	×	×		×	×		×
Stators	×	×	×	×	×	×					
Retaining rings						×					
Motors	×	×	×	×		×			×		
Circuit breakers and switchgear	×		×	×			×		×		×
Transformers	×		×	×	×						×
Transmission lines	×	×						×			×
Underground cables	×		×		×		×				
Hydrogenerators											
Runners and gates							×			×	×
Generators	×	×	×	×	×	×					

since SMS embedded in the insulation of overhead conductors or in concrete structures, for example, cannot be replaced.

Constraints on SMS Development for the Power Industry

All of the foregoing environmental factors, especially the extended nature of the electricity grid and the pervasive presence of strong electromagnetic fields, impose limitations on SMS solutions to power industry challenges. In the first place, SMS from other industries will not, in general, be directly transferable in their original form either because they have not been designed for the specific tasks of the contemplated application or because electrically based SMS cannot function properly in intense electromagnetic fields. Second, the spatial extent of electricity systems dictates that the most urgent job for SMS is to detect damage or deviant conditions and report the location to a central authority. Autonomous actions by SMS are less important. The SMS constituent that matches this requirement is smart sensors. In fact, smart sensors are the aspect of the SMS that has been most thoroughly investigated for utility applications. Here, transference from another industry was vital: without the R&D base built by the communications industry, the technology of fiber-optic sensors might not have matured.

Undoubtedly, a wide variety of power industry problems will someday be addressed by truly smart, autonomous, fully integrated SMS. At this juncture, however, the main thrust for SMS in utility systems has been aimed at developing and applying smart sensors and, to less extent, smart sensor-actuators.

SMART SENSORS

Requirements and Scope

Need for Better Measurements. In recent years, advances in electronics (signal processing and conditioning) have resulted in improved instrumentation and controls for power plants. Many utilities have replaced analog and pneumatic controls with distributed digital control systems (DCSs) in order to reduce operating costs and to increase the capability to respond to system dispatch. A DCS can regulate processes with less than 0.25% uncertainty, compared with 2% to 3% uncertainty for analog controls. However, the conventional sensors used in power plants cannot, without frequent calibration, provide the levels of accuracy appropriate for sophisticated DCSs. Thus, smart sensors offer the means to increase measurement accuracy and thereby realize the benefits of DCSs.

Opportunities for better measurements are numerous in every sector of the power industry. For example, Sachdeva (4) lists 1393 essential sensors in a typical 500 MW coal-fired power plant. Of these, approximately 1150 are devoted to sensing temperature and pressure. In forward-looking utilities that are developing methodologies for early diagnosis of problems, monitoring of critical components has resulted in 4000 sensors in some generating units (5). Likewise, an NSF workshop (6) identified

26 utility problems associated with transmission and distribution that could be alleviated with improved sensors; affected components included lattice steel towers, wooden poles, overhead conductors and insulators, underground transmission and distribution facilities, and substations. Of course, not all of these sensors need to be smart; for many of them the smartness can be in central control. Still, a significant portion must be smart in order to overcome the effects of difficult environments.

Performance in the Environment. The utility industry can be thought of as the agency for three sequential operations: (1) converting thermal energy to mechanical energy, (2) converting mechanical energy to electrical energy, and (3) electricity delivery. The environments characteristic of each sector are summarized in Table 4. For the purposes of this article the second and third functions are similar enough in their environments that they are discussed together as electricity systems in what follows.

The first prerequisite for SMS in the power industry is that they be able to perform reliably in the environment. That is, they must produce accurate signals under normal as well as perturbed conditions for prolonged periods of time. Again, their priority is to inform a central authority that there is a problem or that damage has occurred, and where. Autonomous action can also be part of the SMS mandate; electricity systems are quite redundant, for example, so a smart sensor could initiate close-down of a redundant component that is about to fail. But a report of illness is the primary responsibility. As will be seen, most smart sensors intended for power-industry applications are based on fiber optics because of their immunity to electromagnetic interference (including lightning strikes and ground faults), as well as their small size and light weight.

Thermal Plant Environments

Pressure Sensing. Many pressure transmitters now installed in power plants rely on fill fluids to separate process fluids from the gauge mechanisms. These devices are subject to failures when fill fluids leak, failures that are difficult to detect. In addition, all conventional pressure sensors drift with time, a condition that necessitates disproportionately large efforts to restore accuracy and verify operability. These concerns have prompted an assessment of modern sensor technologies to select candidates for adaptation to use in power plants (7). Among the findings is the conclusion that fiber-optic pressure sensors *could* provide more extensive sampling of process pressures, an improved mode of signal transmission and processing, and freedom from electromagnetic and radio-frequency interference in nuclear and fossil-fueled steam plants.

As a step toward implementing improvements in pressure sensing, a fiber-optic transducer based on the micro-bend attenuation of light transmitted through an optical fiber (Fig. 1) has been developed (8). Since the diaphragm deflects linearly with pressure, process pressure is measured by the diminution of light transmission through an aluminum-coated fiber with core/clad/coating dimensions of 150/180/210 μm . Performance of the transducer was

Table 4. Classification of Environments in the Power Industry

Function	Apparatus	Environment
Convert thermal to mechanical energy	<ul style="list-style-type: none"> • Steam plant • Combustion turbines 	<ul style="list-style-type: none"> • High temperatures • High pressures • Aggressive fluids • Relatively close to central control
Convert mechanical to electrical energy	<ul style="list-style-type: none"> • Generators 	<ul style="list-style-type: none"> • Near-ambient temperatures and pressures • High electrical and magnetic fields • High mechanical forces • Relatively close to central control
Electricity delivery	<ul style="list-style-type: none"> • Transformers • Switchgear • Overhead transmission lines • Underground cable 	<ul style="list-style-type: none"> • Near-ambient temperatures and pressures • High electrical and magnetic fields • High electrical currents through contacts • Often far from central control

characterized in laboratory and field trials at pressures to 22.75 MPa (3300 psig) and temperatures to 438°C (820°F), with measurement error calculated to be 1.2% of full scale. This smart sensor, developed for coal liquefaction service by the U.S. Department of Energy, would be

suitable for pressure measurements in nuclear steam generators.

Strain Measurement. Strain sensing as an indicator of structural health in high-temperature components is

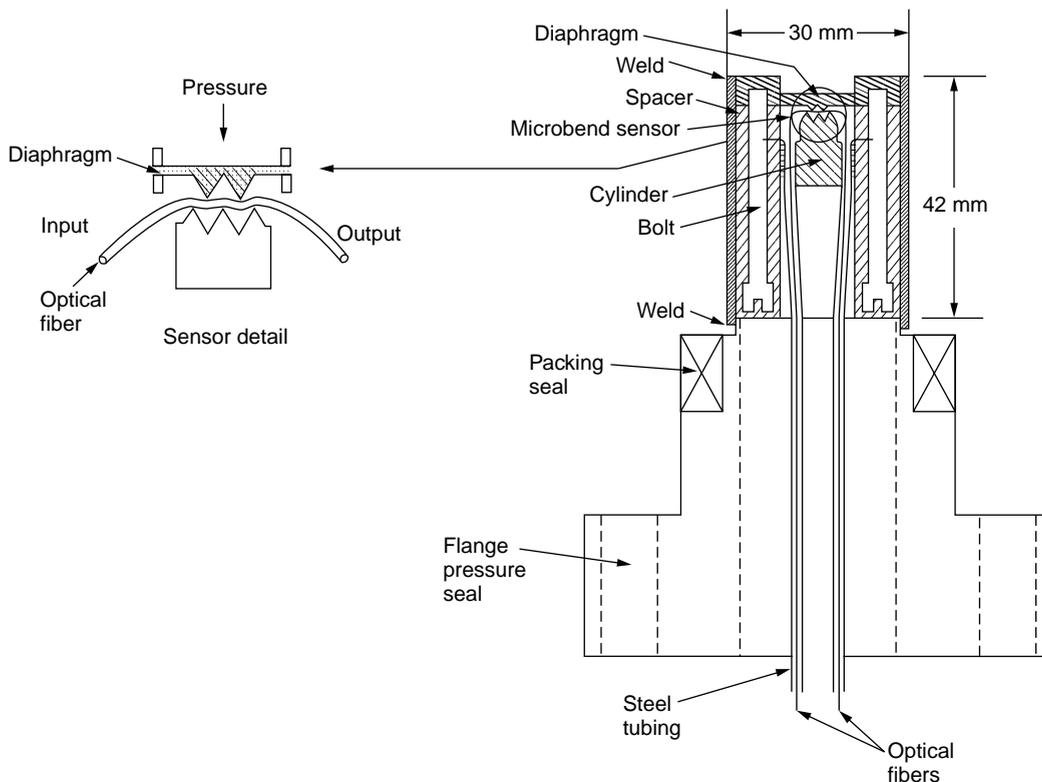


Figure 1. Diagram of the fiber-optic pressure transducer (8).

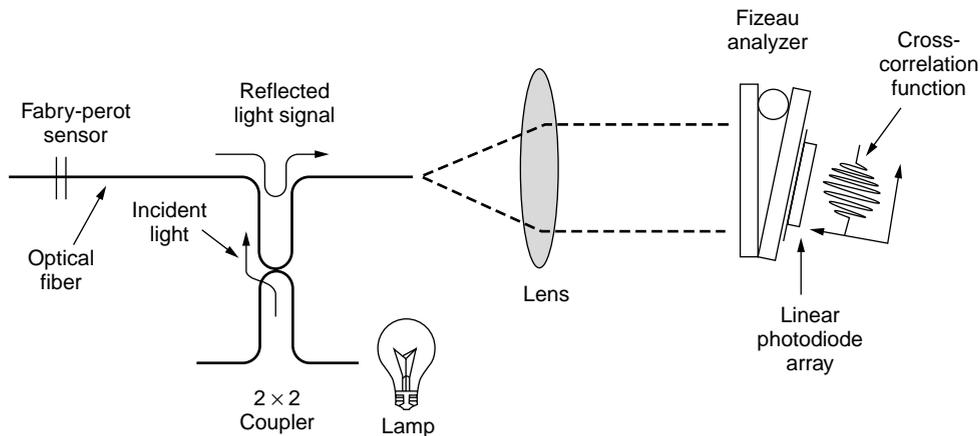


Figure 2. Schematic of interferometric strain sensor (9).

becoming more important as power-generating equipment ages. Conventional foil and wire strain gauges are not reliable at temperatures above 250°C (482°F) for long times, largely owing to the unavailability of adhesives that can withstand those temperatures. Therefore, a fiber-optic sensor to monitor strain in boilers, headers, steam pipes, and other high-temperature structures has been developed for operation at temperatures up to 650°C (1202°F).

Based on the Fabry-Perot interferometric technique, the system is shown schematically in Fig. 2. A broadband optical beam is conducted into a quartz tube containing two fibers, each with a partially reflecting mirror at the end. A small air gap or resonance cavity between the two mirrors forms an extrinsic Fabry-Perot interferometer (Fig. 3). Beams reflected from the two mirrors interfere, travel back toward the detector, and enter an optical wedge (Fizeau analyzer). Reflected light from the Fabry-Perot sensor is transmitted maximally where the optical path length matches the dimension of the Fabry-Perot cavity. Thus, when strain changes the cavity length, there is a corresponding shift in the intensity maximum transmitted through the optical wedge. A linear photodiode array at the back of the optical wedge detects the transmitted beam. Because this is a frequency-modulated sensor, it is insensitive to light attenuation; the signal reaching the photodiode array need only contain enough information for decoding (10).

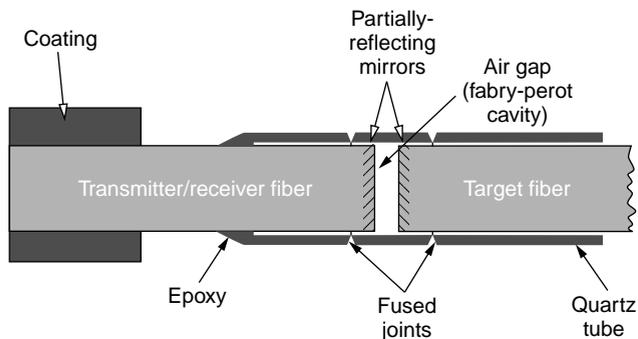


Figure 3. Schematic of Fabry-Perot interferometer (9).

This small, lightweight device has been field tested in two applications at a power station: a spot welded on a main steam line operating at 565°C (1049°F), and in a thermowell inside a reheat steam line at 538°C (1000°F). Feasibility of on-line monitoring of strain in structures at high temperatures was clearly demonstrated.

Measurements with Fiber Bragg Gratings. A potentially more versatile sensor technology is embodied by wavelength-modulated fiber Bragg gratings (FBG), which are created in low-cost, commercially available fibers. Each FBG reflects a characteristic wavelength that changes as the grating periodicity changes with temperature and/or strain. Research sponsored by the power industry has concluded that, in principle, FBGs can serve as generic transducer elements to measure temperature, pressure, strain, vibration, acoustical disturbances, electrical and magnetic field strengths, and the concentrations of certain chemical species (11).

A primary issue in developing FBGs for utility use in high-temperature environments is their stability. Exposures to high temperatures and temperature cycling has demonstrated that FBGs are quite robust: FBGs endured long-time cycling between ambient and 427°C (800°F) without degradation. Use temperature can be extended to 650°C (1202°F) only if very low loads (strains) are imposed on the fiber (11). FBGs for measuring temperature, pressure, and strain are now in prototype development. Although several approaches to chemical sensing with FBGs have been explored in laboratory studies, they are not yet ready for prototyping.

Sensors for Combustion Processes. Accurate and reliable sensors for very high temperature environments would have multiple benefits for thermal plant: avoidance of damage to heat-transfer surfaces, combustion controls for individual burners (not possible at present), reduction of noxious emissions, and structural health monitoring of critical components, to name a few. The chief barrier to developing smart systems based on fiber optics for combustion environments is the temperature constraint on silica fibers. Fiber function is limited by the temperatures at

which (1) the dopants in conventional silica begin to diffuse rapidly enough to affect adversely both signal attenuation and wave-guiding properties, and (2) the silica softens. These temperatures are approximately 800 to 900°C (1472–1652°F).

Since crystalline sapphire has reasonable optical propagation properties, melts above 2000°C (3632°F), and is corrosion resistant, it is the basis for research aimed at developing fiber-optic sensors capable of operating above 1500°C (2732°F). Schemes under investigation include sapphire-fiber-based Fabry-Perot interferometers, FBGs, and intensity-based sensors, as well as ways to modify fiber coatings and claddings that make sensors able to function in combustion gases and superheated water environments (9).

One especially important challenge for high-temperature sensing is management of emissions from combustion. In control systems for nitrogen oxides (NO_x), ammonia (NH_3) or urea ($\text{NH}_2\text{-CO-NH}_2$) is injected into combustion gas to react with NO_x and produce molecular nitrogen and water. Postcombustion NO_x reduction must avoid significant NH_3 in the exhaust, both to comply with emission regulations and to keep from fouling downstream components. A feedback control system is needed, but no reliable, real-time NO_x or NH_3 sensors have been available.

An investigation is exploring measurement of nitric oxide (NO) levels on the basis of radiative emissions from single molecular transitions that are well separated from emission features associated with other constituents in the flow. The system consists of feedback-stabilized, scanning Fabry-Perot interferometers linked with thermoelectrically cooled wavelength detectors. A digital system controls cavity lengths for wavelength scanning. One detector monitors NO upstream of the injection plane and a second monitors NH_3 downstream. Signals from both are fed into the injection system controller, which then determines in real time the quantity of NH_3 or urea to be injected for optimal NO removal. Tests of the prototype system are underway at a utility power station (9).

pH Measurement. Corrosion behavior of steam-plant materials is determined to a large extent by the pH and electrochemical potential of the circulating water. Since the physicochemical properties of water are highly sensitive to temperature, there is strong incentive to develop on-line sensors for pH that can be used at system temperatures, rather than relying on analytical extrapolations from grab samples cooled to ambient, as is the current practice. However, attempts to develop on-line sensors for pH at elevated temperatures must confront two problems: degradation of the sensor materials by hot water; and, in the high-purity water characteristic of cooling loops, interactions of the sensor with the water can affect the pH being measured.

A sensor was developed by incorporating a pH-sensitive organic dye (8-hydroxypyrene-1,3,6-trisulphonic acid) in a polyacrylamide polymer at the end of an optical fiber (Fig. 4). By choosing a dye with two absorption peaks, the sensor indicates the pH as the ratio of the two peaks; long-term leaching of the dye does not compromise the

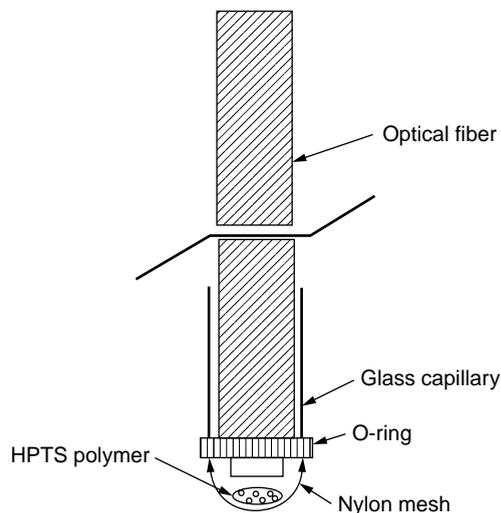


Figure 4. Schematic of the fiber-optic pH sensor (12).

measurements. pH determinations were made successfully over the course of a one-year immersion at 38°C (100°F), and pH changes associated with addition of 1 ppm morpholine were measured consistently (12).

Water at higher temperatures exacerbates stability problems for fiber-optic reflectors. A 250 nm thick, multilayer (titanium-platinum-gold) coating for sapphire was found to be stable to 180°C (356°F). In conjunction with an azo chromophore (dinitrophenyl-azo-naphthol) as an indicator, pH was measured reproducibly in the laboratory for 160 hours at 50°C (122°F), which was the stability limit of the cellulose acetate waveguide (13). In any case, azo-based indicators are stable only to about 100°C (212°F). While none of the pH-sensing systems devised so far are acceptable for use in high-temperature water, preliminary experiments have shown that a copper phthalocyanine chromophore is stable to at least 200°C (392°F). It is expected that the next step, an incremental improvement, will be a fiber-optic pH probe that is usable to 150°C (302°F).

Electricity Systems Environments

Distributed Fiber-Optic Temperature Sensor (DFOTS). Insulator temperature is a key factor in the safe and reliable operation of motor and generator windings, transformers, circuit breakers, underground cables, and overhead transmission lines. Although temporary overload conditions do not normally cause thermal damage to conductors, higher than normal temperatures do have a cumulative effect of shortening insulation life. On-line methods to locate and measure “hot spots” have traditionally not been available, largely owing to the fact that most transducers are electrically based. In other words, conventional sensing devices are usually incompatible with the environment of an electrical system. Furthermore, local temperature measurements (e.g., 0.1 m [4 in] long or less) need to be made throughout an electrical system that may be hundreds of meters (feet) long.

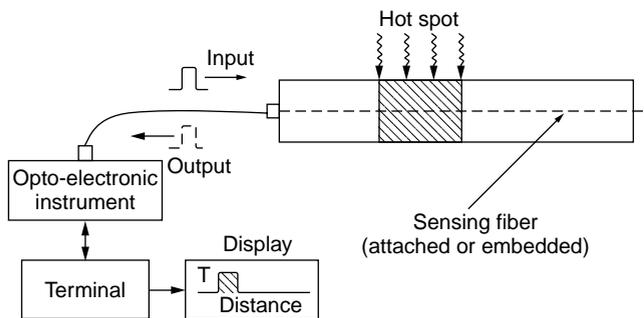


Figure 5. Schematic of the DFOTS system for sensing hot spots in electrical windings (9).

Motors, Generators, and Circuit Breakers. To address this challenge, the DFOTS system has been developed for detecting hot spots in low-temperature ($\leq 150^{\circ}\text{C}$ [302°F]), high-voltage environments. It is based on optical time domain reflectometry (OTDR), which was devised by the telecommunications industry for fault location in fiber-optic telephone lines. A light pulse transmitted by an optical fiber is gradually attenuated by absorption and Rayleigh scattering. The scattered light returns in a direction opposite to that of the injected light pulse. To make a DFOTS, the fiber must be modified along its length such that the local backscattering changes as a function of the local temperature change. This was accomplished by coating the fiber with a UV-curable polymer that changes refractive index reversibly with temperature (14). A change in the local intensity of backscattered light serves as a sensitive indicator of temperature, and the elapsed time of the returned pulse indicates the location of the temperature change, as shown in Fig. 5.

The DFOTS system has monitored winding temperatures in rotor and stator windings of motors and generators in trials at power plants (15) and in switchgear (14). It is accurate to within 5°C (9°F) over the range 0 to 150°C (32 to 302°F). A 100 ps laser pulse is capable of resolving hot spots only 2 cm (0.8 in) long over a fiber length of 40 m (131 ft). In retrofit applications, the DFOTS fiber can be strung between windings bars and slot wedges; in new windings, the sensor can be incorporated in the high-voltage groundwall insulation in direct contact with the conductors. In either application, impending generator problems could be diagnosed rapidly and major winding failures prevented. Improved remaining-life assessments would also ensue from knowing real thermal histories of insulation.

Underground Cables. A similar DFOTS has been devised for temperature monitoring of underground power cables. The power transfer capability of a buried cable is strongly affected by thermal conditions along the length of the circuit: burial depth, ambient earth temperature, soil thermal resistivity, and the like. Power transfer could be optimized if actual temperatures along the cable are known in real time.

When a sharply pulsed laser beam (1050 nm) is injected into a standard multimode fiber, very weak, thermally dependent, molecular vibrations produce reflections

along the fiber length. These reflections, known as Raman backscattering (in contrast to Rayleigh backscattering in the previous example), can be detected by OTDR and the signals processed to determine temperature along the length of the optical fiber. A commercial system (York Sensors, Ltd., DTS-80) based on Raman scattering was installed in a duct containing 66 kV power cables. Ampacity analysis of the measurements enabled the utility to uprate the cable circuit by 8%, without exceeding the 90°C (194°F) maximum design temperature for the cables (16).

Strain in End Windings. The stator winding of an AC generator is comprised of conductors (coil sides or bars) that are housed in slots in the stator core. At the end regions adjoining the active length of the machine, pairs of conductors are linked by end connections to form coils. The complete set of end connections at each end constitutes an end winding. End windings cannot be supported as securely as the conductors, which are in slots; if blocks and lashings become loosened by the forces of starting and stopping or system faults, the end windings can vibrate and progressively damage insulation. Undetected end-winding vibration can lead to a forced outage in a relatively short time. Standard strain gauges (metal foil or wire) cannot function in the strong electromagnetic environment of an operating generator.

A fiber-optic sensor based on microbending has been developed to measure strain in end windings during operation (9). The sensor attaches directly to an end turn and converts deflection to strain. Output of the device is linear over the range $\pm 1000 \mu\text{m/m}$ (1000 microstrain) and 5 to 100 Hz, with a resolution of $\pm 5 \mu\text{m/m}$.

Monitoring Transformers. Substation transformers are large, oil-filled devices and are among the most expensive components in an electric-power network. The cost of a failure, or an outage to repair a unit, can exceed the original cost of the transformer within five days if the cost of replacement power from a less-efficient station is accounted. Deterioration of transformer oil results from excessive temperature, aging, and electrical discharges through the oil. Oil has an effective lifetime of about one million hours at 90°C (194°F) but only about 100 hours at 180°C (356°F).

Winding Temperatures. In principle, temperatures in transformer windings could be measured by DFOTS based on Rayleigh or Raman backscattering. However, it has been found that transformer oil penetrates the fiber jacket and is absorbed by the polymer cladding of Rayleigh-based sensors, and all-silica Raman-based sensors have inadequate spatial resolution (about 5 m [16.4 ft]) at their present stage of development. Point monitoring of winding temperatures in real time is the best that can be done right now by measuring the temperature-dependent fluorescent decay time of a photoluminescent sensor material (manganese-activated magnesium fluorogermanate). Pulses of blue light power the phosphors; fluorescence returning in the all-silica fiber is detected and interpreted in terms of sensor temperature (17). Decay time is a well-characterized, intensity-dependent property of

the sensor material. This is a commercial device (Luxtron Corp.) capable of continuous operation in 60°C (140°F) surroundings.

Partial Discharges. Partial discharges are precursors to full discharge and breakdown of the insulation. If they are not stopped, gas generation leads to explosions and fires in high-voltage transformers. Partial discharges in transformer oil indicate aging caused by thermal degradation, moisture, and acidity.

A partial discharge in the depths of a transformer generates a radio-frequency (RF) signal and an ultrasonic (US) signal. A sensing system has been devised to detect paired RF and US signals: the RF detector responds first, followed by the US detector. The time interval between them depends on the distance from the partial discharge to the integrated detectors. Only a series of paired RF and US signals with the same time interval is accepted by the signal-processing circuitry as an indicator of partial discharge (18). The US sensors are specially fabricated from lead zirconate titanate (PZT) ceramic-epoxy composites that are tailored to have resonant frequencies in the range 200 to 400 kHz, which is above core magnetostriction spectra and below the AM broadcasting band. After suitable encapsulation, the PZT composite is mounted and integrated mechanically with the RF detector, a compact annular metal ring (18). This device is unaffected by contact with transformer oil and functions over the temperature range -10 to 120°C (14–248°F).

Voltage and Current Measurements. Knowledge about electric fields and currents in generators, transformers, and power lines is of obvious importance to the power industry, not only for understanding normal behavior of system elements but for identifying defective or malfunctioning apparatus. Optical sensors for voltage have in the past been based on electro-optical crystal transducers (e.g., bismuth germanate) that exhibit linear birefringence in electric fields (Pockels effect), which can be detected by a suitably polarized light beam (19). Pockels effect sensors have not yet been miniaturized, a step that would be required for distributed sensing.

Another approach to measuring electric fields is through electrostatic forces, which arise from electron rearrangements in conductors subject to external electric fields. Charges induced at conductor surfaces interact with the external field to produce weak electrostatic forces normal to the surface. By constructing a variable-gap Fabry-Perot microsensor as a conductive Faraday cage with a flexible silicon diaphragm (Fig. 6), these forces can be measured quite accurately (20). Dual-wavelength referencing makes the sensor system insensitive to bending and transmission losses in the fiber; electric field strength is thus related to change in the intensity *ratio* at two different wavelengths. dc fields from 0 to 500 kV/m and ac fields for energizing voltages up to 80 kV were successfully measured. This device is a microsensor suitable for emplacement at important locations throughout an electric-power system.

Almost all optical sensors for current rely on the Faraday effect, which is circular birefringence (i.e., a difference in refractive index for left and right circular

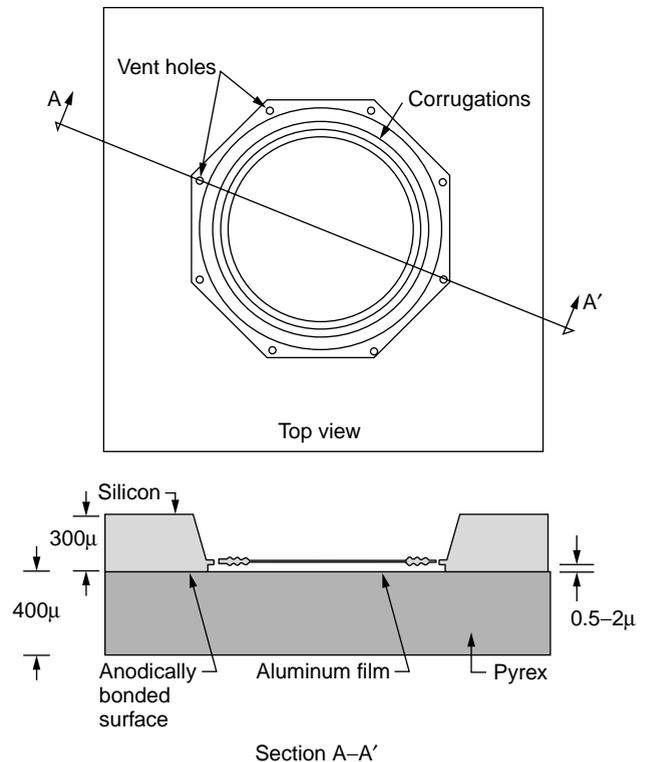


Figure 6. Fabry-Perot microcavity for measuring electric fields: (a) Top view and (b) cross section (20).

polarization) that is proportional to the magnetic field associated with current flow. Commercial fiber-optic current sensors (Square D Company) have been shown to be reliable at utility switchyards: metering accuracy is $\pm 0.2\%$ from 100 A to 1400 A and 0 to 40°C (32–104°F) temperature range (21). As with Pockels-effect sensors, Faraday-rotation transducers are still bulk optical devices.

Chemical Sensing. Ensuring compliance with environmental regulations that govern stack-gas emissions, waste streams, and process streams will require real-time in-situ, monitoring methods. Currently, samples are collected, stored, and transported from the field before they are analyzed in the laboratory. Thus, analyses are neither timely nor always reliable, since samples can change chemically between collection and laboratory. Fiber-optic chemical sensing is one remedy for this problem.

An optical chemical sensor consists of a chemically sensitive indicator and a physical transducer. The indicator interacts with the chemical species of interest (analyte) and undergoes a reversible change in absorbance, refractive index, polarization, and the like, that is a function of change in the analyte. The transducer converts this optical change into usable information. One embodiment of this technology is a fiber-optic hydrocarbon sensor (FiberChem, Inc.). Portions of the conventional (low refractive index) cladding are replaced by thin metal coatings with the desired selective affinity and refractive index. Typical responses to common hydrocarbons are shown in Fig. 7. Sensitivity is in the low ppb range and hydrocarbons can be detected

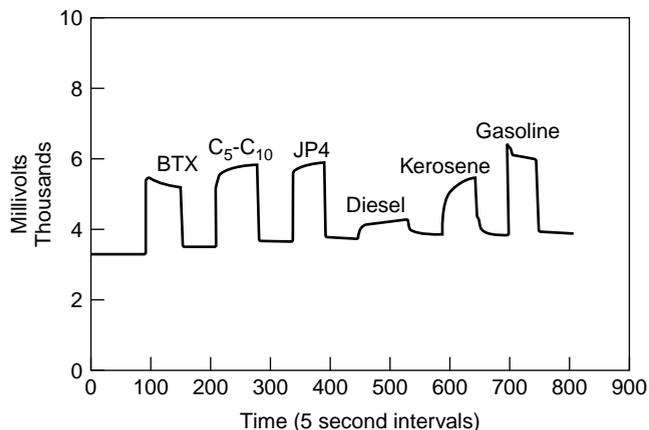


Figure 7. General response of fiber-optic chemical sensor to hydrocarbons.

and quantified in vapors, droplets, thin films on water, dissolved in water, or as water-hydrocarbon emulsions (22).

Structural Integrity of Dams. Dams for hydroelectric stations, whether concrete or rock filled, are very conservatively designed. Nevertheless, seismic events or other unforeseen geologic instabilities can threaten structural integrity. The consequences of failure are so severe that early warning of malfunction or collapse is vital. Embedded fiber-optic sensors are a logical answer.

Feasibility of this approach was demonstrated by installation of over 6.4 km (4 miles) of multiplexed simple and multifunction fiber sensors in a 7.5 MW dam during the construction phase (23). Over 90% of the embedded fibers survived placement, casting, pouring, and curing of the concrete, as well as framework removal and installation of generating machinery. Among the embedded fiber optics were sensors for simultaneous monitoring of water pressure and vibrations at structurally critical locations. Baseline vibration signatures were determined by statistical interpretation of speckle patterns (24). During “shake-down” of the facility, comparison of baseline vibration spectra with dynamic output of sensors within the powerhouse structure identified a vibration spike that was attributed to an out-of-round main gear in the power train. Replacement of the gear eliminated the troubling vibration.

Dams are a likely venue for a more complete smart system: self-repairing concrete. Dry (25) has developed several variations on the theme of injecting hardenable liquids into concrete. One that is completely autonomous involves embedding hollow glass fibers filled with adhesives in the concrete. Overloads on the concrete create microcracks that break the fibers, release adhesives into the cracks, and result in concrete as tough or tougher than the original. Fiber optics will still be needed to detect such occurrences and report their locations to a central authority.

Structural Behavior of Wind Turbine Blades. Windmills are large composite structures, usually located at remote sites, and subject to strong dynamic loads. In-service inspection involves dismantling the blades, a lengthy and

labor-intensive procedure, so windmills are overdesigned to compensate for infrequent inspections. Traditional sensors (e.g., strain gauges and accelerometers) are too bulky to be embedded for real-time monitoring, but fiber optics and piezoelectrics are not.

Thirteen fiber Bragg gratings were installed in a 12 m (39.4 ft) long wind turbine blade during initial layup of glass fiber reinforced polymer, and six piezoelectric micro-sensors were bonded to the blade surface (26). Good agreement up to at least 7000 microstrain was obtained between conventional and fiber Bragg grating sensors in controlled tests, with no fiber-optic failures. Modal frequencies were measured with high efficiency by the piezoelectric sensors. After six months in actual service, the sensors remained functional, with no measurable drift or necessity for recalibration.

SMART SENSOR-ACTUATORS

Several classes of smart materials can function as combined sensor-actuators: piezoelectrics, electrostrictives, magnetostrictives, and shape-memory alloys (SMAs). Eventually, smart materials in all of these categories will find roles in utility systems. Up to the present time, however, interest in applying smart sensor-actuators to power-industry concerns has been mostly focused on SMAs, owing in part to the large reversible deformations and large forces obtainable with SMAs, and in part to the capability of SMAs to perform actuations (shape changes) induced only by the thermal environment, that is, without supervision. In this regard, incorporation of SMAs in utility systems is a step toward autonomous responses, which can offset high-cost maintenance labor to some degree. Examples of applications that have been tested in prototype, mostly in Russia, Ukraine, and Scandinavia, follow.

Compression of Transformer Cores

Electrical coupling in large transformers is improved by compressing the core layers. During conventional manufacture, the core is compressed between the yokes by tightening the bolts, and the entire assembly is then subjected to alternate cycles of vacuum and kerosene spray at 125°C (257°F) to remove air from between the core sheets. As the core layers, compact, the compressive forces decrease, so that when the assembly is taken out of the vacuum chamber, some air reenters the core stack and optimum coupling is lost. Somehow, the maximum compressive pressure has to be maintained during the heating cycles. Making the yoke bolts out of SMA can, in principle, remedy the problem, as depicted in Fig. 8(a). Properly designed SMA bolts, axially prestrained, transform during heating in vacuum and maintain pressure as air is sucked out, as has been demonstrated in Russia (27). A difficulty with this solution stems from the large bolt sizes (50–75 mm [1.97–2.95 in] diameter), which makes homogeneity and heat treatment of the SMA problematic. An alternative has been devised in Sweden that involves conventional steel bolts with a series of precompressed SMA studs between the yoke and core package, Fig. 8(b). As temperature is raised in the

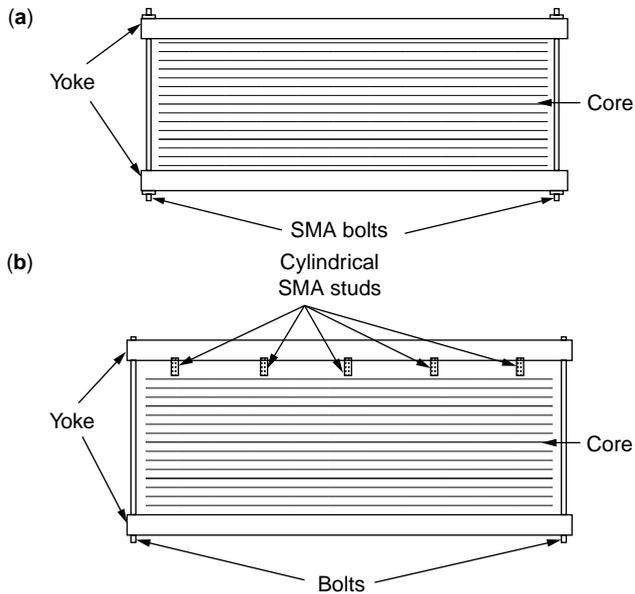


Figure 8. Transformer core compression with SMA elements (27): (a) Yoke bolts; (b) studs.

vacuum chamber, the studs expand axially to maintain well-distributed pressure (27).

Inducing Compressive Stresses in Pipes

The basic idea is to apply circumferential compressive forces on critical regions of piping to counteract tensile stresses caused by welding and service loads. One such arrangement consists of axially extended SMA wire wound on slotted sleeves. The sleeves are slipped on the pipe, positioned over at-risk circumferential welds, and heated above the transformation temperature of the SMA wire (27). Since the operating temperature of the pipe is also above the SMA transformation temperature, stress-corrosion cracking is forestalled and propagation of pre-existing cracks is inhibited by the resulting compressive stresses. In addition, pipe ends are prevented from separating should a break occur. This application, intended mainly for nuclear power plants, has been tested in Sweden.

Sag Control of Overhead Conductors

Aerial power lines sag when they are warmed by high electrical loads, especially in combination with high ambient temperatures. Performance is degraded when overhead lines sag to the extent that ground capacitance becomes excessive. This is a difficult problem, given the many thousands of miles of transmission lines. A solution that is being tested in Canada, Ukraine, Russia, and Japan has several embodiments (27,28); a typical one is to attach an SMA member electrically and mechanically in parallel with the overhead conductor. At low temperatures, the martensitic SMA is easily extended in tension. Upon heating, however, the SMA reverts to the strong austenitic phase and shortens, forming a loop in the overhead conductor and thereby maintaining acceptable ground clearance.

Galloping Control for Transmission Lines

Occasionally, under conditions of line icing combined with high wind speeds, transmission lines experience severe oscillations (galloping) that can be 9 m (29.5 ft) over spans of 185 to 365 m (607–1198 ft). Such occurrences often result in destruction of affected portions of the lines. Among the suggested approaches to galloping control are damping by vertical connectors made of SMA inserted between suspended conductors, and altering tower stiffness with suitably-positioned, externally energized SMA springs (29).

Mitigation of Erosion Caused by Cavitation and Liquid Droplets

Cavitation ensues when the local pressure in a fluid-flow field falls below a critical value that is a function of temperature, surface tension, vapor pressure, and external pressure. Vapor-filled cavities or bubbles form in the liquid at regions of low pressure and collapse as they move into regions of higher pressure. Bubble collapse is associated with very intense local impulses that can be destructive to nearby solid surfaces. Machinery subject to cavitation erosion and closely related liquid-droplet erosion include boiler feed pumps, valves, recirculation pumps in pressurized-water reactors, hydroturbine runners and guide vanes, and last-stage blades in steam turbines.

An investigation conducted for EPRI showed that on the basis of their anomalous resistance to low-cycle fatigue, near-equiatomic alloys of nickel and titanium (NiTi) are very resistant to both cavitation and liquid-droplet erosion. High resistance to fatigue and erosion is ascribable to the ability of NiTi, either as austenite or martensite, to deform reversibly without accumulating much residual damage in the constituent crystallites. Since building large machines entirely out of NiTi is impractical, thin plates of NiTi were explosively bonded to structural steel and the erosion resistance of the resultant clads was demonstrated (30), as indicated by Fig. 9.

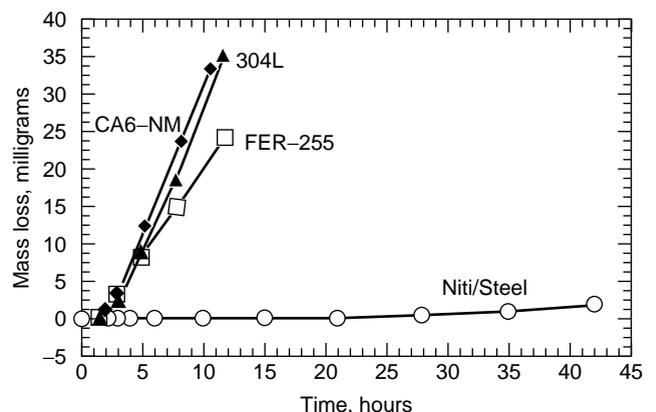


Figure 9. Comparison among cast austenitic steel (CA6NM), wrought stainless steel (304L), duplex stainless steel (FER-255), and NiTi cladding on steel in vibratory cavitation.

Table 5. Other Contemplated Uses of SMAs in the Electric Power Industry

Device	Function/Operational Principle	Reference
Couplings and connectors	SMA tubes are expanded “cold” to fit over members to be joined, then warmed to induce phase transformation and shape recovery. Heat-to-shrink devices apply high clamping forces for pipes and electrical connections.	(28,31,32)
Seals	Similar to couplings, circumferential stresses are induced by shape recovery of pre-strained rings. Hermetic seals are used for electronics and for attaching braid sheathing to terminal plugs.	(31)
Circuit breakers	Traditionally, electricity is activated by bimetal strips capable of only small forces and motions, SMA actuator, in series with locking mechanisms and a vacuum or SF6 chamber, provides much larger actuation that makes a cascade of latches unnecessary.	(28,31)
Fuses and fuse protectors	Similar to circuit breakers, SMA fuses are advantageous in that they can be reset. Quick-acting fuses for lightning protection can blow when heavy line usage causes a temperature rise. An SMA shunt, normally open, closes during heavy usage, thereby avoiding a premature blow.	(31,33)
Thermomarkers	SMA flags bend at a predetermined temperature and signal which phase has blown or which electrical joint is hot.	(28,31)
Switches and valves	Typically, SMA spring operates against a bias spring: under normal conditions the bias spring is stronger than the SMA spring; heat causes the SMA spring to transform, overcoming the bias spring, and operating the switch or valve.	(27,31,33)
Bolted electrical connections	SMA washers in Belleville configuration under bolt heads maintain contact pressure when thermally activated by increased contact resistance. Applied in transformer connectors, aluminum buses, and disconnects.	(27,28)
Thermal actuators for valves, clutches, dampers, and flues	Linear motion devices rely on shape recovery of prestrained SMA elements. They are initiated by externally controlled electrical heating of the SMA. Typically, bias springs reset these devices to their default positions.	(32,34)
Steam pipe hangers	Traditionally, a three-spring device with complicated mechanical linkages provides constant support under thermal conditions. Can be replaced by a single SMA rod or a stack of SMA Belleville washers.	(27)
Steam trap flush	When a trap fills with condensate, a bimetal-actuated mechanism opens the trap. The linear motion of the actuator allows leakage before full opening or closing. SMA Belleville washer is faster and nonlinear, so leakage does not occur.	(33)
Plugs in pipe ends and tubesheets	Pre-strained SMA plugs are inserted cold, then heated to form leak-proof seals.	(32)
De-icing of transmission lines	Externally energized SMA devices apply cyclic loads to overhead conductors to de-ice them.	(28)
Bolt breakers	Explosive bolts are used in safety mechanisms in nuclear power plants and petrochemical plants. They can be accidentally discharged under electrical fault conditions. Pre-strained SMA tubes can serve the same purpose by breaking notched bolts during shape recovery.	(31)

Other Applications of SMAs

A variety of concepts involving SMAs have been suggested as solutions to power-industry problems. Some that have been demonstrated under laboratory conditions or evaluated in engineering analyses are described briefly in Table 5.

CHALLENGES AWAITING SMART MATERIALS SOLUTIONS

The preceding two sections describe applications of smart sensors and smart sensor-actuators to power industry problems. Besides their favorable impact on utility operations, these applications are important steps in the

progression to smart *systems*. At the most sophisticated level, smart systems will feature highly integrated and embedded combinations of sensors, actuators, and processors, which will enable those systems to respond intelligently and *autonomously* to dynamically changing environments. Inasmuch as smart systems with those kinds of capabilities have not yet been implemented in the power industry, it is appropriate to consider the kinds of problems that might motivate development of such smart systems, either as adaptations from other industries, as new ways of organizing existing smart materials, or even as entirely new concepts.

Utility challenges that seem especially suited to solution by smart materials and systems are listed in Table 6, which is meant to be illustrative, not inclusive. Each problem is identified in the first column and a subjective ranking of severity is provided in the second column. The last column differentiates among innovations for which most of the constituents are at least in prototype (near-term), those that are only partially in prototype (midterm), those in early proof-of-concept stages (long-term), or intriguing concepts (vision) perhaps without solution by smart systems. Five of these challenges, chosen because they represent the whole spectrum of power-industry needs and possible solutions with smart materials, are discussed briefly in the balance of this section.

Real-Time Condition Assessment of Equipment

A topic of major concern for utilities is in-service deterioration of critical components (35). Detailed information about equipment condition is an essential part of life-management programs for generating units; such information also provides early warning of structural impairment, and therefore it is crucial for avoidance of forced outages. Obtaining that information is a serious challenge. For example, it has been estimated that about one-half of all forced outages in steam-generating plants are caused by corrosion. Whereas generalized corrosion is readily detected and measured, environmentally induced cracking (e.g., stress-corrosion cracking and corrosion fatigue) is hard to detect and respond to. Damage caused by creep and creep-fatigue is even more difficult to assess.

Another aspect of condition assessment is remaining-life analysis. Nowadays, many utilities opt to keep a component in service after the expiry of its "design life." In order to do so without jeopardizing personnel safety or system reliability, the following quantities must be determined (36): the amount of damage currently in the component, the rate that damage is accumulating, and the degree of damage that will cause failure. At least three factors inhibit those determinations. First, the amount of damage or size of the largest defect (depending on the damage type) may be hard to measure; where to look and what might have been missed are subsets of this difficulty. All traditional inspection technologies suffer from inaccessibility constraints, such as, detecting corrosion of reinforcing bars in concrete or finding creep-fatigue cracks inside steam-turbine rotors. Second, actual material properties can be hard to measure, especially without destroying the component. This is a principal reason for conservative

design practices such as factors of safety, code limits, and use of class minima. Such conservatisms are responsible for overdesign of structures to meet "worst-case" loadings, with attendant increases in equipment size and cost. Third, some damage mechanisms are not well understood or are not time dependent, and therefore they frustrate calculations of remaining life.

In recognition of the importance of this topic, millions of dollars have been spent on characterizing various damage mechanisms, developing models and confirming mitigation options, developing and testing the efficacy of new designs and materials, refining inspection methods, and specifying repair technologies. With such an extensive knowledge base available, improved condition assessment, remaining-life evaluations, self-diagnosis, and even self-repair should be fertile ground for innovations based on smart materials and systems.

Control of Power-Plant Cycle Chemistry and Atmospheric Emissions

The health of water-touched and steam-touched components in fossil-fueled or nuclear-fueled power plants depends critically on the purity of process fluids (37). Dissolved oxygen, chloride ions, and a host of other impurities must be controlled at parts-per-billion levels, which is a tremendous challenge, since a large steam generator circulates millions of kilograms (pounds) of water and steam per hour. Opportunities for impurity ingress, corrosion, and deposition are pervasive, as shown in Fig. 10 for a typical drum boiler cycle. Rapid response to cycle-chemistry upsets is crucial. Penalties for exceeding well-established limits on contaminants are severe: boiler tube failures, localized corrosion, cracked steam-turbine disks and blades, and increased erosion of turbine components by oxide particles spalled from steam-touched tubes and pipes. Any of these ills can cause forced outages of substantial duration. Clear needs exist for smart systems to address this challenge: distributed on-line chemical sensors capable of functioning in high-temperature water and steam, injectors of chemicals to counter pollutants or chemical imbalances, and processors to coordinate multifunctional sensors and actuators.

Similarly, there are sizable financial and environmental benefits from reducing pollutant releases to the atmosphere. Various configurations of sensors, processors, and actuators can be envisaged; an ideal system would be capable of analyzing fuel as it enters a combustor, changing the combustion conditions in real time to minimize formation of pollutant species, and selectively activating control devices that extract residual pollutants from the stack gas. Such systems would also calculate the effects of operational changes on equipment life and on overall unit performance and efficiency.

Systems Issues in Resonance Control

One particularly insidious form of vibration in rotating machinery is that caused by subsynchronous resonance (SSR), which results from a match between the natural frequencies of mechanical components in a generating unit and

Table 6. Power-Industry Challenges and Possible Solutions with Smart Materials

Power-Industry Challenge	Problem Size ^a	Applicable Smart Materials Technologies; Possible Approaches	Time Frame
<i>Generation</i>			
In-service deterioration of machinery and apparatus caused by cyclic loading and high temperatures; need for real-time condition assessments and decision-making tools in order to schedule orderly repairs or replacements.	1	Distributed, embedded sensors for detecting cracks, voids, deformations, and changes in material properties; integrated systems for self-diagnosis, on-line estimates of remaining life, and (eventually) self-repair.	Long-term to vision
Corrosion of power-plant components, particularly by mechanisms that frustrate conventional NDE methods.	1	Sensors that can detect corrosion-induced cracking and identify corrosive species; on-line analysis to locate damage and warn of incipient failure, as appropriate. Eventually, develop systems that release neutralizing chemicals at damage sites.	Midterm to vision
Overdesign of structures to resist hypothetical "worst-case" loads.	2	Detect onset of unusual loads and deploy auxiliary members that supplement load-bearing capability.	Long-term
Integrity of civil structures, especially concrete dams.	2	Determine state of cure and structural defects in fresh concrete; monitor loads and structural responses throughout life; actuate mitigation and repair strategies, e.g., "smart" concrete.	Near-term to midterm
Vibration of rotating machinery, leading to lower efficiencies and premature failure.	1	• Active vibration-control methods from aerospace practice to suppress vibration in individual machines.	Near-term
		• Systems issues, e.g., subsynchronous resonance, requires new technologies to detect onset and actuate adjustments to transmission system parameters.	Long-term
Efficiency of blades in steam turbines and wind turbines.	2	"Smart wing," "adaptive skins," and/or individual helicopter-blade control technologies to change airfoil shapes as a function of operation regime.	Long-term
Monitoring and controlling impurity levels in water and steam.	2	Requires chemical sensors, stable in high-purity water/steam at high temperatures and pressures, to measure pH and dissolved species; processors for coordinating signals and decision-making; actuators for injecting appropriate reagents to maintain or restore proper cycle chemistry.	Mid- to long-term
Optimizing combustion and control of emissions.	1	On-line analysis of fuel entering combustors (by x-ray fluorescence, neutron backscattering, etc.); high-temperature chemical sensors for combustion gases; real-time control of individual burners by integrating sensor signals and calculating combustion adjustments that balance heat rate minimization and stack-gas composition; actuate ammonia/urea injectors, as needed, to react with NO _x .	Long-term
Noise abatement in power plants and substations.	3	• Main source of noise in power plants is ventilation of auxiliary motor drives; new suppression approach needed.	Long-term
		• Noise cancellation by secondary acoustic technology for transformer hum.	Near-term
Protection of dams and nuclear plants during seismic events.	2	Flexible connections with smart dampers and selectively augment structures with smart tendons.	Vision

(cont.)

Table 6. (Continued)

Power-Industry Challenge	Problem Size ^a	Applicable Smart Materials Technologies; Possible Approaches	Time Frame
<i>Generation</i>			
Bolt integrity in critical components, e.g., steam chests and turbine casings; pipe supports in nuclear plants.	3	Fiber-optic sensors inside bolts could measure cyclic stresses (strains) and irreversible deformations, enabling on-line condition assessments or life expenditures.	Near-term
Fish kill in hydroelectric turbines.	2	A serious constraint on hydroelectric generation; no smart materials solution comes to mind.	Vision
Determining fluid levels in buried hazardous-waste tanks.	3	Adapt "smart fuel tank" technology from aerospace sector.	Near-term
Buckling of structural members loaded in compression.	4	Smart sensors detect incipient buckling; lateral stabilization then provided by piezoelectric actuators.	Long-term
Cavitation erosion in pumps and hydroturbines.	3	<ul style="list-style-type: none"> • Concepts from "smart wing" and "adaptive skins" technologies to effect dynamic changes in surface profiles of hydraulic machinery, thus preventing cavity formation by altering flow patterns in real time. 	Long-term
		<ul style="list-style-type: none"> • Implement erosion-resistant claddings with shape-memory alloys. 	Near-term
<i>Power Delivery</i>			
Deterioration of underground cable.	1	Multipurpose fiber-optic sensors could probably be adapted to detect and discriminate among void formation, water treeing, and corrosion of neutrals, i.e., self-diagnosis. Approaches for self-repair are unknown, although aspects of "smart concrete" (release of chemicals) could apply.	Long-term to vision
Real temperatures of high-voltage transmission lines.	2	Operating limits based on thermal capacity are determined by computer models incorporating air temperature, wind velocity, incident sunlight, etc. Knowledge about actual thermal conditions could result in up to 15% higher loading. The requisite fiber-optic technology already exists; embedding fiber optics in overhead conductors is the next step.	Near- to midterm
Condition assessment of transformers, especially life expenditures associated with temperature excursions and partial discharges.	1	Temperature monitoring by distributed, all-silica fiber optics based on Raman backscattering if spatial resolution can be improved to ≤ 0.1 m (4 in); monitoring oil composition with chemical sensors that measure moisture and acidity; partial discharge monitoring with combined ultrasonic (piezoelectric) and radio-frequency sensors.	Mid- to long-term
Condition assessment of wood poles.	3	Although they degrade by reaction with their surroundings, there is no generally accepted method for evaluating wood poles in service. Vibration signature analysis is a likely approach.	Midterm

(cont.)

Table 6. (Continued)

Power-Industry Challenge	Problem Size ^a	Applicable Smart Materials Technologies; Possible Approaches	Time Frame
<i>Power Delivery</i>			
Contamination of insulators on transmission lines.	4	Wind-blown dust and deposits from rain water accumulate on insulators and lead to arcing and system faults. Smart coatings that clean themselves, perhaps activated by sunlight, would solve this problem.	Long-term
Galloping of transmission lines.	3	Ice and wind conditions can induce severe oscillations and destroy line segments. Suggested solutions include damping with vertical shape-memory alloy members between suspended conductors or with SMA springs in the tower supports.	Midterm
Shielding against electromagnetic fields.	?		Vision

^aProblem size scale:

1. Extremely important, widespread throughout industry, costly for each occurrence, or a majority safety concern.
2. Very important, either widespread or very costly per occurrence; may be a major safety concern.
3. Modestly important.
4. Troublesome, but either relatively infrequent or not very costly per occurrence.

resonant electrical frequencies of the interconnected transmission system. When these frequencies match, there is the possibility for uncontrolled interchange of energy and reinforcement of mechanical and electrical perturbations. Where SSR has occurred it has resulted in unit-years of downtime (38).

The power-transfer capability of transmission lines is limited by either thermal characteristics or by electrical stability characteristics of the system. The stable power-transfer limit of transmission lines is given by

$$P = \frac{(V_l)(V_s) \sin \delta}{X}$$

where V_l and V_s are the load-end and source voltages, respectively, δ is the phase angle difference between the two voltages, and X is the reactance of the line. Inductive reactance (X) can be reduced by installing series-connected capacitors, and the transfer capability of the transmission line is thereby increased. Unfortunately, installing series-connected capacitors for compensation can create an electrical resonance circuit and thus establish conditions for SSR. SSR is a phenomenon so catastrophic that situations that might lead to its occurrence must be avoided. Avoidance starts with computer simulations that evaluate the chance of SSR for all likely combinations of transmission circuits and rotating machinery. If there is a possibility

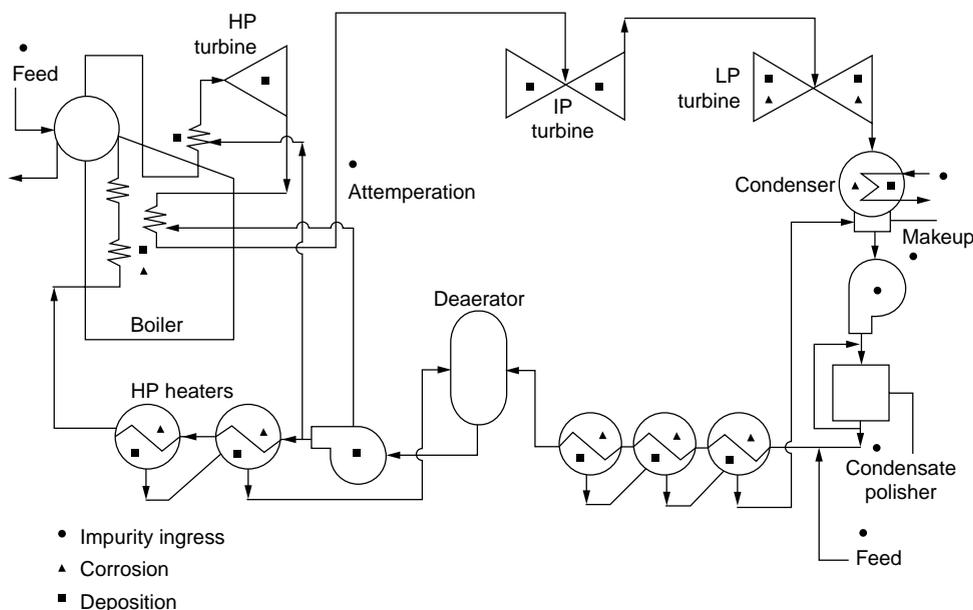


Figure 10. Major unit components and locations of impurity ingress, corrosion, and deposition in drum cycles (37).

for SSR, series-connected capacitors are reduced in size and electrical filters tuned to calculated subsynchronous frequencies are installed. These steps effectively de-tune the circuit to eliminate critical mechanical frequencies. They also reduce the transfer capability of these important power links.

Protection against SSR would seem to be an ideal application of smart material systems to deal with a complex mechanical-electrical control problem. It has been estimated that, under favorable circumstances, the transfer capability of a transmission line can be doubled by incorporating series compensation (39). A smart system for enabling that capability might consist of distributed sensors on all key mechanical and transmission line components to detect incipient SSR, an intelligent processor to analyze signals and coordinate responses, and appropriate actuation of switchable parallel filters and series capacitors.

Temperatures of High-Voltage Transmission Lines

One of the primary constraints on the amount of power that can be pushed down a transmission line is thermal capacity. Operating limits are set by computer models that incorporate factors such as air temperature, wind velocity, incident sunlight, and a derating factor to ensure safe operation. It has been estimated that 5% to 15% higher loading would be possible if actual thermal conditions were known on a continuous basis instead of relying on traditional loading guidelines (40).

Transmission capacity is insufficient in several areas of the United States. Significant benefit would accrue if more power could be transmitted without building new lines. The size of the benefits is estimated as follows: A typical high voltage (345–765 kV) transmission line has a capacity of 1000 MW. A 5% decrease in power equates to 50 MW. If the line is thermally limited 20% of the time, the amount of power lost is $50,000 \text{ kW} \times 8760 \text{ h/yr} \times 0.2 = 87.5 \times 10^6 \text{ kWh/yr}$. At a wholesale price of \$0.02/kWh, the increased revenue that would eventuate from knowing the actual temperature of a line is \$1,750,000 per year.

Temperature measurement in real time is the first element of a smart system for which fiber optics seems admirably suited (41). Huston and Fuhr (42) cite several additional opportunities for the application of fiber-optic systems to electric power transmission lines: detecting overheating, detecting galloping, flaw detection in lines, and control of circuit breakers along a broken transmission line where electromagnetic interference might make conventional control systems inoperable.

Deterioration of Underground Cable

There are now about one million miles of underground distribution lines in the United States, and there is growing pressure to put more of the distribution system underground. This presents serious challenges to utilities, related mostly to difficulty of access for detecting damage or for repairing it. As a consequence, cables are often operated until they fail, an undesirable situation that is becoming intolerable with increasing demand for electricity.

Cable materials, underground conditions, and system age are all factors in maintaining reliable service. Cable is

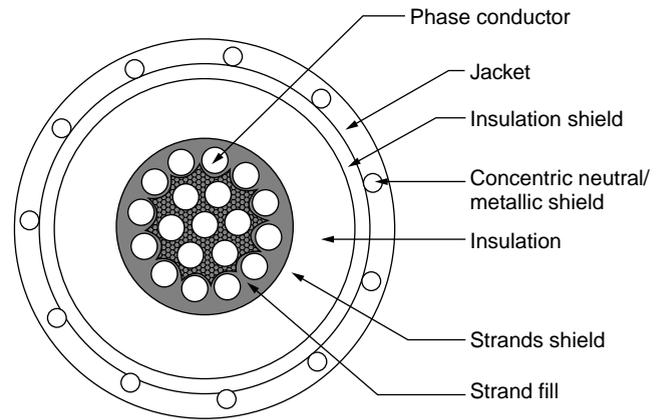


Figure 11. Typical construction of distribution-line cable.

typically composed of phase conductors embedded in concentric insulation (usually a polymer such as polyethylene) that is surrounded by return ground conductors (the neutral, usually copper), as shown in Fig. 11. The sensitivity of cable insulation to temperature was discussed earlier; knowledge of actual temperatures along a cable via a distributed fiber-optic sensor allowed power transfer through the cable to be optimized. Thermal conditions are not the only threat to cable integrity, however. Underground conditions almost always involve the presence of moisture, prolonged exposure to which causes insulation breakdown by void formation and “water treeing” or corrosion of the copper neutral. Water treeing is the development of small semiconductive paths that are believed to originate at particles or voids. Tree growth proceeds by partial discharges and ionization at these electrical stress concentrations. When a tree extends between two conductors at different potentials, the conductivity of the treeing path increases rapidly and culminates in a catastrophic fault (43). These problems are widespread, detection is difficult, and repair practices are unwieldy. Repair generally consists of injecting pressurized silicone into the dielectric in an attempt to drive out water. Routes to solutions are not obvious, but the benefits of self-analysis, and especially of self-repair, would be substantial for either the dielectric or the neutral.

BIBLIOGRAPHY

1. *Electric Power Trends*. Arthur Anderson & Co. and Cambridge Energy Research Associates, Cambridge MA, 1992.
2. *Statistical Yearbook of the Electric Utility Industry/1989*, Edison Electric Institute, Washington, DC, 1990.
3. J.C. Brown, P. Dahl, and J. Sparkman. *Rural Electric Sourcebook*, National Rural Electric Cooperative Association, Washington, DC, 1990.
4. R. Sachdeva. In T. Grandke and W.H. Ko, eds., *Sensors: A Comprehensive Survey, Vol. 1: Fundamentals and General Aspects*. VCH Publishers, New York, 1989, pp. 494–527.
5. P.J. Clelland. In *Proc. 1992 Workshop on Optical Sensing in Utility Applications*. EPRI TR-102349. Electric Power Research Institute, Palo Alto, CA, 1993, pp. 2-1–2-7.
6. *Workshop on Sensors and Sensor Systems for Power Systems and Other Dispersed Civil Infrastructure. Systems*, April 4–5,

1994. National Science Foundation, Directorate for Engineering, Arlington, VA, 1994.
7. R.L. Shepard and L.H. Thacker. *Evaluation of Pressure-Sensing Concepts: A Technology Assessment*. EPRI TR-102177. Electric Power Research Institute, Palo Alto, CA, 1993.
 8. J.W. Berthold. In *Proc. Optical Sensing in Utility Applications*. EPRI TR-100849. Electric Power Research Institute, Palo Alto, CA, 1992, pp. 9-1-9-7.
 9. *Advanced Sensors for Fossil-Fired Power Plant Applications*. EPRI AP-101840-V4P2. Electric Power Research Institute, Palo Alto, CA, 1996.
 10. N. Narandran. *Fiber-Optic Fabry-Perot High-Temperature Strain Measurement System Feasibility Study*. EPRI TR-107212. Electric Power Research Institute, Palo Alto, CA, 1997.
 11. W.W. Morey. *Development of Fiber Bragg Grating Sensors for Utility Applications*. EPRI TR-105190. Electric Power Research Institute, Palo Alto, CA, 1995.
 12. S.M. Angel and M.A. Northrup. *Stability of a Fiber Optic pH Sensor at 100°F*. EPRI TR-101972. Electric Power Research Institute, Palo Alto, CA, 1993.
 13. D. McCrae and E. Saaski. *Optical pH Sensors for High-Temperature Environments*. EPRI TR-106988. Electric Power Research Institute, Palo Alto, CA, 1997.
 14. A.A. Boiarsky and V.D. McGinniss. *Fiber Optic Distributed Temperature Sensor Demonstration*. EPRI TR-101950. Electric Power Research Institute, Palo Alto, CA, 1993.
 15. A.A. Boiarsky and J.P. Kurmer. *Fiber Optic Sensors for Temperature and Strain Monitoring in Motors and Generators*. EPRI TR-101950-V2. Electric Power Research Institute, Palo Alto, CA, 1997.
 16. *Distributed Fiber Optic Temperature Monitoring and Ampacity Analysis for XLPE Transmission Cables*. EPRI, Palo Alto, CA, and Southern California Edison Company, Rosemead, CA, 1998. TR-110630.
 17. K.A. Wickersheim. In *Proc. Optical Sensing in Utility Applications*. EPRI TR-100849, Electric Power Research Institute, Palo Alto, CA, 1992, pp. 4-1-4-16.
 18. J. Unsworth. In C.A. Rogers and G.G. Wallace, eds., *Proc. 2nd Int. Conf. on Intelligent Materials, ICIM '94*, 5-8 June 1994. Williamsburg, VA. Technomic Publishing, Lancaster, PA, 1994, pp. 23-32.
 19. F. Simonelli, S. Franchi Bononi, and P. Boero. In *Proc. Optical Sensing in Utility Applications*. EPRI TR-100849. Electric Power Research Institute, Palo Alto, CA, 1992, pp. 12-1-12-6.
 20. T.F. Morse and A. Mendez. *Fiber Optic Electric Field Micro Sensor*. EPRI TR-103395. Electric Power Research Institute, Palo Alto, CA, 1994.
 21. J. Beatty, T.J. Meyer, and E.A. Ulmer. In *Proc. Optical Sensing in Utility Applications*. EPRI TR-100849. Electric Power Research Institute, Palo Alto, CA, 1992, pp. 10-1-10-16.
 22. S.M. Klainer, J.R. Thomas, D.K. Dandge, K. Goswami, E. Hyman, and A. Jaworski. In *Proc. Optical Sensing in Utility Applications*. EPRI TR-100849. Electric Power Research Institute, Palo Alto, CA, 1992, pp. 15-1-15-12.
 23. D.R. Huston, P.L. Fuhr, T.P. Ambrose, and D.A. Barker. *Smart Mater. Struct.* **3**(2): 129-139 (1994).
 24. W.B. Spillman, Jr., B.R. Kline, L.B. Maurice, and P.L. Fuhr. *Appl. Opt.* **28**: 3166-3176 (1989).
 25. C.M. Dry. *Smart Mater. Struct.* **3**(2): 118-123 (1994).
 26. S. Diaz-Carrillo, C. Pardo, and A. Guemes. In J.H. Jacobs, ed., *Smart Structures and Materials 1999: Industrial and Commercial Applications of Smart Structures Technologies*. SPIE Proc., Vol. 3674. SPIE, Bellingham, WA, 1999, pp. 445-452.
 27. J. Cedarström. In L. McD. Schetky, ed., *Proc. Shape Memory Alloys for Power Systems*. EPRI TR-105072. Electric Power Research Institute, Palo Alto, CA, 1995, pp. 6-1-6-22.
 28. M. Braunovic. L. McD. Schetky, ed., *Proc. Shape Memory Alloys for Power Systems*. EPRI TR-105072. Electric Power Research Institute, Palo Alto, CA, 1995, pp. 7-1-7-12.
 29. A.V. Srinivasan. In L.Mc.D. Schetky, ed., *Proc. Shape Memory Alloys for Power Systems*. EPRI TR-105072. Electric Power Research Institute, Palo Alto, CA, 1995, pp.12-1-12-6.
 30. R.H. Richman and W.P. McNaughton. *J. Mater. Eng. Perform.* **6**(5): 633-641 (1997).
 31. L. McD. Schetky. In L. McD. Schetky, ed., *Proc. Shape Memory Alloys for Power Systems*. EPRI TR-105072. Electric Power Research Institute, Palo Alto, CA, 1995, pp. 4-1-4-11.
 32. R.R. Ionaitis, V.V. Kotov, and I.M. Schookin. L. McD. Schetky, ed., In *Proc. Shape Memory Alloys for Power Systems*. EPRI TR-105072. Electric Power Research Institute, Palo Alto, CA, 1995, pp. 10-1-10-5.
 33. W. Van Moorleghe. In L. McD. Schetky, ed., *Proc. Shape Memory Alloys for Power Systems*. EPRI TR-105072. Electric Power Research Institute, Palo Alto, CA, 1995, pp. 9-1-9-7.
 34. C. Whitehead. In L. McD. Schetky, ed., *Proc. Shape Memory Alloys for Power Systems*. EPRI TR-105072. Electric Power Research Institute, Palo Alto, CA, 1995, pp. 13-1-13-3.
 35. J.D. Parker, A. McMinn, R.H. Richman, W.P. McNaughton, R.J. Bell, J.P. Dimmer, and J.E. Damon. *Condition Assessment Guidelines for Fossil Fuel Power Plant Components*. EPRI GS-6724. Electric Power Research Institute, Palo Alto, CA, 1990.
 36. W.P. McNaughton, R.H. Richman, C.S. Pillar, and L.W. Perry. *Generic Guidelines for the Life Extension of Fossil Fuel Power Plants*. EPRI CS-4778. Electric Power Research Institute, Palo Alto, CA, 1986.
 37. R.B. Dooley, and W.P. McNaughton. *Boiler Tube Failures: Theory and Practice, Vol. 1*. Electric Power Research Institute, Palo Alto, CA, 1996.
 38. R.R. Richardson. In *Workshop Proc. Rotor Forging for Turbines and Generators*. EPRI WS-79-235. Electric Power Research Institute, Palo Alto, CA, 1981, pp. 1-40-1-45.
 39. D.G. Breuer, H.M. Rustebakke, R.A. Gibley, and H.O. Simmons, Jr. *Trans. IEEE Power App. Syst.* **83**(10): 1090-1102 (1964).
 40. D. Douglass. *Electric Light and Power* (Dec.): 25-33 (1994).
 41. B.J. Harrison and I.R. Funnell. In *11th Int. Conf. on Electricity Distribution, Vol. 1, Part 3, AIM-CIRED, Liege, Belgium, 1991*, pp. 1-6.
 42. D.R. Huston and P.L. Fuhr. In E. Udd, ed., *Fiber Optic Smart Structures*. Wiley, New York, 1995, pp. 647-665.
 43. G.J. Toman. In *Proc. Workshop on Power Plant Cable Condition Monitoring*. EPRI EL/NP/CS-5914-SR. Electric Power Research Institute, Palo Alto, CA, 1988, pp. 4-1-4-8.

SELF-DIAGNOSING OF DAMAGE IN CERAMICS AND LARGE-SCALE STRUCTURES

HIDEAKI MATSUBARA

YOSHIKI OKUHARA

HIROAKI YANAGIDA

Atsuta-ku, Nagoya, Japan

INTRODUCTION

The government is always faced with problems of maintaining municipal buildings and infrastructure facilities that support modern urban existence. Buildings, bridges, highways, tunnels, and other large or important structures need to be assessed objectively or scientifically to ensure their safety and long-term utilization. Today new systems and materials are being developed that can assist in such tasks. Currently most tested ideas and materials are too complicated or too expensive to be utilized in real situations. The concept behind the development of new systems is that materials should have the ability to evaluate their own condition after serious loading or relatively long service times. In particular, there are numerous strong demands for a means of monitoring in situ relative to how damage and fracture accumulate in large structures like buildings and tunnels. Until now there haven't been any good methods for judging objectively the amount of damage inflicted upon buildings during earthquakes or accidents in high-speed railway tunnels.

The study of smart or intelligent materials has grown enormously throughout the world. The means of introducing self-diagnosis or self-monitoring functions into modern materials have become various and wide-ranging. For example, piezoelectric ceramics, shape-memory alloys, and optical fibers can all be used to give materials a certain degree of intelligence. In heavy construction, however, where large volumes of material are required, it is essential to avoid high complexity and high cost. The use of electrical conductivity or resistance is perhaps the most straightforward way of obtaining a relatively simple material system as compared with other high-tech materials.

Yanagida, Muto, and co-workers have reported fracture detection using fiber-reinforced plastics (FRP) embedded in structural materials as a reinforcement (1). The electrical resistance change of carbon-fiber glass-fiber reinforced plastics (CFGFRP) indicates the local or partial damage in a composite prior to its fatal fracture. Although the CFGFRP showed appreciable resistance change in the strain range above 0.7% to 1.5% (due to fracture of the carbon fiber), the detectable strain level is too large for diagnosing local damage in structural materials such as concrete (where typical strains are about 0.05%). It is therefore necessary to develop a self-diagnosis material with higher sensitivity in the low strain region if it is possible to achieve a simple health monitoring system.

In our recent work (2–6), composites containing a continuous structure of electrically conductive particles, or a so-called percolation structure, have been designed and fabricated. We have successfully produced the FRP composites and ceramic matrix composites (CMC) with high performances in the low-strain region caused by the “percolation structure” of conductive particles in a matrix phase. These self-diagnosis functions were evaluated from the changes in resistance with applied strain. It was found that the FRP composites with a self-diagnosis function are fairly useful for the measurement of damage and fracture in concrete blocks and piles.

SELF-DIAGNOSIS FUNCTION OF FIBER-REINFORCED COMPOSITE WITH CONDUCTIVE PARTICLES

Electrically conductive composites having a continuous structure of conductive particles, a so-called percolation structure, were designed in an FRP and ceramic matrix composite (CMC). The diagnosis of micro cracks in concrete structures is one of the objectives in using the self-diagnosis FRP. The FRP has proved to be highly sensitivity to small strain levels and at the same time provide the high strength to reinforce concrete structures. CMCs have been attracting attention because of their necessary toughness, which can significantly enhance the reliability of ceramic materials. Hence the research into the self-diagnosis function in CMCs is important for the achievement of higher reliability. This article considers the design of the conductive phases in these composites and investigates their self-diagnosis function during their deformation. These self-diagnosis functions are evaluated by measurements of their resistance changes with applied strain in normal or cyclic loading tests.

Design and Fabrication of Specimens

Figure 1 shows the structural design for two conductive FRPs in the shape of rods, which basically consist of vinyl ester resin (Showa High Polymer Co., Ltd. RIPOXY R-804), and glass fiber (Asahi glass fiber Co., Ltd. ER2220). The resin is hardened at room temperature by adding a curing agent (NOF Co., Ltd. PERCURE VL). In the first composite, the carbon fiber (pitch-based CF, Toho Rayon Co., Ltd. BESFIGHT UM63) was introduced by removing some of the glass fiber. This formed a conductive path and enhanced the strength of the composite in the longitudinal direction. The second composite obtained its conductivity by the formation of a percolation structure consisting of the carbon particles (graphite, SEC Co., Ltd. SPG5) dispersed in part of the matrix. The volume fraction of carbon particles in the conductive phase is 20%. The size of the carbon particle, whose shape is a flake, is 5 μm , on average. The composite containing carbon particles was composed of carbon particles and glass-fiber reinforced plastics (CPGFRP).

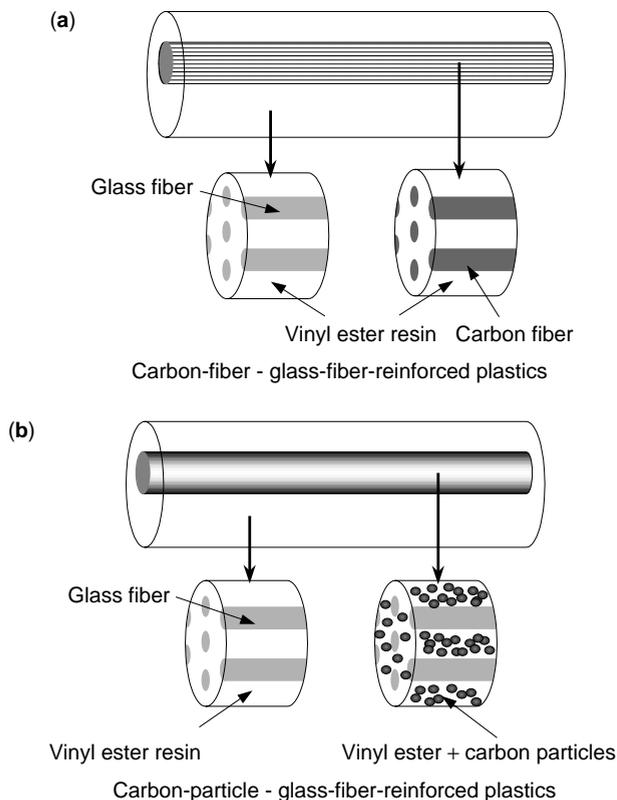


Figure 1. Schematics of the structural design for CFGFRP (a) and CPGFRP (b) in the shape of rod.

The CFGFRP and CPGFRP consisting of unidirectional reinforced fiber have a diameter of 3 mm.

The schematic structural designs for the CMC are shown in Fig. 2. The composites were fabricated by the filament winding method using Si_3N_4 particles (Ube Industries Co., Ltd. SN-COA) as the matrix and SiC fiber (Nippon Carbon Co., Ltd. NL-401) as the reinforcement for strengthening or toughening the composite. A portion of the fibers was replaced with tungsten wire (Nippon tungsten Co., Ltd. ϕ 30 μm). The conductive particles of TiN (Japan New Metals Co., Ltd.) were dispersed in part of the Si_3N_4 matrix. The volume fraction of the conductive phase, which includes 40 vol% of TiN particles, was 0.13%. These conductive phases were formed near the surface (500 μm in

depth) which was the tensile surface in the bending tests. These composites were hot-pressed under 40 MPa at 1773 K in N_2 atmosphere for one hour. The sintered specimens were cut into $3 \times 4 \times 45$ mm bars for bending test pieces.

Self-Diagnosis Function of FRP

Figure 3 presents two scanning electron micrographs of a polished transverse section and of a longitudinal section of CPGFRP (2). The circles in Fig. 3(a) and the white lines in Fig. 3(b) denote glass fibers. The bright gray flakes are the dispersed carbon particles. Note that the carbon particles are sufficiently dispersed in the matrix and that the matrix is well impregnated between glass fibers. This means that a percolation structure consisting of conductive particles has been successfully achieved.

The self-diagnosis functions of these materials were evaluated through simultaneous measurements of stress and electrical resistance change as a function of applied strain in tensile loading tests. The resistance change was defined as relative change in resistance $(R - R_0)/R_0$, indicated by $\Delta R/R_0$ in which R_0 denotes initial resistance before loading. The two types of loading selected were (1) a normal tensile test until specimen fracture and (2) a cyclic loading-unloading test below the maximum stress level. Figure 4 shows the electrical resistance changes and the applied stress for CFGFRP and CPGFRP as a function of the applied strain in the tensile tests. The stresses in both specimens were increased linearly in proportion to the strains until fracture occurred of the carbon fiber or the glass fiber. The CFGFRP indicates a slight change in resistance below a 0.6% strain due to the elongation of carbon fiber and shows a tremendous change around 0.7% strain owing to the fracture of the conductive fiber; namely the resistance of CFGFRP exhibits a nonlinear response to the applied strain as shown in Fig. 4(a). The initial resistance R_0 for CPGFRP was higher than that for CFGFRP because of a slight electrical contact between carbon particles in the percolation structure. As can be seen from Fig. 4(b), the CPGFRP indicates a linear increase in resistance with increasing tensile strain. The response of the resistance to applied strain appears at 0.01% strain (100 μ strain) or lower. The linear increase in the resistance continues until the fracture of the composite. Comparing Fig. 4(a) with (b) illustrates CPGFRP's higher sensitivity at the small strain level and the wider detectable strain range compared to CFGFRP. These results mean that the percolation

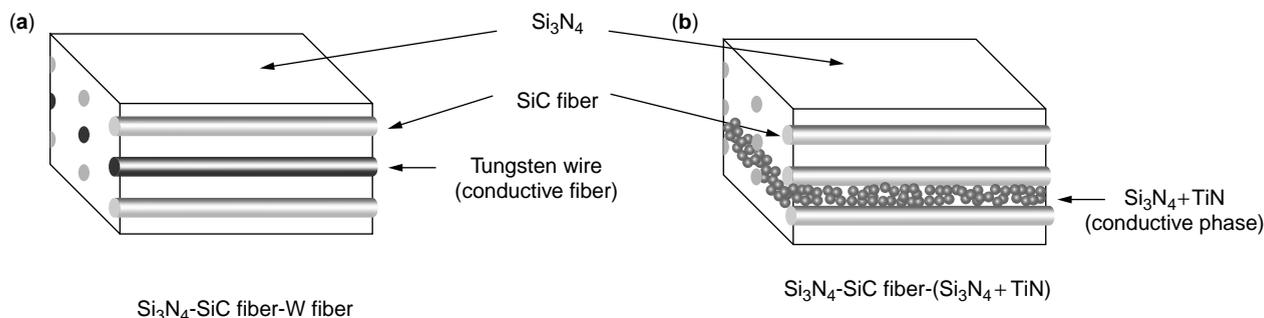


Figure 2. Structural design for CMC containing tungsten wire (a) and TiN particles (b).

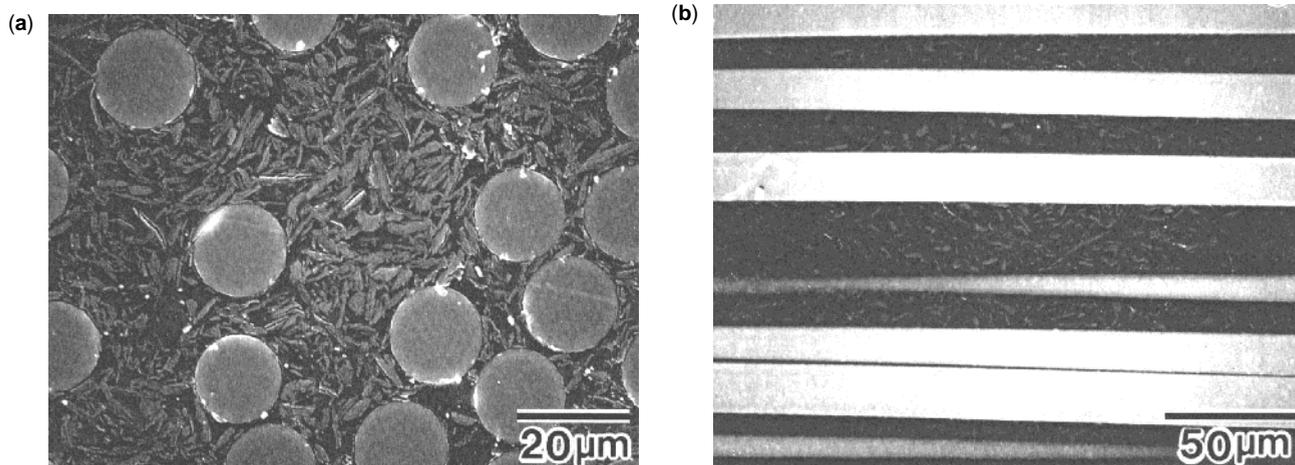


Figure 3. SEM photographs of polished transverse section (a) and longitudinal section (b) of CPGFRP with unidirectional glass fiber.

structure formed with the carbon particle enables more sensitive and adaptable diagnosis of damage than the structure consisting of carbon fiber. The strong response of resistance for CPGFRP was attributed to a local break in electrical contact between carbon particles because of the micro crack formation in the matrix or in the rearrangement in the percolation structure under tensile stress. It should be noted that the dispersion of the carbon particles had no effect on the strength of the composite, since the fracture stress and mode for CPGFRP were similar to those of GFRP without carbon particles.

Figure 5 shows the change of resistance to the applied strain as a function of time in the cyclic loading tests for CFGFRP and CPGFRP. These FRP were loaded and unloaded cyclically under a gradual increase in stress. The resistance of CFGFRP showed poor response below 0.6% strain and a drastic increase above 0.7% strain as shown in Fig. 5(a). From Fig. 5(b), it can be seen that the change in resistance of CPGFRP corresponded well with strain fluctuation (3). It is noteworthy that the resistance decreased

but did not completely return to zero at the unloading state. The residual resistance in CPGFRP appeared after the application of 0.2% strain, and then increased with the increase to the maximum applied strain. The maximum resistance during loading, indicated by ΔR_{\max} , and the residual resistance change after unloading, denoted by ΔR_{res} , were arranged according to the maximum strain applied in the past as shown in Fig. 6. The residual resistance of CFGFRP appeared around the 0.4% strain and increased discontinuously above 0.6%. The appearance of residual resistance for CFGFRP owing to fracture of the carbon fiber was limited in a narrow strain range. The change in residual resistance of CPGFRP correlated closely with previous maximum strain over the wide strain range as shown in Fig. 6(b), suggesting that the CPGFRP has the ability to diagnose the maximum strain based on measurements of past residual resistance at an unloading state (3). A comparison of Fig. 6(a) and (b) shows that the CPGFRP performs a more useful diagnostic function of damage history over the wide strain range than does the CFGFRP.

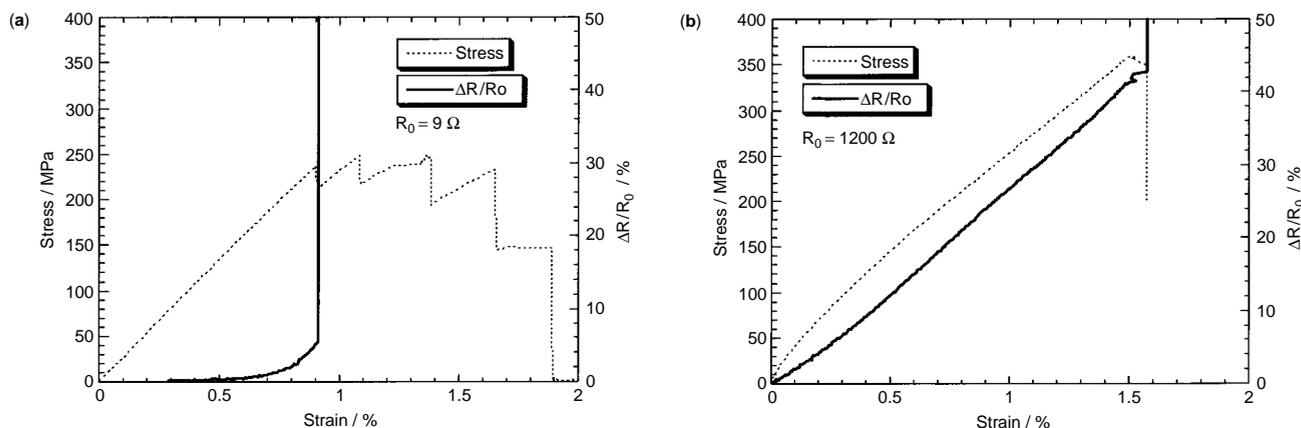


Figure 4. Changes in electrical resistance (solid line) and applied stress (dashed line) as a function of applied strain in tensile tests for CFGFRP (a) and CPGFRP (b).

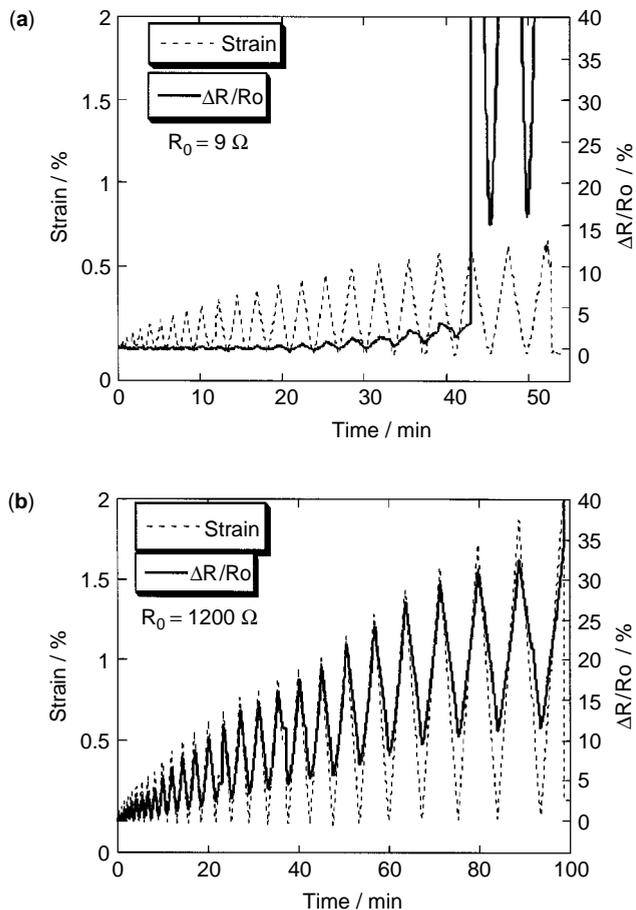


Figure 5. Change in resistance (solid line) and applied strain (dashed line) as a function of time in cyclic loading test for the CFGFRP (a) and CPGFRP (b).

The microstructure of CPGFRP after the loading–unloading cycle induced 0.6% strain and 2.1% strain was observed by scanning electron microscopy (SEM) as shown in Fig. 7 (2). Clearly, the number of micro cracks in the matrix increased with the increase in applied strain. Although the elongation of CPGFRP affected the elasticity after unloading, the percolation structure did not return reversibly

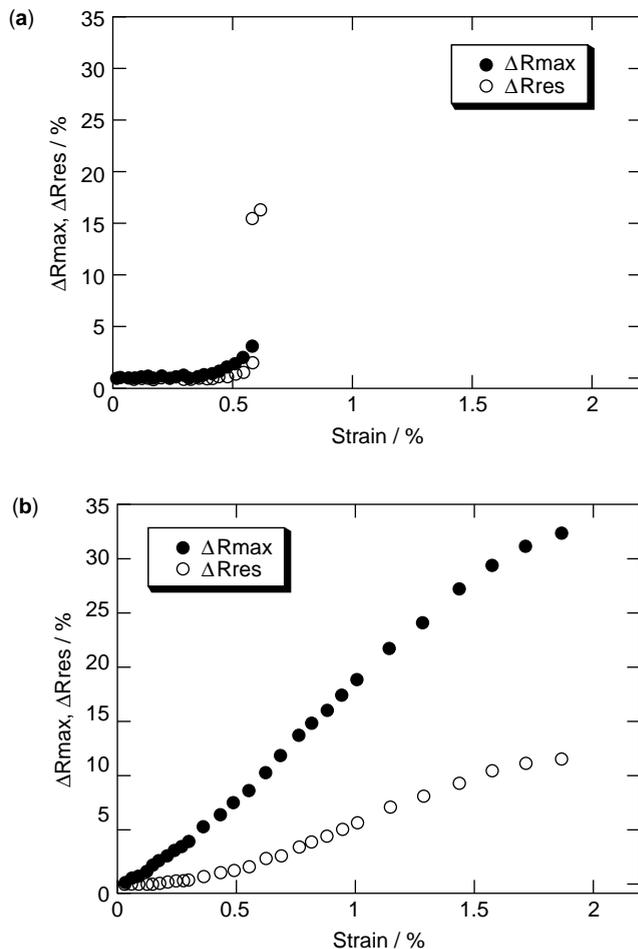


Figure 6. Maximum resistance change at loading state and residual resistance change at unloading state as a function of applied strain in cyclic loading tests for the CFGFRP (a) and CPGFRP (b).

to the initial state because of the micro crack formation in the matrix. The irreversible change in the percolation structure in the conductive phase was partly responsible for the appearance of obvious residual resistance over a wide strain range.

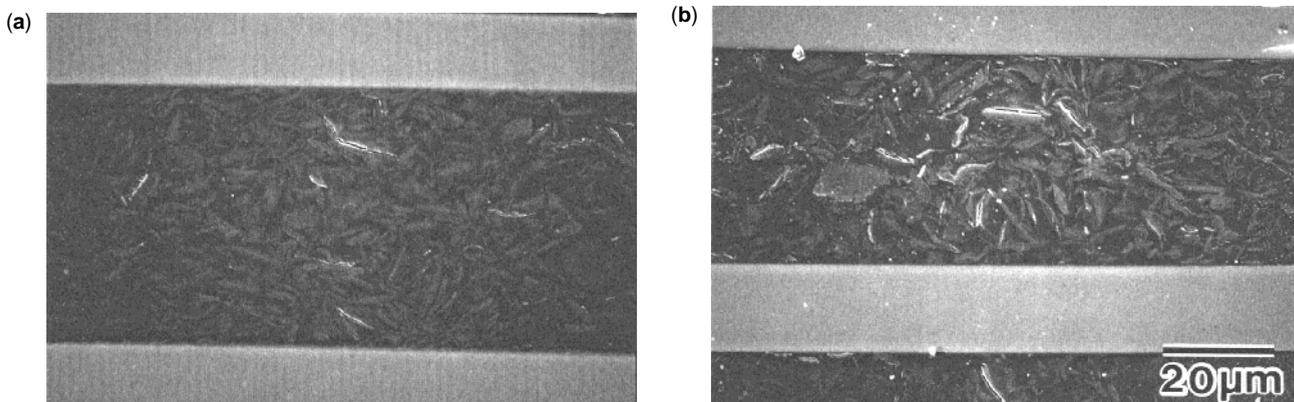


Figure 7. Scanning electron micrographs of CPGFRP after removing 0.6% strain (a) and 2.1% strain (b).

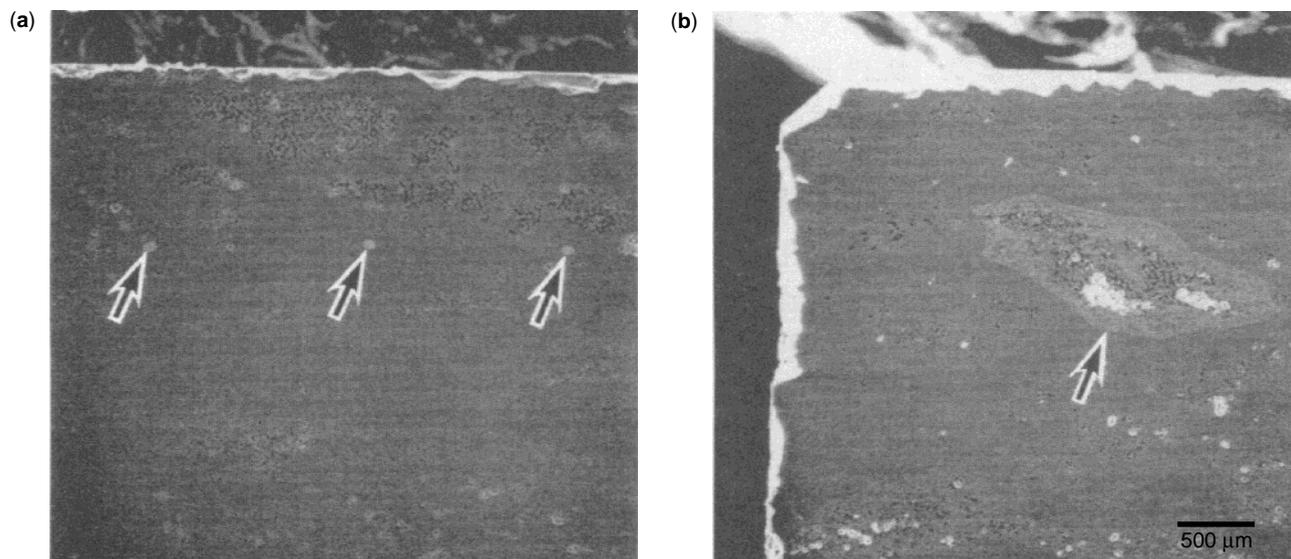


Figure 8. SEM cross sections of polished CMC specimens. The arrows point to the tungsten wire (a) or to the area containing TiN particles (b).

Self-Diagnosis Function of CMC

The conductive phases in the CMC observed by SEM are shown in Fig. 8 (4). Three tungsten wires were embedded near the tensile surface. The conductive phase containing dispersed TiN particles and SiC fibers was observed as the bright area. Some voids (white bareas) appeared in the conductive phase; however, these defects were thought to be insignificant for the damage diagnosis function because the amount was negligible. The interface between these conductive phases and the Si_3N_4 matrix did not show a remarkable reaction and exhibited good adhesiveness.

The self-diagnosis functions of the CMC were evaluated by simultaneous measurements of stress and electrical resistance change ΔR as a function of applied strain in four-point bending tests. The loading was performed two ways: (1) a normal bending test until specimen fracture and (2) cyclic loading-unloading tests below the maximum stress level. The dependence of the applied load and change in resistance on displacement for the CMC is shown in Fig. 9 (4).

Similar fracture behavior peculiar to CMCs was observed in both composites in which a part of the ultimate load was kept after fracture at a displacement of about 0.1 mm. The peculiar load-displacement curve explained from the extraction of SiC fibers from the Si_3N_4 matrix is shown in Fig. 10. The difference in the ultimate load and in the load-displacement curve for both composites was thought to be due to the uneven quality of SiC- Si_3N_4 phase, and not to the difference in conductive phase. The nonlinear response of resistance changes to displacement was exhibited in both composites. The CMC with tungsten wire showed a slight change in resistance in a small deformation, and then a drastic change was accompanied by their own fracture as shown in Fig. 9(a). The CMC containing TiN particles exhibited a distinct change in resistance from a small displacement to the fracture in the composite as shown in Fig. 9(b). These results suggest that the monitoring of resistance for CMCs with percolation structures is advantageous for diagnosing damages to the composites.

Figure 11 shows the hysteresis of resistance change in loading-unloading bending tests under the ultimate

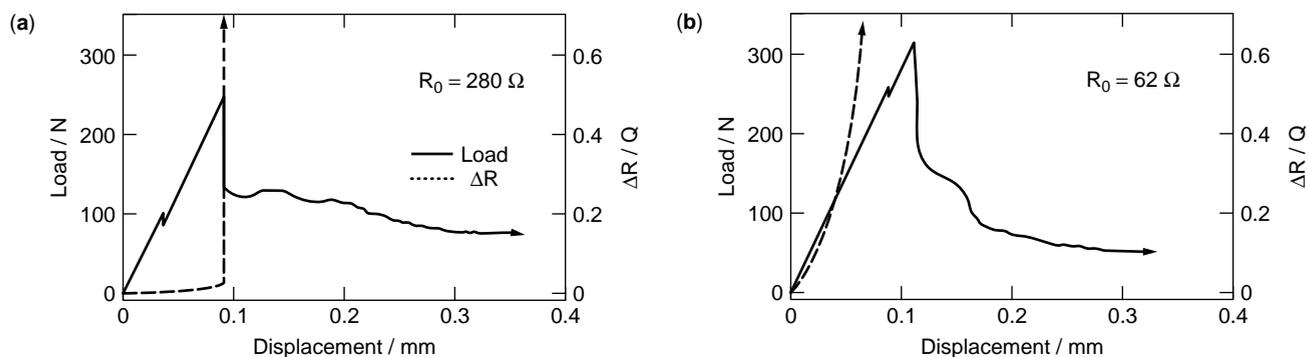


Figure 9. Change in load and resistance as a function of displacement in the four-point bending tests for the CMC containing tungsten wire (a) or TiN particles (b).

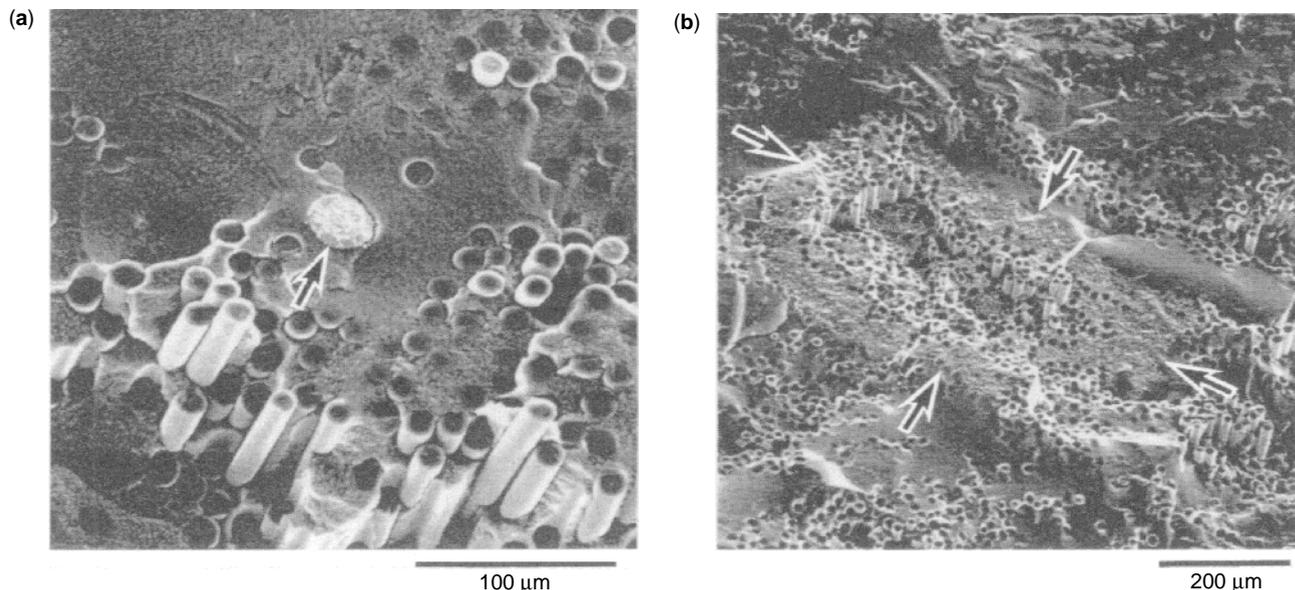


Figure 10. SEM images of fractured surface for the CMC specimens containing tungsten wire (a) or TiN particles (b).

load (4). The resistance of CMCs containing tungsten wire showed no change at the loading and unloading state, which was expected from Fig. 9(a). The applied load of some 50% of the ultimate load induced the increase in the resistance for the CMCs containing TiN particles, and then the increased resistance remained at about 80% of the maximum resistance after unloading. It should be noted that the loading–unloading cycle induced elastic deformation for the CMCs without residual strain. Hence, the residual resistance was thought to be due to irreversible local fracture in the conductive phase. The residual phenomenon in resistance change for the CMCs was more remarkable than that for FRP shown in Fig. 5(b), which was attributed to the brittleness of the ceramic in the matrix.

Figure 12 presents an attempt at repeatedly varying the resistance for the CMC with tungsten wire or TiN

particles in cyclic bending test. The applied load was, however, kept constant at 150 kN. The residual resistance for the CMCs with tungsten wire indicated no change, while that for the composites containing TiN particles after unloading rapidly increased up to 10 cycles. It should be noted that the residual resistance proportionally increased with an increasing number of repetitions after 20 cycles. The linear response of residual resistance was thought to be attributed to the propagation of micro cracking in the conductive Si_3N_4 –TiN phase. This result further confirms that the CMCs containing TiN particles have the ability to diagnose cumulative damage to the composite through measurements of the residual resistance.

The electrical conductive FRP and CMC were designed and produced by adding a conductive fiber or

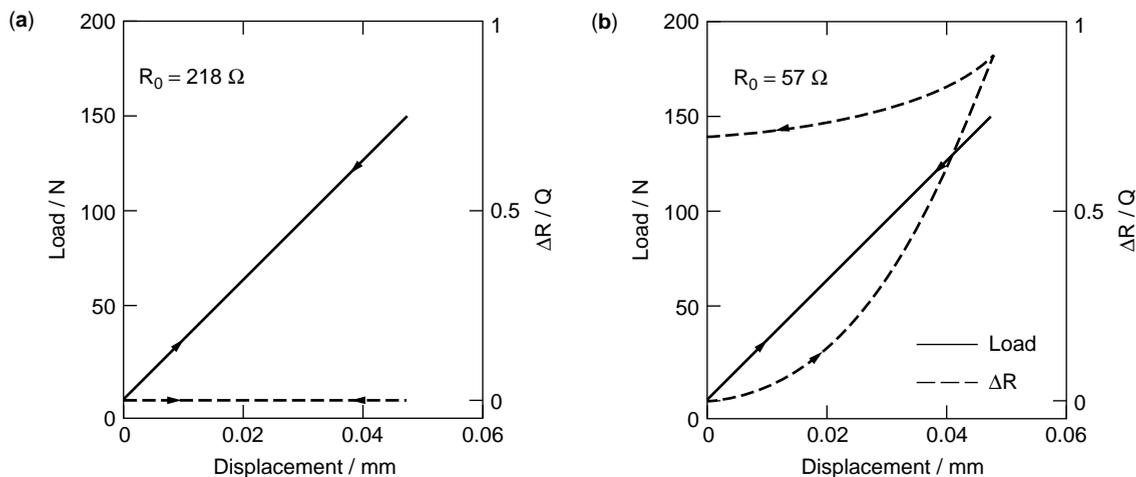


Figure 11. Change in load and resistance as a function of displacement in the loading–unloading tests for the CMC containing tungsten wire (a) or TiN particles (b). The applied maximum load was 150 kN.

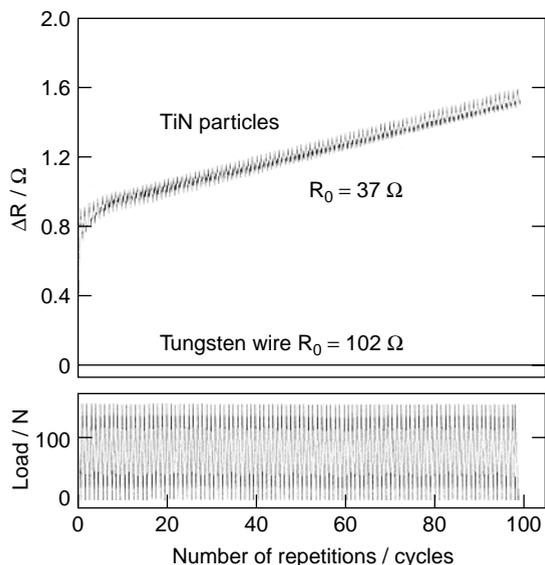


Figure 12. Change in resistance as a function of number of repetitions in the cyclic bending tests for the CMC containing tungsten wire and TiN particles.

particles, and the self-diagnosis functions for these conductive composites were investigated. Compared with the composites that include conductive fiber or wire the composites with the percolation structure consisting of conductive particles were found to be capable of diagnosing deformation or damage in the composites. The composites containing carbon particles appeared capable of diagnosing damage at the sensitivity level of a small strain and in a detectable strain range. Concerning the detectable strain level, the FRP showed an excellent response to the resistance change and to the applied strain. This is a suitable range for the health monitoring of structural materials such as concrete construction. It was also found through measurements of the residual resistance that the FRP composites are capable of memorizing the maximum applied strain or stress. The CMCs with percolation structures consisting of TiN particles exhibited superior resistance to small deformation changes. It should finally be noted that the CMC materials proved capable of diagnosing cumulative damage for the composites by evaluating the residual resistance, and that these self-diagnosis functions are easily obtained by simple measurements of electrical resistance.

APPLICATION OF THE SELF-DIAGNOSIS COMPOSITE TO CONCRETE STRUCTURES

A new type composite was developed that had a self-diagnosis function for health monitoring and damage detection in materials (1–7). The composite, which has electrical conductivity as well as reinforced fibers, provides a signal of electrical resistance change corresponding to the degree of damage in the material. This self-diagnosis composite offers also some advantages in properties, cost, and simplicity, compared with other materials or systems such as the optical fiber and the strain gauge. A concrete

structure is the best application for the self-diagnosis composite because the composite has a good sensitivity to micro cracking in concrete materials, shows high strength in reinforcing concrete material, and provides ease both in its attachment and in the measurement of electrical conductivity. The study was aimed at determining whether the composite was useful for measuring damage and fracture in concrete blocks and piles. Particularly, the application into concrete construction limiting the direct observation of damage or fracture after a serious load has been applied in its utilization. Also investigated, by bending tests and electrical resistance measurements were the function and performance of the composites when embedded in mortar/concrete blocks and concrete piles.

Specimen and Experiment

Two kinds of glass-fiber reinforced plastics composites were fabricated in this study. The first composite included carbon fibers substituted for some of the glass fibers; its electrical conductivity was called CF. The second composite involved carbon powders dispersed in a part of the plastic that formed the percolation structures as a conductive path (CP). The CF and CP composites were embedded into mortar specimens and concrete specimens reinforced by steel bars or rods by the following procedures. Figure 13(a–c) shows the structure and arrangement of the composites in the three concrete specimens types. The first type is a rectangular mortar block specimen with the CP composites. The second type is a rectangular concrete block specimen with the CP and CF composites and two steel bars. The third type is a concrete pile specimen having the CP composites and 16 steel bars. The pile type specimens have been pre-stressed at 14.3 MPa applied by the tension stress of the steel bars, while the block type was free from pre-stress.

Figure 14 illustrates the methods used for bending tests for the block and the pile type specimens with different lengths and distances. The electrical resistance change ($\Delta R/R_0$, where ΔR is an increase of resistance and R_0 is an initial resistance) of the composites was measured simultaneously in the loading tests. The strain gauge measurement attached on the tension-side surface of the specimen was also used. Photographs the actual bending tests for the block and the pile specimens are shown in Fig. 15.

Mortar Block Tests

The CP composite was embedded in the tensile side of the mortar specimens in order to demonstrate the self-diagnosis function. Figure 16 shows the applied load and resistance change of the CP composite as a function of displacement in a bending test. The embedded CP composite was located 8 mm apart from the tensile surface of the mortar. The load–displacement curve indicates discontinuous changes at points A and B, which correspond to the crack formation and propagation in the mortar specimen, respectively. The crack formation and propagation are shown in photographs of the mortar specimen. The resistance of the CP composite begins to increase

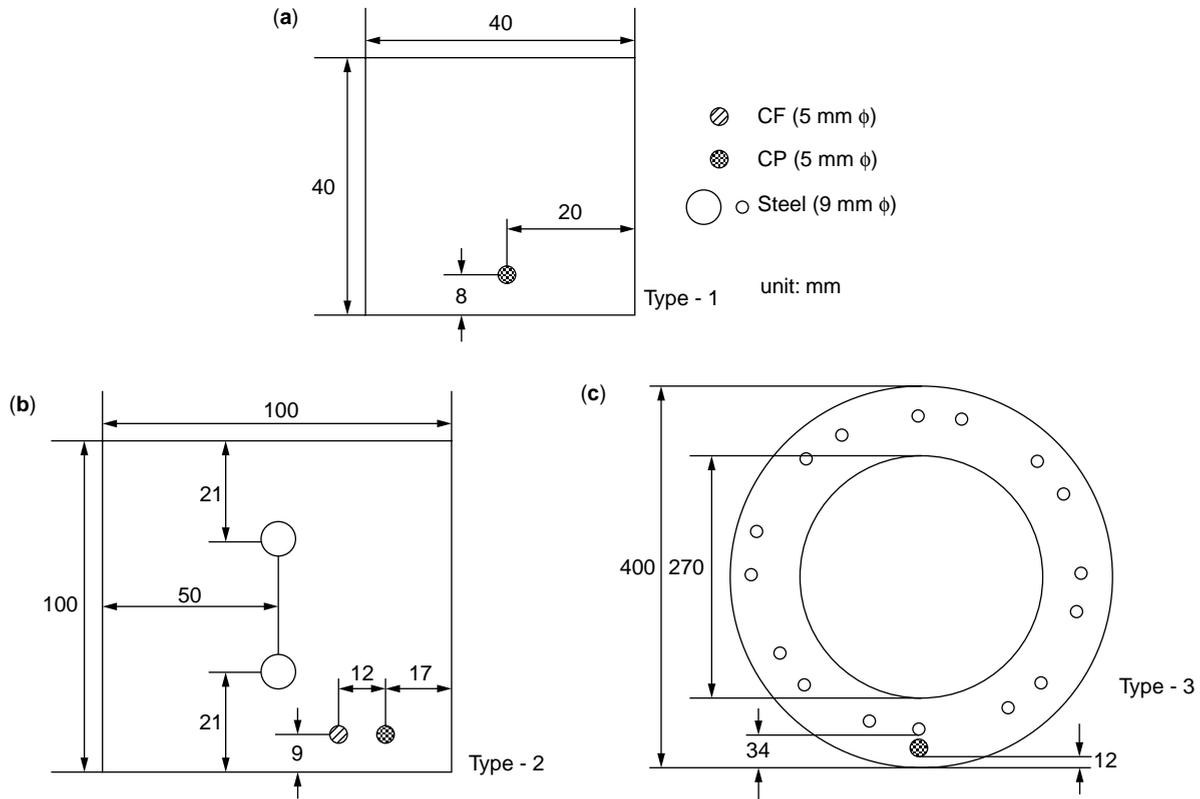
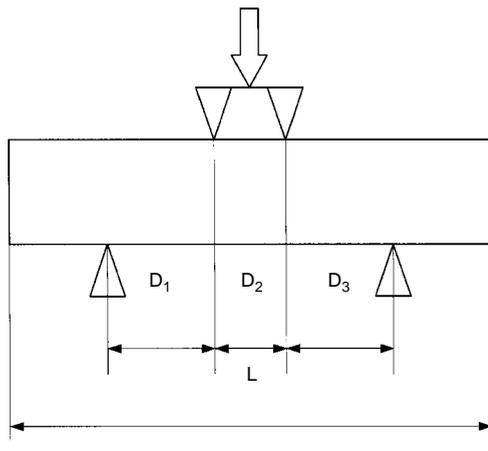


Figure 13. Structure and arrangement of the composites in the three types of concrete specimens. (a) Type-1, a rectangular mortar block specimen with the CP composite. (b) Type-2, a rectangular concrete block specimen with the CP and CF composites. (c) Type-3, a concrete pile specimen with CP composite.



Type	Bending test	L / mm	D ₁ / mm	D ₂ / mm	D ₃ / mm
1	3 points	160	50	-	50
2	4 points	400	100	100	100
3	4 points	8000	3000	1000	3000

Figure 14. Different bending tests for the block-and-pile type specimens with different length and distances corresponding to type-1, type-2, and type-3.

slightly before crack formation. Note that the increase in resistance appears simultaneously with the micro crack formation and that a discontinuous resistance change is generated in response to the crack propagation. The residual resistance was observed in the FRP material after unloading at point *D*. The resistance change of embedded CP composite corresponds well to the propagation of damage inflicted on the mortar specimen. Once again, the results demonstrate that the embedded CP composite has the ability to diagnose micro crack formation/propagation and loading history in cement-based structural materials.

The behavior of residual resistance for the CP composite embedded in a mortar specimen was investigated in detail by cyclic bending tests. Figure 17 presents the hysteresis of resistance changes by cyclic loading–unloading tests under 40% of ultimate load. The application of load caused micro crack formation, and then the crack was closed at an unloading state as shown in Fig. 17. It should be noted that the crack was eliminated, but the behavior of the micro crack induced residual resistance after unloading. The application of higher load (60% of ultimate load) made higher residual resistance after unloading. These results suggested that the CP composite embedded in the mortar specimen has the ability to diagnose the closed micro crack, namely the hysteresis of micro crack formation by evaluation of the residual resistance even after the crack has closed.

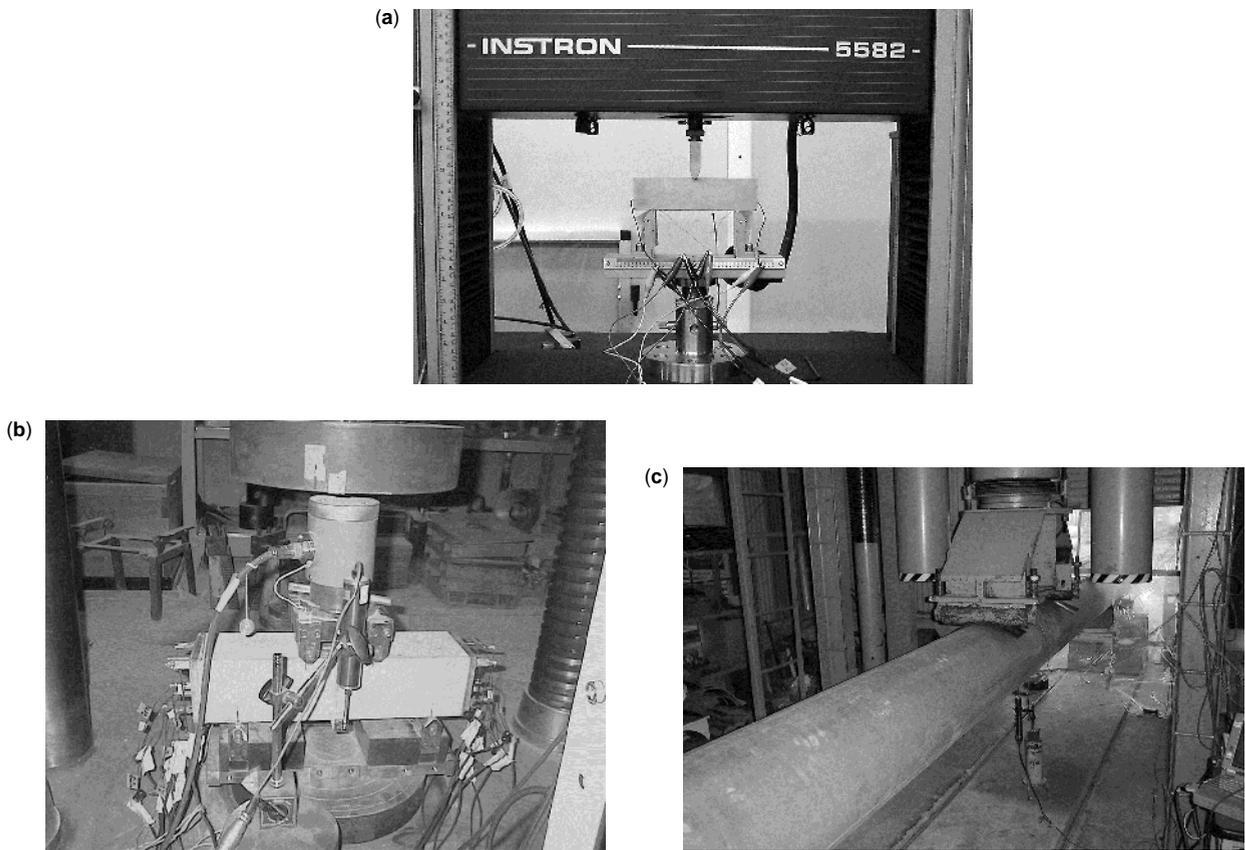


Figure 15. Bending tests in progress for the mortar block specimens (a), the concrete block specimens (b), and the pile specimens (c).

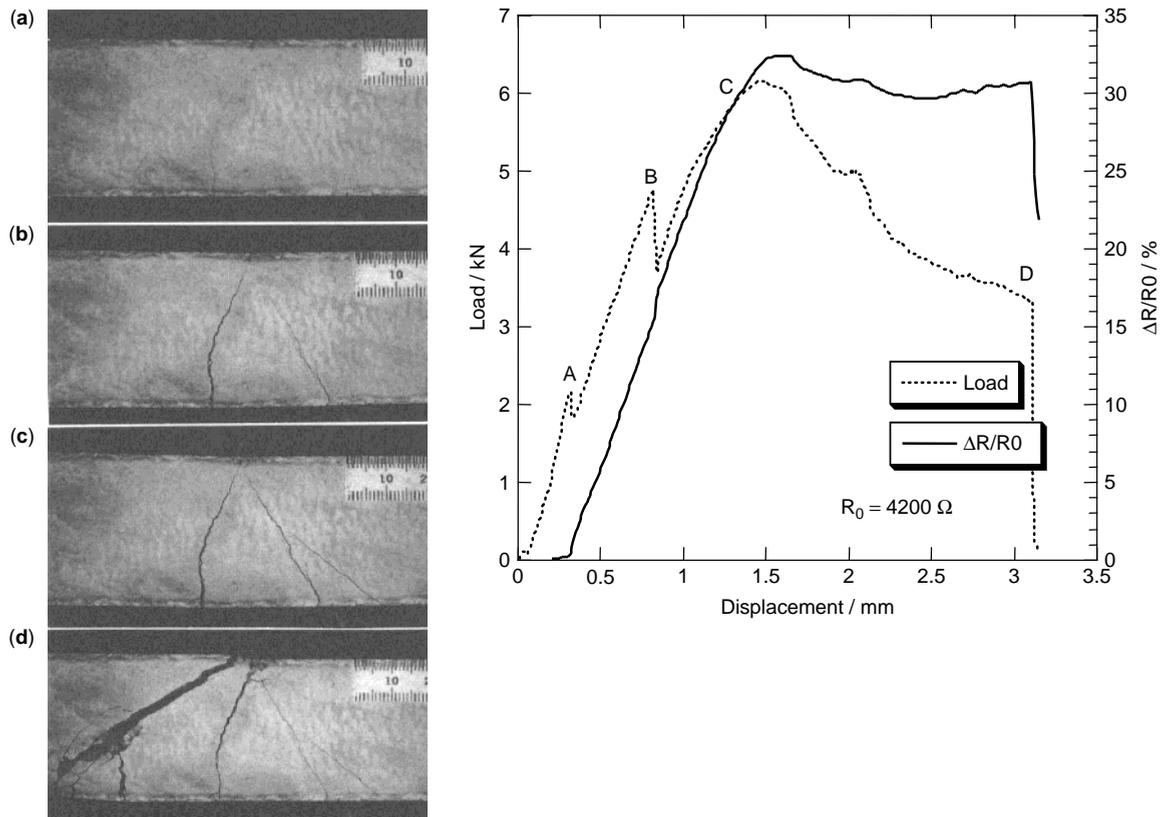


Figure 16. Changes in resistance (solid line) and applied load (dashed line) in a bending test for CPGFRP rod embedded in mortar specimen. These points (A–D) on the graph correspond to the photographs of the mortar specimen.

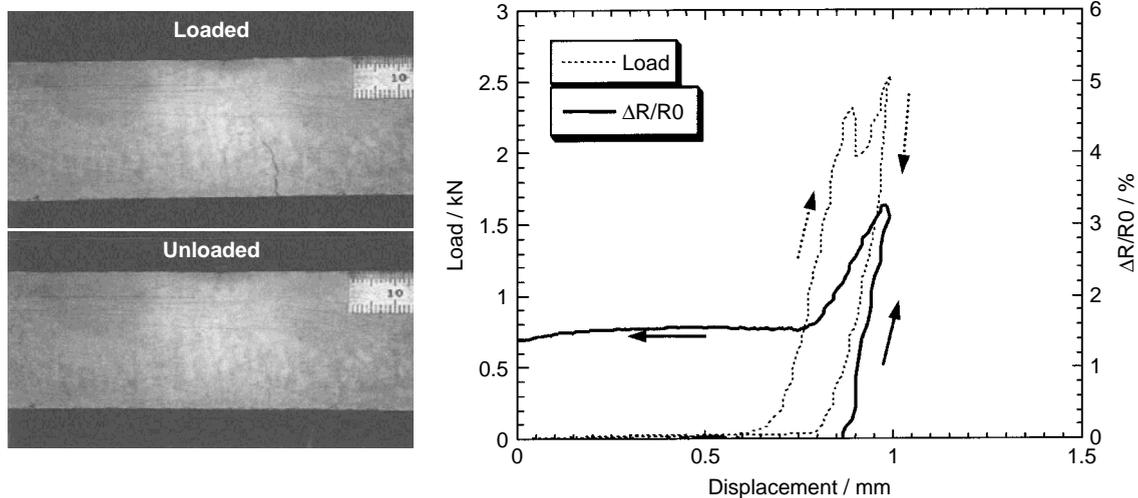


Figure 17. Changes in resistance (dashed line) and applied load (solid line) in the cyclic loading–unloading tests, under 40% of the ultimate load.

Concrete Block Tests

Figure 18 shows the results of load, strain, and $\Delta R/R_0$ of CP and CF composites as a function of time in the bending test for the concrete block (6). The strain change, which followed closely the loading curve, indicates that a micro crack formed at about 200μ strain and the steel bars yielded at about 1000μ strain. The strain gauge was broken in the loading test owing to the crack propagation in the surface of the concrete specimen. The $\Delta R/R_0$ of the CP composite is initiated at about 300 s, which corresponds to the stage of crack formation. The $\Delta R/R_0$ of the CP composite increased with an increased load up to the maximum load at about 1000 s. The $\Delta R/R_0$ of CF is scarcely detected until the high load level when it increases suddenly near the maximum load. Both the CP and CF composites do not break in the test because of their high strength and flexibility. It should be noted that the CP composite shows good sensitivity in the small strain range as well as a continuous response in the wide strain range up to the final fracture of the specimen.

Figure 19 provides the results of a cyclic loading test for the block-type specimen (6). In all, eight cycles of loading and unloading with an increased load level were carried out in this test. The strain change and the $\Delta R/R_0$ of CP composite responded well to the load curve from a lower load level, while the $\Delta R/R_0$ of CF did not act until a higher load was applied. It was also found that the CP composite's residual resistance appeared only after the cycles of the medium load level.

The block specimen is shown in Fig. 20 (a–c) as it appeared in the cyclic bending test (6). The cracks are clearly initiated from the tension-side surface at a low load, and they grow with an increased load level until the specimen finally breaks owing to steel bar fracture.

Concrete Pile Tests

Figure 21 gives the results of the cyclic bending test for the type-3 concrete pile specimen (6). The specimen included

only the CP composite because of the sensitivity it showed under a small load, which was higher than that for the CF composite as confirmed in Figs. 18 and 19. This test aimed to increase the sensitivity of the CP composite, which is arranged near the tension-side surface of the pile specimen. Figure 21(d) is the result from the enlarged $\Delta R/R_0$ axis of the CP composite in Fig. 21(c). The $\Delta R/R_0$ of the CP composite in the pile responds well in a wide range of loading as shown in Fig. 21(c). The CP composite located near the tension-side surface of the pile specimen indicates good sensitivity in the lower load levels as shown in Fig. 21(d). The $\Delta R/R_0$ of the CP composite in the lower load range is very similar to the strain change in Fig. 21(b), which means that the CP composite can signal a smaller strain before the crack forms in the pile surface. In these pile tests there is no clear indication of residual resistance phenomena as detected in the block tests, probably because of the effect of pre-stress in the pile specimens.

The appearances of the pile specimen in the cyclic bending test are shown in Fig. 22(a–c) (6). The crack forms at a low load, its growth occurs with an increased load, and finally the pile fractures after the test has ended.

Performance of the Self-diagnosis Composites

In the bending tests of the concrete block, the CP composite produced good results compared to the CF composite. Remarkably, the electrical resistance of the CP composite increased under a small strain to detect a micro crack formation at about 200μ , and it responded well to small deformations before the crack formation. The CP composite showed continuous resistance change up to a large strain level near the final fracture of the concrete structures reinforced by steel bars. It was also found that the CP composites embedded in mortar/cement block specimens have the ability to diagnose the hysteresis of micro crack formation by the evaluation of the residual resistance even after unloading.

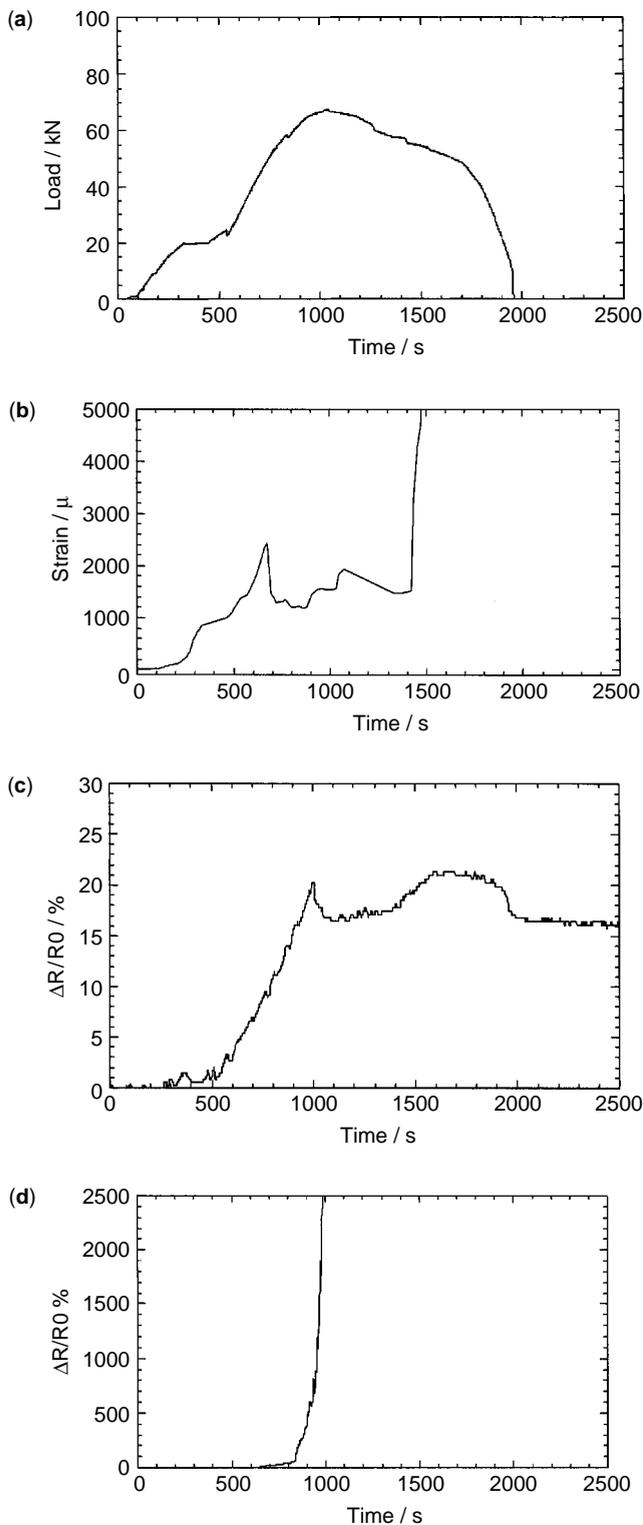


Figure 18. Load (a), strain (b), and $\Delta R/R_0$ (c, d) of CP and CF composites as a function of time in the bending test for the block-type specimen.

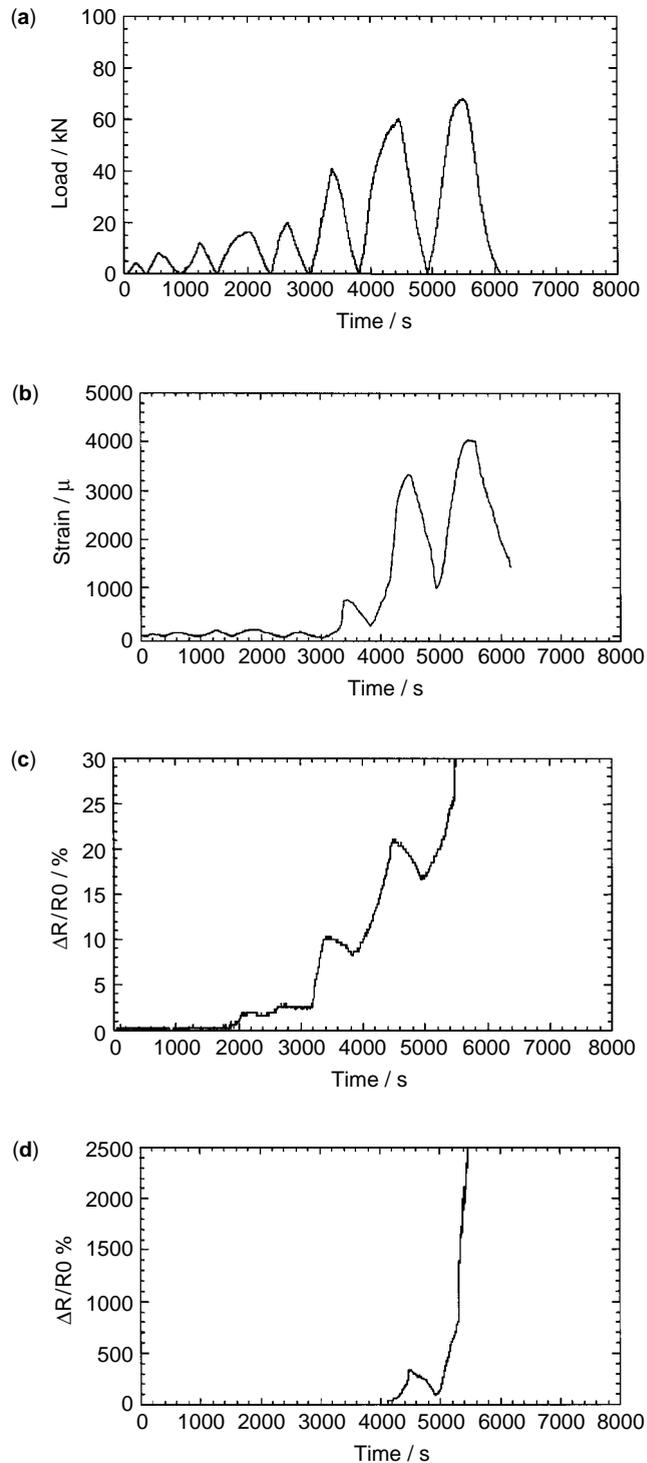


Figure 19. Load (a), strain (b), and $\Delta R/R_0$ (c, d) of CP and CF composites in the cyclic loading test for of the block type specimen.

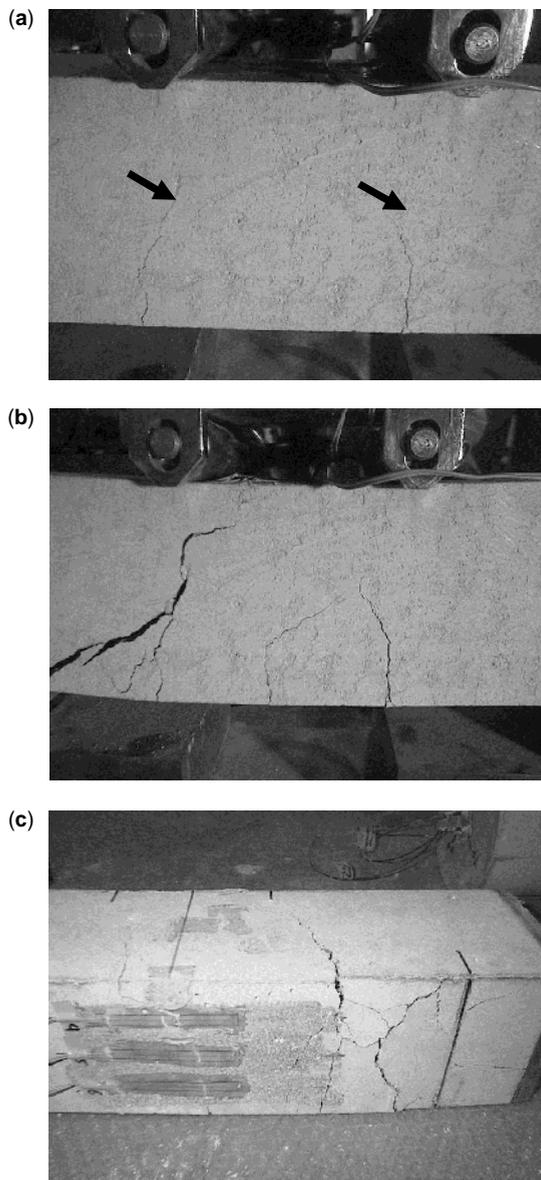


Figure 20. Appearances of the block specimen during the cyclic bending test (a) A low load level; (b) a high load level; (c) after the test.

Such excellent properties can be attributed to the percolation structure of the carbon particles dispersed within a section of the plastic matrix phase. The conductive path in the percolation structure of carbon particles, which is very different from the conductive path in carbon continuous fibers, can react to small strains that are lower than 200μ . This may be due to its flexible structure which is filled with faint gaps and cracks as seen in the microstructures of the carbon particles mixed with plastics. The phenomenon of the residual electrical resistance at the unloading state suggests that the distorted structure at the loading state does not completely return to its original shape at the unloading state. The residual resistance phenomenon has a possibility for the hysteresis function of an applied load.

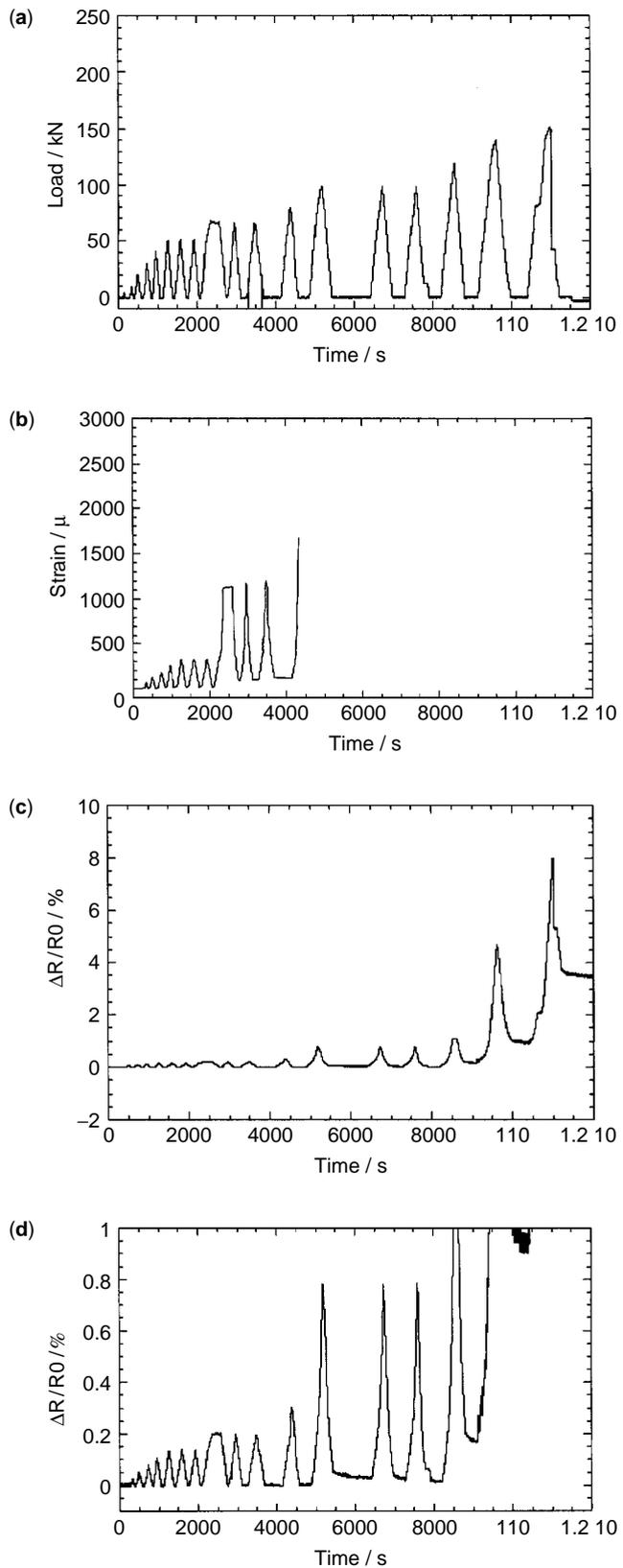


Figure 21. Load (a), strain (b) and $\Delta R/R_0$ (c, d) of CP composite in.

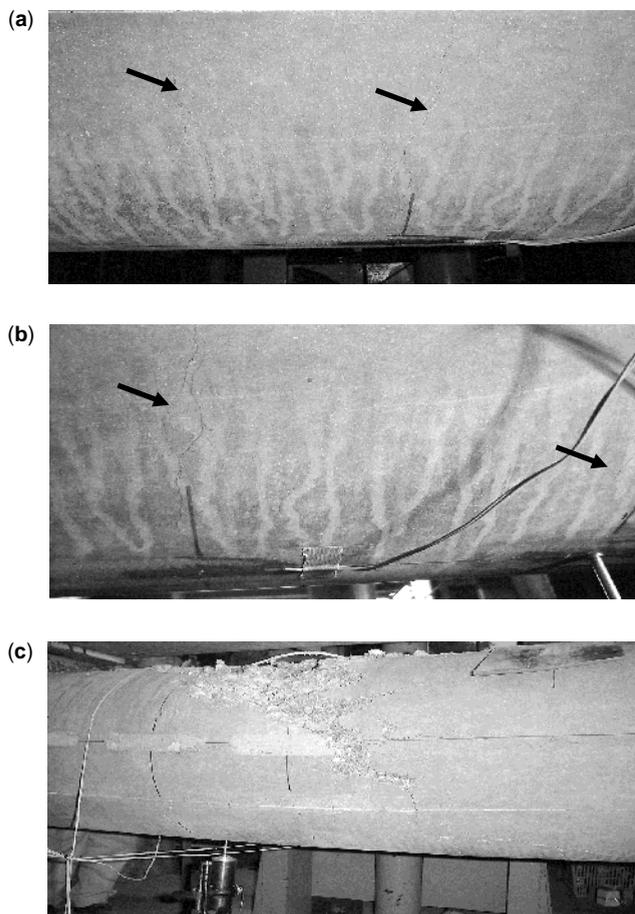


Figure 22. Appearances of the pile specimen in the cyclic bending test of type-3 (a) A low load level; (b) a high load level; (c) after the test.

The continuous change of resistance in the CP composite contributes to the damage detection of concrete structures. The percolation structure in the fiber-reinforced structure can keep its flexible structure up to the final fracture.

It is necessary to arrange the CP composite in concrete specimens to optimize the function. The location near a tension-side surface and far from steel bars is effective in order to obtain a quick response to applied stress and crack formation. The existence of prestress (compression) in concrete structures can dull the sensitivity of the composite. For the CP composite near the tension-side surface in the pile specimen, its clear sensitivity proves that the performance of the composite overcomes the influence of prestress.

Two kinds of glass-fiber reinforced plastic composites with carbon powder (CP) or carbon-fiber (CF) were introduced into the mortar/concrete specimens, with block and concrete pile types and electrical resistance change ($\Delta R/R_0$) of the composites being measured in the bending tests. The $\Delta R/R_0$ of the CP composite in the block specimen showed a good sensitivity in a small strain range to detect crack formation in the mortar/concrete and a continuous change in a large strain range up to the final fracture of the specimen, while the $\Delta R/R_0$ of the CF composite increased

suddenly at a certain strain of the specimen. The CP composite had the good response to cyclic load patterns in the bending test of the block specimen and indicated the residual resistance at an unloading state. The $\Delta R/R_0$ of the CP composite in the pile specimen with prestress showed good results to the loading patterns before and after micro crack formation in the bending test. The arrangement of the CP composite near the tension-side's surface, and far from the steel bars in the pile, effectively improved the sensitivity of the composite. The excellent self-diagnosis function of the CP composite in the concrete structures was considered to be mainly caused by the flexibility in the percolation structure of carbon particles.

BIBLIOGRAPHY

1. N. Muto, H. Yanagida, T. Nakatsuji, M. Sugita, and Y. Ohtsuka. *J. Am. Ceram. Soc.* **76** (4): 875–879 (1993).
2. M. Takada, S.-G. Shin, H. Matsubara, and H. Yanagida. *J. Jpn. Soc. Compos. Mater.* **25**: 225–230 (1999).
3. Y. Okuhara, S.-G. Shin, H. Matsubara, and H. Yanagida. *Trans. MRS-J.* **25** (2): 581–584 (2000).
4. M. Takada, H. Matsubara, S.-G. Shin, T. Mitsuoka, and H. Yanagida. *J. Ceram. Soc. Jpn.* **108** (4): 397–401 (2000).
5. Y. Okuhara, S.-G. Shin, H. Matsubara, H. Yanagida, and N. Takeda. *Proc. SPIE* (2000), in press.
6. H. Nishimura, T. Sugiyama, Y. Okuhara, S.-G. Shin, H. Matsubara, and H. Yanagida, *Proc. SPIE* **3985**, 335 (2000).

SENSOR ARRAY TECHNOLOGY, ARMY

JEFFREY SCHOESS

Honeywell Technology Center
Minneapolis, MN

INTRODUCTION

Today's commercial and military aircraft require significant manpower to provide operational readiness and flight safety. Aging aircraft fleets are much in need of new and innovative health-monitoring methods to prevent catastrophic failure and reduce life-cycle costs. The key needs for characterizing *in situ* structural integrity characteristics of corrosion and barely visible impact damage (BVID) to determine "damage susceptibility" must be addressed. This article presents a new concept for onboard real-time monitoring using conductive polymer sensor array technology.

BACKGROUND

Both commercial and military service personnel currently employ "walk-around" structural inspection as a cornerstone of condition-based maintenance. This means that a hierarchy of inspections is required to ensure that fleet readiness and availability requirements are met. Structural inspection includes daily inspection, phased maintenance based on aircraft operating time, conditional

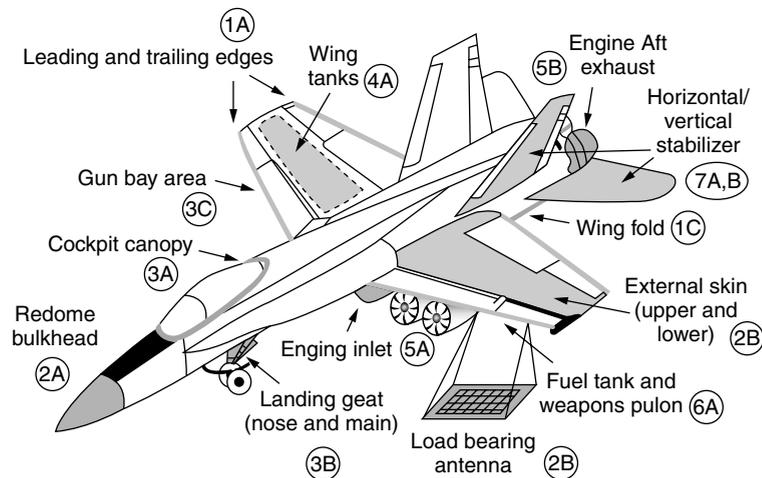


Figure 1. Key sensing locations on aircraft.

inspection based on the mission and location of the aircraft, and calendar-based inspection.

Although condition-based maintenance inspection is mature and is reliable in most cases, its application in future military and commercial systems has significant drawbacks notably high cost and intensive effort. Currently, the cost to maintain a Navy aircraft is up to \$200,000 per year. A 1996 Naval Center for Cost Analysis AMOSC report indicates that the direct cost of maintaining Navy aircraft and ships is at least \$15.0 B per year. As much as 25 to 30% of operating revenue is spent on maintenance for commercial air carriers. According to a 1995 study by the office of the Under Secretary of Defense, 47% of the Navy's active duty enlisted force (173,000 sailors) and 24% of the Marine Corps (37,600 marines) are assigned to maintenance functions. The mandate to reduce manpower while performing duties faster, cheaper, better, and more reliably is a reality in both military and commercial transportation.

In addition to these issues, problem areas exist specifically for maintaining structural integrity, including BVID and hidden and inaccessible corrosion. The increased use of composite materials in aircraft structures introduces the potential for BVID, a maintenance-induced damage effect. At least 30% of all maintenance is related to structural repair due to tool dropping and in-service damage. A significant amount of the loss of structural integrity is due to hidden corrosion as well as corrosion located in inaccessible areas (wheel wells, landing gear areas, and fuel tanks). The practice of applying surface treatments of various types to provide adequate protection, in some cases overcoating the surface with several layers, causes considerable weight increase. This increase results in loss of fuel savings and proved aircraft performance.

TECHNICAL APPROACH

A trade study was performed to identify and assess potential aircraft inspection areas that could benefit from conductive polymer sensor array technology. The trade study involved the identification of seven key areas of a

generic fighter aircraft (F-18 or equivalent). The areas addressed in the study were external wing structure, internal wing and fuselage structure, including landing gear and cockpit canopy, communications, external stores, and empennage structure. The study addressed specific parts of these identified areas and included a problem definition, a proposed sensing layout approach, and a sensing configuration. Figure 1 is a drawing of the F-18 aircraft that shows the functional layout of the seven aircraft sensing areas for possible future technology insertion. The sensing areas are mapped to the aircraft geometry, labeled by area, and keyed to the full-scale trade study chart shown in Table 1. The chart highlights the details of the trade study effort and contains specifics on subassemblies, including a general problem description. It maps the problems using three different types of sensing: "M/C" refers to moisture/corrosion sensing, "ID" refers to impact detection, and LBA refers to "load-bearing antenna." For each sensing approach, three packaging options exist: (1) a conformal sensor array, which would cover a larger surface area such as an external wing area of more than several square feet; (2) a conformal sensor applique to provide sensing coverage in a smaller area (a few square inches, possibly with significant contour shapes); and (3) a conformal boot assembly. The conformal boot design would involve fabricating a preformed structure—a sensory boot that fits the spatial constraints of the aircraft contour. An example of this configuration would be a preformed boot fit over the leading edge or radome bulkhead assembly.

Sensor Development

A conductive polymer sensor array design provides the capability for multifunction conformal sensing. Honeywell has developed polymer sensors to sense moisture (i.e., electrolyte) conditions and the presence of moisture/fluids across an extended surface area. A primary maintenance concern is the need to sense and quantify moisture trapped between the protectant system layer and the aircraft surface that could cause corrosion. Typically, the moisture is an electrolyte, an electrically conducting fluid that has ions in solution. The polymer sensor array has been designed to

Table 1. Aircraft Trade Study Chart

Aircraft Area	Part/Assembly	Problem Definition	Sensing Approach ^a	Sensing Configuration
① Wing external	•Leading edges	•Flap and drive assembly	•M/C	•Conformal array
	•Trailing edges	•Impact (BVID)	•ID	•Conformal boot
		•Corrosion—wing attach fitting		
	•External skin (upper and lower)	•Impact (BVID due to maintenance/repair)	•M/C	•Conformal array
		•Corrosion (fastener area)	•ID	•Conformal applique
	•Wing fold	•Corrosion in hinge area	•M/C	•Conformal tape
		•Wing attachment fatigue		
② Communications support	•Radome bulkhead	•Corrosion (dissimilar ± F-galvanic)	•M/C	•Conformal boot
	•Wing antenna	•Phased-array antenna	•LBA	•Conformal applique
③ Fuselage	•Cockpit canopy	•Corrosion—dissimilar interface (galvanic)	•M/C	•Conformal applique
	•Landing gear	•Corrosion in wheel well area, main landing gear assembly	•M/C	•Conformal applique
	•Gun bay area	•Corrosion—dissimilar interface	•M/C	•Conformal applique
④ Wing internal	•Wing tank	•Fuel leakage in web area (wet bay)	•M/C	•Conformal applique
		•Electrical connector/ground straps		
⑤ Engine	•Engine inlet	•Impact (BVID) from debris/bird strike	•ID	•Conformal applique
	•Aft engine exhaust area	•Corrosion—moisture	•M/C	
⑥ External stores	•Fuel tank pylon	•Corrosion—dissimilar interface	•M/C	•Conformal applique
	•Weapons pylon	•Erosion		
⑦ Empennage	•Horizontal stabilizer	•Pivot shaft corrosion	•M/C	•Conformal applique
	•Vertical stabilizer box	•Corrosion		

^aM/C = moisture/corrosion; ID = impact detection; LBA = load-bearing antenna.

detect the “presence” of an electrolyte, which can be seawater, acid rain, lavatory fluids, fuel, hydraulic fluid, chemicals, or cargo by-products.

The basic design is implemented by printing a specific pattern design on a flexible substrate material, curing it, and layering it using a pressure-sensitive adhesive. A typical pattern developed for electrolyte sensing is a transducer design that has alternating electrode pairs. Figure 2 illustrates the pattern layout for a polymer sensor array. The figure shows a set of dedicated electrode pairs, each of which operates as a sensory element. The sensor is designed to function as a linear 2-D array that measures the “location” where the electrolyte is sensed and the “amount” of electrolyte based on exposure across the sensor array.

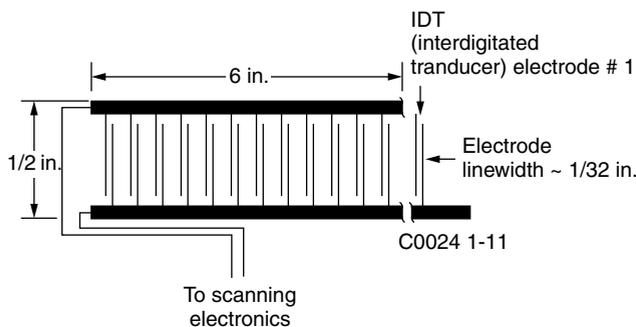


Figure 2. Pattern layout of polymer sensor array.

Detection of Corrosivity. Four conditions must exist before corrosion can occur: (1) the presence of a metal that will corrode an anode; (2) the presence of a dissimilar conductive material (i.e., cathode) that has less tendency to corrode; (3) the presence of a conductive liquid (electrolyte); and (4) an electrical path between anode and cathode. A corrosion cell is formed because of the electrochemical effect, if these four conditions exist, as shown in Fig. 3. In a typical aircraft coating application, paint applied to the surface of the metal acts as a moisture barrier to protect the bare metal from exposure to an electrolyte. The paint film prevents the corrosion cell from functioning by separating the electrolyte from the anodic and cathodic sites on the metal surface. If this paint layer is damaged by erosion, heat exposure, or aging, the cell is activated, and corrosion occurs.

Figure 3 also highlights the concept of using a polymer sensor array to detect corrosivity when a corrosion cell is formed in an aircraft lap joint. As shown, the linear sensor array senses the “conductivity” of the trapped fluid by conducting a current through the fluid that is between IDT electrode pairs. The fluid’s conductive property is, by definition, “the ability to act like an electrolyte and conduct a current, or a measure of its corrosivity.”

The concept of performing corrosive environmental “exposure susceptibility” index monitoring to minimize scheduled inspections and provide direct cost savings is shown in Fig. 4. The basic idea is continuous monitoring of the actual exposure of each aircraft to corrosive

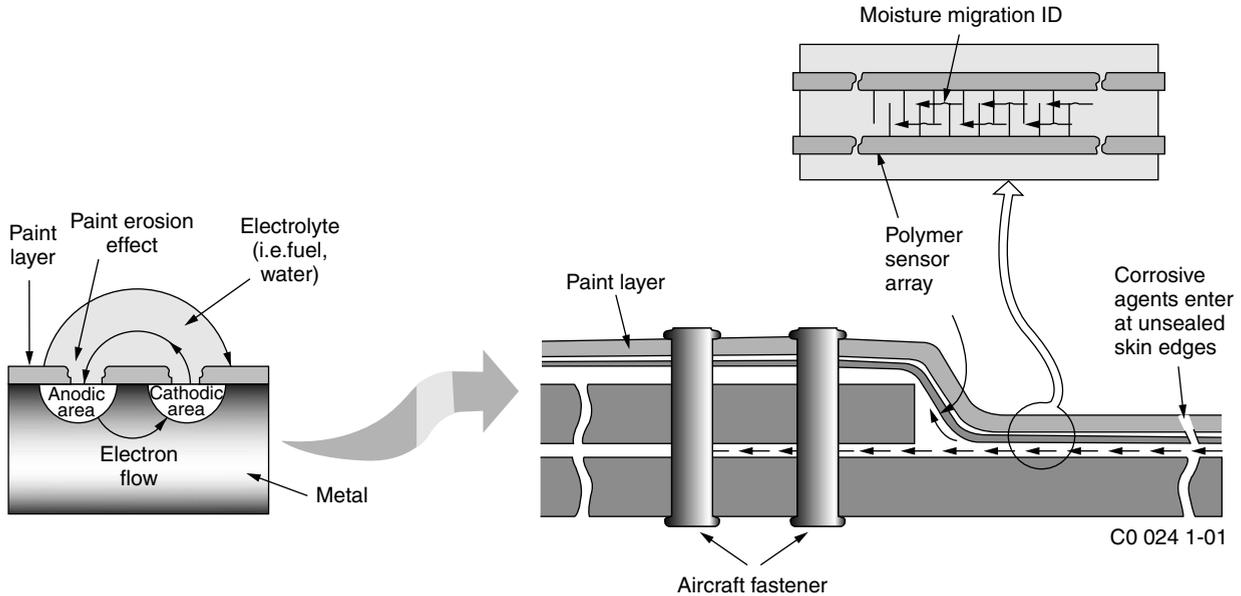


Figure 3. Simplified corrosion cell and lap joint application.

environmental factors (moisture ingress into protective coating, type of corrosive agent, etc.) and then scheduling corrosion inspections based on these measurements, rather than on preset rules that are only loosely related to corrosion. Typical preset rules that an exposure susceptibility index would replace are calendar-based (i.e., inspection every 30 days) or usage-based (i.e., inspection every 10 h of operation) inspections. One can think of the system as a “corrosion odometer” whose a readout steadily increases according to the corrosiveness of the environment to which the aircraft is exposed. Maintenance personnel can intermittently check the odometer and inspect as needed. The exposure susceptibility index provides a reliable method for scheduling corrosion inspections that (1) is based on the true exposure of the aircraft, which leads to a higher degree of susceptibility to corrosion; (2) appropriately reflect variations in exposure due to short-term weather patterns; and (3) can be consistently applied to aircraft of a given type at any location in the world.

The sensor array approach can sense and calculate an exposure index to ingress of an electrolyte (i.e., water) and the “wetness” effect of the electrolyte. The wet/dry cycle of exposure is a strong indicator of how susceptible an aircraft is to corrosion; wetness is a basic requirement for corrosion

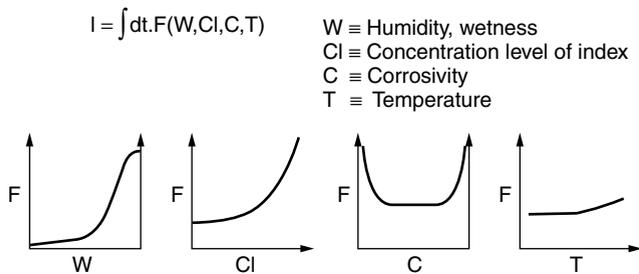


Figure 4. Exposure susceptibility index.

to occur. The wetness exposure index is defined as the integral over time of the function $F_W(W)$. Here W is the time-varying output of a “wetness” sensor (1 = wet, 0 = dry) that quantifies the total corrosive effect of wetness. F_W is a simple function that gives the exposure index on a convenient scale, so an abbreviated inspection is called for each time the index passes through a multiple of 100, for example. Thus, for severe environments such as Puerto Rico, an increase by 100 every 15 days could occur, compared to an increase by 100 every 90 days in Denver.

Further improvement of the exposure susceptibility index can be obtained by adding other environmental factors that influence corrosion, including the concentration of the electrolyte, the temperature, and the conductivity (corrosivity factor).

Figure 5 illustrates the index calculation concept and shows the maintenance cost saving concept in detail. The design approach is set up to collect and analyze the environmental factors related to structural health (moisture ingress, impact forces, etc.) that could lead to loss of structural integrity. These factors are collected and integrated as a “cumulative index” to determine (1) the level of “susceptibility” to failure and (2) whether maintenance is required at a given location in the aircraft. The cumulative index value, it is envisioned, will be represented as a simple

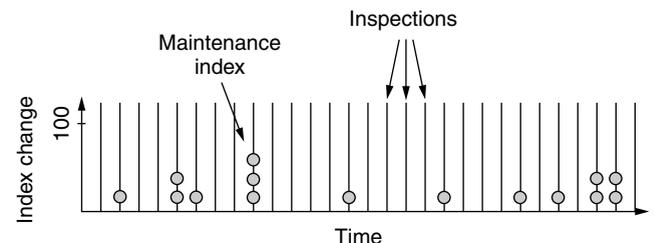


Figure 5. Maintenance cost saving tutorial.

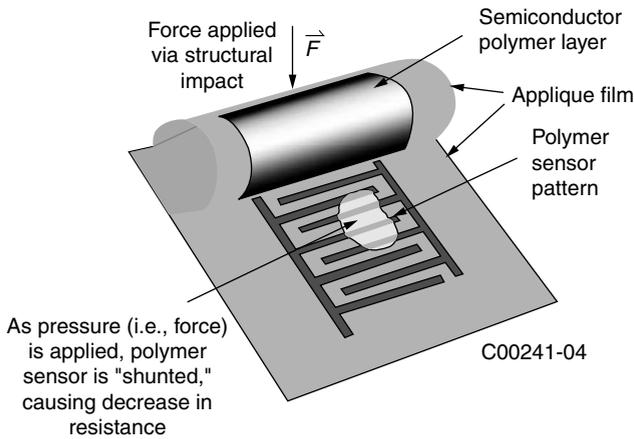


Figure 6. Force-sensing resistor (FSR).

whole number from 0 to 100 (which indicates the level of susceptibility; a higher number indicates that more potential for damage may exist) that could be read out by maintenance personnel from the aircraft maintenance debriefing interface at scheduled inspection intervals. The crew could then decide to perform scheduled maintenance or bypass the action. This would directly reduce the cost of maintaining the aircraft by eliminating or reducing the number of inspections. In addition, reducing the time for a maintenance procedure based on the polymer sensor system's ability to identify the general structural location where the repairs may be needed and the type of repair required (i.e., impact damage vs. corrosion) will result in additional operational cost savings.

Impact Detection. The polymer sensor for moisture/corrosion sensing can also sense impact forces caused by maintenance-induced damage or operational servicing. To provide sensing for impact forces, the polymer sensor array is configured with an additional semiconductor polymer layer, as shown in Fig. 6. The design approach is set up to operate as a force-sensing resistor (FSR). An FSR operates on the principle of converting force applied via

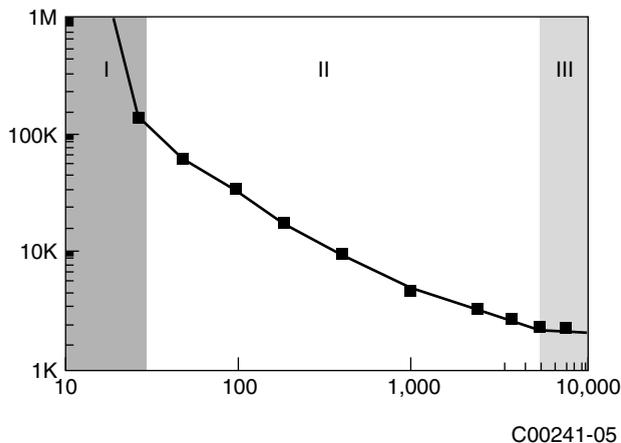


Figure 7. FSR response vs. applied force.

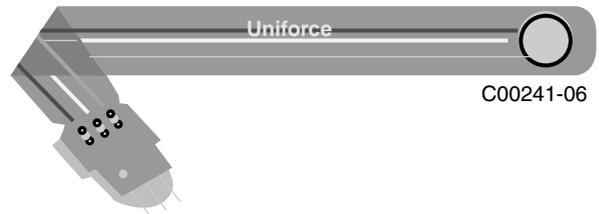


Figure 8. Example of off-the-shelf FSR product.

a structural impact to an equivalent voltage output. As pressure is applied, individual electrode pairs are shunted, causing a decrease in electrical resistance. The measurement of impact force magnitude, impact direction vector along the sensor array, and impact surface area can be quantified, depending on polymer composition, shunt pattern and shunt shape, and the method for applying pressure (hemispherical vs. flat). Figure 7 shows a typical curve of sensor response. The figure is a plot of electrical resistivity versus applied force and has an active sensing region of two to three orders of magnitude from low impedance (kilohms) to high impedance (megohms). The sensor response is approximately a linear function of force across a wide range of applied pressure. The first abrupt transition that occurs is at low pressure. This point is called the "breakover point" where the slope changes. Above this region, the force is approximately proportional to $1/R$ until

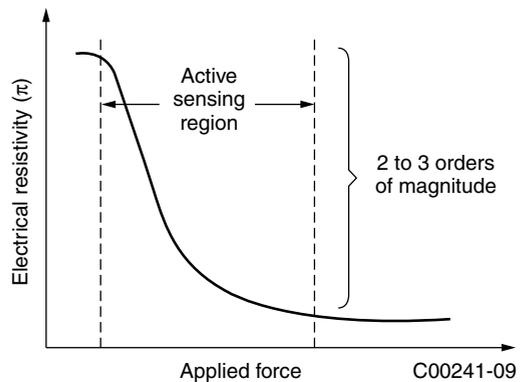
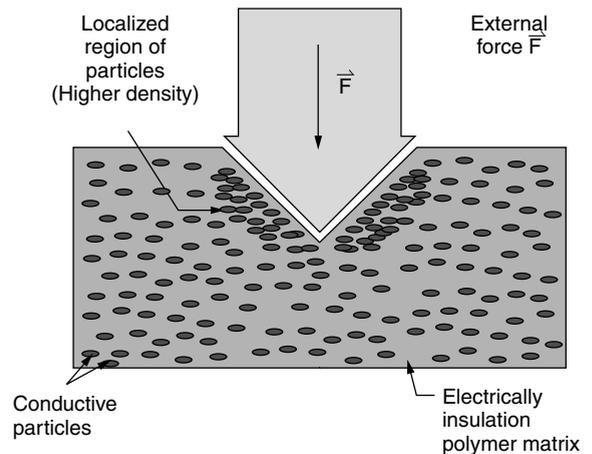


Figure 9. Polymer matrix sensor.

Table 2. PTF Resistor Versus Other Resistor Technology^a

Resistor Type	Gauge Factor (G)	TCR (ppm/°C)	Application Method	Relative Cost
Continuous metallic films	2.0	20.0	Spin cast	High
Thin film	50.0	20.0	RF sputter	High
Semiconductor	50.0	1500.0	Evaporation Diffused	Medium
Thick film (PTF)	10.0	50.0–500.0	Implanted Screen print Stencil Spin cast	Low

^aSource: G. Harsanyi, ed., *Polymer Films in Sensor Applications—Technology, Materials, Devices and Their Characteristics*. Technomic, 1995.

a saturation region is reached. When the force reaches this magnitude, applying additional force does not decrease the resistance substantially.

Figure 8 is a photo illustration of a commercially available off-the-shelf FSR product called Uniforce, which has an operating range of 0–1000 psi.

Another type of conductive polymer sensor is a polymer matrix sensor that consists of electrically conducting and nonconducting particles suspended in a matrix binder. Figure 9 shows a cross-sectional view of a polymer matrix sensor. Typical design construction includes a matrix binder and filler. Matrix binders include polyimides, polyesters, polyethylene, silicone, and other nonconducting materials. Some typical fillers include carbon black, copper, silver, gold, and silica. Particle sizes typically are of the order of fractions of microns in diameter and are formulated to reduce temperature dependence, improve mechanical properties, and increase surface durability. Applying an external force to the surface of a sensing film causes

particles to touch each other and decreases the overall electrical resistance.

Table 2 illustrates the typical performance of polymer thick-film (PTF) resistor technology and other resistor technologies. The table includes a summary for thin films, semiconductor, and continuous metal films. The significant advantage of PTF resistor technology over all other resistor sensing is the cost to fabricate devices. The PTF cost factor is achieved by the ability to print resistive material via stencil, screen printing, and ink-jet printing techniques.

A prime example of using FSR technology for aerospace sensing is structural integrity monitoring. Today’s commercial and aerospace structures incorporate a large amount of composite materials to reduce structural weight and increase load-bearing properties. Composites are susceptible to damage from impact forces experienced in operation, including debris picked up from runways and maintenance-induced damage caused by dropped tools. Figure 10 illustrates the system-level concept of

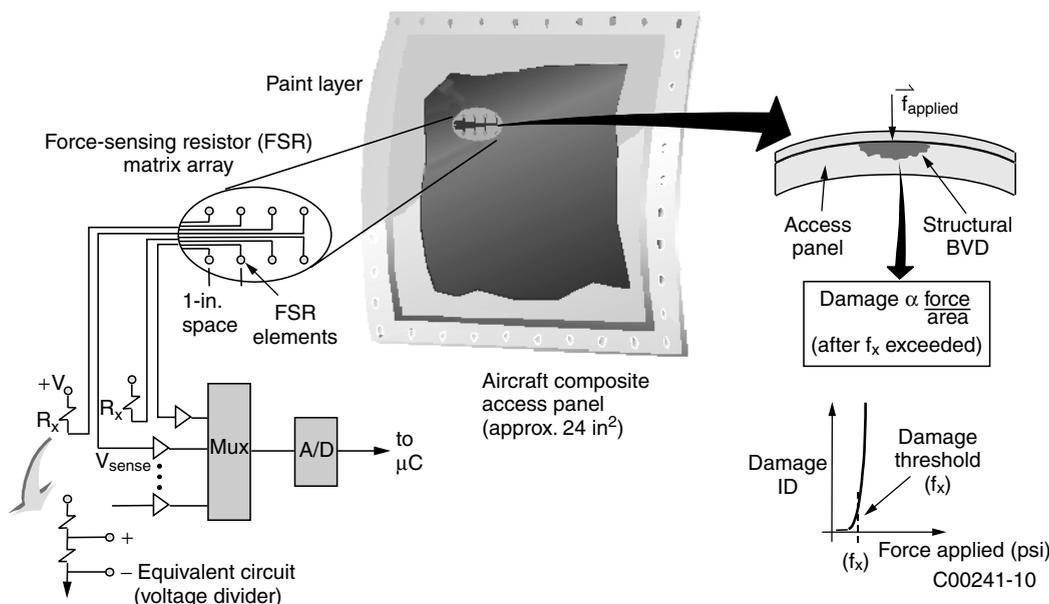


Figure 10. Structural impact damage tutorial.

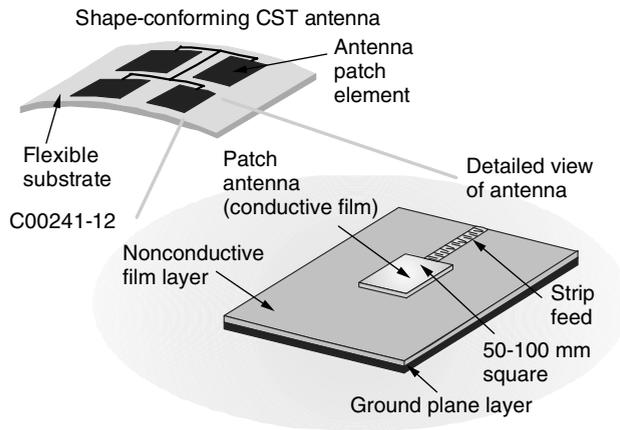


Figure 11. Example of conformal antenna.

impact-damage-detection-based applied force versus damage for a composite aircraft panel. A matrix array of FSR elements integrated into the aircraft panel is shown. Panel construction involves printing FSR elements directly on the panel surface or on a film layer, which is then bonded to the panel using a pressure-sensitive adhesive layer. The polymer patterns incorporated on the panel include a combination of sensor elements and electrical interconnects implemented with conductive polymer materials.

To measure and record impact forces in real time, the output of each FSR element is converted to an equivalent voltage via a simple voltage divider circuit and is provided as input for a dedicated data acquisition system. Each FSR element output is routed to an analog multiplexer.

An analog-to-digital converter sequentially digitizes each FSR value into an equivalent digital word for processing by a dedicated system controller. The illustration on the right-hand side of Fig. 10 shows what happens if structural damage occurs. An external force event (i.e., a tool dropped on the surface) causes an impact. Structural damage usually consists of multilayer delaminating or microcracking of individual composite layers. In composite structure applications, the curve for quantifying structural damage is an exponential relationship and is detected by setting a force threshold value. A value that exceeds the threshold value f_x indicates that barely visible structural damage has occurred. The effects of detected damage can be read out by maintenance personnel periodically to determine if structural repair is needed it or is marked as suspect, and the vehicle is returned to active service. A set of damage identification threshold values could be retained for each major structural component of the aircraft in a 3-D map database to perform maintenance on demand.

Conformal Antennas. A significant feature of polymer sensor array technology is the arrays' ability to operate as a low observable (LO) conformal antenna. The polymer sensor has been tested in laboratory conditions to detect broadband frequencies of several megahertz without any optimization of the polymer circuit pattern. The conformal antenna capability offers a significant benefit of increasing detection of "bad guy" signature threats. Tests performed by aircraft primes have indicated that conformal load-bearing antennas improve detection by a factor of 6 to 14 times. In addition, the conformal polymer construction makes it suitable for phased-array antenna

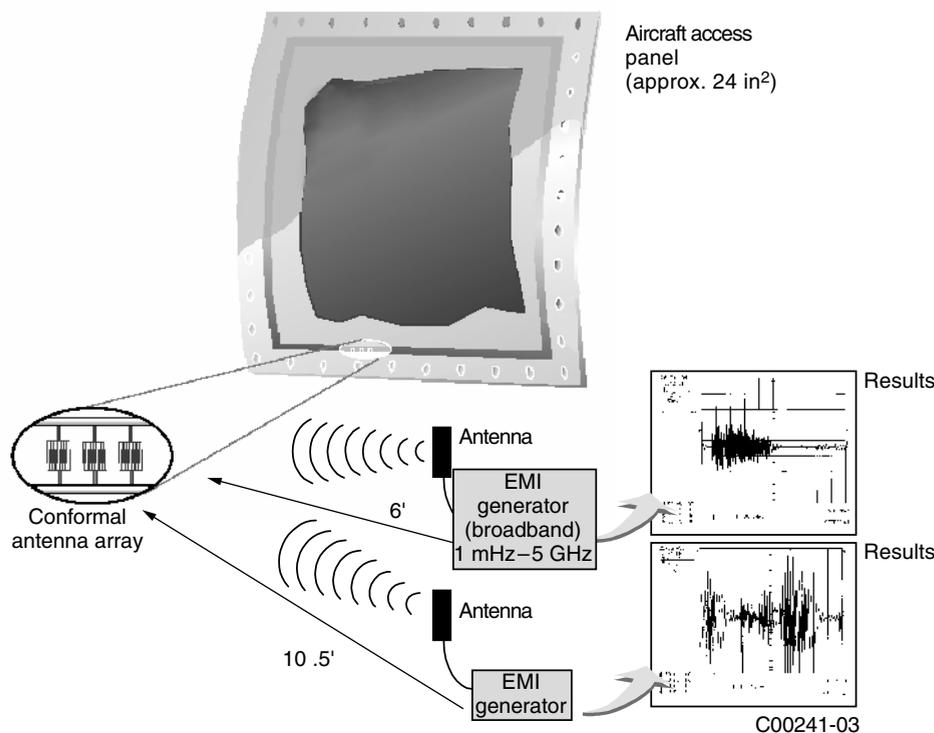


Figure 12. Conformal antenna concept.

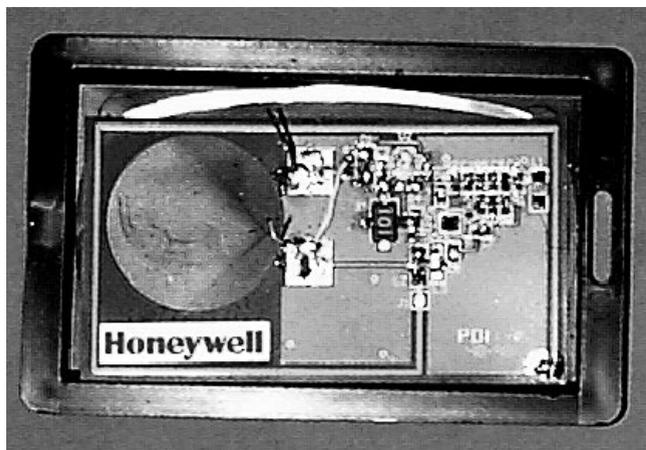


Figure 13. Wireless transceiver module for self-contained communications.

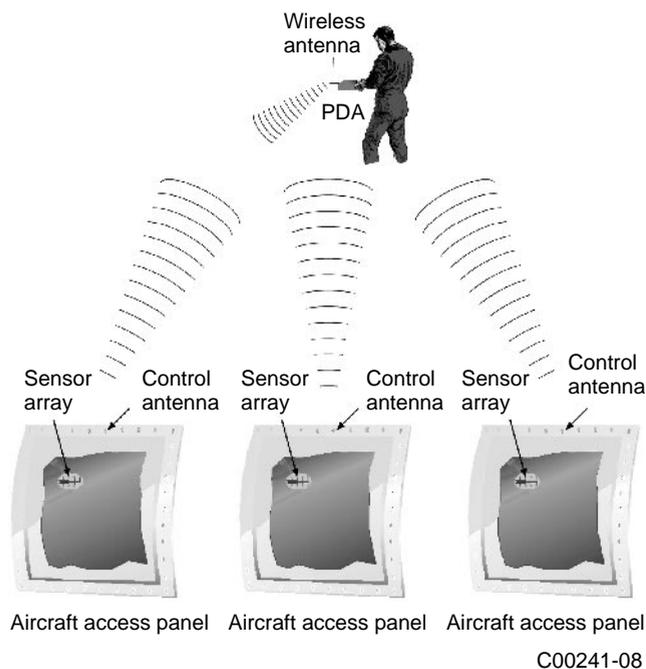


Figure 14. Wireless structural panel sensor web.

design for munitions and guided projectiles. Figure 11 illustrates the feasibility of using the polymer design for antenna functions. Figure 12 highlights the use of a broadband EMI source and detection of electromagnetic wave pickup at increasing distances up to 10 ft from the antenna.

Communications Debriefing

A significant system-level issue is how to obtain data on environmental factors during aircraft inspections without increasing the workloads of maintenance personnel. This can be achieved by providing a wireless link for data access. Figure 13 illustrates a photo of a $2 \times 3 \times 0.125$ in wireless transceiver module for field maintenance communications. The module consists of a low-frequency (128 kHz) receiver

interface, a dedicated high-frequency (315 MHz) transmitter interface, dedicated control logic, and internal RAM memory. The radio-frequency (RF) system can be read at ranges of 6–30 ft and operates at $2 \mu\text{A}$ in standby mode. The RF module is powered by a high-energy-density lithium button-cell battery. Future applications will include an RF module that features a very low profile height of 4 mils and capability for RF power scavenging. This unique capability implies that no batteries will be required to communicate and debrief the sensor suite. Up to 100 RF modules can be read simultaneously by a dedicated wireless RF reader.

Figure 14 illustrates a concept for wireless sensor communications to debrief a suite of aircraft structural components. A field maintenance technician is shown holding a personal data assistant that has a wireless interface. The status of the structural integrity of each component could be assessed by issuing a polling command to search and identify the health status of a designated structural panel. The wireless interface within each structural component would read the poll message, determine if the message request is intended for that component, and the designated panel will then return the health status to the maintenance technician.

SENSORS, SURFACE ACOUSTIC WAVE SENSORS

DAVID W. GALIPEAU
South Dakota State University
Brookings, SD

INTRODUCTION

Surface acoustic wave microsensors are microchips that usually have a sensing film applied to the substrate surface. They differ from silicon-based microcircuits that form the basis for most integrated circuit technology today, in that they are based on piezoelectric versus semiconductor substrates and are usually much less complex. Microsensors are usually defined as sensors of micron dimensions or are made by the same fabrication techniques that are used for integrated circuits. Microsensors have several advantages over older sensor technologies, including small size, low cost, and typically better performance (1). The relatively rapid development of surface acoustic wave (SAW) microsensors has resulted from the demand for low-cost, high-performance sensors to measure such things as hazardous gases, biological pathogens, chemical and biological weapons, automotive emissions, and indoor air quality. Additionally, there is strong interest in developing an electronic nose for both industrial and laboratory applications. As a result of these advantages and the strong demand, the development of SAW microsensors has grown rapidly during the last twenty years.

BACKGROUND

SAW devices were originally developed and are still widely used as high-performance signal processing elements such as filters and delay lines in electronic circuits (2,3). The

term acoustic wave refers to the class of waves that displaces particles of the solid, liquid, or gas medium in which they propagate. Therefore, acoustic waves are considered mechanical waves compared to electromagnetic waves, which can propagate in a vacuum because they do not require a medium or have related particle displacements. The term surface acoustic wave (SAW) usually refers to the class of acoustic waves that propagates at a solid surface, versus bulk waves, which propagate within a solid. The first type of SAW was discovered by Rayleigh in 1885. It has longitudinal particle displacements (in the propagative direction) and transverse particle displacements (perpendicular to the propagative direction) that are normal to the substrate surface. This type of SAW is called a Rayleigh wave. There are several other types of SAWs that are distinguished primarily by their wave particle displacements and are usually allowed only for certain crystallographic orientations. For example, a wave that has transverse particle displacements in the plane of the substrate and propagates just below the surface is called a surface skimming bulk wave (SSBW). It occurs on ST-cut quartz. Bulk waves are also classified as longitudinal or transverse (shear) based on particle displacements. Classic reviews of acoustic waves in solids are provided by Auld (4) and Kino (5).

A SAW microsensor normally consists of two metal interdigital transducers (IDTs) fabricated on a piezoelectric substrate. The IDTs are patterned from a thin metal film (usually aluminum) that has been deposited on the substrate. The patterning is done by using standard photolithographic techniques. Figure 1 illustrates a single delay line (channel) SAW sensor. The term delay line is used for this design because it can be used for this application in signal processing. The operation of a SAW microsensor is as follows. SAWs are launched onto the delay path (Fig. 2a) via the reverse piezoelectric effect when an RF signal at the microsensor's operating frequency is applied to the input IDT. These SAWs travel across the delay path (Fig. 2b) to the output IDT where they are converted back into electrical signals via the piezoelectric effect (Fig. 2c). The velocity and amplitude of the SAWs are the sensor outputs.

The acoustic velocity V of any material is a function of the elastic constant c and density ρ of the material. For the simplest (isotropic) case, the velocity is given by $V = (c/\rho)^{1/2}$ (5). The relationship between the SAW velocity, frequency f , and wavelength λ , is given by $V = f\lambda$. The attenuation of the waves is primarily a function of the viscosity

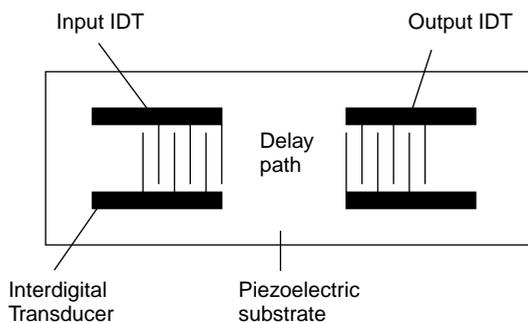


Figure 1. Diagram of a single channel (delay line) SAW microsensor.

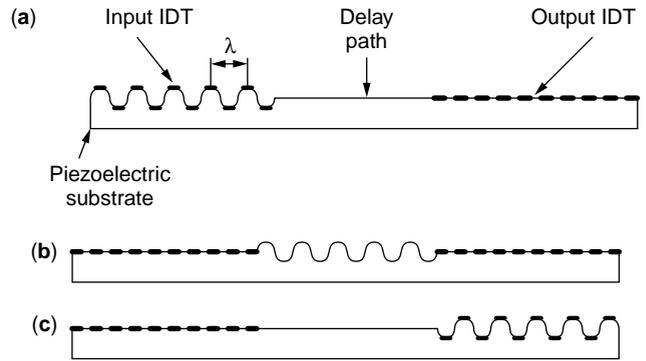


Figure 2. The generation and propagation of SAWs on a sensor surface.

of the material in which the waves propagate. The SAW velocity or amplitude can be changed by a film or deposit on the sensor surface or other disturbances at or near the sensor surface because the density, elastic constant, and viscosity (viscoelasticity) of the film or deposit are usually different from those of the SAW substrate. The SAW velocity or amplitude can also be affected by the conductivity or permittivity of the deposit because the SAW also has an electrical component on a piezoelectric substrate. Secondary physical parameters that affect the previously mentioned (primary) parameters also affect the SAWs' characteristics. They include temperature, stress, pressure, and electric and magnetic fields. A change in the SAW velocity due to mass on the surface is commonly referred to as "mass loading." Mass loading and amplitude attenuation are the most commonly used sensing mechanisms for SAW sensors and are the primary focus of this review article.

The change in SAW velocity ΔV_R has been related to the mass of a thin nonviscous (lossless) film on the sensor surface by Wohltjen (6).

$$\frac{\Delta V_R}{V_R} = (k_1 + k_2) f h \rho', \quad (1)$$

where V_R is the SAW velocity, k_1 and k_2 are substrate material constants, f is the SAW frequency, and h and ρ' are the height and density of the thin film layer, respectively. Therefore, the change in SAW velocity due to a layer depends on $h\rho'$ of the layer. Because the units of $h\rho'$ are kg/m^2 , this is also the layer surface density ρ_s . The change in SAW velocity is determined experimentally by measuring the phase shift, $\Delta\phi$ or the frequency shift Δf of the SAWs that are related to the change in SAW velocity by (6)

$$\frac{\Delta V_R}{V_R} = \frac{\Delta f}{f_o} = -\frac{\Delta\phi}{\phi_o}. \quad (2)$$

The relationship between the attenuation of the SAWs and the viscosity of the mass or film on the sensor surface has been described by Martin et al. (7) for acoustically thin ($< 1 \mu\text{m}$ thick) hard polymer films as

$$\Delta\alpha = \omega^2 \frac{n}{V_R^3} C G'', \quad (3)$$

where $\Delta\alpha$ is the attenuation change, ω^2 is the angular frequency ($2\pi f$), C is a SAW-film coupling constant, and G'' is the loss modulus (complex part of the shear modulus) of the film which is directly related to its viscosity. The term acoustically thin denotes a film that does not resonate at the SAW microsensor's operating frequency. As film thickness increases, it can resonate at the operating frequencies. A macroscopic analogy for a resonant film would be a plate of Jello which, if shaken at a certain frequency, will also resonate when the Jello is high enough.

The earliest uses of SAW devices as microsensors were reported by Das in 1978 (8) for measuring pressure (physical) and by Wohltjen in 1979 (9) for measuring thin film properties (chemical). These sensor applications resulted from the observed high sensitivity of SAW signal processing devices to "external" physical parameters such as temperature changes and package stress, as well as "internal" properties of the films deposited on the SAW substrate. A major application of SAW sensors has been highly sensitive mass detectors (microbalances). Wohltjen stated that SAW sensors have a potential mass sensitivity 200 times greater than the better known quartz crystal microbalance due to their higher operating frequency (6). The effect of frequency on SAW velocity is illustrated by Eq. (1). SAW devices can operate at frequencies higher than 1 GHz, compared to about 10 to 50 MHz for the quartz microbalance which operates by using bulk (shear) acoustic waves. However, because noise in measurement electronics increases as frequency increases, the practical frequency limit for SAW sensors may be closer to 500 MHz. SAW and bulk wave devices have significantly different geometries and fabrication techniques. SAW devices are fabricated by standard microelectronic fabrication methods, whereas bulk wave devices are manufactured individually as small disks that have thin metal film electrodes on each side. Thus, SAW devices are typically less expensive and have a much wider range of designs than bulk devices. The following are additional advantages of SAW sensors. (1) They can be configured in "smart" designs by using two sensing channels on the same substrate where one is a reference. This allows the sensor to be self-compensating for interfering environmental parameters such as temperature (10). (2) Because SAW microsensors are sensitive to several parameters, they can provide an amplified sensor response via multiple detection mechanisms. (3) They are easy to use in wireless sensing applications in both the active mode, as the frequency control element in a transmitter (8), and the passive mode, as an energy reflector (11). The passive mode is particularly interesting because the sensor does not need a power source but is read using a special FM radar type system. SAW devices are also considered one of the earliest types of microelectromechanical systems, or MEMS devices. MEMS devices are usually defined by having both mechanical and electrical components or functions in a single unit and are fabricated using microelectronic fabrication techniques. SAW devices fit this description because they have acoustic waves (mechanical) that are launched and detected electrically.

Commercial SAW microsensor based systems are currently available for gas and biological sensing, gas-chromatography vapor sensing, and chilled-surface,

dew-point hygrometry. All of these systems capitalize on the high sensitivity of the SAW microsensor to small mass changes. Systems for chemical and biological sensing have been developed by Microsensor Systems (now a subsidiary of Sawtek) (12). These include "Vaporlab," which uses an array of SAW microsensors coated with proprietary films and pattern recognition to identify the vapor, and the SAW "Minicad," which uses the same techniques to detect chemical warfare agents. The SAW gas-chromatography system was developed by Electronic Sensor Technology (13). This system uses a single bare SAW microsensor and can be used as an electronic nose in several gas sensing applications that have been validated by the EPA. The SAW dew-point hygrometer was developed by Microconversion Technologies Co. (14). This hygrometer, the "Ultra DP5," also uses a bare SAW microsensor, in this case to provide precision measurements of water vapor concentration.

This article presents a fairly wide range of SAW microsensor applications that are based on the personal research and development experience of the author. These applications include detection of water vapor and other gases; thin polymer film characterization, including adhesion, surface properties, and curing; chilled-surface dew-point measurements; measurement of surface energy and cleanliness; and temperature measurement. A review of acoustic wave biosensors has been provided by Andle and Vetelino (15). Additional SAW microsensor applications are reviewed in books by Ballantine et al. (16) and Thomson and Stone (17). These books also provide more comprehensive descriptions of SAW microsensor theory, design, and applications. SAW device design procedures can also be found in the literature (2,3).

EXPERIMENTAL PROCEDURES FOR SAW SENSING

The two most common methods for measuring SAW velocity are the *phase* and *frequency* techniques (9,16). The experimental setup for the phase technique requires applying an RF signal (from a signal generator) to the input IDT of the SAW sensor. A vector voltmeter is then used to monitor both the phase and amplitude of the SAWs, as shown in Fig. 3. The experimental setup for the frequency technique requires using the SAW sensor as the frequency control element in an oscillator circuit. A frequency counter is then used to monitor the oscillatory frequency, as shown in Fig. 4. The advantages of the phase technique are ease of use, stability, and easily obtainable amplitude information.

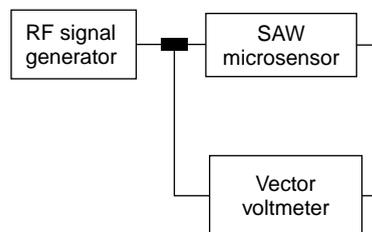


Figure 3. The experimental setup for the phase (vector voltmeter) technique.

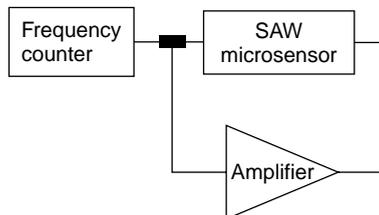


Figure 4. The experimental setup for the frequency (oscillator) technique.

However, this technique typically involves relatively expensive laboratory equipment. The advantages of the frequency technique are high sensitivity and less expensive equipment; however, oscillatory stability can be a problem, and additional circuitry is required to obtain the amplitude information.

DISCRIPTION OF APPLICATIONS AND EXPERIMENTAL RESULTS

Gas Detection: Water Vapor and Hydrocarbons

One of the most widely studied applications of SAW microensors has been measuring gas concentration. The most commonly used configuration consists of a very thin film ($<1\ \mu\text{m}$) applied to the SAW microsensor. These films are carefully selected or designed to provide both high sensitivity and selectivity to the gas of interest and also long-term reliability. To study the response of the SAW sensor to gas concentration, the sensors are placed in a chamber in which the atmosphere is controlled by a gas delivery system.

The measurement of water vapor concentration and relative humidity have been of high interest for many years because of their effect on human comfort and health. More recently, the measurement of water vapor has become very important in several other fields, including meteorology, agriculture, and manufacturing due to the effects of water vapor on weather forecasting, product quality, and the large energy costs of drying processes. Polymer-coated SAW devices have been studied as an improved means to measure relative humidity. Polymers are good candidates for sensing films due to their ease of processing, widely customizable properties, and relatively low cost. Polyimide is a readily available polymer that is widely used in microelectronic applications. It has the advantages of durability at high temperatures, low dielectric constants, and ease of application. Therefore, it was chosen for this work.

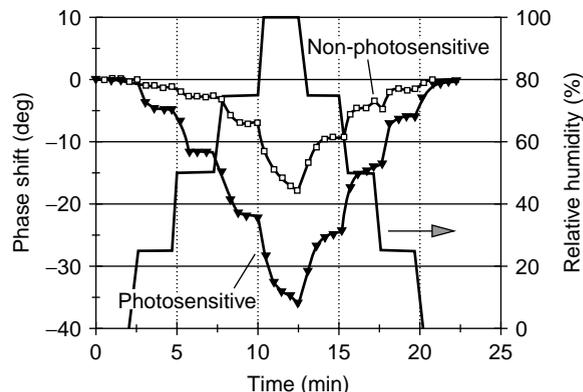


Figure 5. Water vapor response of photosensitive and nonphotosensitive polyimide films (15).

In addition, photosensitive polyimides have recently been developed that reduce the number of processing steps required for patterning. The major drawback of polyimide in microelectronic applications is that it typically absorbs more than 2% by weight of water vapor when placed in high humidity. However, this property allows its use as a humidity sensing film.

Figure 5 (18) shows the response to water vapor of two SAW microensors coated by different polyimide films, one that was photosensitive and one that was nonphotosensitive. The photosensitive polyimide had about twice the sensitive to water vapor as the nonphotosensitive polyimide, as indicated by the maximum phase shifts (at 100% relative humidity) of 35° and 18° , respectively. The difference in the responses was attributed to the more open molecular structure of the photosensitive polyimide (which comes in precured form). The higher sensitivity suggests that the photosensitive polyimide would be preferred for a relative humidity sensor. However the long-term stability of the film needs to be studied. A comparison of a SAW humidity sensor with other types of low cost humidity sensors indicated that SAW sensors have the potential for the highest sensitivity at low relative humidities but would probably be more expensive than capacitive or resistive types when signal conditioning circuitry is considered (19).

Polyimide has also been used to measure hydrocarbon and alcohol vapors. Selective hydrocarbon measurement is of high interest to the petroleum refining industry. Table 1 summarizes the results of studies that used SAW microensors coated by nonphotosensitive and photosensitive polyimides to detect three different hydrocarbons, methyl ethyl ketone (MEK) vapor, and water vapor. The maximum SAW

Table 1. Maximum SAW Phase Shift of Polyimide-Coated SAW Microensors for Various Gases^a

Polyimide Type		Water	<i>n</i> -Heptane	<i>n</i> -Octane	Iso-Octane	MEK
Non photosensitive	Initial	6°	5°	2°	2°	140°
	Aged	18°	3°	2°	3°	200°
Photosensitive	Initial	18°	6°	5°	6°	400°
	Aged	37°	27°	17°	25°	1050°

^aRef. 18

microsensor phase responses are shown for polyimide films before and after 4 months of aging (18). The responses were similar in shape to those for water vapor (Fig. 5) but differ in magnitude. Neither polyimide was selective among isooctane, *n*-octane, and *n*-heptane, but the nonphotosensitive polyimide had good selectivity between water vapor and MEK (large responses) versus *n*-octane, *n*-heptane, and isooctane (small responses). Aging had a significant effect on water and MEK responses for both polyimide types. However, only aging significantly affected the heptane and octane responses of the photosensitive polyimide. These results suggest that the structure of the photosensitive film may become more open or that its viscoelastic properties changed due to additional curing, as the polyimide film aged. Other investigators have used a similar polyimide as a light guide (20) and have shown that polyimide film can select between *n*-heptane and isooctane gas molecules. This selectivity was attributed to the different cross-sectional areas of these molecules. The poor selectivity of the SAW microsensor between *n*-heptane and isooctane was attributed to a different molecular structure of the film caused by the differences in film processing or thickness from that of the light guide work. These results illustrate some of the key difficulties in developing appropriate films for chemical sensors, namely, poor selectivity and long-term stability. The most promising approach to the selectivity problem for many gases appears to be the use of a sensor array using pattern recognition such as that used by Microsensor Systems or a chromatography system that uses pattern recognition such as that of Electronic Sensor Systems. There is also room for SAW sensor-based systems designed for specific gases or applications such as the Microconversion Technologies Co. hygrometer.

Polymer Film Characterization: Surface Treatments and Adhesion

The SAW microsensor has been used to characterize the effects of surface treatments on thin polyimide films and as a nondestructive indicator of film adhesion. Surface treatments are of high interest because they are commonly used to modify film properties, particularly surface energy. The surface energy is important because it is directly related to the adhesion of additional layers to the film and to the film's ability to absorb vapors. This is particularly important in the microelectronics industry. The characterization method consisted of measuring changes in the water vapor response of the films as a function of the film parameter of interest.

The effects of plasma and chemical surface treatments on the water uptake of polyimide films are illustrated in Table 2 (21) which shows the maximum water vapor response (100% relative humidity) for polyimide films that were untreated, sputtered, exposed to KOH, and coated by Teflon-AF. The maximum phase shift for untreated film was about 40°. This compares to the smallest response of about 5° for Teflon-AF treated film, to about 12° for sputtered film, to a maximum response of about 80° for KOH treated film. These results indicate that the surface treatments significantly affect the water uptake of polyimide

Table 2. Maximum Water Vapor Response for Polyimide Film Subjected to Various Surface Treatments

Surface Treatment	Phase Change (degrees)
Teflon-AF	5
Argon sputtered	12
None	40
KOH	80

^aRef. 21.

film and that these changes can easily be measured by a SAW microsensor. A small water vapor response may be desirable when polyimide is used as a protective coating, and the large response would be desirable when the polyimide is used as a sensing film.

Adhesion of thin films is directly related to film reliability. Therefore, a method that can measure the adhesion of thin films nondestructively would be extremely useful. The water uptake response of thin polyimide films was examined as a possible nondestructive indicator of film-substrate interfacial characteristics and adhesion. The water uptake response was measured for two polyimide films which were identical except for the surface treatment used to prepare the substrates before film application. For this work, a dual channel SAW microsensor was used because it can directly measure the response *difference* between two films. The experimental setup used for this study is a slightly modified version of the vector voltmeter (phase) setup previously described (Fig. 3). The modifications include applying the signal generator output to both SAW microsensor channels by using a splitter and putting one of the vector voltmeter probes at each of the two output IDTs, as shown in Fig. 6. The *difference* in the water uptake responses of two polyimide films, one applied over silane adhesion promoter and one applied without promoter is shown in Fig. 7 (22). The positive phase shift indicates that less water was absorbed in the film that used promoter. Because the two films were identical except for the interfacial region, these results suggest that the adhesion promoter prevented water from entering the interface and that a significant amount of water was present at the interfacial region of the film/substrate without promoter.

This agrees with neutron scattering studies of water adsorption at similarly treated polyimide/silicon interfaces (23) and suggests that the SAW technique may provide a simple and nondestructive indication of adhesion that could be used in process control.

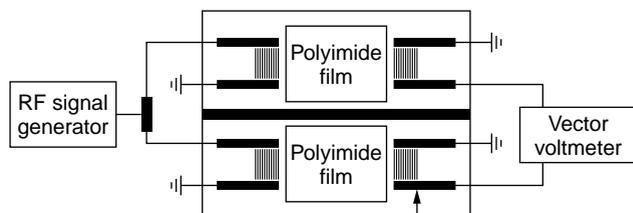


Figure 6. A dual delay line SAW microsensor that has two polyimide film samples and the experimental setup for the comparative phase technique.

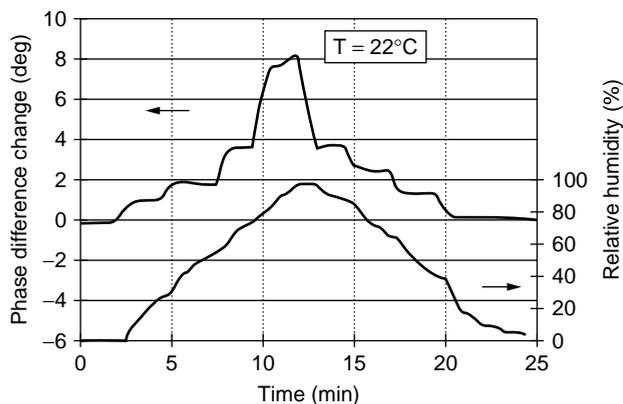


Figure 7. The difference in the water uptake of polyimide films applied with and without silane adhesion promoter (22).

Note that this dual channel design has been widely studied as a method of adjusting SAW microsensors for undesirable effects and can be considered a “smart design.” For example, the most common application of the dual channel design has been for temperature compensation which is necessary because the SAW velocity for many SAW substrates is sensitive to temperature. When the dual channel design, is used in conjunction with the vector voltmeter setup in Fig. 6, it results in canceling the temperature response of the substrate because it is the same for both channels. However, because only one channel can be coated by a sensing film, the temperature response of the film itself is not compensated for.

Polymer Film Characterization: Curing and Glass Transition Temperature

Polymer films are widely used in microelectronics as replacements for more traditional materials such as inorganic coatings on integrated circuits and ceramic printed circuit boards (PCBs). This is due to their low cost, ease of fabrication, and the ability to modify their properties easily to ensure compatibility with fabrication processes. The increased use of thin polymer films in microelectronic applications has resulted in the need for new characterization methods because these films are much smaller and thinner than polymer films used previously and are therefore not always compatible with existing characterization techniques. For example, the curing processes of some new high-temperature polymer films are not fully understood. Of particular interest are changes in mass and viscoelasticity during curing. Thermogravimetry, a common method used to study curing, is the measurement of mass changes caused by outgassing of solvents and other chemical changes in polymers during curing. It involves heating the sample while simultaneously weighing it on a precision balance. The balances currently used can measure mass changes of the order of micrograms. The mass changes in thin polymer films are in the parts per million range, so, a relatively large amount of the polymer must be tested to obtain mass changes that are measurable by these balances. This results in measuring the bulk properties of the polymer which can be significantly different from the

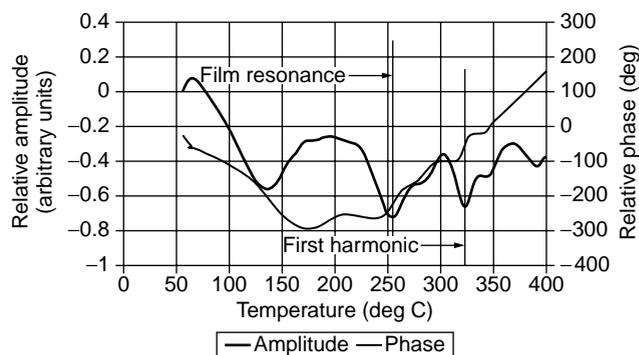


Figure 8. The temperature-compensated phase and amplitude response of a polyimide film during cure (24).

thin film properties. Therefore, a highly sensitive technique is needed to monitor mass and viscoelastic changes in thin polymer films during curing. A surface acoustic wave (SAW) system was developed that can measure the mass lost due to water outgassing during the cure of thin polymer films in a temperature range of 20 to 400°C. It can also measure the apparent glass transition temperature of acoustically thin films and film resonance of acoustically thick films. The principle limitations of the system were the limited accuracy of temperature compensation and the limited ability to separate mass loss effects from viscoelastic effects.

The SAW sensor used was similar to that in Fig. 1, and the polymer film to be tested was applied to the delay path. A sensor test chamber contained the SAW sensor and a heater and allowed dry nitrogen gas purging to prevent water sorption by the polymers. The temperature compensation was done by curve fitting the temperature response data of an uncoated sensor because this provided much better compensation than the dual delay line technique at this high temperature range.

Figure 8 (24) shows the temperature-compensated amplitude and phase responses for a 1.2- μm thick polyimide film. Both the phase and amplitude initially decreased with increasing temperature, indicating that the polymer was softening, until a minimum in amplitude was reached at about 135°C. Because the phase continued to decrease at this temperature, this corresponds to the apparent glass transition temperature (a function of the sensor operating frequency) described by Martin et al. (7). The first film resonance point is indicated by the second amplitude minimum at 255°C because it corresponds to a sharp increase in phase. There was also a phase increase of 43° between 175 and 210°C which was attributed to water outgassing caused by the reaction of the polyamic acid to form polyimide monomers. This agrees reasonably with the predicted 65° phase change based on the expected mass lost due to water outgassing. It also agrees with work done by others (25) which showed that the water outgassing of polyimide during cure occurs between 175 and 225°C. The difference in the measured and theoretical phase change may be accounted for by partial imidization during the soft bake of the polymer application process or by further softening of the polymer. A second resonance point was also

observed at the third amplitude minimum (and a concurrent phase increase) at 335°C. This is the first harmonic of film resonance. This was predicted by others (17), but it was not previously observed because their studies did not exceed 150°C. The further large increase in phase at higher temperatures may indicate the curing reaction of the polyimide in which polyimide monomers combine to form polymer chains. This would cause hardening of the polymer and therefore, an increase in phase.

In summary, this system can measure the mass lost due to water outgassing during cure of thin polymer films to 2% of total polymer mass in a temperature range of 20 to 400°C. It can also measure the apparent glass transition temperature of acoustically thin films, and film resonance including the first harmonic of acoustically thick films. The principle limitations of the system are the accuracy of temperature compensation, how well mass losses can be separated from viscoelastic effects, and how well the glass transition temperature can be separated from acoustic resonance. When water outgasses, the apparent glass transition temperature and film resonance occur at different temperatures as they did in this polyimide study, and the responses are distinguishable. However, should a polymer be tested in which the mass loss occurs at the same temperature as film resonance, the mass loss is most likely to be masked by the larger film resonance response. Therefore, this system can provide a powerful technique for thin polymer film analysis, but the user must have some idea what to expect and the system's limitations. Improvements to this system could include identifying SAW substrate materials by linear temperature responses using lithium niobate or lithium tantalate.

Dew-Point Measurement

The SAW microsensor has proven to be a very useful tool for studying water vapor condensation and measuring the dew point. Dew-point measurement provides one of the most accurate and widely used methods for measuring the absolute water vapor content of a gas. This has traditionally been done by using an optical chilled-mirror, dew-point hygrometer (25). These instruments detect condensation and dew-point by measuring changes in the reflectivity of a condensing surface (26). Although dew-point hygrometers perform better than polymer-film based and other types of resistive and capacitive humidity sensors, they have some drawbacks, including instability due to mirror contamination, inability to detect the frost-point transition, limited resolution, and high cost. A SAW microsensor-based, dew-point hygrometer offers a chilled-surface technique for dew-point measurements that has improved performance at lower cost. The SAW microsensors' small size also suggests its application for dew-point measurements inside small structures such as microelectronic packages where water vapor can affect yield and device reliability. A description of the SAW dew-point hygrometer and some examples of the SAW dewpoint hygrometer's advantages follow.

To study condensation using the SAW microsensor, two changes were made to the previous test setups. First, the frequency technique was used for some of the studies due to its expectedly higher sensitivity. The principal design

difference between a frequency device and the phase device shown in Fig. 1 is a much shorter delay path. This allows only one mode of oscillation. Sensitivity comparisons, based on theory, indicate that a surface density of $1 \mu\text{g}/\text{cm}^2$ results in a 10° SAW phase shift for a 80 MHz phase device versus a 1.43 kHz frequency shift for a 50 MHz frequency device. Assuming a 0.1° phase resolution and 1 Hz frequency resolution for the phase and frequency systems, respectively, the frequency system would provide about 10 times better resolution. The second change was adding a temperature control system which was required to lower the temperature of the SAW sensor until water (dew) condensed on its surface. The temperature at which this occurs is defined as the dew point (26). A thermoelectric cooler and a PC-based data acquisition and control system were included that detected the amount of condensation on the SAW sensor and then maintained the desired amount of condensation so that an accurate reading of the temperature could be made. The SAW velocity change, which corresponds directly to the condensation density on the sensor, was used as the feedback parameter to maintain the predetermined condensation density that was specified via the control software. The desired condensation density was determined by balancing the need for fast response time (less condensation) and minimizing dew-point measurement error (more condensation). Dew-point error is discussed in the section on surface energy and cleaning. The temperature of the sensor was measured by using a resistive temperature device (RTD). A diagram of this system is shown in Fig. 9, and includes an optical microscope that was used to image the condensation and a LED-phototransistor setup that is discussed later.

The ability of the SAW microsensors to measure condensation density accurately was examined by correlating SAW microsensors measurements of condensation with optical microscope images (27). The condensation density (g/cm^2) was determined from the optical images by estimating the total mass of all water drops in a specified area of the image using drop diameter and contact angle information. The contact angle of the water drops was used to obtain drop height. The SAW frequency was linearly related to the condensation density, the sensitivity was about $1.5 \text{ Hz}/\text{ng}/\text{cm}^2$, and the minimum mass resolution was $18.5 \text{ ng}/\text{cm}^2$. This mass resolution is more than an order of magnitude smaller than optical techniques.

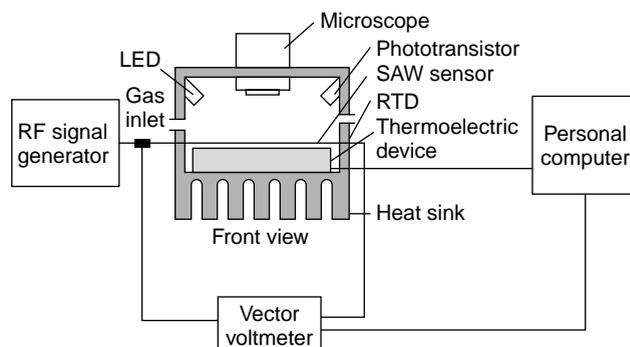


Figure 9. Diagram of the hybrid SAW/optical dew-point measurement system.

A novel hybrid SAW-optical system was developed to obtain meaningful comparisons between the performance of the SAW and optical chilled-mirror, dew-point measurement techniques. This system permitted simultaneous measurement of condensation using both the SAW and optical methods. The design of this system involved adding a light emitting diode (LED), a phototransistor light detector, and an aluminum-mirror film to the SAW sensor delay path. Condensation on the aluminum mirror changed the phototransistor output voltage which was used in a feedback control system, similar to that used for SAW velocity, to maintain the mirror at the dew point. This setup is also shown in Fig. 9. The hybrid SAW/optical system allowed direct comparisons between the SAW and optical chilled-mirror techniques of sensitivity and the effects of surface contamination and the frost-point transition on dew-point measurements. One of the key findings was that the SAW system could maintain a constant condensation density on the sensor without the dew coalescing (small dew drops combining into fewer large drops). This was observed by microscope while controlling the condensation density by using the SAW velocity. Conversely, the optical technique could not maintain a constant condensation density, and coalescence of the dew deposit resulted along with an increase in the SAW phase. Figure 10 (27) shows the reflection voltage and condensation density plotted as a function of time as the dew point was lowered from -10 to -20°C . These data indicate that the condensation density increased by approximately 100% while the reflection voltage was held constant (to within 0.1%). Microscopic imaging indicated that the condensate was dew (supercooled) and not frost. The condensation density change for a constant reflection voltage indicated that reflection voltage was not proportional to condensation density. This lack of a direct relationship can adversely affect control system stability and result in coalescence of the dew which, in turn, can further adversely affect control system stability. These effects typically result in the need for a dry-off cycle when using optical techniques. Therefore, an advantage of the SAW velocity technique is direct measurement of condensation surface density which results

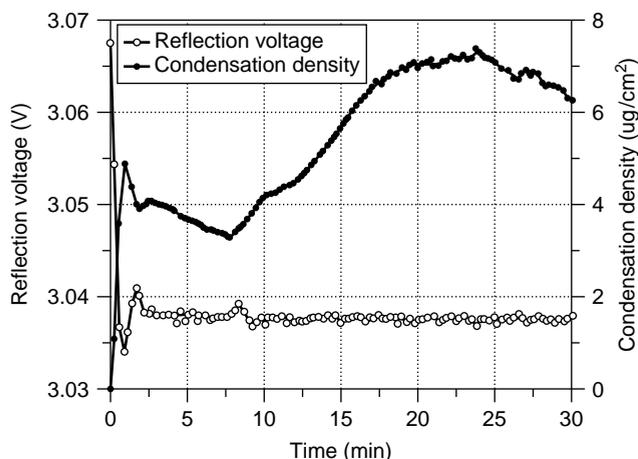


Figure 10. Condensation density variation when controlled by optical detector voltage (27).

in a more stable system than the indirect measurements provided by the optical and SAW attenuation techniques.

The effect of surface contamination on dew-point measurements was also examined. The amount of condensation which caused the same (60 mV) change in the optical detector voltage for clean and contaminated surfaces, respectively, was studied by using SAW phase and microscopic images (27). The clean surface had $5.5 \mu\text{g}/\text{cm}^2$ of condensation versus more than $30 \mu\text{g}/\text{cm}^2$ for the contaminated surface. By comparison, when the SAW sensor was used to measure the dew point, the condensation density (measured with microscopic images) did not change as the surface became contaminated. This indicates that a SAW dew-point sensor can provide more stable and accurate dew-point measurements in dirty environments.

The dew to frost-point transition and its effect on SAW velocity and amplitude and optical detector voltage were also examined (27). The frost-point transition observed by a microscope began at about -23.5°C , as indicated by a mixture of dew and ice crystals on the sensor surface. Therefore, supercooled dew was present before frost formed. The dew became completely frozen at -24.2°C . This transition occurred during a period of about six minutes. Detection of the frost-point transition around -23°C is in agreement with that of other investigators who reported frost-point transition temperatures ranging from -18 to 28°C , depending on the condensation density and gas flow rate (28,29). As the dew deposit froze, it also coalesced, whereas the SAW velocity simultaneously returned (increased) to approximately the same value as that when no condensation was present. This increase was attributed to the different acoustic velocities of ice versus water and the observed coalesced state of the frost deposit, which resulted in a much smaller contact area with the sensor surface. This resulted in significant instability of dew-point measurements during the frost-point transition. Similar instability occurred when controlling the condensation density by using the optical reflection voltage. However, it was found that SAW amplitude could be used to maintain constant condensation density during the frost-point transition, thus providing continuous dew-point measurements, as well as an indication of the frost-point transition.

The resolution of the SAW dew-point hygrometer (MCT) was compared with those of EG&G (now Edgetech) Model 2000 and General Eastern Model Hygro-M3 optical dew-point hygrometer. Figure 11 (13) shows a constant dew point measured by the three hygrometers. The resolution of the SAW microsensors was about $\pm 0.02^{\circ}\text{C}$ versus $\pm 0.2^{\circ}\text{C}$ for the EG&G and $\pm 0.5^{\circ}\text{C}$ for the General Eastern. The superior performance to the SAW hygrometer was attributed to its direct and more precise measurement of condensation density. The accuracy differences were expected to be resolved by recalibration.

The Measurement of Surface Energy and Effects of Surface Cleaning

A SAW microsensors technique based on the dew-point system was also examined as a novel method for measuring surface energy and the effects of surface cleaning. The

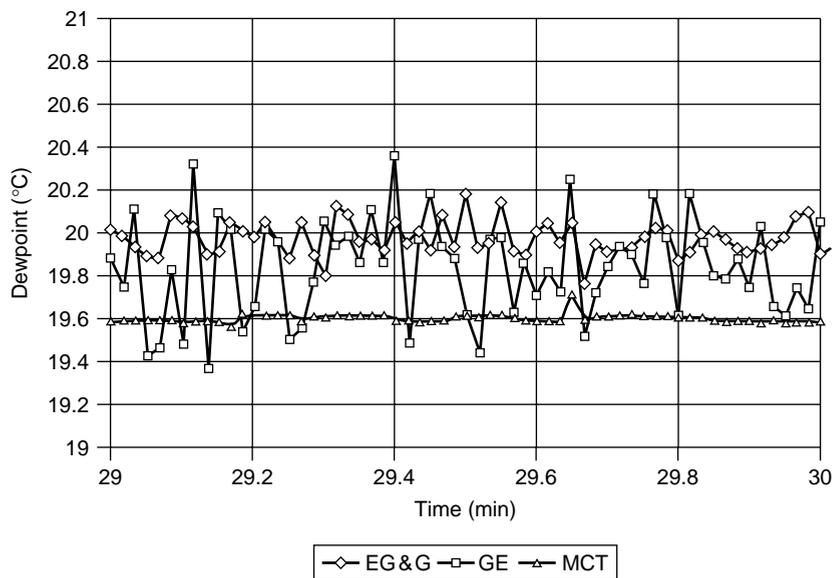


Figure 11. The dew point measured by the SAW-based hygrometer and two commercial optical dew-point hygrometers (EG&G Model 2000 and General Eastern Model Hygro-M3) (14).

measurement of the surface energy of thin metal and polymer films is of high interest in the microelectronics industry due to its relationship to film cleanliness and adhesion and ultimately, to microelectronic device reliability (31). Surface energy is also important for determining the biocompatibility of materials and for developing a better understanding of gas sorption on sensing surfaces. Surface energy is defined as the energy spent to create a surface and is closely related to the reactivity or wettability of that surface. Wettability indicates how well a liquid spreads across a solid surface (32). The presence of organic films or contamination on a surface can also be measured because they usually change the surface energy of a material. Plasma cleaning is commonly used to remove organic and mobile ion contamination from integrated circuits and multichip modules (MCMs) before encapsulation, as an alternative to solvents and detergents that typically contain chlorofluorohydrocarbons (CFCs) and other hazardous materials. Plasma treatments have also been used to improve the adhesion of conductors and insulators by changing the surface structure and energy of the metal or polymer layer (33). Improved methods for measuring the results of plasma treatments and cleaning are important for optimizing these processes.

The traditional method for obtaining information about contamination, surface energy, and wettability of solid materials and the effects of plasma treatments is contact angle measurements. The contact angle is the angle created at the liquid–solid–vapor interface when a drop is placed on a solid surface. A surface that has a small contact angle wets better and has a higher surface energy than a surface that has a large contact angle. Small and large contact angles and the effects of surface treatments on them can easily be observed by examining the interface of water drops on the hood of a car before and after it is waxed, respectively. Contact angle is commonly measured by dispensing a fluid on the surface in question and observing the drop by using a goniometer. The advancing angle is measured as the drop volume is increased, and the receding angle is measured as

the drop volume is decreased. The difference between the advancing and receding angle is known as contact angle hysteresis. Contact angle hysteresis is caused by surface microroughness and heterogeneity and by chemical interactions between the surface and the fluid (32). The principal limitations of contact angle measurements are hysteresis, difficulty in measuring angles less than 20° , and poor reproducibility resulting from variations in operator technique. Detailed reviews of contact angle and surface energy are provided in (32,34).

The purpose of these SAW microsensor studies was to determine if a technique for measuring the surface energy and the effects of surface treatments for common electronic materials could be developed that did not suffer from the errors typical of contact angle measurements. The experimental technique used was to measure the dew point while maintaining extremely small condensation densities on the sensor surface. This resulted in a measured dew point that was higher than the true dew point, dewpoint error, which can then be compared for various surfaces. Alternatively, the condensation density on various surfaces at a specific temperature above the dew point can also be compared. The ability of the SAW microsensor to measure a dewpoint error was first demonstrated by coating a SAW sensor with known highly hydrophobic (low surface energy and high contact angle) or hydrophilic (high surface energy and low contact angle) materials. Films of Teflon-AF and polyimide $1\ \mu$ thick were used, respectively. Figure 12 shows the deviation from the dew point (dew-point error) as a function of condensation density for quartz, polyimide, and Teflon-AF surfaces (27). The dew-point error was greater for the hydrophilic material (polyimide) than for the hydrophobic material (Teflon-AF).

For example, at a condensation density of $0.2\ \mu\text{g}/\text{cm}^2$, the dew-point error for polyimide was approximately 2°C versus 0°C for the Teflon-AF. This inversely correlates with the contact angles of 30° and 90° for polyimide and Teflon-AF, respectively, as expected. Noted that the temperature of the SAW sensor was at the true dew point for all surfaces

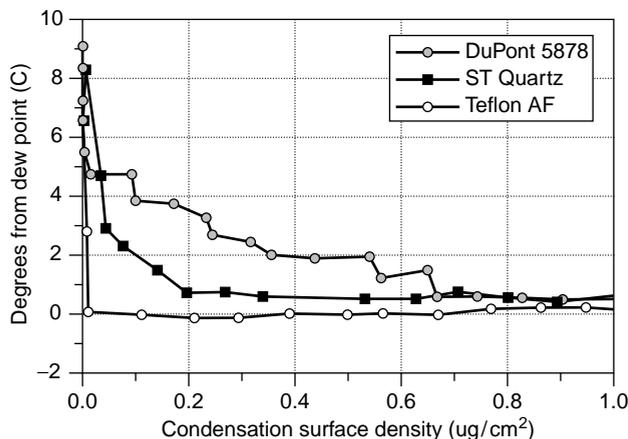


Figure 12. Deviation from the dew point (dew-point error) as a function of condensation density for quartz, polyimide, and Teflon-AF surfaces (27).

when the condensation density exceeded $0.8 \mu\text{g}/\text{cm}^2$. A dew-point measurement error does not occur when large dew densities are present because condensation forms on existing drops of water.

The ability of SAW dew-point error measurements to measure surface energy and the effects of various surface treatments on quartz is illustrated in Table 3 (35), which shows contact angles and dew-point errors at $0.1 \mu\text{g}/\text{cm}^2$ for several sensor surface treatments listed in order of decreasing contact angle. There appears to be reasonably good correlation between decreasing contact angle and increasing dew-point error, considering the previously mentioned limitations of contact angle measurements. The most significant deviation was HCl which had the largest contact angle but not the smallest dew-point error. Contamination between experiments is a possible source of error for all measurements. These results suggest that dew-point error may provide a measurement of surface energy which does not suffer from hysteresis, however, more work is necessary in this area to establish a more definitive relationship.

Temperature Measurement

Acoustic temperature sensors (ATSs) use the temperature dependence of the acoustic velocity in a piezoelectric substrate to measure temperature. Their principle advantages

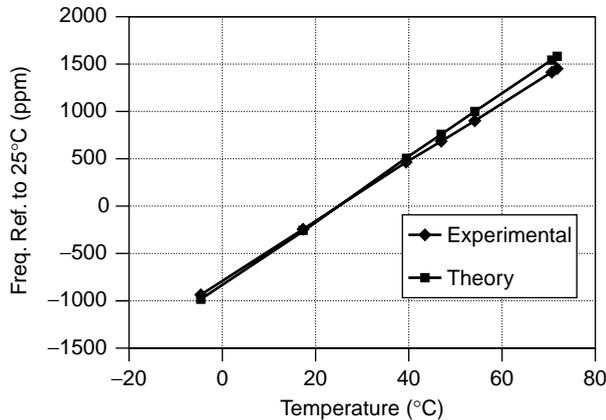


Figure 13. The frequency versus temperature response for an SSBW ATS (36).

over other types of temperature sensors are ease of integration with other acoustic sensors, low self-heating, robustness, wide temperature range, a digital form of output, low noise susceptibility when used in an oscillator, relatively low cost, high resolution, and short response times. Their main disadvantage is that they require hermetic packaging to prevent inaccuracies caused by sensor contamination. This slows the response time and increases cost. Surface skimming bulk wave (SSBW) devices have been found much less sensitive to surface contamination than other acoustic devices because the waves have horizontal displacements and can propagate just beneath the surface of the substrate. Therefore, they may not require hermetic packaging. However, these devices have not been studied as temperature sensors. The objectives of this work were to study the temperature characteristics and contamination sensitivity of an SSBW ATS. The introductory section provides a description of this wave.

Figure 13 shows the frequency versus temperature response for an ATS characterized in a Styrofoam chamber (36). A second-order curve fit yielded a first-order temperature coefficient of frequency (TCF) of approximately $31.5 \text{ ppm}/^\circ\text{C}$. The temperature deviation from the second-order curve fit indicated a resolution of $\pm 0.22^\circ\text{C}$ across a 78°C temperature range. The TCF of $31.5 \text{ ppm}/^\circ\text{C}$ could provide a theoretical temperature resolution of better than 0.0003°C if a 1 Hz frequency resolution is available. Figure 14 shows the test setup used to compare the

Table 3. Contact Angles and Dew-Point Error for Several SAW Sensor Surface Treatments^a

Treatment	Advancing Contact Angle (degrees)	Receding Contact Angle (degrees)	Dew-Point Error (degrees phase)
HCl	88	59	1.8
Chromic acid	71	46	0.3
KOH	54	29	1.2
Microsolution	26	15	8.7
Argon sputtering	25	14	7.0
Oxygen-argon sputtering	10	6	4.5

^aRef. 25.

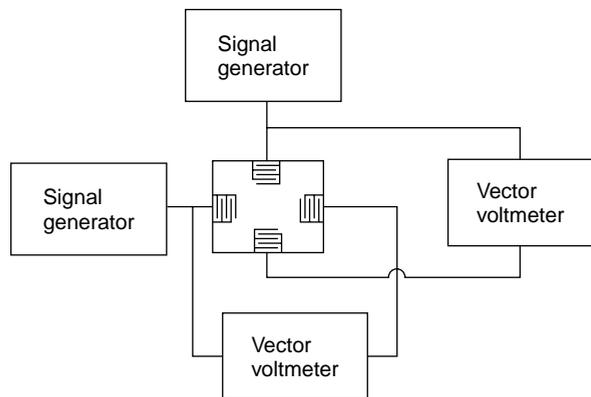


Figure 14. The test setup used to compare the mass sensitivities of SSBW and SAW microsensors (36,37).

mass sensitivities of SSBW and SAW acoustic delay lines (36). A novel dual delay line design was used where a SAW channel was aligned at 90° to the SSBW channel (37). The phase shift of the SSBW due to mass loading was determined by maintaining a fixed surface density of condensed water by using the SAW device (as described in the previous section) while monitoring the phase shift of the SSBW device using a vector voltmeter. Figure 15 (36) shows the phase shift due to mass loading for the SSBW and SAW delay lines. The SSBW phase shift was more than an order of magnitude less than that of the SAW for the same mass loading. Assuming a TCD of $32 \text{ ppm}/^\circ\text{C}$, $5 \mu\text{g}/\text{cm}^2$ of contamination would result in a temperature error of 16°C for SAW versus 0.4°C for SSBW ATs, respectively. The SSBW response was attributed more to the effect of the water on IDT capacitance than mass loading, and therefore, it is most likely that it can be significantly reduced by a protective coating. These results suggest that SSBW ATs may have applications in acoustic chemical sensing where temperature information is also desired. Note that ATs do require calibration and that the procedures are not simple because SAW microsensor calibration can be affected by the electronic circuitry used.

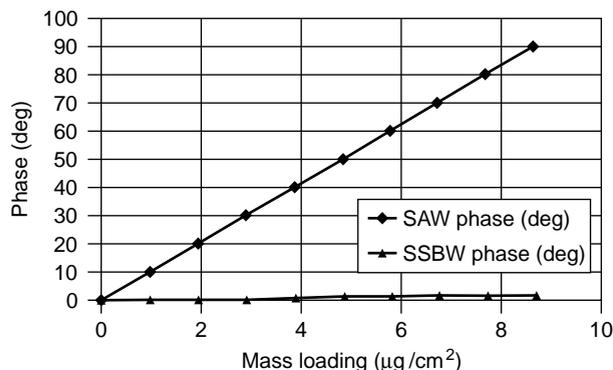


Figure 15. The SSBW phase versus SAW phase due to mass loading (36).

CONCLUSIONS

It has been shown that the SAW microsensor provides a highly sensitive and versatile sensing platform for both commercial applications and as a laboratory tool. Of the systems and applications described in this article, the SAW microsensor-based dew-point hygrometer is the only one that is currently commercialized. It offers clear advantages over existing chilled-mirror technology, including significantly better dew-point measurement resolution and stability. This results from the SAW microsensor's high sensitivity and its ability to measure condensation density directly. This ability also prevented ripening and coalescence of the condensate which makes the instrument more stable. The SAW technique can also detect the frost-point transition and is much less susceptible than optical systems to instability caused by surface contamination.

The ATS is the application most likely to be commercialized next. It will be most useful when used in conjunction with other SAW-based chemical sensors due to cost considerations. Commercial applications for coated SAW microsensors beyond those described in the introduction hinge on the other key chemical sensor requirements of selectivity and long-term reliability. Further advances here are most likely to come from improved film chemistry which has been a relatively slow process. The high sensitivity of the SAW microsensor to condensation at temperatures slightly above the dew point and the lack of hysteresis of the dew-point measurement suggest that it may have advantages over contact angle for measuring surface energy and for characterizing certain polymers, metal surfaces, and surface treatments, if further developed. Possible applications of this SAW microsensor system include *in situ* monitoring of surface and interfacial treatments for process control in microelectronics manufacturing. Similarly, the use of a SAW microsensor for characterizing thin polymer films is also promising but requires further development.

BIBLIOGRAPHY

1. S.M. Sze, *Semiconductor Sensors*. Wiley, NY, 1994.
2. C. Campbell, *Surface Acoustic Wave Devices and Their Signal Processing Applications*. Academic Press, Boston, 1989.
3. D.P. Morgan, *Surface-Wave Devices for Signal Processing*. Elsevier, Amsterdam, 1991.
4. B.A. Auld, *Acoustic Fields and Waves in Solids*. Krieger, Malabar, FL, 1973, 1990.
5. G.S. Kino, *Acoustic Waves: Devices, Imaging, and Analog Signal Processing*. Prentice-Hall, NJ, 1987.
6. H. Wohltjen, *Sensors and Actuators* **5**: 307–325 (1984).
7. S.J. Martin, G.C. Frye, and S.D. Senturia, *Anal. Chem.* **66**: 2201–2219 (1994).
8. P. Das, C. Lanzl, and D. Barone, *IEEE Ultrasonics Symp.*, Sept. 1978, Piscataway, NJ, pp. 458–463.
9. H. Wohltjen, *Anal. Chem.* **51**(9): 1458–1475 (1979).
10. A. Bryant, D.L. Lee, and J.F. Vetelino, *IEEE Ultrasonics Symp.*, 1981, pp. 171–174.

11. X.Q. Bao, W. Burkhard, V.V. Varadan, and V.K. Varadan, *IEEE Ultrasonics Symp.*, Denver, CO, Oct. 1987, pp. 583–585.
12. Sawtek, Orlando, FL 32860, www.microsensorsystems.com.
13. Electronic Sensor Technology, Newbury Park, CA www.estcal.com.
14. Microconversion Technologies Co., Brookings, SD 57006, www.brookings.net/~mctco/.
15. J.C. Andle and J.F. Vetelino *IEEE Ultrasonics Symp.*, 1995, pp. 451–460.
16. D.S. Ballantine, R.M. White, S.J. Martin, A.J. Ricco, E.T. Zellers, G.C. Frye, and H. Wohltjen, *Acoustic Wave Sensors*. Academic Press, San Diego, 1997.
17. M. Thompson and D.C. Stone, *Surface-Launched Acoustic Wave Sensors*. Wiley, NY, 1997.
18. P.R. Story, D. Nesthus, R.D. Mileham, C. Feger and D.W. Galipeau, *Soc. Plast. Eng. Annu. Tech. Conf.* Indianapolis, IN, 1996, pp. 2148–2151.
19. P.R. Story, D.W. Galipeau, and R.D. Mileham, *Sensors and Actuators B* **25**: 681–685 (1995).
20. R.P. Podgorsek, H. Franke, and C. Feger, *Opt. Lett.* **20(5)**: 501–503 (1995).
21. D.W. Galipeau, P.R. Story, C. Feger, and K.-W. Lee, *J. Adhesion Sci. Technol.* **8**: 1–13 (1994).
22. D.W. Galipeau, J.F. Vetelino, and C. Feger, *J. Adhesion Sci. Technol.* **7**: 1335–1345 (1993).
23. W.-L. Wu and M.S. Kent, *Mater. Res. Soc. Symp. Proc.* **338**: 565–569 (1994).
24. R.D. Mileham, J.D. Sternhagen, and D.W. Galipeau, *IEEE Trans. Ultrasonics Ferroelectrics, and Frequency Control* **45**: 1273–1280 (1998).
25. J.J. McAndrew and D. Boucheron, *Solid State Technol.* 55–60 (Feb. 1992).
26. R.G. Wylie, D.K. Davies, and W.A. Caw, in *Humidity and Moisture*, Vol. 1, A. Wexler and R.E. Ruskin, eds., Reinhold, NY, 1965, pp. 125–134.
27. K.A. Vetelino, P.R. Story, R.D. Mileham, and D.W. Galipeau, *Sensors and Actuators B* **35–36**: 91–98 (1996).
28. L.C. Paine and H.R. Farrah, in *Humidity and Moisture*, Vol. 1, A. Wexler and R.E. Ruskin, eds., Reinhold, NY, 1965, pp. 174–188.
29. T.I. Inamatsu, C. Takahashi, and T. Furuya, *Proc. 1985 Int. Symp. Moisture Humidity*, Washington, DC, 1985, pp. 379–382.
30. EG&G Moisture and Humidity Systems, EG&G Dewpoint Error Analysis, EG&G Applications Data, EG&G Moisture and Humidity Systems, Burlington, MA, 1976.
31. K.A. Vetelino, P.R. Story, B.L. DeClercq, and D.W. Galipeau, *Int. J. Microcircuits Electron. Packag.* **19**: 212–218 (1996).
32. R.J. Good, in *Contact Angle, Wettability and Adhesion*, K.L. Mittal, ed., VSP, Utrecht, The Netherlands, 1993.
33. C.P. Wong, *Polymers for Electronic and Photonic Applications*. Academic Press, San Diego, 1993.
34. S. Wu, *Polymer Interface and Adhesion*. Marcel Dekker, NY, 1982.
35. L.D. Eske and D.W. Galipeau, *Colloids Surf. A*, **154**: 33–51 (1999).
36. C.E. Wold, J.D. Sternhagen, R.D. Mileham, K.D. Mitzner, and D.W. Galipeau, *IEEE Ultrasonics Symp.*, 1999.
37. US Pat. Pending, 1999, D.W. Galipeau and R.D. Mileham.

SHAPE MEMORY ALLOYS, APPLICATIONS

CAROLYN RICE
Cordis-NDC
Fremont, CA

INTRODUCTION

Shape-memory alloys have been engineered for applications and devices since the first discovery of the shape-memory effect in the 1930s. The majority of this design activity was initiated by the discovery of Nitinol (nickel–titanium alloy) in 1962, and since then more than 10,000 patents have been issued for applications using shape-memory alloys (1). This article reviews a number of these applications, discusses aspects of design, and illustrates representative examples.

TYPES OF SHAPE-MEMORY ALLOYS

General Theory

Shape-memory alloys (SMAs) are known primarily for one fundamental and unique property—the ability to remember and recover from large strains without permanent deformation. Unlike most conventional metals that recover less than 1% strain before plastic deformation, SMAs undergo a diffusionless, thermoelastic martensitic phase transformation that enables the material to deform via a twinning process rather than the conventional dislocation slip mechanism and allows complete recovery of strains as large as 8%. The metallurgical phenomena that explain these martensitic transformations are detailed in many sources (2–5). The discussion to follow includes only a brief summary of SMA behavior to review its basic properties.

Shape-Memory Effect. The most well-known form of transformation behavior exploited in SMAs is thermally induced shape change, often labeled the shape-memory effect (SME). A material component may be deformed, or strained, at low temperatures, and when heated, it reverses this strain and remembers its prestrained shape. The low-temperature, deformable martensite phase transforms to a more stable austenite phase at higher temperatures. This transformation occurs across a temperature range, known as a transformation temperature range (TTR). This range for Nitinol (Ni–Ti) is approximately 30 to 50°C, and is also known as temperature hysteresis. The A_s (austenitic start) temperature is the beginning of the transformation to austenite upon heating, the A_f (austenitic finish) is the finish of the transformation to austenite, M_s (martensitic start) is the beginning of the martensitic transformation upon cooling, and M_f (martensitic finish) is the finish of the transformation to martensite. A typical SMA stress–strain curve, depicted in Fig. 1a, demonstrates shape-memory behavior at temperatures below the material M_f temperature. A schematic example of a shape-memory application is shown in Fig. 1b.

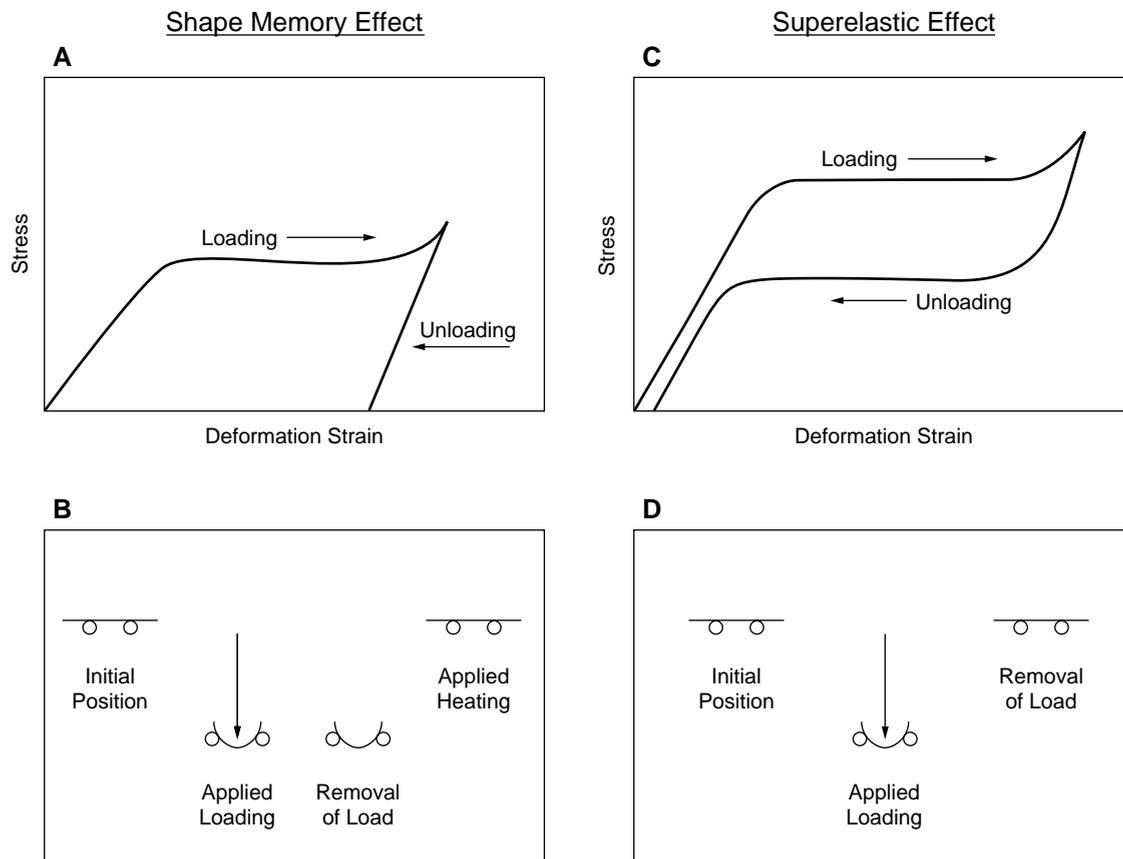


Figure 1. Typical SMA behavior in tensile tests and bending applications: (a) stress-strain curve of shape memory (martensite) material, (b) schematic of a shape memory application, (c) stress-strain curve of superelastic (austenite) in tension, (d) superelastic behavior in a bending application.

Shape-memory alloys may also be trained to exhibit a two-way shape memory effect. Similar to the thermal shape-memory effect, two-way shape memory (TWSM) requires special thermomechanical processing to impart shape memory in both martensitic and austenitic phases. A trained shape in the austenitic phase reverts to a second trained shape upon cooling, allowing the material to cycle between two different shapes. This TWSM is theoretically ideal for many shape-memory applications; however, practical uses are limited due to behavior instability and complex processing requirements.

Superelastic Effect. This effect, known also as pseudoelastic, describes material strains that are recovered isothermally to yield mechanical shape-memory behavior. The phenomenon is essentially the same as the thermal shape-memory effect, although the phase transformation to austenite (A_f) occurs at temperatures below the expected operating temperature. If the austenitic phase is strained by an applied load, a martensitic phase is induced by stress, and the twinning process occurs as if the material had been cooled to its martensitic temperature. When the applied load is removed, the material inherently prefers the austenitic phase at the operating temperature, and its strain is instantly recovered. A typical stress-strain

curve is depicted in Fig. 1c, and a schematic example of a superelastic application is shown in Fig. 1d. The stress-strain curve indicates a difference in stress levels during loading and unloading, that is known as superelastic stress-strain hysteresis.

Alloys

Several alloys have been developed that display varying degrees and types of shape-memory behavior. The most commercially successful have been Ni-Ti, Ni-Ti-X and Cu-based alloys, although Ni-Ti and ternary Ni-Ti-X alloys are used in more than 90% of new SMA applications (6). Ni-Ti alloys are more expensive to melt and produce than copper alloys, but they are preferred for their ductility, stability in cyclic applications, corrosion resistance, biocompatibility, and higher electrical resistivity for resistive heating in actuator applications (6).

The most common Cu-based alloys, Cu-Al-Ni and Cu-Zn-Al, are used for their narrow thermal hysteresis and adaptability to two-way memory training. Ni-Ti ternary alloys are used to enhance other parameters. Examples include Ni-Ti-Nb for wide thermal hysteresis, Ni-Ti-Fe for extremely low TTR, Ni-Ti-Cr for TTR stability during thermomechanical processing, and Ni-Ti-Cu for narrow thermal hysteresis and cyclic stability (7).

Material Forms

SMA are manufactured in many of the conventional forms expected of metal alloys: drawn round wire, flat wire, tubing, rolled sheet, and sputtered thin films. Additional forms include shaped components, centerless ground tapered wires and tubing, alternate core wire (Ni–Ti filled with a conductive or radiopaque material), PTFE coated wire, stranded wire, and embedded composites. At present, Ni–Ti–X alloys are the most readily available in all of these forms.

The processing of SMA material is critical for optimizing shape-memory behavior. Many adjustments can be made to optimize the properties of a material form for a particular application; however, most efforts are made to optimize a balance of strain recovery, ductility, and tensile strength. SMAs such as Ni–Ti are melted using extreme purity and composition control, hot worked to bars or plates, cold worked to their final form, and subjected to specialized thermomechanical treatments to enhance their shape-memory properties.

DESIGNING WITH SHAPE MEMORY ALLOYS

Shape-memory alloys have intrigued engineers and inventors for more than 30 years. One might conclude from the large number of SMA patents that have been issued and the knowledge that relatively few of the ideas have been commercially successful that the majority of these designs have not fully accounted for the unique behaviors, limitations, and constraints of SMAs. The focus of this section is to highlight the properties best used in SMA applications and to discuss SMA design considerations.

Functional Properties

SMA applications are often categorized in terms of the specific material property used. The majority of these properties are either thermal shape memory or mechanical shape memory (superelastic), but some unique properties are only indirectly related to these shape-memory effects. General categories of applications are classified according to these properties.

Shape Memory. The thermally activated ability of a shape memory material to change shape yields several types of applications that can be summarized in three distinct categories: applications that use the shape change to display motion, those that actuate, and those that harness stresses produced from constraining the recovery of the shape-memory material.

Displayed motion, also referred to as free recovery, describes applications that exploit the pure motion of thermal shape memory (8). An example of this application, a moving butterfly, is displayed in Fig. 2. These butterflies, produced by Dynalloy, Inc., use a specially processed form of Ni–Ti wire to move wings back and forth for thousands of cycles without significant signs of fatigue. This processed wire, known as Flexinol™, changes shape via cyclic heating by electric current. The small mass of the butterfly body is sufficient to extend the Ni–Ti wire when cooled, but the

Ni–Ti wire can contract and close the wings when heated to its stronger austenitic shape.

Actuation applications are designed to perform work. A simplified example is a mass suspended from a shape-memory tension spring. When cooled, the weaker martensitic phase deforms, and the spring is extended by the mass. When heated to austenite, the spring recovers its shape with forces sufficient to lift the weight, resulting in actuation that performs work.

Constrained recovery applications use the change in material strength from martensite to austenite to produce a stress that can be harnessed as a clamping force. A popular example of a constrained recovery application is a shape-memory coupling which is expanded at low temperatures, then heated to shrink and clamp to join two pipes.

Superelasticity. Unlike thermal shape-memory applications, which can be categorized into several types, applications that exploit this mechanical shape memory are defined as those that require high strain recovery at

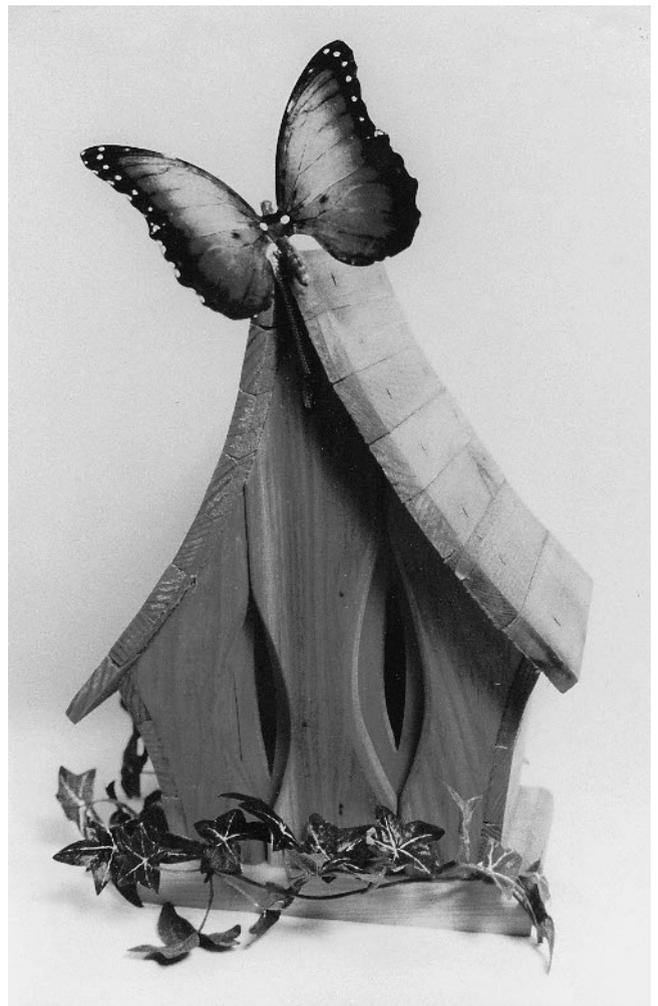


Figure 2. Photograph of a Flexinol™ actuated butterfly (courtesy of Dynalloy, Inc.).

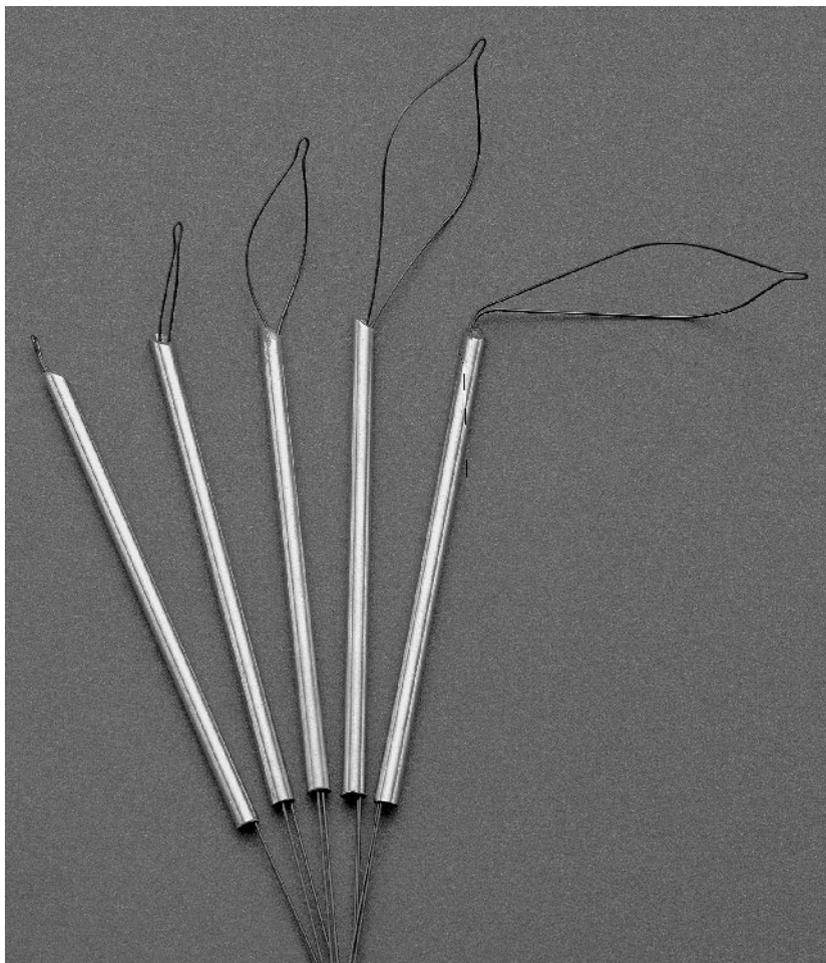


Figure 3. Suture retrieval loops designed to recover their shape once deployed from a 6 fr. cannula (courtesy of Shape Memory Applications, Inc.).

operating temperatures. Many examples of applications that use superelasticity are found in the medical industry (Fig. 3), but one of the most well known is found in consumer eyeglass frames marketed as Flexon[®] by Marchon Eyewear, Inc. (Fig. 4).

Energy Absorption for Vibrational Damping. An energy absorbing ability found in both the martensitic and austenitic phases of SMAs is indirectly related to their shape-memory behavior. The energy absorption of SMA materials has demonstrated excellent vibrational damping characteristics, which can be harnessed for use in various damping applications. The types of devices that exploit this property are classified in three categories of damping: martensitic, martensitic transformation, and superelastic.

Martensitic damping devices operate by using only the martensitic phase of SMAs. Energy is absorbed by the martensite during its twin reorientation process, and accommodates large strains for high-amplitude, low-frequency

loading. They offer the best damping characteristics of the three categories, and although they cannot recover large strains without subsequent heating, they provide excellent damping properties across a broad temperature range.

Martensitic transformation damping elements are designed to operate near martensitic transition temperatures for peak performance in vibrational attenuation. This peak is due to a sharp increase in internal friction during the martensitic phase transformation. These damping elements offer ideal properties for low-amplitude, high-frequency vibrations within a small operating temperature range (9). This type of device could be used in ski materials to damp vibrations when the ski is in contact with snow (6).

Superelastic damping devices use the plateau hysteresis portion of the stress-strain curve for properties similar to those of a rubber band. Superelastic SMA materials are pretensioned to reach this stress-strain plateau, and any additional strains are accommodated easily by changes in

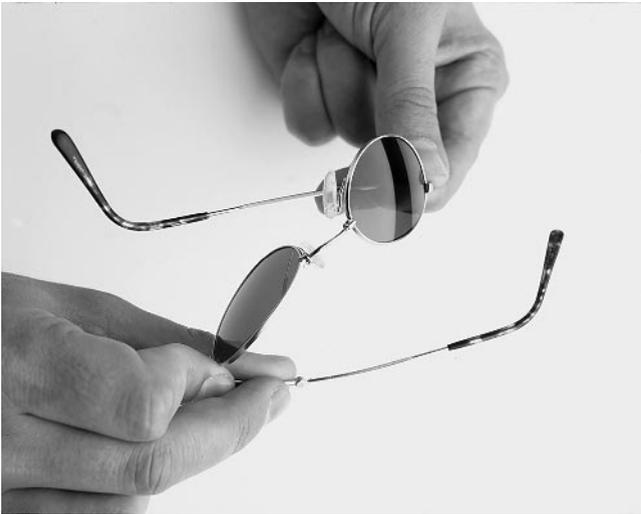


Figure 4. Deformation resistant eyeglass frames (courtesy of Marchon Eyewear, Inc.).

the applied load. This property optimizes a combination of damping capacity, shape recovery, and temperature range of operation (9). Unlike martensitic damping elements, superelastic devices recover their original shape when vibrational loading is removed. Suggested superelastic tensioning devices are presented in a U.S. patent regarding hysteretic damping (10); one example is shown in Fig. 5 (9). These SMA tension elements cycle through a superelastic stress-strain hysteresis to dissipate energy and serve as a damping mechanism. Vibrations due to environmental impacts such as violent winds and earthquakes deform the tensioned elements, and when the vibrational impact is lessened, the elements recover their shapes.

Cavitation-Erosion Resistance. Cavitation erosion is a phenomenon that affects equipment and machinery in many industries. Small bubbles explode with large

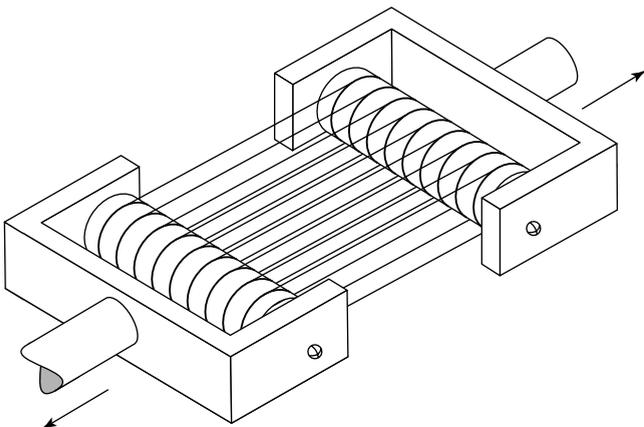


Figure 5. Schematic of a superelastic damping device, using loops of SMA wire in tension. Reprinted with permission from D.E. Hodgson and R.C. Krumme, *Damping in Structural Applications, SMST Proceedings*, 1994.

impacts, causing pitting and erosion in metallic surfaces and reducing the service life of expensive equipment. Both the martensitic and austenitic phases of SMAs have displayed cavitation-erosion resistance; they recover from impact and minimize material loss when exposed to vibratory cavitation. Studies that explored the performance of Ni-Ti on stainless steels have indicated that both martensitic and austenitic Ni-Ti have significant potential for covering and protecting equipment that suffers wear from cavitation erosion. Ni-Ti cladding could be used in applications such as machinery, hydraulics, large hydroelectric generator turbines, and ship propellers (11).

Low Elastic Modulus. The martensitic phase of SMA materials is soft and pliable, in contrast to the stiff, springy characteristics of the austenitic phase. This softness, or low effective (nonlinear) elastic modulus, is often used in applications that require deformability and excellent fatigue characteristics. This property is exploited alone or in conjunction with a shape-memory effect in fatigue-resistant applications.

An example of a low elastic modulus application is shown in Fig. 6: a martensitic tool developed by St. Jude Medical, Inc., is used by surgeons during open heart surgery to orient a tissue-restraining device. During this procedure, surgeons must make adjustments to optimize the tool geometry for each patient, and the use of SMAs allow surgeons to bend the martensitic handle to an appropriate angle. Upon completing the operation, the tool is sterilized in an autoclave where it is exposed to elevated temperatures and reverts to its trained, austenitic shape. Due to its ability to recover large strains repeatedly, these tools are marketed for both fatigue resistance and shape-memory properties.

Design Constraints and Considerations

When assessing a potential design challenge, designers are often anxious to develop a solution that uses the unique and exciting properties of SMAs. It is critical, however, for designers to understand the complexity of SMA behavior. As a general rule, if conventional materials and designs can be applied to yield an acceptable and desirable result, the use of SMAs to provide an alternative solution will increase complexity and cost. SMAs are best used when their unique properties are necessary for design success—when conventional materials cannot meet the demands of the application.

The design of SMA applications requires more than traditional design techniques and textbook methods. Due to the many unique properties of SMA materials, several considerations specific to SMA design must be addressed and accounted for. This section discusses the majority of issues that should be addressed before designing an application using SMAs.

General Guidelines

Recoverable Strain. The expected recoverable strain of SMA material must be within the limitations of the alloy

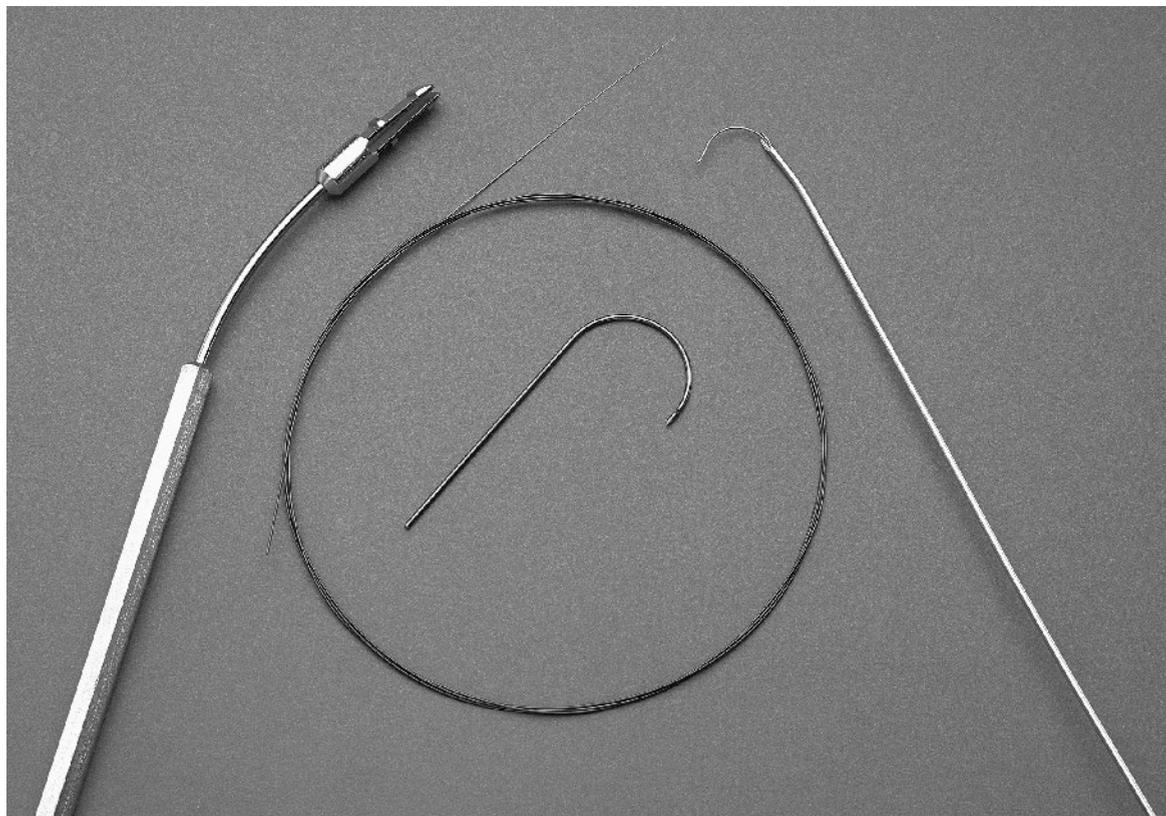


Figure 6. Medical tools and devices. *Left to right:* Flexible martensitic, Ni-Ti handle developed by St. Jude Medical for open heart surgery procedures; superelastic, tapered guidewire core; shaped Ni-Ti tubing component; retractable, superelastic component within a small diameter cannula (courtesy of Shape Memory Applications, Inc.).

chosen. For example, Ni-Ti may recover 8% strain for a single cycle application, but less than 4% for higher cycle applications. Recommended strain limits are 6% for Ni-Ti and 2% for Cu-Zn-Al for lower cycle applications and 2% and 0.5% for higher cycle applications, respectively (3). A maximum strain recovery of about 2% is expected for applications that require two-way shape memory (12).

High Temperature Stability. Alloy stability must be considered when an application requires or will be subject to high operating temperatures. Ni-Ti alloys tend to be the most stable of all SMAs at elevated temperatures and can withstand exposure to temperatures up to approximately 250°C before previously memorized shaping is deleted. For Cu-Zn-Al, this maximum temperature is around 90°C.

Fatigue. SMA fatigue can be defined as degradation of any or all of its functional properties. SMAs affected by application cycle quantity, frequency, temperature range, stress, and strain may fatigue by fracture, decreasing recoverable strain, shifting transformation temperatures, or decreasing recovery stress.

Manufacturability. SMA materials are infamously difficult to machine. Tool wear is rapid for conventional

machining methods such as turning, milling, drilling, and tapping (2). Currently, the most successful machining techniques include surface grinding, abrasive cutting, EDM, and laser cutting. Component shaping must be considered as well; the memories of SMA shapes are trained at high temperatures, typically around 500°C (932°F). Unlike most conventional materials that may be cold formed, SMAs must be rigidly clamped in a desired shape and exposed to these elevated temperatures.

Designing for assembly is also an important manufacturing consideration. Fastening SMAs to other materials by bonding and joining presents additional challenges. Because SMAs are designed to exhibit strains up to 8% and other materials have a strain limit of less than 1%, when the two are rigidly joined, the conventional material may break during operation. This often causes problems in using plated or painted SMA materials because the coating on the SMA will often crack and flake during the strains of operation.

Cost. Most SMAs are inherently more expensive than conventional materials due to the higher cost of both raw material components and processing methods. The positional control necessary for the raw forms of SMA

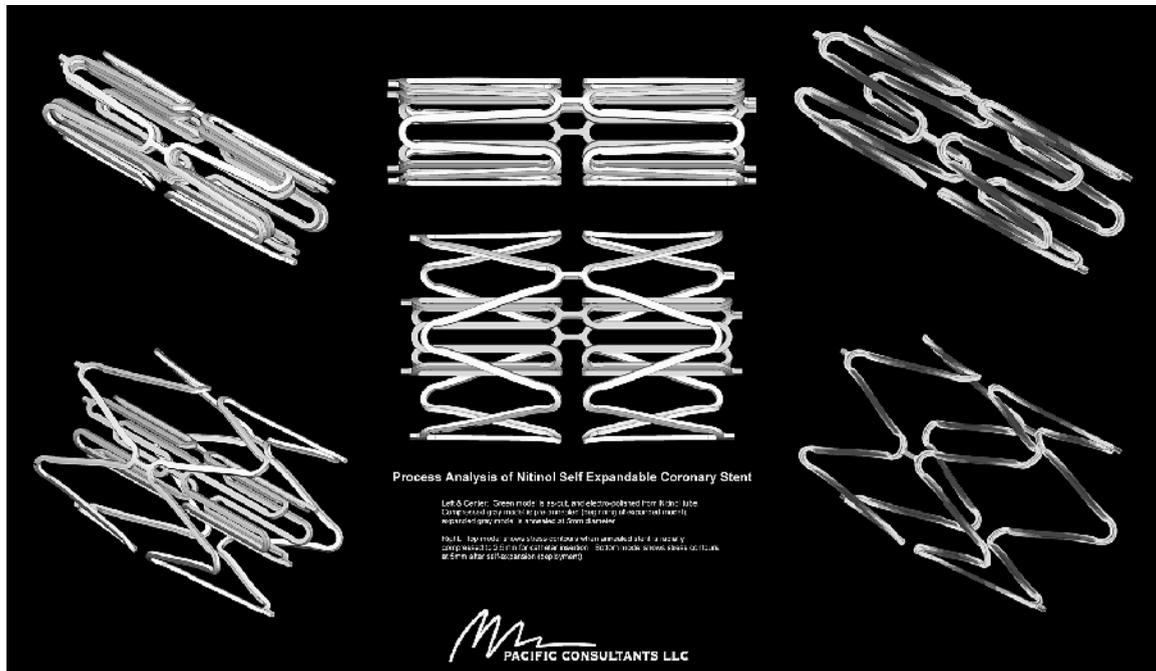


Figure 7. Finite element analysis model of a self-expandable Ni-Ti stent: displays a quantified mapping of stress and strain amplitudes in both expanded and compressed positions (courtesy of Pacific Consultants, LLC).

material requires special furnaces and processes, the sequence of cold working and annealing to ensure optimal SMA properties is extensive, and the special tooling and fixturing required for producing the many forms and shapes of the materials increase the cost of using SMAs.

Computer Modeling Capability. Finite element analysis (FEA), often used in conventional material design, has also been used to model the behavior of SMAs. The analysis is difficult, however, and should not depend on standard material templates and subroutines because the functional properties of SMAs rely on nonstandard factors, such as composition and processing history. Highly specific and complex modeling techniques must account for the state of the SMA material once formed in its trained shape, and then must incorporate the nonlinearity of the stress-strain curve, the property dependence on temperature, and the difference between loading and unloading stress behavior. Figure 7 is an illustration of a FEA model of a superelastic coronary stent that was achieved by using custom modeling subroutines to predict mechanical properties.

Shape Memory Applications

Temperature Cycling. Thermally activated SMA applications require temperature control to optimize the effect of shape memory. To harness the unique properties obtained from martensitic transformations, temperatures usually cycle between the extremes of the SMA temperature hysteresis. Depending on the alloy selected, this hysteresis might be smaller than 1°C (1.8°F) or larger than 100°C (180°F). Applications must incorporate a method of heating and cooling SMA components through their

hystereses; the rate of cycling for a shape-memory device is limited by the rate of temperature cycling. The butterfly example of Fig. 2 has a hysteresis of about 30 to 50°C. The Flexinol™ actuator wire is heated to a temperature above 90°C to contract, but must be cooled to approximately 40°C to transform to martensite. This application uses ambient air for cooling—only a few seconds are required for the wire to cool and stretch with the mass of the butterfly body. A few alloys have been developed to reduce or to increase this temperature hysteresis, as mentioned earlier. In addition, a secondary martensitic phase transformation found in many SMAs, called an R-phase transformation, can be exploited for its hysteresis of less than 2°C. Although recoverable strains are limited to about 0.5% by this transformation, it may be ideal for certain applications (13).

Power. SMAs require thermal energy for memory activation that is most often delivered as direct heat or applied current for resistive heating. Applications that use thermal shape memory must account for the power requirements of the material, the connection of the power source to the SMA, space requirements to house the source, and any safety mechanisms necessary to prevent overheating.

Force Requirements. SMA applications typically exploit the strength differential between the martensitic and austenitic phases of the material. In many actuating devices, an SMA element is paired with a conventional material element to serve as a biasing mechanism. The conventional element, such as a steel coil spring, displaces the SMA element when martensitic, but yields to the SMA

when heated to its stronger austenitic phase. This enables a one-way shape memory element to behave as a two-way shape-memory device. The forces delivered at each of the temperature extremes must be considered in SMA design.

Superelastic Applications

Operating Temperature Range. The temperature range of operation for a superelastic application must be above the A_f temperature for optimal superelasticity, but must also be below its M_d temperature (the temperature at which martensite can no longer be stress induced). This range is typically 50 to 80°C above the A_f (14).

Force Requirements. As operating temperatures increase above the SMA A_f temperature, loading and unloading stresses increase as a function of the Clausius–Clapeyron equation (14). Due to variations in the latent heat of transformation among alloy types, the increase in stress, as temperature changes, ranges from 2.5 MPa/°C to more than 15 MPa/°C (15).

SMA APPLICATIONS

Applications that use each of the unique properties of SMAs have been designed, prototyped, and marketed throughout the world. This section provides examples of these applications and includes some discussion of design choices, material limitations, and SMA behavior. These examples are categorized by industry to demonstrate the varied and widespread use of SMA applications.

Aeronautics/Aerospace

Many of the initial product ideas and applications that incorporate SMAs were pioneered in the fields of aeronautics and aerospace. SMA materials are used in these industries to take advantage of properties such as high power-to-mass ratios and ideal actuating behavior in zero-gravity conditions. Designs that use these properties replace heavier, more complex conventional devices because of reduced weight, design simplicity, and reliability.

Cryofit® Hydraulic Pipe Couplings. SMA couplings were the first successful commercial application of shape-memory alloys (Fig. 8). In 1969, Raychem Corporation introduced shrink-to-fit hydraulic pipe couplings for F-14 jet fighters that were built by Grumman Aerospace Corporation. This coupling is fabricated from a Ni–Ti–Fe alloy whose martensitic transformation temperature is below –120°C. It is machined at room temperature to an inner diameter approximately 4% smaller than the outer diameter of the piping it is designed to join. When cooled below –120°C by liquid nitrogen, the coupling is forced to a diameter 4% greater than the pipe diameter for an overall

internal strain of about 8%. When warmed above its TTR, the coupling diameter decreases to form a tight seal between the pipes (16).

This shape-memory application of constrained recovery continues to be a commercial and financial success. Despite the difficulties of cooling the couplings to liquid nitrogen temperatures for expansion and storage, the aerospace industry has welcomed their many advantages over traditional pipe-joining techniques such as welding or brazing. Installation is simple, less costly, and does not rely on high levels of operator skill. The replacement of couplings and hydraulic lines is straightforward, and the possibility of annealing and damaging the hydraulic lines as in welding or brazing is eliminated (2).

Frangibolt® Release Bolts. Shape-memory bolts were developed by the TiNi Alloy Company to replace conventional exploding bolt devices in aerospace release mechanisms. The bolts are used to attach spacecraft accessories during launch and to release them after launch by using an activated heating element (17). A martensitic, shape-memory cylinder is compressed and assembled to a notched bolt. When activated by an electrical heater, the cylinder increases in length and delivers a force greater than 22 kN (5000 lbs) to fracture the bolt at its notch (18). These release bolts were used successfully aboard the spacecraft Clementine in 1994, and have improved upon designs for conventional explosive mechanisms by eliminating the risks of off-gassing, accidental activation during shipment, and potential spacecraft damage during explosions.

Mars Sojourner Rover Actuator. An SMA wire was used to actuate a glass plate above a small solar cell on the Rover unit during the recent Pathfinder/Sojourner mission to Mars. A material adhesion experiment performed during the mission used the actuator to replace large, heavy motors and solenoids. A small, simple length of Ni–Ti wire heated and contracted when the Rover applied power and pulled a glass plate away from the solar cell to allow comparison of sunlight intensity with and without the plate. The rate of dust collection was then determined, and the resulting data will be used to design cleaning methods for future missions to Mars (19).

Self-Erectable Antenna. A prototype space antenna was constructed by Goodyear Aerospace Corporation. Designed to fold compactly at room temperature, the device would unfold into a large, extended antenna shape when heated by solar energy (2). Although this did not become a commercial success, the concept is feasible, and the prototype has served as a model for similar designs pursued within the aerospace industry.

Smart Airplane Wings. Composite structures that have SMA wires embedded can be used to change the shape of an airplane wing. The embedded wires may be activated



Figure 8. Shape memory devices. *Clockwise from top left:* memory card ejector mechanism for laptop computers; Cryofit[®] hydraulic pipe couplings; Cryocon[®] electrical connector; fire safety lid release for public garbage receptacles (courtesy of Shape Memory Applications, Inc.).

to constrict and improve the vibrational characteristics of the wing, heated to change their effective modulus to reduce vibration, or activated to alter the shape of the wing for optimal aerodynamics. All of these properties can be used to produce an adaptive airplane wing that alters as environmental conditions change to improve efficiency and reduce noise.

Space System Vibrational Damper. Vibrational dampers comprised of composite materials using pre-strained, embedded SMA wire or ribbons can reduce unwanted motion in various space systems. A sensor detects vibration in the system and sends a signal to activate the embedded composite, which then alters the structural dynamics to damp or cancel the existing vibration (1).

Consumer Products

SMA devices and components have been used in high-volume consumer products for more than 20 years. Although many consumers who use these products are unaware of their SMA components, there is a growing public

awareness of SMAs due to recently marketed items that advertise their merits.

Flexon[®] Optical Frames. Superelastic eyeglass frames marketed by Marchon Eyewear, Inc., are one of the most widely known uses for SMAs. They are frequently advertised in television commercials and can be found at most optical frame retailers. The components of eyeglass frames that are most susceptible to bending, the bridge and temples, are wire forms of Ni-Ti, the remainder of the frame is comprised of conventional materials for adjusting purposes and cost savings. Due to the high strain recovery capability of Ni-Ti, these frames are highly deformation- and kink-resistant (Fig. 4). Marchon is aware of the high strain and high cycle fatigue limitations of Ni-Ti materials, as demonstrated by the marketing brochures that appropriately suggest bending and twisting limits that are within the design guidelines for the material.

Portable Phone Antennae. The growing demand for portable phones has resulted in a high-volume application for superelastic Ni-Ti material because most cell

phone antennae produced today are Ni–Ti wires coated with polyurethane. The superelasticity resists permanent kinking and withstands the abuses of user handling during the lifetime of portable phones.

Greenhouse Window Opener. An SMA that has a small temperature hysteresis is used as an actuator to open and close greenhouse windows at predetermined temperatures for automatic temperature control. The opener is a spring-loaded hinge that has a Cu–Zn–Al shape-memory spring and a conventional metal biasing spring. The SMA spring is compressed by the biasing spring at temperatures below 18°C, and the window is closed. The SMA spring activates around 25°C, overcomes the force of the bias spring, and opens the window (21). This actuator design relies on reduced thermal hysteresis using a biasing force. As the SMA spring cools to 18°C, although not sufficiently cool to completely transform to its softer martensitic phase, it is transformed enough to accommodate deformation via stress-induced martensite.

Recorder Pen Mechanism. A shape-memory pen driver was designed by The Foxboro Company in the early 1970s to replace conventional pen-drive mechanisms, which used a galvanometer to actuate a pen arm. The replacement used Ni–Ti wires pretensioned in a driver unit and actuated by heat from an induction coil in response to input signals. The new design reduced the number of moving parts, improved reliability, and decreased costs. The new recorder pen units were first introduced in 1972; by 1980 more than 500,000 units were produced (16,21).

Nicklaus Golf Clubs. Superelastic SMA golf club inserts were developed by Memry Corporation for a line of Jack Nicklaus golf clubs. The damping properties of the inserts hold the golf ball on the club face longer and provide more spin and greater control for golfers (22).

Brassiere Underwires. Superelastic Ni–Ti shapes that conform to the user's body are ideal for underwire applications, because they are unaffected by the temperatures and external forces from repeated washings. Wires are shaped in predetermined configurations, using either round wire or flat ribbon. The product is a commercial success in Asia, but the increased cost compared to that of conventional underwires has prevented the product from entering markets in North America and Europe.

Residential Thermostatic Radiator Valve. SMA actuators have been used to regulate the temperature of residential radiators. An actuator expands when the room temperature increases, overcomes a biasing spring force, and closes a radiator hot water valve. Assisted by the biasing spring, the SMA temperature hysteresis can be as low as 1.2°C (21). The thermostatic valve can be adjusted via a knob that alters the compression of the biasing spring—the more compression it exerts, the higher the temperature required for the SMA actuator coil to activate and close the hot water valve (16).

Rice Cooker Valve. SMA valve mechanisms have been successfully employed to improve the performance of rice cookers. The mechanism, comprised of an SMA spring and a bias spring, is inserted into the top lid of a rice cooker. The valve is open while rice cooks and steam is generated, but when the rice is finished cooking, the SMA spring cools and the bias spring closes the valve to keep the rice warm. A Ni–Ti–Cu alloy is used for the SMA spring because of its low strain, high cycle fatigue properties. Although its recovery force decreases due to repeated cycling, this application has demonstrated repeatability for more than 30,000 cycles, which corresponds to several daily operations for 10 years (23).

Robotic Doll. SMA actuator wires were designed to move the arms and legs of a doll to display human characteristics. The application is technically feasible, and prototypes were successful; however, the power required to activate the wires was too great. The battery changes required were sufficiently frequent to limit market acceptance of the product.

Miscellaneous Products. Furukawa Electric Co. Ltd. of Japan produced SMA-actuating air-conditioning louvers to deflect air up or down, depending on temperature. They also manufactured coffee makers that use temperature-control valves to initiate the brewing process when water starts to boil (24). Other products include superelastic fishing lures, superelastic SONY Eggo™ headphones for the minidisk Walkman®, and novelty items, such as a magic teaspoon that has a memory. The teaspoon is given to someone to stir a hot drink, and when the spoon is exposed to the hot liquid, it is immediately transformed to a bent position.

Commercial/Industrial Safety

Many safety devices for temperature sensing and actuation have been successfully used in actual operation. The following examples have all been sold in consumer or industrial markets.

Antiscald MemrySafe® Valve. An SMA valve was designed to shut off a faucet's hot water source when water temperatures become too high (above 50°C). The valve reopens when the water cools to safer temperatures and protects the user from scalding water. ShowerGard®, BathGard®, and Flow-Gard® are similar products, and all have been marketed in retail hardware stores.

Firechek® Valve. A safety device that employs an SMA actuator is often used in industrial process lines to shut off a gas supply in the event of fire. Exposure to high temperatures activates a valve and cuts off the pneumatic pressure that controls flammable gas cylinders and process line valves (25).

Circuit Breakers. SMAs have been used in circuit breakers to replace conventional bimetals. Due to the high forces required in large circuit breakers, a series of levers must be employed to amplify the forces available from bimetals. Cu–Al–Ni alloys have been used in this application

for their high temperature activation and low hysteresis. Simple cantilever beam designs increase force and stroke and eliminate the need for levers.

Proteus[®] Safety Link Device. A chain link has been fabricated from Cu–Zn–Al to change shape at high temperatures and act as a release mechanism. The release may activate sprinkler systems or trigger fire doors to close, depending on the application (26).

Telecommunication Line Fuses. Cu–Zn–Al shunts are coupled with high-sensitivity fuses throughout Europe to protect communication systems from lightning strikes. During normal operation, the fuse heats up more rapidly than the SMA, and the shunt remains inactive. Under heavy usage, however, the shunt increases in temperature and activates to bypass the fuse and protect it from a critical burnout temperature (26).

Safety Trash Lid Mechanism. An SMA device has been designed to smother accidental fires in public trash receptacles. The device holds a trash lid in the open position at normal operating (ambient) temperatures, but when heated by a fire within the trash can, the shape-memory component releases a latch and the lid drops to extinguish the fire (Fig. 8).

Medical

The medical industry is rapidly accepting the use of SMAs in a wide variety of applications. From simple pointed needles to complex components implanted in the bloodstream, Ni–Ti has been adopted by the industry for its ability to offer unique and ideal solutions to traditional medical challenges. Well known for its excellent biocompatibility and corrosion resistance, Ni–Ti has been used in many successful medical devices and is now widely accepted throughout the medical industry.

The majority of SMA medical applications use the superelastic property of Ni–Ti, and many of them are in the expanding field of minimally invasive surgery. Due to the high strain recovery of Ni–Ti, components can withstand extreme shape changes for minimal profiles during delivery and then expand to larger devices within the body. Many of these SMA devices have eliminated the need for open heart surgery and thereby reduce patient risk and decrease hospital recovery periods.

Orthodontic Dental Arch Wires. Dental arch wires, one of the first medical applications that used SMAs, were first introduced in 1977 to replace stainless steel arch wires for straightening teeth. The wires were initially used in the martensitic condition, cold worked, and deformed around the teeth. They exhibited sufficient springback properties for this application, although later superelastic (austenitic) forms of Ni–Ti wire were introduced to improve product performance. The superelastic arch wire is now designed to exploit the plateau region of SMA's stress–strain curve, which provides nearly-constant stress on the teeth as the wire recovers its shape and straightens the teeth.

Mitek Homer Mammalok[®]. Mitek Surgical Products, Inc., introduced a superelastic needle wire localizer in 1985 which is used to locate and mark breast tumors to make surgical removal less invasive. The needle is used as a probe to pinpoint the location of a breast tumor first identified by mammography. Surgeons find it difficult to discern the tumor from surrounding tissue, so the probe highlights the correct location for the surgical procedure (27).

Mitek Suture Anchors. Mitek anchors, fabricated from a titanium or NiTi body that has two or more arcs of superelastic NiTi wire, are secure, stable suture holders used to reattach tendons, ligaments, and soft tissues to bone. The anchors are placed in a hole drilled into a patient's bone and are locked in place by Ni–Ti arcs. In 1989, Mitek Surgical Products, Inc., introduced these anchors for use in shoulder surgery to fasten sutures to bone. Since then, the firm has expanded its product line for use in many other orthopedic applications, such as ligament anchors used for reattaching the anterior cruciate ligament (ACL) of the knee (28).

Guidewire Cores. Ni–Ti wires are snaked through the tortuous pathways of the human body to guide and deliver other tools and devices for interventional procedures. These superelastic guidewires are optimal for use in minimally invasive surgery where procedures are performed through a small portal in a major artery, and offer superior flexibility, kink resistance, and torquability for optimal steering and ease of operation (29).

Stents. SMA stents are becoming increasingly popular in the medical industry. These structural, cylindrical components, designed to prop open and support human blood vessel walls, ducts, and other human passageways, are implanted to prevent collapse or blockage and to patch lesions. Ni–Ti materials are used in place of more conventional metals for coronary artery stenting but are most often used in a peripheral location such as the carotid artery, esophagus, or bile duct. Several shapes and forms of Ni–Ti stents are displayed in Fig. 9.

Stents currently on the market and in development use various functional properties of SMAs: the superelasticity of austenite, the thermal shape memory, and the low effective modulus of martensite. Many of these SMA stents use a combination of superelastic and shape-memory properties. For example, a stent may be chilled in ice water for transformation to martensite, compressed in the martensitic state, covered with a protective sheath for a minimal profile, and then delivered into the body through a small portal. When in place, the sheath is retracted, and the stent warms to body temperature to recover its original shape. Once recovered, or transformed to austenite, the superelastic properties of the stent result in gentle and constant radial forces on the vessel wall. Stents that are martensitic at room and body temperature must be compressed on a delivery balloon for expansion once delivered into the body. The Paragon Coronary Stent developed by Vascular Therapies is an example of a martensitic stent that is marketed for its even, symmetrical expansion and its flexibility during delivery.

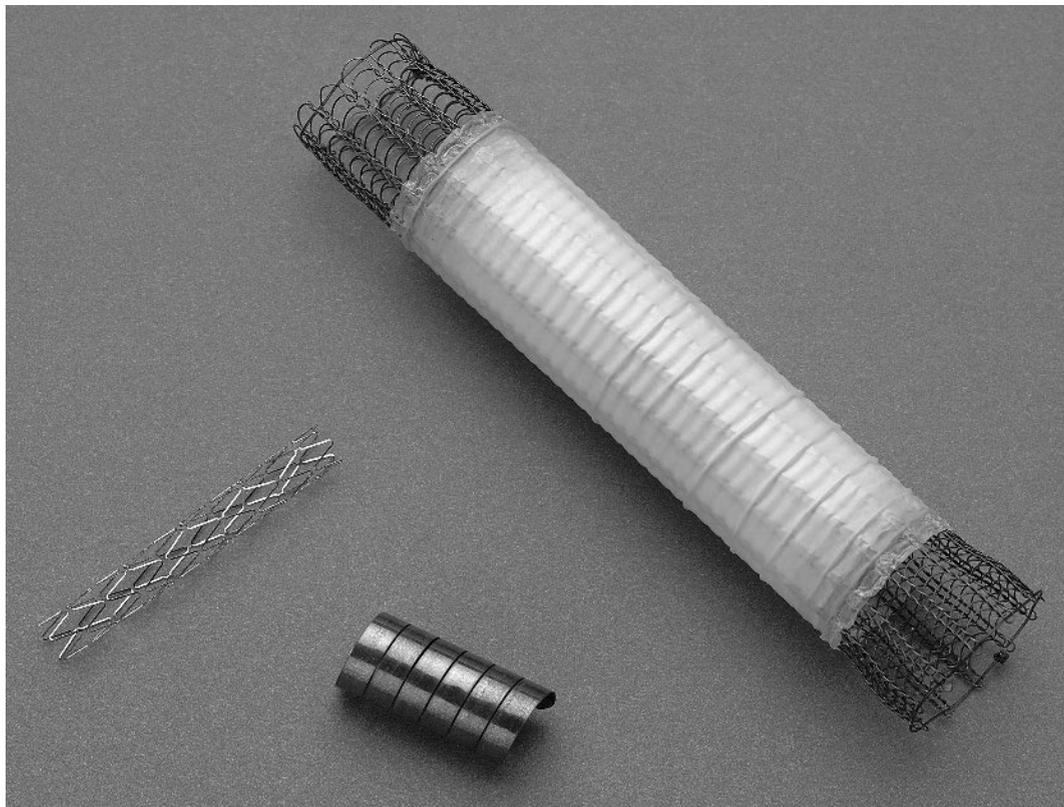


Figure 9. Nitinol stents. *Left to right:* Superelastic stent (5 mm OD \times 40 mm long), laser-cut from Ni-Ti tubing; Ni-Ti ribbon set in a coil configuration; Ultraflex™ esophageal knitted stent with fabric covering (courtesy of Shape Memory Applications, Inc.).

Simon Nitinol Filter®. An SMA vena cava blood clot filter was invented by Dr. Morris Simon of Nitinol Medical Technologies, Inc. The design is a Ni-Ti wire form shaped like an umbrella frame or wire basket. For delivery in the body, the filter is chilled below its transformation temperature and collapsed into a small insertion tube. The filter is cooled by a flow of cold saline solution while inserted into the patient and then expands when exposed to body temperature. Its recovered, umbrella-shaped form is designed to catch blood clots in the patient's bloodstream to prevent a pulmonary embolism (30).

AMPLATZER® Septal Occluder. Occlusion devices are designed to serve as Band-Aids to cover and heal holes in the heart without requiring open heart surgery and are typically fabricated from traditional materials. The AMPLATZER Occluder is an SMA device, comprised of a Ni-Ti wire frame that is woven and shaped into two flat caps connected by a short tubular section. The device is deployed through a portal in the femoral artery of a patient and placed at the center of the hole to be patched. A sheath is then retracted, and the two flat caps spring to shape and clamp on either side of the

hole, thus closing the hole and providing a tight seal (31).

Orthopedic Devices. SMA materials are used for their superelasticity and shape-memory properties in a variety of orthopedic devices designed to accelerate bone and cartilage formation under constant compressive stresses. Pre-strained SMA plates are used to treat bone fractures; when attached on both sides of a fracture and screwed together, they provide a compressive force to heal the fracture area. Staples are used to heal fractures as well. A Ni-Ti staple in the martensitic state is positioned so that its legs can be driven into bone sections across the fractured area to be closed. When heated, the legs of the staple bend inward and pull the bone sections together, creating a compressive force (32). Ni-Ti spacers are also used to assist spinal reinforcement in surgical procedures. In this procedure, a spacer is inserted as a compressed ring and then is heated to expand and force open a gap in the vertebrae. Bone chips are placed in the gap and fusion occurs over time to create solid bone. SMAs in solid rod form have been used to treat spinal curvature. Martensitic rods are pre-deformed to conform to a patient's original spinal curvature, wired to the spine, and then gradually recovered to move the spine to its corrected position (1).

Miscellaneous Instruments and Devices. Thousands of other medical devices have been developed to exploit the unique behavior of SMAs. A few additional, typical SMA products include catheters, endodontic files, aneurysm clips, retrieval baskets, surgical needles, and retractable grippers.

Automotive

SMA applications for the automotive industry are challenging for two primary reasons: the extreme range of operating temperatures expected during use and the market demand for low-cost components. Most automotive devices are expected to perform during exposure to the temperature extremes of climates throughout the world. The success of SMA automotive devices has been limited due to the deterioration of shape-memory properties over time caused by exposure to high temperatures (21). The successful SMA devices have exploited the benefits of lightweight, simple solutions. For example, SMA actuators are used as replacements for thermostatic bimetals and wax actuators and provide single metal components in place of complex systems. The simple SMA solutions further improve design performance by activating more quickly because they can be completely immersed in a gas or liquid flow (24).

Pressure Control Governor Valve. A shape-memory governor valve developed by Raychem Corporation and Mercedes-Benz AG was introduced in 1989 Mercedes cars to improve the rough cold weather shifting of automatic transmissions. The valves employ Ni-Ti coil springs to counteract the effects of increased oil viscosity. A Ni-Ti coil is immersed in the transmission fluid. At low temperatures, it is martensitic and is forced by a steel bias spring to move a piston. This activates a mechanism to reduce pressure and ease shifting. At higher temperatures, the NiTi spring is much stronger and pushes the bias spring in the opposite direction to optimize shifting pressure at the warmer temperatures (24).

Due to the ideal performance parameters of the application this governor valve design is one of the few technical and economic successes in the automotive industry. Operating temperatures are within the limits of the material, required force output is low, and expected life is less than 20,000 cycles. These conditions reduce the possibility of fatigue and degradation of SMA properties during the life of the product (33).

Toyota Shape-Memory Washer. Ni-Ti Belleville-type washers were developed by Toyota Motor Corporation and were used in Sprinter/Carib cars to reduce vibration and rattling noise at elevated temperatures. Automotive assemblies such as gearboxes are often combinations of many dissimilar metals. During the temperature increase in standard operation, a difference in the thermal expansion rates of the metals causes assemblies to loosen and rattle.

The washers were designed to change shape at high temperatures using forces up to 1,000 N (225 lb), which is sufficient to tighten the assembly and reduce the undesired rattling noise (24).

Shock Absorber Washer. This component, an application similar to the governor valve, is also designed to counteract the high viscosity of oil at colder temperatures. SMA washers are placed in shock absorber valves to alter their performance effectively in cold and hot temperatures (24).

Automotive Clutch Fan. An SMA device, developed as a selective switching mechanism for air cooled engines, requires the activation of a Cu-based SMA coil to control the operation of a clutch fan. The coil is activated at a temperature close to 50°C and engages a clutch to power the engine fan. The fan speeds up to cool the engine until the temperature is reduced, at which point the SMA coil is forced to disengage the clutch via a set of four steel leaf springs that serve as a biasing force. The device is designed to reduce engine noise and fuel consumption when the car is idle (21).

Industrial/Civil Engineering

Many SMA solutions have been designed and implemented to satisfy some of the rigorous, large-scale demands of civil engineering projects and miscellaneous industrial applications. Although the constraints of SMA properties often restrict use in industrial applications (the limited acceptable temperature range of operation, for example), many creative SMA solutions and design alternatives have successfully improved or replaced traditional industrial designs and devices.

Pipe Couplings. Cryogenic couplings developed by Raychem Corporation for the aerospace industry have been adapted for use in deep sea operations and are also used in the chemical and petroleum industries. The advancement of a Ni-Ti-Nb alloy for its wide hysteresis has helped to expand the use of these shape-memory couplings. When the components are machined at room temperature and then chilled in liquid nitrogen for expansion, they will not recover their shape until temperatures reach approximately 150°C. This allows storing and transporting expanded couplings without using liquid nitrogen. Once applied to the piping and heated to transform to austenite, they maintain their strength at temperatures colder than -20°C (33).

Structural Elements. Superelastic SMA materials can be used to increase strength and energy dissipation in a building. A project is currently underway to reinforce the Basilica of St. Francis in Assisi, Italy, after severe damage

during earthquakes in 1997. SMA wires will be placed in series with horizontal conventional steel ties to connect the walls to the Basilica's roof. The superelastic behavior of the wires will allow ductile, high strain motion to occur without breakage during an earthquake and will recover the strain by using a gentler, lower force stroke (exploiting the lower plateau of the stress-strain curve). It has been estimated that the SMA structural design will withstand an earthquake at least 50% stronger than if the Basilica were reinforced with conventional steel bars (34).

Rock Breakers. SMAs are used to replace explosives in demolition, which is similar in function to the Frangibolt[®] release mechanisms used in the aerospace industry. A prestrained SMA cylinder is placed within a crevice of a structure, electrically heated to expand, and the recovery forces produced are sufficient to destroy rocks and cement structures. This concept has been employed in Russia to yield demolition forces greater than 100 tons of force (1). Rock breakers comprised of Ni-Ti rods are also used in Japan. Nishimatsu Construction Co., Ltd. uses rods 29 mm long and 15 mm in diameter, which are compressed and inserted into boreholes in rock using an assembly of wedge-shaped platens. The NiTi rods are heated by attached electric heaters, causing them to expand and break the rock apart. Tests have demonstrated forces as high as 14 tons when they are heated to 120°C (23).

Power Line Sag Control. SMA materials are often used to prevent a sag or droop in overhead power lines. Using thermal shape-memory alloys, sag control in power lines has been successfully tested in Canada, the Ukraine, and Russia (35). When temperatures increase because of ambient temperatures or high load along the power lines, the lines tend to sag. Nitinol wires are attached to the lines and deform at the colder, high tension state, but contract when warm and remove the slack from the lines.

Steam Pipe Sag Control. Similar to the power line application, pipe hangers made from Ni-Ti are used to reduce the sagging of large steam pipes heated by the steam. This reduces the load variation in the system, rather than counteracting the shifts in geometry, as in power lines (35).

Transformer Core Compression. SMAs assist in compressing transformer cores, a critical aspect of transformer design. Ni-Ti bolts are prestrained axially to couple core sheets in large transformers, then heated to contract and provide high compressive forces on the sheets. This SMA solution improves on traditional techniques of tightening with nuts and bolts, where core sheets are placed in a vacuum to withdraw air from between sheets, then removed from the vacuum to install the bolts. Although technically a design improvement, the Ni-Ti solution requires 2-3 inch diameter bolts, which is larger than ideal for SMA products (35). The cost to process and machine the bolts may offset the benefits gained in the assembly process.

Multiwire Tension Device. SMA devices are used to increase piping integrity. They prevent crack initiation and propagation by compressing areas of a pipe using the

concept of multiwire tension (MWT) developed by the ABB Nuclear Division in Sweden. A split-sleeve coupling is wrapped with pretensioned SMA wire, the wire ends are fixed, and the coupling is placed over a weld point on a pipe. Then, the assembly is heated so that the wire contracts and the coupling then clamps to a tight fit. MWT techniques are used to improve stress in welded areas, preventing stress corrosion cracking and connecting the ends of the pipe should the weld point fail and break (35).

Indicator Tags. SMAs can be used to indicate high temperature points in a system. Wires used as tags are bent manually at typical operating temperatures, and the tags straighten when temperatures reach a critical level (the material's A_f temperature). Operators note the site of the high-temperature source and can fix the problem (35).

Sentinel[®] Temperature Monitoring System. SMAs are used on the low-voltage side of step-down transformers to indicate maximum temperatures and to close a switch at critical temperatures. This SMA mechanism is used to provide information for monitoring, so that operators can prevent overheating (35).

Injection Molding Mandrels. Centerless ground Ni-Ti round bar and wire are often used to replace conventional, deformation-prone mandrels. Plastics are molded in the shape of a superelastic Ni-Ti mandrel, and during rough handling as the mandrel is withdrawn from the cured polymer, the superelastic material recovers its original shape.

Heat Engines. SMA elements are used in heat engine designs to convert thermal energy to mechanical energy via thermal shape-memory behavior. Thermodynamic analyses of ideal, theoretical SMA heat engines have resulted in a wide range of calculated efficiencies, although the most thorough calculations yield maximum thermal efficiencies of only 2-4% (36). A great number of prototype engines has been constructed using SMA elements that change shape when they pass between hot and cold reservoirs. However, due to loss factors, such as friction, hysteretic effects inherent in the material, and energy input required to maintain a reservoir temperature differential, the practical efficiencies of these heat engines are much too low to serve as low-cost, high-volume energy converters.

The pursuit of a revolutionary SMA heat engine typifies a common occurrence in SMA application design. Intrigued by the potential of SMAs to provide unique and dramatic design solutions, inventors often pursue creative solutions before completing cost/benefit analyses. Although suitable for small-scale demonstrations, many SMA applications (as in heat engines) do not provide the necessary efficiencies—cost or energy—to warrant replacing conventional designs.

Electronics

The electronics industry has adopted SMA materials primarily for connection mechanisms. Applications in nearly all industries use the electronic activation capability of SMA materials to exploit thermal shape-memory behavior

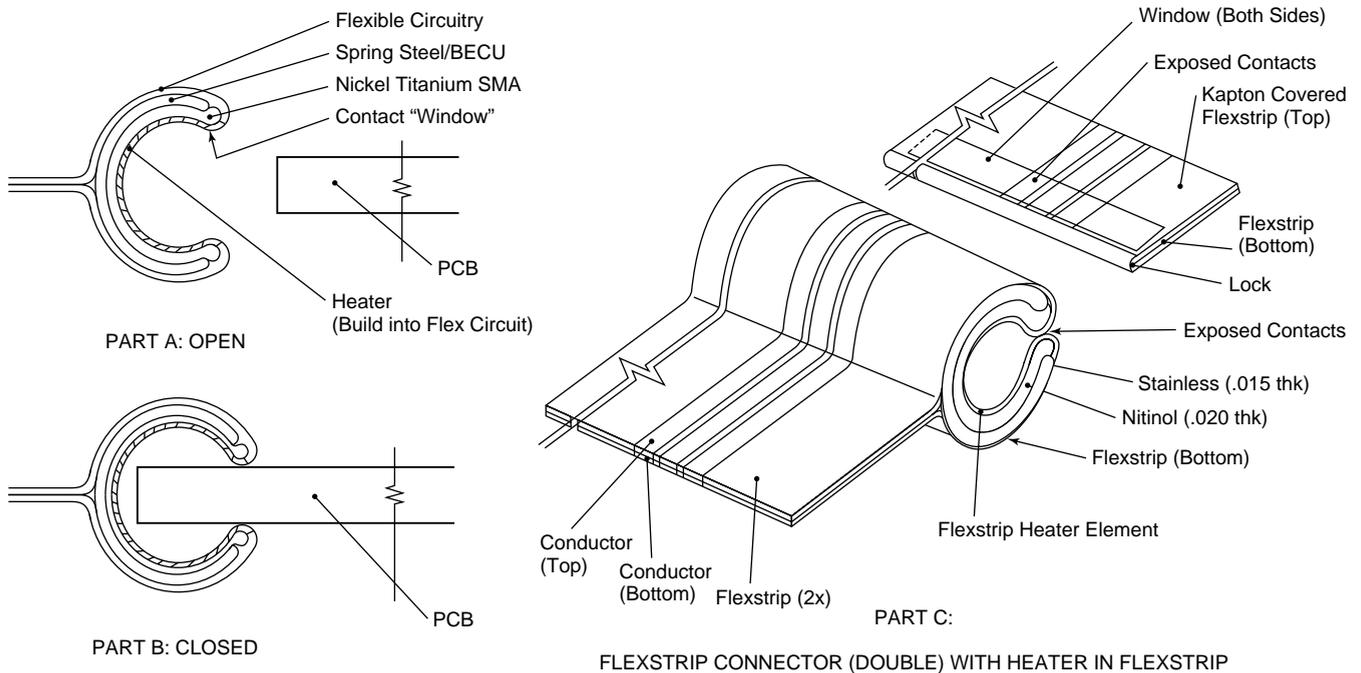


Figure 10. Schematic diagram of a zero insertion force connector. Reprinted with permission from J.F. Krumme, *Electrically Actuated ZIF Connectors Use Shape Memory Alloys*, *Connection Technology*. Copyright 1987, Lake Publishing Corporation.

(including robotics); however, a few examples are highly specific and unique to the electronics industry.

Cryocon® Electrical Connector. An SMA connector has been designed to attach the braid sheathing of an electrical cable to a terminal plug. The connector is a shape-memory ring sheath on a split-walled, collet-shaped tube whose diameter expands when chilled due to the radial forces of the tube. When warmed to room temperature, the ring recovers its smaller diameter shape and clamps the tubing collet prongs together to form a tight electrical connection (3,37).

ZIF Connector. A Zero insertion force electrical connector was designed by Beta Phase, Inc., to simplify the installation and increase the quality of circuit board connections (Fig. 10). A U-shaped strip of Ni-Ti is martensitic at operating temperatures and is forced to grip the boards via the bias force of a conventional closing spring. When heated by electrical current, the Ni-Ti strip overcomes the force of the closing spring and opens the radius of its U shape, allowing insertion or removal of the boards without force. This combines installation simplicity and maintains optimal, high force electrical contact between mother and daughter circuit boards (38).

Microactuators

SMAs have been successfully processed as thin film actuators for microactuating devices such as tiny valves, switches, and microelectromechanical systems (MEMS) (39). Ni-Ti films are sputter deposited on silicon substrates, and actuators are fabricated by chemical milling and lithographic processes. Devices one millimeter in

diameter and three microns thick have been fabricated, resulting in tiny actuators such as microvalves for fluid and pneumatic control. The Ni-Ti film actuators can provide up to 3% strain recovery (40).

FUTURE TRENDS

SMA applications continue to gain acceptance in a variety of industries throughout the world. SMAs are introduced to an increasing number of students who study engineering and metallurgy and to the general public via the growing number of SMA products available to consumers. Although property values and design techniques for SMAs are not as readily available or as thoroughly standardized as those of conventional materials, current trends indicate a steady course toward complete characterization of SMA materials. The medical industry has already prompted the standardization of material production, testing, and mechanical behavior; current efforts include the formation of an ASTM standard for using Ni-Ti in medical devices. Spearheaded by supplier representatives of both the medical device and SMA material supply industries, the standard is scheduled to be completed and approved before this article is published. This new standard represents a significant milestone in the continual effort to demystify shape-memory alloys and augment their use in engineering and design.

BIBLIOGRAPHY

1. L. McD. Schetky, *Proc. Shape Memory Alloys Power Syst.* Palo Alto, CA, 1994, pp. 4.1-4.11.

2. C.M. Jackson, H.J. Wagner, and R.J. Wasilewski, NASA Report SP-5110, Washington, DC, 1972, pp. 74, 78, 79.
3. H. Funakubo, ed., *Shape Memory Alloys*. Gordon and Breach, NY, 1987, pp. 201, 206.
4. T.W. Duerig, K.N. Melton, D. Stoeckel, and C.M. Wayman, *Engineering Aspects of Shape Memory Alloys*. Butterworth-Heinemann, London, 1990.
5. D.E. Hodgson, M.H. Wu, and R.J. Biermann, *Shape Memory Alloys*, Vol. 2, 10, (1990), pp. 897–902.
6. J.V. Humbeeck, *Manside Project Workshop Proc.* Rome, Italy, 1999, pp. II-1, II-2, II-22, II-28.
7. M.H. Wu, *Proc. Shape Memory Alloys Power Syst.* Palo Alto, CA, 1994, pp. 2.1–2.2.
8. T.W. Duerig and K.N. Melton, *MRS Int. Meet. Adv. Mater.*, Tokyo, Japan, 1988, Vol. 9, p. 583.
9. D.E. Hodgson and R.C. Krumme, *Proc. 1st Int. Conf. Shape Memory Superelastic Technol.* Pacific Grove, CA, 1994, pp. 371–376.
10. Hysteretic Damping Apparatus and Methods, US Pat. 5, 842, 312, Dec. 1, 1998, R.C. Krumme and D.E. Hodgson.
11. R.H. Richman, C.A. Zimmerly, O.T. Inal, D.E. Hodgson, and A.S. Rao, *Proc. 1st Int. Conf. Shape Memory Superelastic Technol.*, Pacific Grove, CA, 1994, pp. 175–180.
12. J. Perkins and D. Hodgson, in *Engineering Aspects of Shape Memory Alloys*, Butterworth-Heinemann, London, 1990, p. 204.
13. K. Otsuka, in *Engineering Aspects of Shape Memory Alloys*, Butterworth-Heinemann, London, 1990, pp. 36–45.
14. C.M. Wayman and T.W. Duerig in *Engineering Aspects of Shape Memory Alloys*, Butterworth-Heinemann, London, 1990, pp. 3–20.
15. K.N. Melton, in *Engineering Aspects of Shape Memory Alloys*, Butterworth-Heinemann, London, 1990, pp. 21–35.
16. L.McD. Schetky, *Sci. Am.* **241**(5): 79, 81 (1979).
17. <http://www.sma-mems.com/aero.htm>.
18. J.D. Busch, *Proc. 1st Int. Conf. Shape Memory Superelastic Technol.* pp. 259–264, Pacific Grove, CA, 1994.
19. <http://www.robotstore.com/mwmars.html>.
20. Marchon Company Brochure, 1998.
21. C.M. Wayman, *J. Met.* pp. 129–137 (June, 1980).
22. <http://www.techtran.smfc.nasa.gov/new/memmetal.html>.
23. H. Tamura, *Proc.: Shape Memory Alloys Power Syst.* Palo Alto, CA, 1994, pp. 8–3, 8–5.
24. D. Stoeckel, *Adv. Mater. Process.* pp. 35, 38 (Oct, 1990).
25. <http://www.techtran.smfc.nasa.gov/new/memmetal.html>.
26. W.V. Moorlegghem, *Proc.: Shape Memory Alloys Power Syst.* Palo Alto, CA, 1994, pp. 9–1, 9–3.
27. J.P. O'Leary, J.E. Nicholson, and R.F. Gatturina, in *Engineering Aspects of Shape Memory Alloys*, Butterworth-Heinemann, London, 1990, p. 477.
28. Mitek Surgical Products, Inc. Company Brochure, 1995.
29. J. Stice, in *Engineering Aspects of Shape Memory Alloys*, Butterworth-Heinemann, London, 1990, p. 483.
30. <http://www.nitinolmed.com/products/>
31. <http://www.agamedical.com/patients/index.html>
32. J. Haasters, in *Engineering Aspects of Shape Memory Alloys*, Butterworth-Heinemann, London, 1990, pp. 426–444.
33. D. Stoeckel, *Proc.: Shape Memory Alloys Power Syst.* Palo Alto, CA, 1994, pp. 1–9.
34. <http://www.alphagalileo.org/> (9/20/99).
35. J. Cederstrom, *Proc.: Shape Memory Alloys Power Syst.* Palo Alto, CA, 1994, pp. 6–2 to 6–7.
36. P. Wollants, M. De Bonte, L. Delaey, and J.R. Roos, *Proc. NITINOL Heat Engine Conf.* Silver Spring, M, 1978, pp. 6.1–6.22.
37. E. Cydzik, in *Engineering Aspects of Shape Memory Alloys*, Butterworth-Heinemann, London, 1990, pp. 149–157.
38. J.F. Krumme, *Connection Technol.* Lake, 1987.
39. http://www.sma-mems.com/t_film.htm
40. A.D. Johnson and J.D. Busch, *Proc. 1st Int. Conf. Shape Memory Superelastic Technol.* Pacific Grove, CA, 1994, pp. 299–304.
41. D. Stoeckel, *Shape-Memory Alloys, Adv. Mater. Process.* pp. 35, 38 (Oct, 1990).

ADDITIONAL READING

- A. Pelton, D. Hodgson, S. Russell, and T. Duerig, *Proc. 2nd Int. Conf. Shape Memory Superelastic Technol.* (SMST-97), Pacific Grove, CA, 1997.
- Jan Van Humbeeck, Non-Medical Applications of Shape Memory Alloys, *Mater. Sci. Eng.* **A273–275**: 134–148 (1999).
- T. Duerig, A. Pelton, and D. Stoeckel, An Overview of Nitinol Medical Applications, *Mater. Sci. Eng.* **A273–275**: 149–160 (1999).

SHAPE-MEMORY ALLOYS, MAGNETICALLY ACTIVATED FERROMAGNETIC SHAPE-MEMORY MATERIALS

R.C. O'HANDLEY
SAMUEL M. ALLEN
Massachusetts Institute of Technology
Cambridge, MA

INTRODUCTION

For years it was speculated that the large strains associated with the thermoelastic shape-memory effect, such as in NiTi alloys, could be captured by application of a magnetic field in certain martensites that are also ferromagnetic (1–3); see Fig 1(a). Ferromagnetic shape-memory alloys (FSMAs) moved from a hypothetical new class of active materials, to join piezoelectric and magnetostrictive materials, upon observation of a 0.2% magnetic field induced strain in a single crystal of Ni₂MnGa in 1996 (4). Figure 1(b) shows the field-induced strain measured at 265 K in the martensitic phase of a Ni₂MnGa single crystal. The crystal contracts in the direction of the applied field. The strain components conserve volume and are of even symmetry in the field. There is some hysteresis in the strain with change in sweep direction of the field, and there is some unrecovered strain after the first field cycle. By way of comparison, piezoelectrics show strains of order 0.1% (5) and the leading magnetostrictive material, Terfenol-D (Tb_{0.33}Dy_{0.67}Fe₂), shows a field-induced strain of about 0.24% (6,7).

The strain-versus-temperature (ϵ - T) curves of thermoelastic martensite, Fig. 1(a), bear little resemblance to the ϵ - H curves of Fig. 1(b). The former typically show a

2. C.M. Jackson, H.J. Wagner, and R.J. Wasilewski, NASA Report SP-5110, Washington, DC, 1972, pp. 74, 78, 79.
3. H. Funakubo, ed., *Shape Memory Alloys*. Gordon and Breach, NY, 1987, pp. 201, 206.
4. T.W. Duerig, K.N. Melton, D. Stoeckel, and C.M. Wayman, *Engineering Aspects of Shape Memory Alloys*. Butterworth-Heinemann, London, 1990.
5. D.E. Hodgson, M.H. Wu, and R.J. Biermann, *Shape Memory Alloys*, Vol. 2, 10, (1990), pp. 897–902.
6. J.V. Humbeeck, *Manside Project Workshop Proc.* Rome, Italy, 1999, pp. II-1, II-2, II-22. II-28.
7. M.H. Wu, *Proc. Shape Memory Alloys Power Syst.* Palo Alto, CA, 1994, pp. 2.1–2.2.
8. T.W. Duerig and K.N. Melton, *MRS Int. Meet. Adv. Mater.*, Tokyo, Japan, 1988, Vol. 9, p. 583.
9. D.E. Hodgson and R.C. Krumme, *Proc. 1st Int. Conf. Shape Memory Superelastic Technol.* Pacific Grove, CA, 1994, pp. 371–376.
10. Hysteretic Damping Apparatus and Methods, US Pat. 5, 842, 312, Dec. 1, 1998, R.C. Krumme and D.E. Hodgson.
11. R.H. Richman, C.A. Zimmerly, O.T. Inal, D.E. Hodgson, and A.S. Rao, *Proc. 1st Int. Conf. Shape Memory Superelastic Technol.*, Pacific Grove, CA, 1994, pp. 175–180.
12. J. Perkins and D. Hodgson, in *Engineering Aspects of Shape Memory Alloys*, Butterworth-Heinemann, London, 1990, p. 204.
13. K. Otsuka, in *Engineering Aspects of Shape Memory Alloys*, Butterworth-Heinemann, London, 1990, pp. 36–45.
14. C.M. Wayman and T.W. Duerig in *Engineering Aspects of Shape Memory Alloys*, Butterworth-Heinemann, London, 1990, pp. 3–20.
15. K.N. Melton, in *Engineering Aspects of Shape Memory Alloys*, Butterworth-Heinemann, London, 1990, pp. 21–35.
16. L.McD. Schetky, *Sci. Am.* **241**(5): 79, 81 (1979).
17. <http://www.sma-mems.com/aero.htm>.
18. J.D. Busch, *Proc. 1st Int. Conf. Shape Memory Superelastic Technol.* pp. 259–264, Pacific Grove, CA, 1994.
19. <http://www.robotstore.com/mwmars.html>.
20. Marchon Company Brochure, 1998.
21. C.M. Wayman, *J. Met.* pp. 129–137 (June, 1980).
22. <http://www.techtran.smfc.nasa.gov/new/memmetal.html>.
23. H. Tamura, *Proc.: Shape Memory Alloys Power Syst.* Palo Alto, CA, 1994, pp. 8–3, 8–5.
24. D. Stoeckel, *Adv. Mater. Process.* pp. 35, 38 (Oct, 1990).
25. <http://www.techtran.smfc.nasa.gov/new/memmetal.html>.
26. W.V. Moorleghem, *Proc.: Shape Memory Alloys Power Syst.* Palo Alto, CA, 1994, pp. 9–1, 9–3.
27. J.P. O'Leary, J.E. Nicholson, and R.F. Gatturina, in *Engineering Aspects of Shape Memory Alloys*, Butterworth-Heinemann, London, 1990, p. 477.
28. Mitek Surgical Products, Inc. Company Brochure, 1995.
29. J. Stice, in *Engineering Aspects of Shape Memory Alloys*, Butterworth-Heinemann, London, 1990, p. 483.
30. <http://www.nitinolmed.com/products/>
31. <http://www.agamedical.com/patients/index.html>
32. J. Haasters, in *Engineering Aspects of Shape Memory Alloys*, Butterworth-Heinemann, London, 1990, pp. 426–444.
33. D. Stoeckel, *Proc.: Shape Memory Alloys Power Syst.* Palo Alto, CA, 1994, pp.1–9.
34. <http://www.alphagalileo.org/> (9/20/99).
35. J. Cederstrom, *Proc.: Shape Memory Alloys Power Syst.* Palo Alto, CA, 1994, pp. 6–2 to 6–7.
36. P. Wollants, M. De Bonte, L. Delaey, and J.R. Roos, *Proc. NITINOL Heat Engine Conf.* Silver Spring, M, 1978, pp. 6.1–6.22.
37. E. Cydzik, in *Engineering Aspects of Shape Memory Alloys*, Butterworth-Heinemann, London, 1990, pp. 149–157.
38. J.F. Krumme, *Connection Technol.* Lake, 1987.
39. http://www.sma-mems.com/t_film.htm
40. A.D. Johnson and J.D. Busch, *Proc. 1st Int. Conf. Shape Memory Superelastic Technol.* Pacific Grove, CA, 1994, pp. 299–304.
41. D. Stoeckel, *Shape-Memory Alloys, Adv. Mater. Process.* pp. 35, 38 (Oct, 1990).

ADDITIONAL READING

- A. Pelton, D. Hodgson, S. Russell, and T. Duerig, *Proc. 2nd Int. Conf. Shape Memory Superelastic Technol.* (SMST-97), Pacific Grove, CA, 1997.
- Jan Van Humbeeck, Non-Medical Applications of Shape Memory Alloys, *Mater. Sci. Eng.* **A273–275**: 134–148 (1999).
- T. Duerig, A. Pelton, and D. Stoeckel, An Overview of Nitinol Medical Applications, *Mater. Sci. Eng.* **A273–275**: 149–160 (1999).

SHAPE-MEMORY ALLOYS, MAGNETICALLY ACTIVATED FERROMAGNETIC SHAPE-MEMORY MATERIALS

R.C. O'HANDLEY
 SAMUEL M. ALLEN
 Massachusetts Institute of Technology
 Cambridge, MA

INTRODUCTION

For years it was speculated that the large strains associated with the thermoelastic shape-memory effect, such as in NiTi alloys, could be captured by application of a magnetic field in certain martensites that are also ferromagnetic (1–3); see Fig 1(a). Ferromagnetic shape-memory alloys (FSMAs) moved from a hypothetical new class of active materials, to join piezoelectric and magnetostrictive materials, upon observation of a 0.2% magnetic field induced strain in a single crystal of Ni₂MnGa in 1996 (4). Figure 1(b) shows the field-induced strain measured at 265 K in the martensitic phase of a Ni₂MnGa single crystal. The crystal contracts in the direction of the applied field. The strain components conserve volume and are of even symmetry in the field. There is some hysteresis in the strain with change in sweep direction of the field, and there is some unrecovered strain after the first field cycle. By way of comparison, piezoelectrics show strains of order 0.1% (5) and the leading magnetostrictive material, Terfenol-D (Tb_{0.33}Dy_{0.67}Fe₂), shows a field-induced strain of about 0.24% (6,7).

The strain-versus-temperature (ϵ - T) curves of thermoelastic martensite, Fig. 1(a), bear little resemblance to the ϵ - H curves of Fig. 1(b). The former typically show a

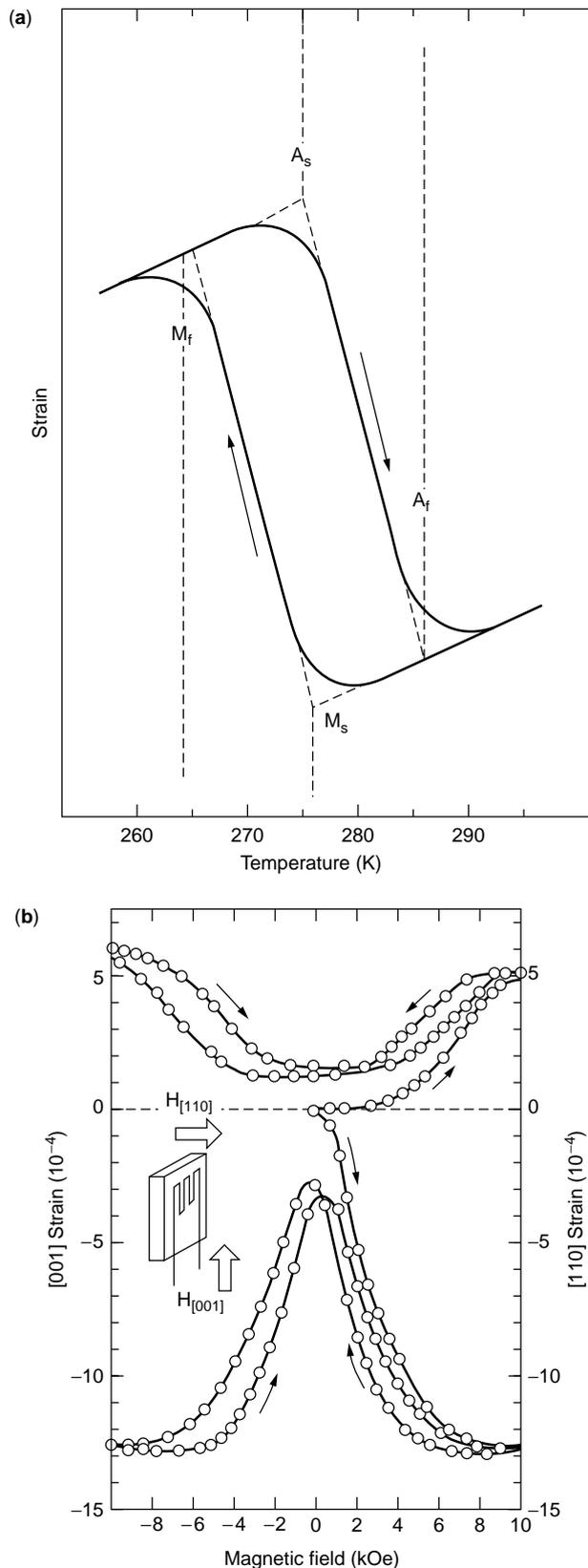


Figure 1. Contrast between (a) the thermally induced strain in the shape memory alloy, NiTi and (b) the magnetic-field-induced strain in a ferromagnetic shape memory alloy. Here, strain versus applied field is measured at 265 K in the martensitic phase of Ni₂MnGa. Inset: Relative orientation of sample, strain gauge, and applied field (4).

large thermal hysteresis, with martensite and austenite start and finish temperatures defined by the sharply curved points of the ε - T curves; the change in sample strain through the thermoelastic hysteresis can be several percent. The ε - H curves of Fig. 1(b) bear some resemblance to those of a negative magnetostriction material: the strain is of even symmetry in the applied magnetic field and increases superlinearly in H before saturating. However, the FSMA strain effect differs in many ways from magnetostriction as will be shown in detail later. The strain shown in Fig. 1(b) represents only a small fraction of the 6% or 7% transformation strain expected to be accessible by field-induced twin selection in Ni-Mn-Ga FSMAs. In fact, FSMAs have recently exhibited the full field-induced strain associated with their crystallographic distortion by application of a field of 320 kA/m (4 kOe) to a single-variant sample of an off-stoichiometry crystal of Ni-Mn-Ga (8).

The mode of actuation shown in Fig. 1(b), namely field-induced strain within the fully martensitic state, is to be contrasted with application of a magnetic field to the austenitic (A) phase of an FSMA to induce the transformation to the single-variant martensitic (M) phase, Fig. 2(a). Here the A-M phase boundary is shown as a surface in H - σ - T space. As is the case with most martensites, a *shear* stress applied to the A phase at a temperature just above the M-start temperature can initiate the A \rightarrow M transformation; see vertical path ε - σ in Fig. 2(a) (1,9,10). The A \rightarrow M transformation is accompanied by the release of a heat of transformation measured to be $Q = 40 \text{ MJ/m}^3$ (288 cal/mole) (11) and $Q = 98 \text{ MJ/m}^3$ (706 cal/mole) (12). Kanomata et al. (13) find that *hydrostatic* pressure stabilizes the A phase (the transformation temperature *decreases* with increasing *hydrostatic* pressure). This occurs despite the smaller volume of the martensitic phase: $a_A = 0.582 \text{ nm}$, $a_M = 0.592 \text{ nm}$, and $c_M = 0.557 \text{ nm}$. In contrast, a shear stress stabilizes the M phase (14). The M phase should also be stabilized by application of a magnetic field to an FSMA; see the horizontal ε - H path in Fig. 2 (a) (15). However, the Clausius-Clapeyron equation indicates that a magnetic field of order 10^7 A/m ($B = 12 \text{ T}$) would be needed to induce the transformation in a sample held about 1°C above the martensite start temperature (14). Experiments on Ni_{2.19}Mn_{0.81}Ga bear this out (16). The field required for transformation could be reduced by simultaneous application of an appropriate stress. Such a field-induced phase transformation would release the full transformation strain accompanied by a stress comparable to the yield stress of the detwinned martensite. Stress-assisted, field-induced transformations have been investigated by Vasil'ev et al. (17). On transforming back to A, the heat of transformation is absorbed by the material (11). Such field-induced and stress-induced transformations can show significant hysteresis because phase boundary motion is involved.

The present article focuses on the application of a magnetic field to a twinned M phase that is ferromagnetic, Fig. 2(b). It is necessary that the magnetic anisotropy of the M phase be large compared to the energy required for twin boundary motion and, further, that the preferred direction of magnetization changes across the twin boundary. When this is the case, application of a magnetic field results in a difference in Zeeman energy, $\Delta(\mu_0 \mathbf{M}_s \cdot \mathbf{H})$,

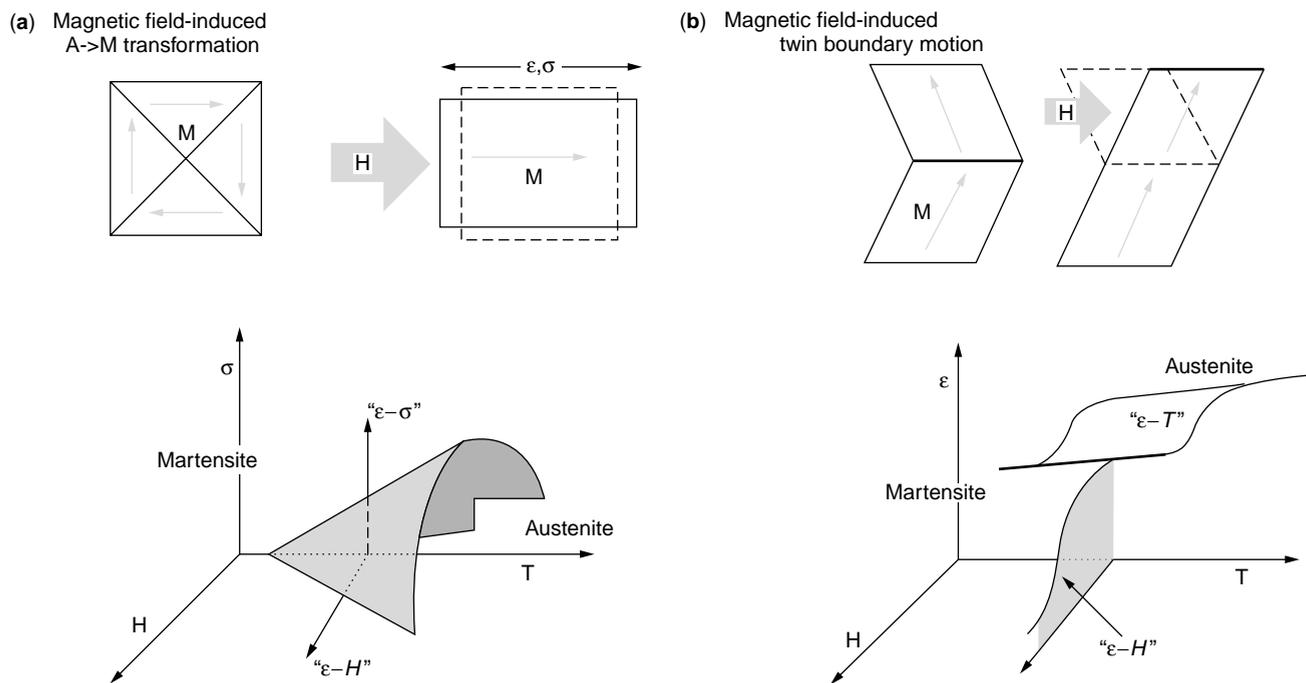


Figure 2. (a) Magnetic field-induced transformation of an austenitic sample to a single-variant martensitic structure. The A-M phase boundary is shown below in field-stress-temperature space; the arrow indicates how an increasing field can induce the transformation (horizontal ε - H path). This process may be assisted by application of a suitable stress (vertical path ε - σ path). (b) A field may be able to move twin boundaries in a twinned martensitic FSMA, changing the state of strain of the martensite. Compare the strain-versus-temperature hysteresis loop in lower panel with Fig. 1(a). The added field axis shows how application of a field in the martensitic state (bold line on ε - T loop) can alter the state of strain of the sample.

across the twin boundary. This energy difference exerts a pressure on the twin boundary so as to grow the twin variants having the more favorably oriented magnetization. The resulting field-induced twin-boundary motion produces a large strain, fully within the martensitic state of an FSMA.

This article describes the crystallography and magnetism of Ni-Mn-Ga in order to explain the very large strains produced by field-induced twin-boundary motion in martensite. Examples of field-induced strain by twin boundary motion in Ni-Mn-Ga FSMA samples having different twin structures are given. Martensitic Fe₇₀Pd₃₀ has also shown field-induced strains of 0.5% (18), and efforts are under way to develop other iron-base FSMA (19–21). These other materials will not be covered in depth. The state of theoretical modeling of strain and magnetization in FSMA is reviewed. FSMA field-induced strains are compared and contrasted with the thermoelastic shape-memory effect and magnetostriction.

FIELD-INDUCED STRAIN IN FSMA

Crystallography and Mechanical Properties

The chemically ordered austenitic phase of Ni₂MnGa has the $L2_1$ Heulser structure ($Fm\bar{3}m$) with room temperature lattice constant $a = 0.582$ nm (22). Figure 3(a) shows the chemical ordering favored below 800°C, and the {101} slip planes are shaded in. The eight Ni atoms are located

at the center of each of the eight cubic sub-units; Ni orders to these sites in the high-temperature cubic phase near or above the melting temperature. The Mn and Ga atoms order on the remaining sites (as shown by the two sizes of spheres in Fig. 3a) below 800°C (23). In the martensitic state, this structure is tetragonally distorted with $a = b = 0.592$ nm, $c = 0.557$ nm. The c axis contraction is $c/a = 0.94$. The appropriate unit cell of the tetragonal martensite ($I4/mmm$, having $c' = c$ and $a' = b' = a/2$) is shown in Fig. 3(b). The twin plane identified in the structure at left as {101} has Miller indexes (112) when referred to the tetragonal unit cell axes. The c'/a' ratio in the unit cell has the magnitude $\sqrt{2} \cdot c/a \approx 1.33$.

In this article, discussions are referenced to the structure at left ({101} twin planes) rather than the true body-centered unit cell, in order to bring out the fact that the martensitic phase of Ni₂MnGa is contracted along its preferred direction of magnetization, relative to the austenitic phase.

The off-stoichiometry phases of Ni-Mn-Ga generally used in engineering studies may have more complex structures. The most useful phase is a five-layered, tetragonal martensite (22). Slight variations in composition near stoichiometry can result in different tetragonal structures having very different magnetic anisotropy, c/a ratios and responses to applied magnetic fields. For example, Sozinov et al. (24) report that while Ni_{49.2}Mn_{29.6}Ga_{21.2} is an easy axis ($M \parallel c$ -axis), tetragonal five-layered martensite with $c/a = 0.94$, and field-induced strains of several

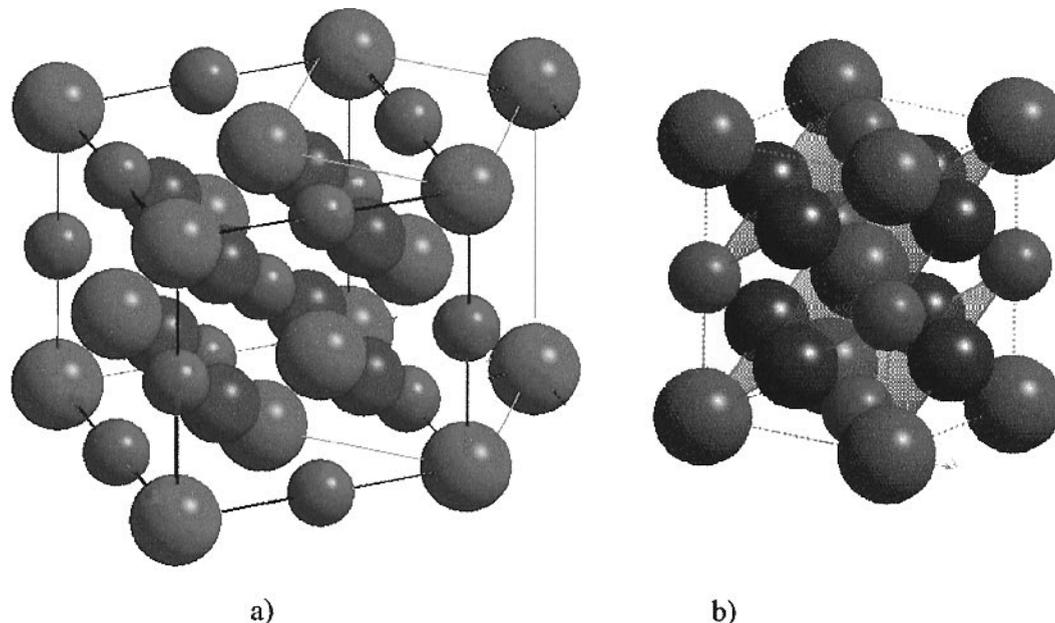


Figure 3. (a) Model of the cubic Heusler, $L2_1$, structure of austenitic Ni_2MnGa . In the martensitic state this structure contracts along its c -axis ($c/a < 1$), and it may twin along $\{101\}$ planes. (b) The $14/mmm$, body-centered tetragonal ($c'/a' > 1$) martensitic unit cell is shown with the same twinning planes, now identified as $\{112\}$.

percent, $\text{Ni}_{52.1}\text{Mn}_{27.3}\text{Ga}_{20.6}$ is an easy plane ($\mathbf{M} \perp c$ -axis), tetragonal martensite with $c/a = 1.2$, and negligible field-induced strain. Some Ni–Mn–Ga samples (both polycrystalline) show orthorhombic structures (16,25). While orthorhombic structures have more possible twin-plane orientations than tetragonal structure, the models described below for the field-induced strain still seem to apply.

The mechanical properties of FSMAs in the martensitic phase are similar to those of conventional SMAs (Fig. 4). With increasing stress, the material first shows a modulus, C_0 , characteristic of the single-variant state. Above a critical stress, σ_0 , at which deformation by twin-boundary motion initiates, the modulus decreases to C_{tb} and the material may strain to its full transformation strain, ε_0 , after which it is mechanically detwinned. The transformation

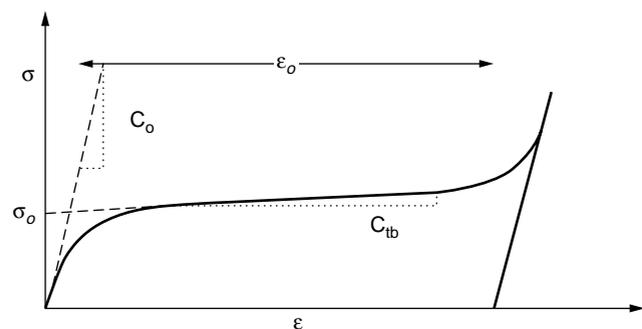


Figure 4. Schematic of the stress–strain behavior of martensite. Above a critical stress, σ_0 , the modulus drops to C_{tb} as the mode of deformation becomes twin-boundary motion. Depending on the twin-variant structure at σ_0 the material may deform as much as ε_0 , the transformation strain, before it is detwinned and reverts to the stiffness of the single-variant state, C_0 .

strain is the crystallographic distortion in a martensitic transformation: $\varepsilon_0 = 1 - c/a$ referred to the austenitic unit cell. For still larger stresses, the modulus reverts to its single-variant value.

Values of the parameters in Fig. 4 measured for Ni_2MnGa crystals are typically of order $\sigma_0 = 1 - 10$ MPa, $C_0 = 2$ GPa, $\varepsilon_0 = 0.06$, and $C_{tb} = 18 - 30$ MPa (26,27). The small value for C_0 relative to the stiffness of the parent phase, which is of order 70 GPa, probably reflects a small degree of twin-boundary motion in the martensite before the initiation of abundant twin-boundary motion at σ_0 . Hence 70 GPa should be regarded as more typical of the martensitic stiffness at constant strain.

Magnetization

Ni_2MnGa shows a saturation magnetization at 265 K (in the martensitic phase) of $M_s = 484$ kA/m (484 emu/cm³) (4). Given its mass density of 8.3 g/cm³, this translates to a saturation flux density of $B_s = 0.6$ T ($4\pi M_s = 6$ kG). The Curie temperature is about 350 K. This is greater than the martensite transformation temperature, which varies strongly with composition and stress and is typically about 273 K (1,11,27). The strength of the uniaxial anisotropy of tetragonal Ni_2MnGa has been measured from single-crystal magnetization curves and found to be about 1.2×10^5 J/m³ at -8°C , favoring magnetization parallel to the c axis (4). Tickle and James (29) made measurements on single crystals of a $\text{Ni}_{51.3}\text{Mn}_{24.0}\text{Ga}_{24.7}$ constrained to be in the single-variant state during magnetization. They find $K_u = 2.45 \times 10^5$ J/m³ at -17°C . The samples cited in this article, close in composition to $\text{Ni}_{50}\text{Mn}_{28}\text{Ga}_{22}$, show a magnetic anisotropy in the single-variant constrained M phase at room temperature that ranges from 1.6 to

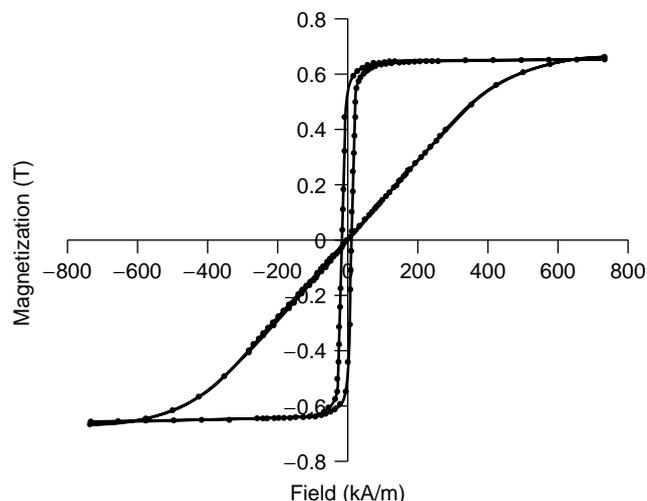


Figure 5. $\mu_0 M$ - H measured on a sample of $\text{Ni}_{49.8}\text{Mn}_{28.5}\text{Ga}_{21.7}$ constrained to be in the single-variant state. The two curves are measured with the field parallel (square loop) and perpendicular (linear M - H process) to the c -axis of the variant. (27).

$1.9 \times 10^5 \text{ J/m}^3$ (27). Figure 5 shows M - H loops taken on a small crystal constrained to be in a single-variant state at room temperature. Loops are shown for field-applied parallel (\parallel) and perpendicular (\perp) to the c -axis of the variant (shape effects have been removed from both M - H loops). The coercivity of the H_{\parallel} loop suggests that the parallel magnetization process takes place by magnetic domain-wall motion. The zero-coercivity, linear H_{\perp} loop indicates magnetization by a rotation of the moment from the c axis to the field direction. The area between the H_{\parallel} and H_{\perp} curves, $1.6 \times 10^5 \text{ J/m}^3$, is the strength of the uniaxial magnetocrystalline anisotropy, K_u . A key factor that allows FSMAs to be deformed by application of a magnetic field is the relatively large field required to rotate the magnetization from the c -axis.

Field-Induced Twin Rearrangement

The magnetic driving force behind field-induced twin boundary motion is well illustrated by considering a NiMnGa sample showing large twins. Figure 6 shows an off-stoichiometry single-crystal sample, $\text{Ni}_{49.4}\text{Mn}_{29.7}\text{Ga}_{20.9}$, in the single-variant state and with a sharp kink (bend) at the position of a twin boundary introduced by application of a field at room temperature (30). The kink at the twin boundary spanning the cross section of the material can be moved along the sample length by varying the field strength and direction. In this mode of operation, the material always expresses its full transformation shear strain, γ_0 , across a twin boundary. The shear strain across the twin boundary is given by $\gamma_0 = (a/2c)(1 - c^2/a^2)$. Thus, for $c/a = 0.93, 0.94, \text{ and } 0.95$, $\gamma_0 = 7.3\%, 6.2\%, \text{ and } 5.1\%$, respectively.

When a single-variant sample, cut as that in Fig. 6(a), is placed in a magnetic field, the long sample axis does not align with the field as magnetostatic considerations alone would dictate. Instead, it aligns at about 45° to the field and is stable for only one orientation about its shape

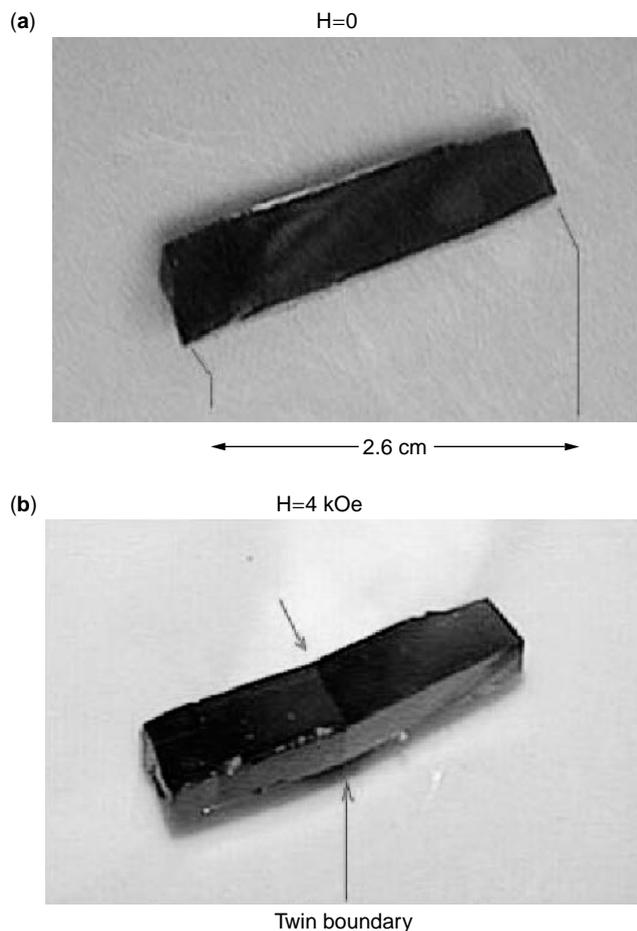


Figure 6. a) Photo of single-variant sample of $\text{Ni}_{49.4}\text{Mn}_{29.7}\text{Ga}_{20.9}$ at room temperature in zero external field. b) The same sample with a 6° kink at a twin boundary introduced by application of a field of 320 kA/m (4 kOe). This kink defines a 5% shear strain relative to the unchanged variant at the left end of the sample (30).

axis. The same orientation is stable if the field is reversed. This indicates that this sample is characterized by a uniaxial magnetic anisotropy with the easy axis of magnetization oriented at about 45° to the sample length. X-ray diffraction shows this magnetic easy axis to lie along the crystallographic c -axis.

Given this orientation of the crystal structure, Fig. 7 illustrates how the crystal shown in Fig. 6 can be cut to optimize axial strain. The orientation of the twin boundary observed in Fig. 6 and sketched in Fig. 7(b), coincides with a $\{101\}$ plane (a $\{112\}$ plane in the unit cell of Fig. 3b). Note that the orientation of the c axis changes across the twin boundary.

Because the c -axis changes direction across the twin boundary, the preferred direction of magnetization also changes. The directions of magnetization shown in Fig. 7(b) have been confirmed by scanning the four long sample surfaces with a small Hall probe (30). The field normal to the sample surface, arising from the component of its magnetization perpendicular to the surface, was found to map very closely with the twin structure, changing sign across the twin boundary and from the front to the back

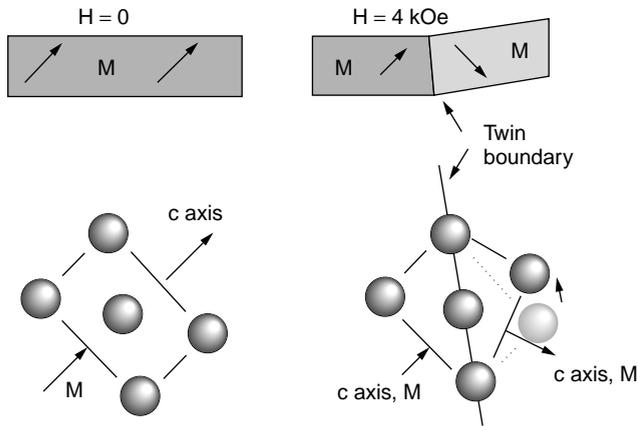


Figure 7. At the top are shown schematic views of the sample in Fig. 6 with the directions of magnetization shown for each variant. Below is sketched the orientation of the martensitic crystal structure, left, and the twinned martensite, right. The atomic displacements increase further to the right of the twin boundary shown, maintaining a constant shear angle.

of the sample. This change in the preferred direction of magnetization provides the mechanism for field-induced motion of the atoms that constitutes twin-boundary motion.

Application of a field orthogonal to the magnetization of a single-variant sample may cause a new variant to nucleate (perhaps at a surface defect) and grow. The twin boundary in Fig. 6(b) can be moved along the sample length by sliding the sample over the edge of a suitably oriented permanent magnet (30). The field at the corner of the magnet is approximately 320 kA/m (4 kOe). The magnetic field exerts a pressure on the twin boundary largely by virtue of the difference in Zeeman energy, $\Delta\mu_0\mathbf{M}_s\cdot\mathbf{H}$, between the two variants. For this pressure to exist, it is necessary that the anisotropy energy density, $K_u = 1.8 \times 10^5 \text{ J/m}^3$, be comparable to or greater than the Zeeman energy density difference, $K_u > \mu_0\mathbf{M}_s\mathbf{H}$. (If, on the other hand, the Zeeman energy is much greater than the anisotropy energy, the magnetization vectors in the two variants align with the field, and the Zeeman energy difference vanishes.) If the magnetic pressure on the twin boundary is greater than the energy density u_{tb} associated with twin-boundary motion, $\mu_0\mathbf{M}_s\mathbf{H} > u_{tb}$, then field-induced twin-boundary motion may result. The energy u_{tb} can be estimated from the stress-strain data such as that depicted in Fig. 4. A critical mechanical stress σ_0 is needed to initiate twin-boundary motion in martensite (30,31). Martynov and Kokorin (26) show that σ_0 can be as small as a few MPa. Using $\varepsilon_0 \approx 0.06$ at this critical strain, $u_{tb} \approx \varepsilon_0\sigma_0$ is of order 10^5 J/m^3 . The detailed interplay of these comparable energy densities, K_u , $\mu_0\mathbf{M}_s\mathbf{H}$, and u_{tb} , as well as the magnetostatic energy (which is a strong function of sample shape and is of order 10^4 to 10^5 J/m^3 for Ni_2MnGa) is best understood using quantitative models. These are described in the next section.

Because the strain across a twin boundary is a pure shear strain, samples having the active twin planes at 45° to the sample end faces can show extensional strain under twin boundary motion (as observed in Fig. 1). When the

active twin planes are at 45° to the sample surfaces, the FSMA behaves more like a magnetostrictive material such as Terfenol-D, $[(\text{Tb}_{0.3}\text{Dy}_{0.7})\text{Fe}_2]$ or a piezoelectric; it extends or contracts in the field direction and conserves volume to first order.

QUANTITATIVE MODELS OF TWIN-BOUNDARY MOTION

Three of the models describing the field-induced motion of twin boundaries in FSMA are reviewed. They are the numerical micromagnetic model (2,18), and the analytic, thermodynamic models of O'Handley (32) and Likhachev and Ullakko (31,33,34). Thermodynamic models developed by Vasil'ev et al. (17) to describe the composition dependence of T_C and T_m and by Lvov et al. (35) to describe the magnetization versus temperature behavior through the martensitic transformation are not described. A thermodynamic model of the relative stability of limiting states in a twinned FSMA (36) will be discussed later.

These models generally include the Zeeman energy, magnetic anisotropy energy, and an external stress. They may also include a magnetostatic energy (which tends to restore \mathbf{M} to zero when the field vanishes), an internal elastic energy (which tends to restore the field-induced strain to zero when the field is removed), and energies associated with the parameters σ_0 and C_{tb} shown in Fig. 4. The action of the external stress depends on its orientation relative to the field-induced deformation. Here, the stress is assumed to be oriented to oppose the field-induced strain. When the anisotropy is very large, the models generally predict a stress-strain product that increases linearly with applied field times the saturation magnetization. (If the external stress is increased, less strain-per-unit-field results.) When the anisotropy is not sufficiently strong to keep the magnetization along the crystallographic easy axis in the unfavorably oriented twins (i.e., those variants in which \mathbf{M} is perpendicular to \mathbf{H}), the Zeeman energy difference across the twin boundary is decreased by application of a field and the achievable strain is limited.

Micromagnetics

James and Wuttig (2) adapted for FSMA a numerical micromagnetic theory originally developed to describe the magnetostriction of Terfenol-D (37). Their model includes Zeeman energy, external stress, and magnetostatic energy in a twinned sample at the micromagnetic level:

$$g(\langle\varepsilon\rangle, \langle\mathbf{H}\rangle) = -\mu_0 H_0 \langle\mathbf{M}\rangle - \sigma_0 \langle\varepsilon\rangle + \left(\frac{\mu_0}{2}\right) \langle\mathbf{M}\rangle \mathbf{D} \langle\mathbf{M}\rangle. \quad (1)$$

The strain $\varepsilon(\mathbf{x})$ and magnetization $\mathbf{M}(\mathbf{x})$ are position dependent, and the energy density is calculated from their average values in each twin variant or domain, respectively. The external magnetic field and stress are given by H_0 and σ_0 , respectively. \mathbf{D} is the demagnetization tensor of the twin or domain. This energy is minimized subject to the constraint that the strains be determined by the allowed twin systems in the martensite. In addition to predicting linear ε - H_0 characteristics (18), detailed magnetization

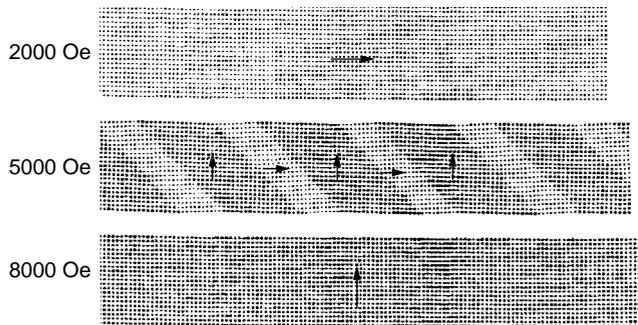


Figure 8. Crystallography and variant magnetization directions calculated for martensitic Ni₂MnGa from Eq. (1). Increasing applied field (vertical in this figure) stabilizes variants having smaller lattice constant in the magnetization direction (38).

distributions in the twins can be plotted. Figure 8 shows the results of numerical minimization of Eq. (1) with the field applied orthogonal to the initial single-variant magnetization direction. The variant magnetization prefers to lie along the tetragonal c -axis. Increasing the field strength results in stabilization of new twin variants having magnetization more closely aligned with the field direction. Note the mirror symmetry in the crystal structure across the twin boundaries. In addition to a net contraction of the material in the direction of the applied field, the surfaces become ridged at the twin boundaries.

When $\mu_0 M_s H \approx K_u$, the external field can cause the moment in unfavorably oriented variants (M nearly horizontal in Fig. 8) to rotate away from the local c -axis. As the magnetization vectors in two adjacent variants become parallel to each other, the Zeeman energy difference across the twin boundary decreases and the field-induced twin boundary motion is diminished. When magnetocrystalline anisotropy is included in the micromagnetic model (39), the rotation of the domain magnetization by the external field can be mapped from variant to variant as an applied field alters the variant distribution. Plots similar to those

in Fig. 8 have been used to interpret the rich twin/domain images observed by magnetic force microscopy in NiMnGa samples (40).

These numerical micromagnetic models can include micromagnetic effects not tractable in analytic models. However, numerical models do not offer some of the insights afforded by analytic models.

Analytic models

O'Handley (32) wrote a free energy expression for an FSMA comprised of two twin variants separated by a single mobile twin boundary. Zeeman energy, $\mu_0 \mathbf{M}_s \cdot \mathbf{H}$, magnetic anisotropy energy, $K_u \sin^2 \theta$, and an internal restoring elastic energy, $C_{\text{eff}} \varepsilon_0^2 / 2$, were included initially; an external stress-induced energy, $\sigma \cdot \mathbf{e}$, was added later (8,41):

$$g_i = -\mu_0 \mathbf{M}_i \cdot \mathbf{H} + K_u \sin^2 \theta_i + \left(\frac{1}{2} \right) C_{\text{eff}} e^2 + \tilde{\sigma} \cdot \tilde{\varepsilon}. \quad (2)$$

The subscript i corresponds to variant 1 or 2 and the directions of the c axis and the local magnetic moment define the angle θ_i as in Fig. 9. The elastic energy density expresses energy stored in unresponsive variants; C_{eff} is the stiffness of the martensite in the presence of mobile twin boundaries (Fig. 4). The two-dimensional representation is justified because the deformation and magnetization changes occur only in the plane defined by the c -axes of adjacent variants. It applies to multivariant samples having more than two c -axis directions only with modifications (33).

The field-dependent strain is expressed purely as a function of the volume fraction, f_i , of each variant: $\varepsilon(H) = \varepsilon_0 \delta f(H)$, where ε_0 is the transformation strain and $\delta f = f_1 - \frac{1}{2}$ is defined in Fig. 9. The equivariant state, $f_1 = f_2 = \frac{1}{2}$, is defined here as $\varepsilon = 0$. In contrast, the expression for magnetization includes contributions from both twin-boundary motion and magnetization rotation within the twin variants (32).

In the strong anisotropy limit, the sample is magnetized only by twin-boundary motion rather than magnetization

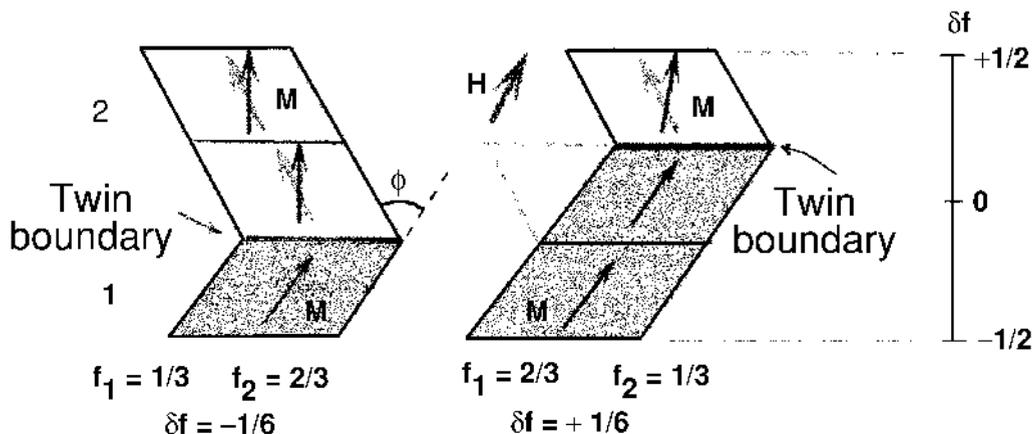


Figure 9. Two-dimensional representation of field-induced twin-boundary motion. The parameters f_i ($i = 1, 2$) are the volume fractions of variants 1 and 2 and $\delta f = f_1 - \frac{1}{2}$ describes the displacement of the twin boundary from $\varepsilon = 0$ as shown at the far right.

rotation. The external stress is applied in a direction that opposes the action of the field; for Ni_2MnGa , that would be a compressive stress parallel to the c -axis of the twin variant that is unfavorably oriented with respect to the field. The total energy density in Eq. (1) can be minimized with respect to the twin-boundary displacement, δf , as well as with respect to the angles θ_i . In the special case $H \parallel M_1$, $\theta_1 = 0$ and only the θ_2 minimization need be done. The latter minimization gives $h = \sin \theta_2$, where $h = \mu_0 M_s H / (2K_u)$, and the two minimizations combine to give

$$\begin{aligned} \varepsilon(H) = \varepsilon_0 \delta f &= \frac{\mu_0 M_s H (1 - h) + K_u h^2 - \sigma \varepsilon_0}{C_{\text{eff}} \varepsilon_0} \\ &= \frac{2K_u h (1 - h/2) - \sigma \varepsilon_0}{C_{\text{eff}} \varepsilon_0}. \end{aligned} \quad (3)$$

For large K_u (small h), Eq. (3) indicates that $\varepsilon(H) = (\mu_0 M_s H - \sigma \varepsilon_0) / C_{\text{eff}} \varepsilon_0$. That is, the low-field strain response of the system is governed by the balance between the Zeeman energy and the external stress. This equation describing strain by twin variant rearrangement as positive *linear* in H is to be compared with the *quadratic* form for conventional magnetostrictive strain in an elastically isotropic, magnetically uniaxial material, $\varepsilon = (3/2) \lambda_s (x^2 - 1/3)$ where $x = \cos \theta = M/M_s = H/H_a$ for a hard-axis magnetization process (42). The field-induced strain in Eq. (3) can be shifted rigidly toward negative values by an external compressive stress. Further, the strain scales inversely with the internal elastic restorative stress, $C_{\text{eff}} \varepsilon_0$, due to the noncompliant parts of the sample. If this restorative stress is small (i.e., the twin boundaries move easily), a twinned sample strains easily ($d\varepsilon/dH$ approaches infinity), and even for large K_u , it magnetizes easily (dM/dH

approaches infinity) once the critical field for twin-boundary motion is exceeded.

On approaching rotational magnetic saturation ($h = 1$), $\varepsilon(H)$ scales with $(K_u - \sigma \varepsilon_0) / C_{\text{eff}} \varepsilon_0$ and is no longer linear in H . That is, the maximum field-induced strain is limited by the balance between the anisotropy and the external stress. The FSMA can show positive or negative strain depending on the relative magnitudes of anisotropy and applied stress.

These conclusions are plotted numerically in Fig. 10 for six different combinations of anisotropy-energy density and internal strain energy (32). At the upper left, for large anisotropy ($K_u = 3 \times 10^6 \text{ J/m}^3$) and small internal stress-induced energy ($0.5 C_{\text{eff}} \varepsilon_0^2 = 2 \times 10^5 \text{ J/m}^3$), the sample can be magnetized only by twin-boundary motion, not by magnetization rotation. Thus, the reduced magnetization, m , and the reduced strain, ε_y , increase together, linearly in applied field. They reach their saturation values in relatively small fields. Moving across the top row of panels, the anisotropy remains strong, but the internal restoring stress increases. The field dependence of m and ε remain correlated and close to linear even up to $\mu_0 H = 1 \text{ T}$ but the magnetic and strain response achieved in these fields are reduced because twin-boundary motion faces greater internal elastic opposition. The lower row of panels begins with both weak anisotropy and weak internal restoring force. The weak anisotropy allows the field to rotate the magnetization in the unfavorably oriented variant, increasing m . However, the rotation of the magnetization reduces the Zeeman pressure on the twin boundary. As a result, the magnetization can increase (by rotation into the field direction) while the strain lags behind because it depends only on twin-boundary motion. Further, the

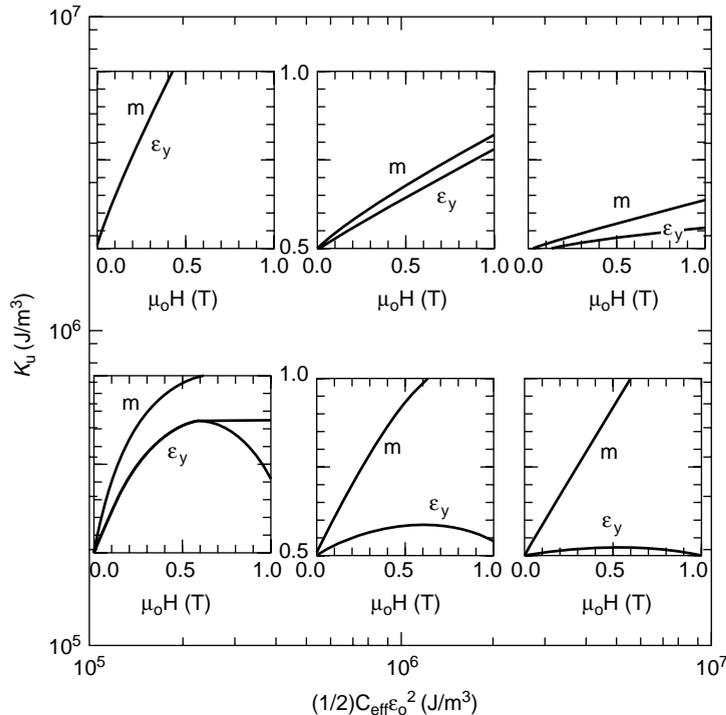


Figure 10. Graphs of reduced magnetization, $m = M/M_S$, and reduced strain, $\varepsilon_y = \varepsilon/\varepsilon_0$ from Eq. (3) with $\sigma = 0$ in the intermediate anisotropy case, $h \approx 1$ for $\mu_0 M_S = 1.25 \text{ T}$ ($M_S = 1000 \text{ emu/cm}^3$). The twin geometry and field orientation are chosen to be as shown in Fig. 9. Curves are plotted for six combinations of magnetic anisotropy and internal elastic energy densities given by the bold ticks on the outer coordinate axes. Note that in all cases the remanence is 0.5 because of the geometry chosen. A nonzero applied stress would simply shift the strain values negative, including the value at $H = 0$ (32).

field-induced strain gained by twin-boundary motion at weak fields may be reclaimed by the material (depending on the irreversibility of the twin-boundary motion) as the Zeeman pressure on the twin boundaries decreases with increasing field, allowing the internal stored stress to relax. These effects become even more pronounced on moving to stronger internal stress as shown in the bottom row of inset figures. The magnetization curves calculated here differ from the hard-axis, magnetization-rotation process shown in Fig. 5 because these are the result of twin boundary motion as well as magnetization rotation; that in Fig. 5 is for magnetization rotation only as the sample there is constrained to the single-variant state. It should be noted that when M_2 rotates toward H in the weak anisotropy limit, the Zeeman energy difference decreases like $\sin \theta_2$ and the anisotropy energy difference increases like $\sin^2 \theta_2$. The net magnetic pressure reaches a maximum value of K_u when $h = 1$.

This model also indicates that the field component responsible for driving twin-boundary motion is that parallel to the twin plane. The data of Fig. 1(b), as well as $M(H)$ for that sample, are well described by this model with $\sigma = 0$, $K_u = 2.2 \times 10^5 \text{ J/m}^3$, and $C_{\text{eff}\varepsilon_0} = 2.6 \text{ MPa}$ (32).

The consequences of the parameters σ_0 and C_{tb} (Fig. 4) are to increase the irreversibility of the strain-versus-field loops. The twin boundary yield stress, σ_0 , adds a term to the applied field in Eq. (3) with a sign that always opposes the applied field: $H \rightarrow H \pm \sigma_0 \varepsilon_0 / \mu_0 M_s$. This addition to the applied field is a twin-boundary coercivity, the field needed to initiate twin-boundary motion. If one were to take account of the threshold field needed to initiate twin-boundary motion, the curves in Fig. 10 would be shifted on the field axis to show hysteresis with the increasing-field (decreasing-field) curves displaced to the right (left) by H_c .

Likhachev and Ullakko (31,33,34) have taken a more general analytic approach to the problem in three dimensions. They integrate the Maxwell relation

$$\frac{\partial \sigma}{\partial m} = -\frac{\partial h}{\partial \varepsilon} \quad (4)$$

to arrive at the field-dependent strain:

$$\varepsilon(H) = \left(\frac{\partial \sigma}{\partial \varepsilon} \right)_{\varepsilon=0}^{-1} \left[\frac{\partial}{\partial \varepsilon} \int_0^H \mu_0 M dH \right]_{\varepsilon=0}. \quad (5)$$

They assumed that from the equivariant state, one-third of the material is easily magnetized (the variant fraction that has its c -axis parallel to the field); H_{\parallel} is the anisotropy field for such *axial* variants. H_{\parallel} should be the shape anisotropy of the sample; about 0.1 MA/m in Fig. 5). Also, two-thirds of the material (those variants with c perpendicular to H) will have to be magnetized in their hard direction with a transverse anisotropy field, $H_{\perp} > H_{\parallel}$ (compare Fig. 5). As the equivariant state is upset by twin-boundary motion under the application of a field, the magnetization and strain can be expressed as functions of x , the volume fraction of axial variants, which plays a role similar to δf in the 2-D model (Fig. 9). The parameter x is eliminated between

the expressions for $\varepsilon^{\text{FSMA}}(H, x)$ and $M(H, x)$, giving

$$\varepsilon^{\text{FSMA}} = \frac{2}{3} \left(\varepsilon_0 \frac{d\sigma}{d\varepsilon} \right)_{\varepsilon=0}^{-1} \int_0^H \mu_0 \{M_{\parallel}(H) - M_{\perp}(H)\} dH. \quad (6)$$

This result indicates that the field-induced strain depends inversely on $\varepsilon_0 d\sigma/d\varepsilon = C_{\text{tb}\varepsilon_0}$. Further, the strain varies linearly with the magnetic field energy stored in the material because the integral in Eq. (6) is the anisotropy energy (the area between the two $M(H)$ curves in Fig. 5) that must be overcome by the applied field in order to saturate the material. Experimental $M(H)$ and $\sigma(\varepsilon)$ data are used as input to Eq. (6) to allow prediction of $\varepsilon(H, \sigma)$ (31). At saturation Eq. (6) reduces to

$$\varepsilon_{\text{Sal}}^{\text{FSMA}} = \frac{1}{3} \left(\varepsilon_0 \frac{d\sigma_0}{d\varepsilon} \right)_{\varepsilon=0}^{-1} (H_{\perp} - H_{\parallel}) \mu_0 M_s \approx \frac{\mu_0 M_s \Delta H}{3C_{\text{tb}\varepsilon_0}}. \quad (7)$$

The term $\mu_0 M_s H (\Delta H/H)$ expresses succinctly and elegantly the dual importance of both the Zeeman energy, $\mu_0 M_s H$, and the anisotropy, $\Delta H/H$, which make up the energy difference across a twin boundary. This interplay of both Zeeman and anisotropy energy is represented by the first two terms in the numerator of Eq. (3). The model of Likhachev and Ullakko suggests that the field-induced strain is limited by the internal stress, $\sigma_0 = C_{\text{tb}\varepsilon_0}$, needed to nucleate and move mobile twin boundaries in the martensitic phase. This term has a different meaning than $C_{\text{eff}\varepsilon_0}$ in Eq. (3) although the magnitude of both terms is similar. In both models, a value for σ_0 of order a few MPa provides a good fit to data in Fig. 1.

The analytic models suggest that large values of the difference $H_{\perp} - H_{\parallel}$ (or large K_u), large saturation magnetization, and low-threshold stress, σ_0 , are critical parameters for achieving large field-induced strain by twin-boundary motion.

Finally, it is important to be able to extend the results of these single-crystal models to polycrystalline materials that may be used in some applications. Bhattacharya and Kohn (43) have given a Taylor (*lower*) bound to the strain that can be achieved in a random polycrystalline shape memory material. They find that for cubic-to-tetragonal martensites, the lower bound to the strain is zero. However, for lower symmetries, some fraction of the transformation strain can be induced by stress because in lower symmetry structures there are more planes along which the system can twin to accommodate an external stress.

Marioni et al. (44) have taken FSMA twin variants, described in a model similar to that in O'Handley (32), and assembled them with various textures to simulate noninteracting grains in a polycrystalline sample. This model gives an *upper* bound for the maximum field-induced strain because grain-to-grain elastic interactions are treated only in a mean-field way by the parameter $C_{\text{eff}\varepsilon_0}$. They find that for tetragonal FSMA, a random polycrystalline sample could give a field-induced strain up to 20% of ε_0 . For a texture in which twin plane normals are distributed about the field direction at a common angle (a texture which may be achieved by uniaxial compression), the field-induced strain can be as much as 34% of ε_0 . In addition, this model predicts a threshold field below which there is little strain

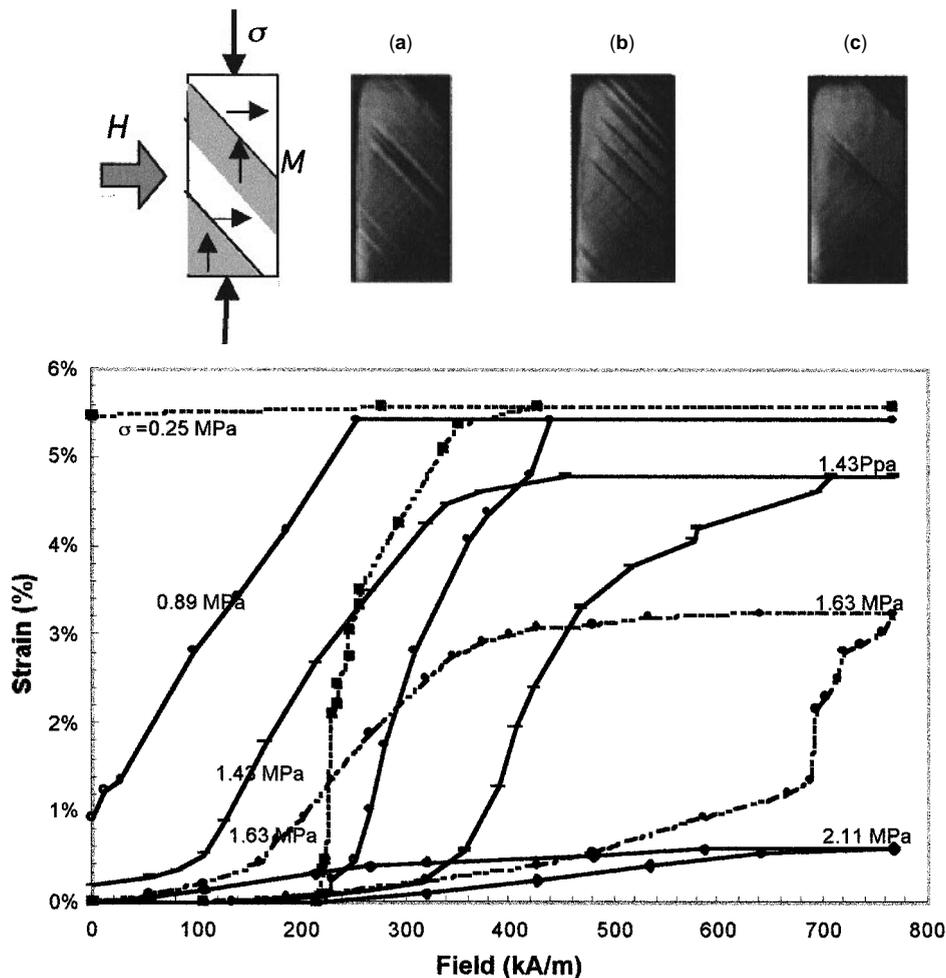


Figure 11. Above left: Orientation of M , H , and σ relative to the twinned sample. High-speed video frames (a), (b), and (c) show the sample in the initial vertically compressed ($\sigma \approx 1$ MPa) state ($H = 0$), in an intermediate state, and in the final magnetically saturated (and fully strained vertically by 6%) state, respectively. Below: Selected field-induced strain curves at various external opposing stresses at room temperature for a $\text{Ni}_{49.7}\text{Mn}_{28.5}\text{Ga}_{21.7}$ single crystal (8).

response because the component of the applied field parallel to the twin planes of a given orientation is less than the twin-motion coercivity. For fields above this threshold, the strain increases rapidly toward the linear value predicted by the thermodynamic models. The physical origin of this threshold field in polycrystalline samples is different from that associated with σ_0 and the initiation of twin-boundary motion.

FIELD-INDUCED STRAIN UNDER LOAD

With the background developed so far, it is now possible to describe and analyze in more detail the measurements of field-induced strain under load that form the basis for use of FSMAs as actuator materials.

dc Actuation under Static Stress

Figure 11 shows the results of field-induced strain measurements in a $\text{Ni}_{49.8}\text{Mn}_{28.5}\text{Ga}_{21.7}$ single crystal at room temperature for various axial external stresses that oppose the field-induced strain. The sketch at the upper left shows

the orientation of the magnetization, magnetic field, and external stress relative to the twinned sample. The three photographs are frames from a high-speed (1200 frames/s) video taken on the sample close to the initially stressed state (approximately 1 MPa) at $H = 0$ (frame a), at about 15 ms into the actuation (frame b), and at saturation after about 23 ms (frame c). The structure in frame (a) is dominated by the dark twin bands (M parallel to stress). In frame (b) the lighter twin bands (M parallel to H) fill more than half the area on the front surface. In frame (c) the sample is essentially filled by the light-colored twin bands (except for the thin twin band that apparently remains pinned).

The graph in Fig. 11 shows the ε - H loops with fairly abrupt strain changes of several percent, occurring over a narrow field range. On returning the field to zero, significant hysteresis is evident. With increasing external stress, the threshold field for strain actuation increases and the strain at saturation decreases. At low external stress, the field-induced strain does not reset to $\varepsilon = 0$ upon removal of the field. This suggests that C_{eff} is small in this case. At strain levels in excess of 0.7 MPa, the sample resets

Table 1. Mechanical and Magnetic Parameters measured for Single-Crystal Ni-Mn-Ga and Used in Eq. 3 in the Model Results of Fig. 12

Parameter	Range of Measured Values	Values used in Fig. 12
σ_0 (MPa)	0.8–1.5	1.0
C_0 (GPa)	70	—
C_{tb} (MPa)	15–25	18
ε_0 (%)	5.7–6.1	5.7
$\mu_0 M_s$ (T)	0.60–0.63	0.60
K_u (MJ/m ³)	0.16–0.19	0.18
H_c (MA/m)	0.1–0.16	0.093

to condition a) when $H = 0$. The hysteresis appears non-monotonic in applied stress. Nearly the full transformation strain is achieved in this sample for stresses less than 0.5 MPa. Some samples have shown strains of 6.1% at saturation (8,45).

The data of Fig. 11 can be modeled with Eq. 3 using the measured Ni-Mn-Ga parameters (see Table 1). The hysteresis is accounted for in an ad-hoc manner by adding to the applied field a coercive (offset) field $H_c = \pm \sigma_0 \varepsilon_0 / (\mu_0 M_s) = 93.3$ kA/m (1.17 kOe). The model results displayed in Fig. 12(a) and (b) show that these parameters give a reasonable reproduction of the major trends in the experimental data, namely the shape of $\varepsilon(H)$, as well as the increase in threshold field and decrease in strain with increasing external stress. The model does not account for the observed change in coercivity with external stress. The predicted decrease in saturation strain with increasing stress, Fig. 12 (b), is consistent with observations.

The limitations of the model in fitting the data may be the result of neglecting magnetostatic effects in the model. It is not the result of a stress-induced anisotropy ($3\sigma\lambda/2$) adding to K_u as external stress is increased. Based on the measured magnetostriction of the martensitic phase, $\lambda_s = -145 \times 10^{-6}$ (28), the magnetoelastic anisotropy at $\sigma = 2$ MPa is more than two orders of magnitude smaller than K_u .

Alternating-Field Actuation under Dynamic Load

The quasi-static, field-induced strain measurements shown in Fig. 11 have also been carried out at actuation frequencies up to 330 Hz. To perform these measurements, the static load was replaced with a spring against which the sample extends under a transverse field (Fig. 13).

Figure 14 shows a set of field-induced strain curves taken at 1 Hz drive, 2 Hz actuation in the system shown in Fig. 13 (46). The sample is a single crystal of Ni-Mn-Ga measuring $1 \times 5 \times 7$ mm, with the field applied normal to the 1×7 mm face and the strain measured along the 7 mm direction. The saturation strain at any given stress level increases with increasing stress, reaching a maximum value near 1.4 MPa. Note the much smaller hysteresis in this case compared to the quasi-static situation shown in Fig. 11. For larger stresses, the saturation strain decreases, and the hysteresis as well as the threshold field for actuation increase.

When the sample is driven to higher frequencies than in Fig. 14, the saturation strain is unchanged up to an

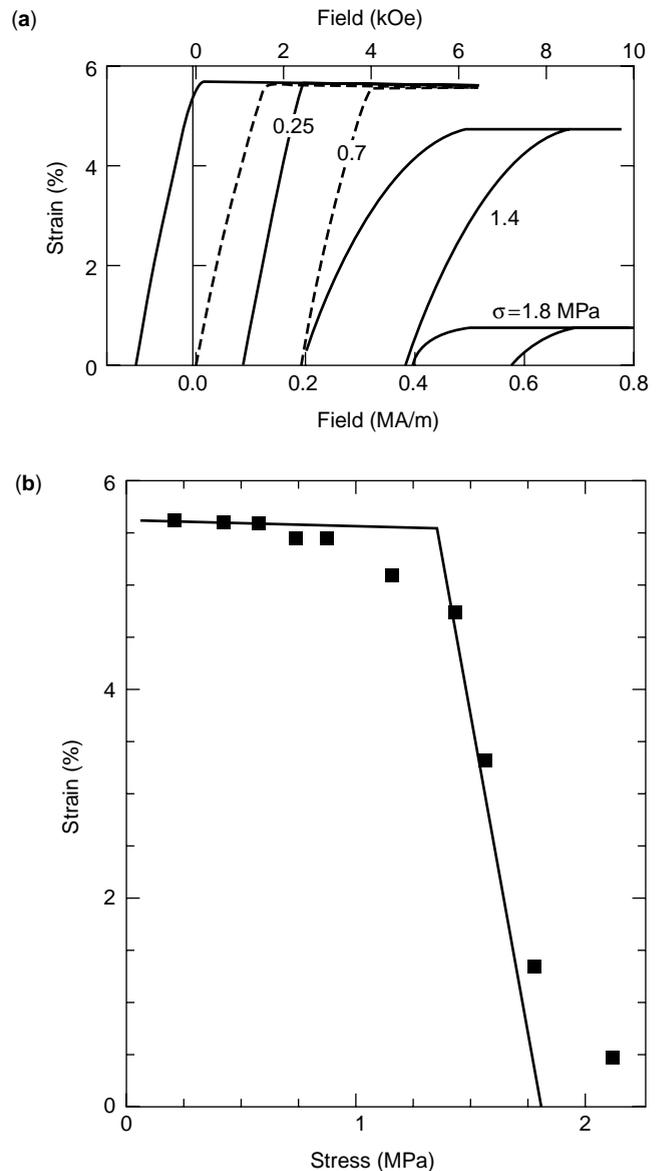


Figure 12. a) Calculated strain versus applied field curves from Eq. (3). b) Calculated strain at peak field versus stress with overlaid experimental points from the full set of data depicted in Fig. 11 (8).

actuation frequency of about 100 Hz. Beyond that value, the peak output strain drops off sharply. However, it is clear from the data in Fig. 15 that the drop-off in response is due to the reduction in the applied field. Above a drive frequency of about 50 Hz, the inductive reactance of the field coils becomes sufficiently large that the power supply can no longer deliver the current needed to generate a field sufficient to saturate the strain. Power supplies must be designed to match the impedance of the load over the operating bandwidth. Pulsed-magnetic-field measurements with a drive-field rise time of 1 ms indicate that single-crystal samples of Ni-Mn-Ga can strain at a rate that keeps up with the rise time of the pulse (47). This implies a bandwidth of at least 1 kHz for these materials.

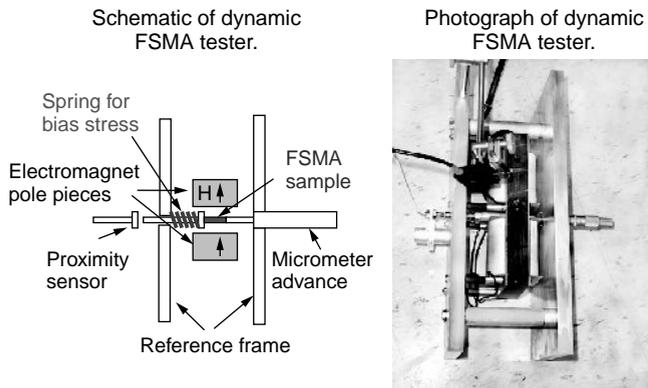


Figure 13. Test system, schematic (left) and photograph (right) The FSMA sample is subjected to a bias stress along one axis and an ac magnetic field along an orthogonal axis. A micrometer advances the sample into the spring to establish the desired bias stress level. The sample elongation under applied field is measured with an eddy-current proximity sensor. The apparatus measures approximately $12 \times 15 \times 30$ cm (46).

DISCUSSION

Engineering Parameters

James and Wuttig (18) have observed the rearrangement of twin boundaries in martensitic $\text{Fe}_{70}\text{Pd}_{30}$ accompanying a field-induced 0.5% extensional strain. Similarly, Ni–Mn–Ga crystals show extensional strains under quasistatic excitation of $\varepsilon > 4\%$ at room temperature (8,45). AC strains in excess of 3% in Ni–Mn–Ga crystals at room temperature have been reported (46). The response of active magnetic materials is generally described by the magnetostrictivity defined as $\mathbf{d}_{ij} = \partial \varepsilon_i / \partial H_j$, where the subscripts refer to directions in Cartesian coordinates. Magnetostrictive materials such as Terfenol-D are often operated under field bias. In the case of the FSMA data shown in Fig. 14, application of a bias field of about 2 kOe and an ac field of ± 1 kOe about that bias, would result in actuation at the drive frequency with an output strain of about 2% peak-to-peak. The value of \mathbf{d}_{31} under such actuation is about 1% per kOe or 12.5×10^{-8} m/A. This value compares

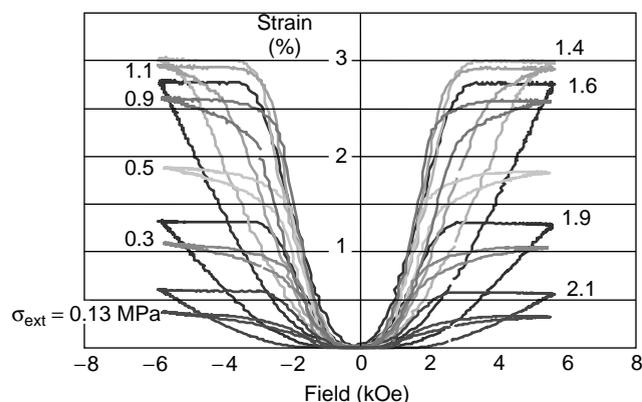


Figure 14. Field-induced strain data for several values of applied stress at an actuation frequency of 2 Hz. The sample elongates against the spring for both positive and negative field cycles, giving an actuation frequency twice the drive frequency.

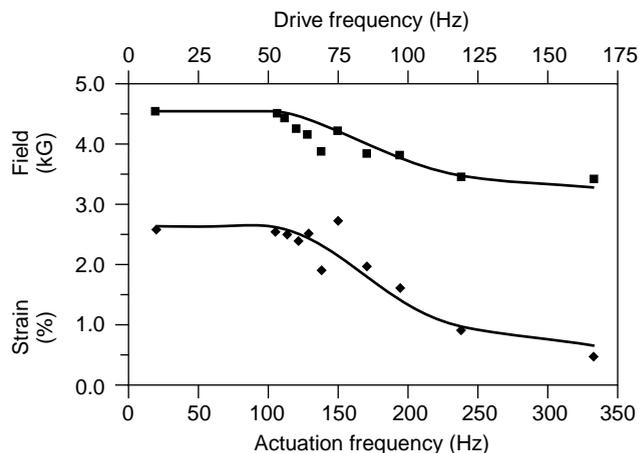


Figure 15. Frequency dependence of peak field generated in the system shown in Fig. 13 (upper curve) and the strain at peak field. The decrease in actuation strain above a drive frequency of 50 Hz has been identified as due to the decrease in field applied to the sample (46).

favorably with the value of \mathbf{d}_{33} for Terfenol-D in Table 2. Table 2 compares the magnetostrictivities, \mathbf{d}_{ij} , for two active FSMAs and the leading magnetostrictive material, Terfenol-D, $\text{Fe}_2(\text{Tb}_{0.3}\text{Dy}_{0.7})$.

The negative (positive) sign of \mathbf{d}_{33} (\mathbf{d}_{31}) for Ni–Mn–Ga reflects the fact that it contracts along the axis in which the magnetization increases and expands along the original axis of \mathbf{M} . The magnetomechanical coupling coefficient, k , is defined for a magnetically driven actuator by the ratio of the output mechanical energy to the total input energy (magnetic plus mechanical). For purposes of determining the coupling coefficient, the following relations based on clamped and free permeabilities or free and clamped elastic moduli, can be derived: $\mu^\varepsilon = (1 - k^2)\mu^0$ or $C^H = (1 - k^2)C^M$ (6). From Fig. 4 and the data in Table 1, the latter relation suggests that k approaches unity for these materials, that is, they couple magnetic energy to a mechanical load with near-perfect efficiency (48). (The free modulus, C^H is taken as C_{tb} in Table 1 and the clamped modulus, C^M is given by C_0 in Table 1.)

Stress Dependence of $\varepsilon(H)$

The introduction of an external stress in the thermodynamic model when $C_{\text{eff}} \neq 0$ predicts a strain that decreases linearly with applied stress, Eq. (3), and Fig. 16(a).

Table 2. Comparison of Currently Achieved Field-Induced Strain and Magnetoelastic Coupling Coefficients \mathbf{d}_{33} and \mathbf{d}_{31} in FePd and Ni–Mn–Ga FSMAs with the Magnetostrictive Material, Terfenol-D

Active Magnetic Material	ε (%)	H-field (kA/m)	\mathbf{d}_{33} (10^{-8} m/A)	\mathbf{d}_{31} (10^{-8} m/A)	k
Fe₇₀Pd₃₀	0.5	800	+0.63	−0.63	—
Ni₅₀Mn₂₈Ga₂₂	6.1 ^a	400	−25	+25	—
	3.1 ^b	400	−13	+13	0.8 (1.0)
Terfenol-D	0.2	40	+6	−3	0.8

^a Quasi-static actuation.

^b ac actuation.

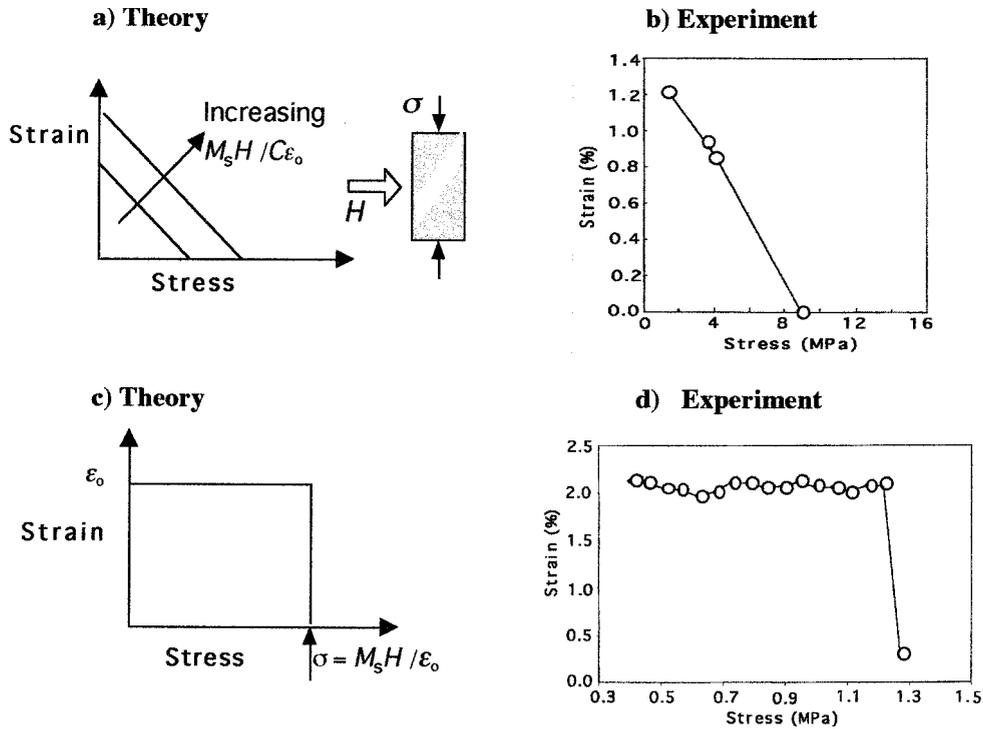


Figure 16. (a) Field-induced strain versus stress predicted by single-variant thermodynamic model (32) (b) Field-induced strain versus stress observed in Ni_2MnGa at -15°C (38). (c) Field-induced strain versus stress calculated with no restoring force (30). (d) Field-induced strain versus stress observed in Ni–Mn–Ga at room temperature (36).

Figure 16(b) shows one set of strain versus stress data for Ni_2MnGa at -15°C and $H=12\text{ T}$ (29). The saturation strain achieved here is less than ε_0 , suggesting that many of the twin variants are not responding to the applied magnetic field. These unresponsive twins may present an elastic resistance ($C_{\text{eff}} \neq 0$) to the deformation caused by motion of the active twin boundaries. The observed blocking stress of 9.2 MPa (σ at which $\varepsilon = 0$) is calculated from Eq. (3) to be 5 MPa (using $K_{\text{u}} = 2.45 \times 10^5 \text{ J/m}^3$ and $\varepsilon_0 = 0.05$). In contrast, Murray et al. (36) have noted that different Ni–Mn–Ga crystals may respond to an applied field with little or no restoring force, namely $C_{\text{eff}} \approx 0$. When the stored elastic energy is omitted from Eq. (2), the free energy cannot be minimized for $|\delta f| < 0.5$. Instead, an instability arises in which the twin boundary moves completely ($\delta f = \pm 0.5$) in the direction favored by the field if $\mu_0 M_s H > \sigma \varepsilon_0$ and in the opposite direction if $\mu_0 M_s H < \sigma \varepsilon_0$. There is no internal elastic opposition to this motion when there are no unfavorably oriented twin planes. In such situations, the strain under load does not decrease linearly with stress but rather maintains a constant value until a critical stress is reached, $\sigma_c = \mu_0 M_s H / \varepsilon_0$, at which point the strain vanishes abruptly as in Fig. 16(c). Recent data on Ni–Mn–Ga at room temperature, Fig. 16(d), support this instability model (36). In addition, the field-induced strain in these FSMAs is more bistable (like a Barkhausen jump), whereas the FSMAs whose response is shown in Fig. 16 (b) show smoother, more reversible $\varepsilon(H)$ as depicted in Fig. 1 (36). Likhachev et al. (49) have recently shown data for the strain dependence of $\text{Ni}_{48}\text{Mn}_{30}\text{Ga}_{22}$ that

fall between these two limits and are well described by their model.

It thus appears that a range of $e(H, \sigma)$ responses may be able to be achieved with FSMAs. Smaller output strain with larger blocking stress may be achieved in some crystals, Fig. 16 (b), or larger output strain with smaller blocking stress may be observed in other crystals, Fig. 16 (d). The reasons behind these different types of response are not yet well understood. In the present case, the different temperatures and compositions (and hence different magnetic anisotropies, magnetizations, and mechanical properties) may be factors. Another difference between the two samples contrasted in Fig. 16 is that the one in panel (b) shows a much finer twin structure (measured in tens of microns) with multiple twin systems present. The sample represented in panel (d) shows a much coarser twin structure (twin spacing of order 0.5 – 1 mm), and only one family of twin boundaries is present. It may be possible that some of the variously oriented twin systems in the former sample may not respond to the applied field and hence provide a mechanism for storing energy elastically ($0.5C_{\text{eff}}\varepsilon_0^2$) as the active twins respond to the field.

Comparison with Shape-Memory Effects

Here, we compare the field-induced strain observed in FSMAs with (1) the thermoelastic shape memory effect (pseudoplasticity) and (2) stress-assisted martensitic transformations (superelasticity).

First, in the thermoelastic shape memory effect, a *twinned, martensitic* material is macroscopically deformed in a manner that appears to be plastic. In fact, the deformation is not the result of dislocation formation and motion, but rather twin-variant rearrangement (twin-boundary motion). Upon heating to the austenitic phase, the macroscopic deformation is erased by the structural transformation of the martensite to austenite. This is the one-way shape memory effect if cooling back to the martensitic state does not restore the macroscopic deformation. In some cases, cooling back to the martensitic state can restore the macroscopic deformation. This is called the “two-way shape-memory effect.”

This thermoelastic shape-memory effect achieves a shape change by structural transformation of the material between a twinned phase and a different untwinned phase. By contrast, the shape changes so far observed in FSMA are induced by a magnetic field fully within the martensitic state. It involves the field-induced motion of twin boundaries. Thus, the effect in FSMA may be faster and more efficient compared to the thermoelastic shape-memory effect where the need for heat transfer limits the kinetics.

Second, when a material showing the shape-memory effect is subjected to a stress at a temperature just above the martensite start temperature, the stress can facilitate the transformation to the martensitic phase. Once twinned martensite is formed, the stress can result in a large (several percent) macroscopic deformation of the material (superelasticity). Upon removal of the stress, the material re-transforms to the austenitic phase and the large deformation is erased. This effect can be much faster than thermally induced shape changes associated with the martensitic transformation.

FSMA have been shown to exhibit stress-induced martensite that then responds to a magnetic field with an additional strain (19). When the external stress is removed, the material reverts to the austenitic phase and the large field-induced strain decreases to the smaller value typical of the austenitic phase.

Comparison with Magnetostriction

The field-induced strain observed in FSMA is similar in some ways to the magnetostriction generally observed in ferromagnetic materials.

1. In both cases, the strains conserve volume to first order. Thus, the strain measured from the demagnetized or equi-twin-variant state in a direction perpendicular to the field will be $\varepsilon_{\perp} \approx -\varepsilon_{\parallel}/2$, where ε_{\parallel} is the field-induced strain parallel to the field. Non-uniform initial distributions of domain magnetizations or twin variants can upset this relation (as in field-biased or pre-stressed samples such as that shown in Fig. 11 where $\varepsilon_{\perp} \approx -\varepsilon_{\parallel}$).
2. The bending effect across the twin boundary (which is also a 90° domain wall), shown in Fig. 6 for Ni–Mn–Ga, would also occur in an appropriately cut ferromagnetic crystal such as Fe ($\langle 100 \rangle$ at 45° to bar axis) if a single 90° domain wall could be

isolated. (Because the magnetostriction coefficient of Fe in the $[100]$ direction, λ_{100} , is positive, the domain magnetization would be orthogonal to those pictured in Fig. 6 or the bending would be in the other direction.) However, in Fe the bend angle across the 90° domain wall would be only 0.002° corresponding to a shear strain of 2×10^{-5} .

The differences in field-induced-strain between FSMA and magnetostrictive materials are more important than the similarities.

1. Field-induced strain in FSMA is due to twin-boundary motion, which brings with it a change in the direction of magnetization. The FSMA strain is tied to the crystallography, not to the direction of \mathbf{M} . That is, it is possible to rotate \mathbf{M} with no FSMA strain, only conventional magnetostriction, in FSMA that are characterized by relatively weak anisotropy. In magnetostrictive materials, on the other hand, field-induced strain is a result of magnetization rotation relative to the crystallography; the strain is tied to \mathbf{M} , and not to the crystallographic orientation.
2. In the ferromagnetic martensitic phase, the increased magnetocrystalline anisotropy relative to austenite means that saturation of the magnetization requires stronger magnetic fields than in austenite. If $T_C > T_{\text{mart}}$, there is no large FSMA field-induced strain between T_{mart} and T_C because the material is in the austenitic phase and twins are not present. The FSMA strain shows a peak on heating through T_{mart} (28,49). If $T_C < T_{\text{mart}}$, there is a static strain in each variant of the martensitic phase above T_C , but it cannot be controlled by a field because $\mathbf{M} = 0$. It can be controlled by an applied stress. Magnetostrictive strain becomes possible below T_C with a *second-order magnetic* transformation; it has a temperature dependence governed by $[M(T)/M(0)]^{l(l+1)/2}$ (51). (Here, l defines the symmetry of the lowest-order crystal field term: $l = 2$ is uniaxial and $l = 4$ is cubic.) The crystal strain in an FSMA, on the other hand, appears in the martensitic phase by a *first-order structural* transformation below T_{mart} [James and Wuttig (2)].
3. Field-induced strains in FSMA decrease as the strength of the magnetocrystalline anisotropy of the martensitic phase decreases below $\mu_0 M_s H$. For weak anisotropy martensite, the field may rotate \mathbf{M} without moving the twin boundaries, and there is no change in macroscopic strain. The field-dependence of strain in FSMA—that do not show discontinuous $\varepsilon(H)$ versus σ behavior in Fig. 16(b) and (d)—is predicted to be linear in H below saturation in the strong anisotropy limit (18,32). Reduced anisotropy can introduce strong nonlinearities in $\varepsilon(H)$ (18,31–34). On the other hand, the magnetostrictive strain accessible in a given field $H < H_a$ will be greater, the smaller the anisotropy. The field dependence of magnetostrictive strains in a hard-axis magnetization process is quadratic in H or M below saturation (42).

SUMMARY

Ferromagnetic shape-memory alloys have shown field-induced strains at room temperature greater than those of any magnetostrictive, piezoelectric, or electrostrictive material. This strain is due to the motion of twin boundaries in the martensitic phase. A variety of $\varepsilon(H, \sigma)$ characteristics appears to be possible in FSMAs. Some samples tend to show abrupt twin-boundary motion as the sign of the quantity $2K_u h(1 - h/2) - \sigma \varepsilon_0$ changes. In the other limit of FSMA behavior, twin-boundary motion appears to be opposed by an internal elastic restoring energy, possibly associated with unfavorably oriented twin boundaries. In the latter case, the field-induced strain is smaller and more linear in the applied field for $h < 1$, and the blocking stresses can be greater.

Micromagnetic and analytic thermodynamic models are able to describe the main features of the magnetization patterns in the twinned FSMAs and the forms of $\varepsilon(H, \sigma)$, and $M(H)$ in single-crystal FSMAs, respectively. Field-induced strains in FSMAs show some incidental similarities to magnetostrictive strains but are essentially different, arising from the field-induced motion of twin boundaries in a martensitic phase that is strongly distorted by a first-order transformation not connected to T_C . The field-induced strains occur at smaller fields as the stress required to nucleate twin boundary motion, σ_o , decreases. Decreased magnetocrystalline anisotropy or increased external stress limits the magnitude of the field-induced strains. Unlike the thermoelastic shape-memory effect, large magnetic-field-induced strain in FSMAs so far is observed fully within the martensitic state.

ACKNOWLEDGMENTS

The authors acknowledge fruitful discussions with R.D. James and M. Wuttig. The work at MIT, described in this review was carried out largely by S.J. Murray, M. Marioni, and C.P. Henry. It has been supported by the Finnish Ministry of Science and Technology (TEKES) with a consortium of Finnish companies, by a subcontract from Boeing Corporation on a DARPA contract, by grants from the Lord Corporation, Midé Technologies, and the Office of Naval Research, as well as by contracts from DARPA, ACX Corporation, and Midé Technologies. The crystals used in our study were grown by Dr. V.V. Kokorin, Institute of Metallurgy, Kiev (Fig. 1) and by Dr. Tom Lograsso of Ames Laboratory, Department of Energy (Figs. 6, 11, and 14).

BIBLIOGRAPHY

1. V.A. Chernenko, E. Cesari, V.V. Kokorin, and N. Vitenko. *Scripta Metall. Mater.* **33**: 1239 (1995).
2. R.D. James and M. Wuttig. *SPIE* **2715**: 420 (1996).
3. K. Ullakko. *J. Mat. Eng. Perform.* **5**: 405 (1996).
4. K. Ullakko, J.K. Huang, C. Kantner, V.V. Kokorin, and R.C. O'Handley. *Appl. Phys. Lett.* **69**: 1966 (1996).
K. Ullakko, J.K. Huang, V.V. Kokorin, and R.C. O'Handley. *Scripta Mater.* **36**: 1133 (1997).
5. G.H. Haertling. *J. Am. Ceram. Soc.* **82**: 797 (1999).
6. A.E. Clark. In E.P. Wohlfarth ed., *Ferromagnetic Materials*, Vol. 1. North-Holland, Amsterdam, 1980, p. 531.
7. J.R. Cullen, A.E. Clark, and K.B. Hathaway. In K.H.J. Buschow, ed., *Handbook of Materials Science*. VCH Publishers, Amsterdam, 1997, ch. 16.
8. S.J. Murray, M. Marioni, S.M. Allen, R.C. O'Handley, and T. Lograsso. *Appl. Phys. Lett.* **87**: 886 (2000).
9. V.V. Kokorin, V.V. Martynov, and V.A. Chernenko. *Scr. Metall. Mater.* **26**: 175 (1992).
10. V.V. Martinov and V.V. Kokorin. *J. Physique III* **2**: 739 (1992).
11. A.N. Vasil'ev, A.D. Bozhko, V.V. Hkovailo, I.E. Dikshstein, V.G. Shavrov, V.D. Buchelnikov, M. Masumoto, S. Suzuki, T. Takagi, and J. Tani. *Phys. Rev.* **B59**: 1113 (1999).
12. V.A. Chernenko, C. Seguí, E. Cesari, J. Pons, and V.V. Kokorin. *Phys. Rev.* **B57**: 2659 (1998).
13. T. Kanomata, K. Shirahawa and T. Kaneko, *J. Magn. Magn. Mtls.* **76-82**, **65** (1987).
14. A.A. Gonzalez-Comas, E. Obradó, L. Manosa, A. Planes, V.A. Chernenko, B.J. Hattink, and A. Labarta. *Phys. Rev.* **B60**: 7085 (1999).
15. M. Taya. Unpublished results (1998).
16. K. Inoue, K. Enami, Y. Yamaguchi, K. Ohoyama, Y. Morii, Y. Matsuoka, and K. Inoue. *J. Phys. Soc. Jpn.* **69**: 3485 (2000).
17. A.N. Vasil'ev, S.A. Klestov, R.Z. Levitin, V.V. Snegirev, V.V. Kokorin, and V.A. Chernenko. *Sov. Phys. JETP* **82**: 524 (1996).
18. R.D. James and M. Wuttig. *Phil. Mag.* **A77**: 1273 (1998).
19. R. Hayashi. S.M. Thesis. Massachusetts Institute of Technology June 1998.
20. S.J. Murray, R. Hayashi, M. Marioni, S.M. Allen, and R.C. O'Handley. *SPIE Conf.* **3675**: 204 (1999).
21. R. Hayashi, S.J. Murray, M. Marioni, M.J. Farinelli, S.M. Allen, and R.C. O'Handley. *Sensors and Actuators* **A81**: 219 (2000).
22. P.J. Webster, K.R.A. Zieback, S.L. Town, and M.S. Peak. *Phil. Mag.* **B49**: 295 (1984).
23. R.W. Overholser, M. Wuttig, and D.A. Nuemann. *Scripta Mater.* **40**: 1095 (1999).
24. A. Sozinov, A.A. Likhachev, and K. Ullakko. *SPIE Conf. Proc.*, March (2001).
25. K. Ullakko, Y. Ezer, A. Sozinov, G. Kimmel, P. Yakovenko, and V.K. Lindroos. *Scripta Mater.* **44**: 475 (2001).
26. S.J. Murray. Ph.D. Thesis Massachusetts Institute of Technology, January (2000).
27. S.J. Murray, M. Farinelli, C. Kantner, J.K. Huang, S.M. Allen, and R.C. O'Handley. *J. Appl. Phys.* **83**: 7297 (1998).
28. R. Tickle and R.D. James. *J. Magn. Magn. Mater.* **195**: 627 (1999).
29. S.J. Murray, M. Marioni, A. Kukla, J. Robinson, R.C. O'Handley, and S.M. Allen. *J. Appl. Phys.* **87**: 5744 (2000).
30. A.A. Likhachev and K. Ullakko. *Phys. Lett.* **A275**: 142 (2000).
31. R.C. O'Handley. *J. Appl. Phys.* **83**: 3263 (1998).
32. A.A. Likhachev and K. Ullakko. *EuroPhys. J.* **B2**: 1 (1999).
33. A.A. Likhachev and K. Ullakko. *EuroPhys. J.* **B14**: 263 (2000).
34. V.A. L'vov, E.V. Gomonaj, and V.A. Chernenko. *J. Phys.: Condens. Matter* **10**: 4587 (1998).
35. S.J. Murray, R.C. O'Handley, and S.M. Allen. *Proc. MRS Conf.* **604**: 279, (2000).
36. R.D. James and D. Kinderlehrer. *Phil. Mag.* **68**: 237 (1993); *J. Appl. Phys.* **76**, 7012 (1994).
37. R. Tickle, R.D. James, T. Shield, M. Wuttig, and V.V. Kokorin, *IEEE Trans. Magn.*, **35**: 4301 (1999).
38. R.D. James. Unpublished Manuscript (1999).
39. Q. Pan and R.D. James. *J. Appl. Phys.* **87**: (2000).

40. R.C. O'Handley. *SPIE Conf. Proc.* March, Newport Beach, CA, unpublished (1999).
41. R.C. O'Handley, S.J. Murray, M. Marioni, H. Nembach, and S.M. Allen. *J. Appl. Phys.* **87**: 4712 (2000).
42. R.C. O'Handley. *Modern Magnetic Materials, Principles and Applications*. Wiley, New York, 2000.
43. K. Bhattacharya and R.V. Kohn. *Acta Mater.* **44**: 529 (1996).
44. M. Marioni, S.M. Allen, and R.C. O'Handley. Accepted for pub., *J. Appl. Phys.* (2001).
45. O. Hezcko, A. Sozinov, and K. Ullakko. *IEEE Trans.* **MAG-36**: 3266 (2000).
46. C.P. Henry, J. Feuchtwanger, D. Bono, M. Marioni, P.G. Tello, M. Richard, S.M. Allen, and R.C. O'Handley, *SPIE Conf. Proc.*, March 2001, in press.
47. M. Marioni, unpublished.
48. A.E. Clark Personal communication (2001).
49. A.A. Likhachev, A. Sozinov, and K. Ullakko. *SPIE Conf. Proc.*, March (2001).
50. G.H. Wu, C.H. Yu, L.Q. Meng, J.L. Chen, F.M. Yang, S.S. Ai, W.S. Zhan, Z. Wang, Y.F. Zheng, and L.C. Zhao. *Appl. Phys. Lett.* **75**: 2990 (1999).
51. E. Callen. *J. Appl. Phys.* **39**: 516 (1968).

SHAPE MEMORY ALLOYS, TYPES AND FUNCTIONALITIES

J. VAN HUMBEECK
K.U. Leuven-MTM
Katholieke Universiteit Leuven
Heverlee, Belgium

R. STALMANS
Flexmet
Aarschot, Belgium

SHAPE-MEMORY ALLOY SYSTEMS

Many systems exhibit martensitic transformation. Generally, they are subdivided into ferrous and nonferrous martensites. A classification of the nonferrous martensites was first given by Delaey et al. (1) (Table 1), and ferrous alloys that exhibit a shape-memory effect were first reviewed by Maki and Tamura (2) (Table 2).

Of the systems mentioned in both tables, only one system became industrially successful: Ni-Ti(X,Y) in which X,Y are elements that replace Ni or Ti. Besides the Ni-Ti system, a lot of attention had been given in earlier times to Cu-based alloys (3) and to Fe-Mn-Si alloys (4). Furthermore, in recent years, special attention has been given to high-temperature shape-memory alloys (HTSMA) (5).

The aim of this article is mainly an introduction to industrially applicable shape-memory alloys; the following alloy systems will be reviewed:

- Fe-based alloys
- Cu-based alloys
- Ni-Ti alloys
- HTSMA
- other systems

Fe-Based Alloys (2,4,6)

The austenite (fcc- γ phase) in ferrous alloys can be transformed to these three kinds of martensites, depending on composition or stress: γ - α' (bcc), $\gamma \rightarrow \varepsilon$ (hcp) and $\gamma \rightarrow$ fct martensite.

Although a shape-memory effect has been observed in all three types of transformation, most attention in developing a commercial alloy has been given to the alloys that have a $\gamma \rightarrow \varepsilon$ transformation. These alloys have a low stacking fault energy in austenite (Fe-Cr-Ni, Fe-high Mn alloys). The austenite to ε -martensite transformation proceeds by the a/6 [112] Schockley partial dislocations that trail a stacking fault ribbon on every {111} austenite plane and change the crystal structure to martensite. The shape-memory effect, which is of the one-way type, results mainly from reverse motion of the Schockley partial dislocations during heating.

A complete shape-memory effect has been reached in both single-crystal (7,8) and polycrystalline Fe-Mn-Si alloys (9,10) that contain suitable amounts of Mn and Si. A 9% shape-memory strain in single crystals (8) and 5% in polycrystals (9) have been reported.

Any factors that impede the reversibility of the motion of partial dislocations lead to incomplete recovery and in turn a poor shape-memory effect.

Table 1. Classification of Nonferrous Martensites^a

Group	Alloy System
1. Terminal solid solutions based on an element that has allotropic phases	1. Cobalt and its alloys 2. Rare-earth metals and their alloys 3. Titanium, zirconium, and their alloys 4. Alkali metals and their alloys and thallium 5. Others such as Pu, Ur, Hg, and alloys
2. Intermetallic solid solutions that have a bcc-parent phase	1. β -Hume-Rothery phases of Cu-, Ag-, and Au-based alloys 2. β -Ni-Al alloys 3. Ni-Ti-X alloys
3. Alloys that show cubic to tetragonal trans. (incl. Quasi-martensite)	1. Indium-based alloys 2. Manganese-based alloys (paramagn. \leftrightarrow antiferromagn.) 3. A15 compounds 4. Others: Ru-Ta, Ru-Nb, Y-Cu, LaCd, LaAg _x -In _{1-x}

^aRef. 1.

Table 2. Ferrous Alloys That Exhibit a Complete or Nearly Complete Shape-Memory Effect^a

Alloy	Composition	Crystal Structure of Martensite	Nature of Transformation ^b
Fe–Pt	≈ 25at%Pt	bct (α')	T.E.
	≈ 25at%Pt	fcc	T.E.
Fe–Pd	≈ 30at%Pd	fcc	T.E.
Fe–Ni–Co–Ti	23%Ni–10%Co–4%Ti	bct (α')	—
	33%Ni–10%Co–4%Ti	bct (α')	T.E.
Fe–Ni–C	31%Ni–0.4%C	bct (α')	Non-T.E.
Fe–Mn–Si	30%Mn–1%Si	hcp (ϵ)	Non-T.E.
	28~33%Mn–4~6%Si	hcp (ϵ)	Non-T.E.

^a Ref. 2.^bT.E.: Thermoelastic martensite, non-T.E.: Nonthermoelastic martensite.

The internal factors that hamper recovery include alloy composition, Néel temperature, transformation temperature, and lattice defects. External factors are applied stress and strain, deformation, recovery annealing temperature, and thermomechanical treatment.

For example, Murakami et al. (11) showed that Fe–Mn–Si alloys that contained 28–33% Mn and 4–6% Si exhibit a nearly perfect shape-memory effect. But alloys whose Mn content is less than 20% have also been developed successfully. Cr (less than 20%) and Ni are added to improve the corrosion resistance of commercial Fe-based alloys.

So far, Fe-based alloys are not successful SMA. They exhibit only a (limited) one-way shape-memory effect after a labor-intensive thermomechanical treatment. No significant two-way effect or pseudoelastic properties have been reported, whereas only moderate damping capacity might have some interest. Therefore the only reported successful applications of these Fe-based alloys are couplings. This type of application is based on the one-way effect. The recovery stresses are moderate but sufficient (12).

Cu-Based Alloys [(1,3,13–16)]

Copper-based shape-memory alloys are derived from Cu–Zn, Cu–Al, and Cu–Sn systems. The composition range of these alloys corresponds to that of the well-known β -Hume–Rothery phase. In most shape-memory alloys, this phase has a disordered bcc structure at high temperatures but orders to a B2, D0₃, or L2₁ form at lower temperatures. The shear elastic constant of the β phase exhibits anomalous behavior as temperature decreases, that is, it is lowered till the lattice instability with respect to $\{110\} \langle 1\bar{1}0 \rangle$ shears at some temperature and transforms β to martensite. The temperature of the transformation to martensite, M_s , varies with the alloy composition. The elastic anisotropy of the β phase is much higher compared to normal metals and alloys and increases further as the martensitic transformation is approached.

Cu–Zn and Cu–Al martensites are of three types α' , β' or γ' : the subscript 1, 2, or 3 is added to indicate the ordering schemes in β , namely, B2 (2) or D0₃ (1) or L2₁ (3). Some conversion from one martensitic structure to another, for example $\beta' \rightarrow \gamma'$, may also take place. The net result is a coalescence of plates within a self-accommodating group and even coalescence of groups. Heating this deformed martensitic microstructure transforms it to the β phase,

and the shape-memory effect accompanies the structural change.

Copper-based shape-memory alloys presently used are derived from Cu–Zn and Cu–Al systems, and elements are added for various metallurgical reasons. The working martensite in these alloys is only or predominantly the $\beta'_{1,2 \text{ or } 3}$, type where γ' martensite is the minor constituent in the latter case. Alloys that have α' martensites have so far not been used. Therefore, alloys of $\beta'_{1,2 \text{ or } 3}$ martensite are the subject in this part.

Two criteria should be taken into account when selecting an alloy composition to obtain a complete β microstructure that transforms to martensite: (1) The β phase must be stable across as wide a temperature range as possible. The less wide this temperature range, the faster the cooling rate required to retain the β phase without diffusional decomposition. (2) Transformation temperatures must fall within a range that satisfies the requirement for the shape-memory application (–150 to 200°C). The three alloy systems in Table 3 satisfy these criteria. They are used nowadays, but in limited amounts. Apart from composition, transformation temperatures are also strongly influenced by other factors.

The Influence of Chemical Factors on the Transformation Temperature

The Influence of Composition. Several authors have attempted to quantify the M_s –composition relationship for several Cu-based alloys. An overview is given in (3). Different authors weight the same element differently. The main reason for this discrepancy might be that the samples measured have different thermomechanical histories, that is, one has probably not measured “identical samples.” Indeed, composition is not the only chemical factor that affects the M_s temperature. The type and degree of order of the β and the martensite lattice also affect the M_s . Thermal treatments can, therefore, influence the transformation, as discussed in the following sections.

Quenching and the Order State of the β Phase. The transformation temperatures of Cu-based alloys are very sensitive to minute changes of the degree of order in the β phase. Such changes are easily brought about by quenching from intermediate and high temperatures in the form of dilute disorder in an otherwise well-ordered material. The effect is noticeable in both Cu–Zn–Al (17,18) and

Table 3. Industrial Cu-based Alloys

Base Alloy	Composition (wt%)	M_s ($^{\circ}\text{C}$)	Hyst. ($^{\circ}\text{C}$)	Other Alloying Elements in Solution (%)	Current Grain Refining Elements Producing Precipitates	Remarks on the Base Alloy
Cu–Zn–Al	5–30 Zn 4–8 Al	–190 to +100	10 (β')	Ni (5%) Mn (12%)	Co (CoAl); B (AlB ₂); Zr (?); B, Cr (Cr _x B _y)	Good ductility and reproducibility; prone to martensitic stabilization; poor β stability ($T > 200^{\circ}\text{C}$)
Cu–Al–Ni	11–14.5 Al 3–5 Ni	–140 to +200	10 (β') 40 (γ')	Mn (5%)	Ti [(Cu, Ni) ₂ TiAl]; B (AlB ₁₂); Zr (?)	Low ductility; low martensitic stabilization; good β -stability
Cu–Al–Be	9–12 Al 0.4–1 Be	–80 to +80	6 (β')	Ni (5%)	B (AlB ₂ or AlB ₁₂) Ti (Cu ₂ TiAl)	Poor reproducibility; excellent β stability ($T > 200^{\circ}\text{C}$)

Cu–Al–Ni (19,20) alloys and manifests as a suppression of the transformation temperatures thereby stabilizing the β phase relative to the martensitic. The suppression is temporary, but it is easily recoverable in Cu–Zn–Al alloys by aging in the β condition at as low a temperature as room temperature. However, the recovery of Cu–Al–Ni alloys tends to be more sluggish and requires higher aging temperatures. For example, Cu–Al–Ni alloys aged at 300°C for 1 hour can have transformation temperatures up to 60°C higher than the as-quenched alloys.

Aging and the Order State of Martensite. Aging a Cu–Zn–Al alloy in the martensitic condition can appreciably shift the reverse transformation temperatures of the martensite to the β phase (21). This shift to higher temperatures stabilizes the martensitic relative to the β phase. This stabilization is brought about by a thermally activated diffusional process and, as it is presumed, alters the ordered state inherited by the martensite from the β to a relatively disordered state (21,22). The effect is more pronounced in the presence of excess vacancies retained after a prior quench from higher temperatures. A quench to a temperature above the M_s followed by a hold at the same temperature (step quenching) to rid the alloy of excess vacancies reduces the problem considerably (21). But even then, stabilization of martensite can recur during subsequent aging, and the effect is worse, the higher the aging temperature in the martensitic condition.

Manganese or nickel addition to Cu–Zn–Al, it has been shown, too lessens the problem of stabilization. This happens possibly through a slowing of diffusion in the martensite in the presence of the added elements. More interestingly, the effects it has been shown are inhibited, even in the absence of these elements, by dislocations introduced into the β phase during hot rolling (23) or through transformation cycling (24). Further understanding of the role of these dislocations in such inhibition might provide the information needed to improve the stability of these alloys for use at higher temperatures. Stabilization of Cu–Al–Ni martensite is much slower compared to Cu–Zn–Al (19,25). The former alloys thus are more thermally stable than Cu–Zn–Al and are more suited for use at higher temperatures.

Influence of Other Factors on the Transformation Temperature. Besides the chemical factors such as composition and order; certain nonchemical factors may also influence the M_s temperature. Among the latter are contributions from defects such as vacancies, dislocations, grain boundaries, and precipitates.

Influence of Nonequilibrium Precipitates. Precipitates like the γ phase can be formed in Cu–Zn–Al by flash heating to an intermediate temperature after prior dissolution followed by quenching (17). These precipitates may shift the transformation temperatures with respect to their nominal values and also may produce variations in the transformation temperature range and the hysteresis that accompanies the transformation. The exact changes depend on the coherency, size, and distribution of the precipitates. The variations are brought about by an alteration in the chemical, stored, elastic, and frictional energies of the transformation because of the presence of the additional phase. Stored elastic energy plays a dominant role when the precipitates are small and coherent and when their presence does not appreciably change the composition of the matrix. This usually leads to a suppression of the M_s and to minor changes in the hysteresis, providing the precipitates are deformed in the transformation. Larger semi- or incoherent precipitates that substantially alter the composition of the matrix and impede the growth of martensitic plates lead to changes in the M_s that depend on the partitioning of the elements and an enlarged hysteresis (26).

Precipitation and concomitant changes in transformation temperatures can be disadvantageous if they are produced inadvertently during service, but they can be incorporated in the heat treatment schedule to fine-tune the transformation temperatures or when wider hysteresis is required.

The Influence of Grain-Refining Elements that Form Precipitates. Copper-based shape-memory alloys exhibit rapid grain growth at higher dissolution temperatures. When grain sizes are of the order of millimeters and the elastic anisotropy in the β phase is high, they suffer

intergranular cracking and plastic deformation during quenching. The problem has been solved by adding grain refining elements to the two shape-memory alloys. Zr (0.4–1.2 wt%), Co (0.4–0.8%), Ti (0.5–1.0 wt%), and B (0.4–0.2 wt%), have been added to Cu–Zn–Al alloys to reduce the grain size to the 100- μm level. Titanium is also effective in refining the grain size in Cu–Al–Ni alloys to the 50–100 μm range (27,28).

Refining is brought about by the formation of insoluble particles that aid nucleation of the grains or retard their growth. These grain refining elements have four direct or indirect effects on transformation temperatures: (1) By forming intermetallics, they deplete the original β lattice of alloying elements that change the transformation temperatures. (2) Part of these elements remain in solution within the β matrix. Depending on the atom size, this can give rise to solid-solution hardening that decreases the M_s and eventually the other transformation temperatures (29). (3) They can also have a chemical contribution, which means that the global composition determines the transformation temperatures on a purely thermodynamic basis. (4) The precipitates limit grain growth during annealing, which influences the transformation temperatures, as discussed in the next section.

The Influence of Grain Size. Several authors have shown that small grain size results in stabilizing the parent phase and depressing the transformation temperatures up to 40°C (30,31). This effect is observed in alloys with and without the special addition of grain-refining elements, which indicates the restraining effect of grain size itself on the transformation. Lowering of transformation temperatures is attributed to the increasing grain restraint as grain size decreases. This is the conclusion of most authors (32,33) and is also consistent with Hornbogen's argument that the increase in yield stress σ_y is proportional to the stress required to start the transformation (20). Hornbogen's important assumption is that matrix strengthening increases the undercooling $\Delta T (= T_0 - M_s)$ but does not influence necessarily the T_0 temperatures. T_0 is the temperature at which the free energy of the β phase is equal to the free energy of the martensitic phase.

Adnyana (30) and Jianxin (33) found a linear relationship between the M_s temperature and the yield stress derived via the classic Hall–Petch relationship for Cu–Zn–Al alloys. The restraining effect of grain size is, however, also influenced by the grain size (gs) to thickness (t) ratio. At high gs/t ratios, the contribution of the free surface becomes important and M_s no longer changes linearly with gs, as observed by Wood (34). This is consistent with the conclusion of Mukunthan and Brown (35) who showed that the flow stress in all specimens decreases as specimen thickness decreases when the value of t/d becomes smaller than a critical value. These authors showed further that this critical value of t/d increases as both grain size and stacking fault energy decrease. These elements that contribute to high stacking fault energy have an effect similar to a small grain size.

Influence of Defects. Often it is not only the effect of the grain size or the grain size thickness ratio that

accounts for changes in transformation temperatures. Annealing a sample at higher temperatures can give rise to grain growth but will also reduce the amount of defects and thus the nucleation sites. In Cu-based alloys, the situation is again complicated by the quenched-in vacancies and the size of the antiphase domains, which can also be regarded as strengthening the matrix. An increase in the energy of the β phase due to a higher defect concentration such as foreign elements in solid solution, precipitates, internal strain fields (e.g., coherency strains) causes a lowering of M_s (32). Moreover, if the defect density is proportional to that of the nucleation sites, a higher defect density gives rise to much smaller martensitic plates. A Hall–Petch type relationship is also found between martensitic plate thickness and fracture stress (36).

Specific defect configurations can be introduced by thermal cycling and also by two-way memory training. The influence of such defects, notably dislocations, has been discussed in some recent literature. It has been suggested that the changing character of the same dislocation in the β and martensitic phases alters the relative phase stability of the two phases.

Ni–Ti Alloys

Ni₅₀–Ti₅₀ and near equiatomic Ni–Ti alloys are the best explored system of all shape-memory alloys and occupy almost the whole SMA market. Ni₅₀–Ti₅₀ is an intermetallic phase that has some solubility at higher temperature.

The science and technology of Ni–Ti is overwhelmingly documented. The influence of composition and thermomechanical processing on functional properties is well understood and described in the literature. Therefore, we refer to some very interesting and relevant publications such as (37–41).

The basic concept of processing Ni–Ti alloys is that the martensitic and β phases have to be strengthened to avoid plastic deformation during shape-memory or pseudoelastic loading. This occurs by classic methods: strain hardening and during cold deformation, solution hardening, and precipitation hardening. Ni–Ti alloys have the significant advantage that these techniques can be easily applied due to excellent ductility and a very interesting but complicated precipitation process (42).

The compositions of Ni–Ti SMA are approximately between 48 and 52 at% Ni and the transformation temperatures of the B2 structure to the martensitic phase that has a monoclinic B19' structure are very sensitive to the nickel content (a decrease of about 150°C for an increase of 1 at% Ni). Transformation temperatures can be chosen between –40 and +100°C.

Ni–Ti alloys have the best shape-memory behavior of all SMA. Even in the polycrystalline state, 8% shape recovery is possible, 8% pseudoelastic strain is completely reversible above A_f , and the recovery stress is of the order of 800 MPa.

In some cases, the martensitic transformation is preceded by the so-called R-phase transition. The R transition is a B2 \leftrightarrow rhombohedral transformation that also has second-order characteristics (43).

The most specific characteristics of this R-phase transition are that it shows clear one- and two-way memory

effects of the order of 1% recoverable strain and that the hysteresis of the transformation is very small, only a few degrees which creates possibilities for accurately regulating devices.

Note that further cooling transforms the R phase into B19' martensite. During heating, generally only the reverse martensitic transformation is observed. It has been shown that the appearance of the R phase depends on composition, alloying elements, and thermomechanical processing (39). The major common point is that all effects that depress the martensitic forward transformation below room temperature favor the appearance of the R-phase transition that is quite stable near 30°C.

Ternary Ni–Ti Alloy Systems

Adding third elements opens even more possibilities for adapting binary Ni–Ti alloys to more specific needs of applications. Adding a third element implies a relative replacement of Ni and/or Ti. Therefore, it must be always very well indicated which metal, Ni or Ti or both, is replaced by the third element.

Alloying third elements influences the transformation temperatures and also affects hysteresis, strength, ductility, shape-memory characteristics, and the B2→(R)→B19' sequence. The influence of several elements has been described in (44–48).

Although more application oriented, one can distinguish four purposes to add third elements:

1. to decrease (Cu) or increase (Nb) hysteresis,
2. to lower transformation temperatures (Fe, Cr, Co, Al),
3. to increase transformation temperatures (Hf, Zr, Pd, Pt, Au), and
4. to strengthen the matrix (Mo, W, O, C).

Some ternary alloys have been developed for large-scale applications. We will summarize only the two most well developed: Ni–Ti–Cu and Ni–Ti–Nb.

Ti–Ni–Cu. Ternary Ti–Ni–Cu alloys in which mainly Ni is substituted by Cu are certainly as important as binary Ti–Ni. Increasing the Cu content decreases the deformation stress in the martensitic state and also decreases the pseudoelastic hysteresis without affecting the M_s temperature significantly (49). However, addition of more than 10% Cu embrittles the alloys and hampers formability.

It should also be remembered that Ti–Ni transforms from a B2 into a monoclinic phase, but Ti–Ni–Cu that contains more than 15 at% Cu transforms from a B2 into an orthorhombic phase. Ti–Ni–Cu that has less than 15 at% Cu transforms in two stages (37).

A disadvantage of most Ti–Ni–Cu alloys is that the transformation temperatures do not decrease below room temperature. Cr or Fe can be alloyed to obtain pseudoelastic alloys at room temperature that have small hystereses. An Ni_{39.8}–Ti_{49.8} Cu₁₀ Cr_{0.4} alloy was developed that has small hysteresis (130 Mpa), one-fourth compared with Ni₅₀–Ti₅₀, and an M_s below room temperature (50).

Ti–Ni–Nb (51,52). The inherent transformation hysteresis of Ni–Ti–Nb is larger than that of binary Ni–Ti alloys. By using a large dispersed volume fraction of deformable β -Nb particles, the hysteresis can be further widened by an overdeformation of stress-induced martensite, generally between M_s and M_d . Originally, Ni–Ti–Nb (more specifically Ni₄₇–Ti₉₄–Nb₉) was developed by Raychem Corp. for clamping devices. The large shift of reverse transformation temperatures from below to above room temperature by deformation, allows room temperature of storage open couplings.

Recently, pseudoelastic Ni–Ti–Nb alloys have also been developed that have three significant differences from binary alloys (52):

1. Stress rate is much lower.
2. σ^{P-M} stresses are much higher.
3. The superelastic window is much larger.

High-Temperature Shape-Memory Alloys (5)

Actual shape memory alloys (SMA) are limited to maximal A_f temperatures of 120°C: M_s is generally below 100°C. However, because market demands for SMA have expanded greatly, the need for SMA that transform at higher temperatures than presently available is increasing. The main application areas of interest are actuators in the automobile and oil industries and in safety devices.

There is also an interest in robotics because shape-memory alloys that have high transformation temperatures allow faster cooling, which would significantly increase the bandwidth in which the robot can operate.

Although many alloy systems have high transformation temperatures, no large-scale applications have been developed. A major breakthrough has not been reported yet mainly due to the following problems: (much) lower performance than the successful Ni–Ti alloys, stabilization of martensite, decomposition of the martensitic or parent phase, and brittleness due to high elastic anisotropy or due to the presence of brittle phases or precipitates.

Another condition for a good shape-memory effect is that the stress to induce martensite or the stress to reorient martensite is (much) lower than the critical stress for normal slip. Because the critical stress for slip generally decreases as temperature increases, this condition is quite difficult to fulfil, especially at high temperatures. Thus, a HTSMA should be designed at such a composition and/or thermomechanical treatment that strengthening mechanisms are incorporated to increase the critical stress for slip.

Table 4 summarizes the systems under investigation. For references to this table, see (5).

OTHER TYPES OF SHAPE MEMORY ALLOYS

β -Ti Alloys

In spite of the good biocompatibility of NiTi-alloys, doubts remain on the long-term stability or on the danger of bad surface treatment leading to Ni leaching. Since Ni is

Table 4. High-Temperature Shape-Memory Alloys^a

Base System	Type of Martensite	Other Alloying Elements		$A_s \leq$
		Name	Reason	
Fe–Mn–Si	Nonthermoelastic $\gamma \leftrightarrow \varepsilon$	Co, Ni, Cr	To improve corrosion resistance	150–200°C
Cu–Al–Ni	Thermoelastic $\beta 1 \leftrightarrow 18 R (2H)$	Mn, Ti, B, Zn	To improve machinability, control of transformation temperatures; Grain refinement to improve ductility.	100–200°C
(Ni–X)–Ti	Thermoelastic $\beta_2 \leftrightarrow B19', B19$	X = Pt, Pd, Au, Rh	Based on B2–Ti–X intermetallic compounds that have martensitic transformation at very high temperatures	150–500°C
		B	To reduce the grain size and improve the strength	
Ni–(Ti–X)	Thermoelastic $\beta_2 \leftrightarrow B19', B19$	X = Hf, Zr	Based on Ni–X intermetallic compounds that form a pseudobinary with Ni–Ti.	120–350°C
Ni–Al	Thermoelastic $B_2 \leftrightarrow 3R (7R) (L1_0 \text{ structure})$	Cu, Co, Ag	To increase transformation temperatures	100–600°C
Ni–Mn	Thermoelastic (?) $B2 \leftrightarrow \theta (L1_0 \text{ structure})$	Fe, Co, Mn, B Al, Ti, Cu for Ni	To improve ductility To decrease M_s and to improve shape-memory characteristics	500–750°C
Zr-based intermetallics	B19' monoclinic	Mg, Al, Si, Ti, V, Sn, Cr, Co, Fe, Mo for Mn Ti, Ni	To increase M_s and to improve shape-memory characteristics To improve ductility	200–900°C
Cu–Zr	Nonthermoelastic			
Zr ₂ –Cu–Ni	Nonthermoelastic			
Zr ₂ –Cu–Co	Thermoelastic			

^aRef. 5.

known for his high allergic reaction, Ni-less shape memory alloys could be attractive. Such alloys might be developed based on the allotropic transformation in Ti, a highly biocompatible material. Pure titanium shows an allotropic transformation from β (bcc) to α (hexagonal) phase at 1155 K. Transition elements (TM) stabilise the β -phase. Thus the temperature of the $(\alpha + \beta)/\beta$ transition decreases with increasing concentration of the alloying element.

β -phase Ti alloys can be martensitically transformed if they are quenched from the stable β -phase. Two types of martensite, respectively α' and α'' can be formed, depending on the composition and the solution treatment conditions (53).

The α' -martensite is hexagonal, while α'' has an orthorhombic structure (54). It is the α'' -martensite that shows the shape memory effect. The shape memory effect was first studied in detail by Baker in a Ti-35 wt% Nb alloy (127). Since then several observations of SME especially in Ti-Mo base alloys have been reported (57,58, 59,60,61). A systematic work on the influence of different alloying elements on the shape memory effect can be found in (62), a patent deposited J. Albrecht, T. Duerig and D. Richter.

The authors come to the conclusion that α'' -martensite can be obtained when the following condition is fulfilled:

$$-1100 \leq \varepsilon(A_i X_i + B_i X_i^2) \leq -700$$

where X_i is the atomic percentage for each element, A_i and

B_i are constants given in the patent for each element (V, Al, Fe, Ni, Co, Mn, Cr, Mo, Zr, Nb, Sn, Cu). Ta was not claimed although it also offers its contribution to SME as described in (61).

Generally, a shape recovery in the order of 3% can be obtained based on strain-induced martensite and recovery stresses up to 170 MPa have been reported (60). The disadvantage is that those alloys are very prone to stabilisation and decomposition due to the fact that the β -phase is retained after quenching in its metastable state and competes with ω -phase during quenching. Also spinodal decomposition of α'' -martensite in Ti-Mo and Ti-Nb has been observed (53). The sensitivity to decomposition at moderate temperatures is less, if not, important at room temperature. Therefore pseudoelastic β -Ti alloys could offer an interesting alternative to Ni-Ti alloys for example for orthodontic wires. Such an alloy has recently been developed by Lei et al. (63). Ti-11Mo-3Al-2V-4Nb was selected for optimization. Good pseudoelasticity of the order of 3% was obtained after cold working and heat treatment.

Magnetic-Field-Induced Martensitic Transformation

T. Kakeshita et al. (64) defined a magnetoelastic martensitic transformation: when a magnetic field is applied (above A_F) to an alloy that exhibits a thermoelastic martensitic transformation, martensite variants may be induced while a magnetic field is applied and revert to the parent phase when the magnetic field is removed. This has been observed in Fe_{31.9}-Ni_{9.8}-Co_{4.1}-Ti (at%) (64,65). Apart from

Fe-based alloys, Ni–Mn–Ga near the Ni₂MnGa compound, which is a ferromagnetic Heusler ordered alloy, is one of the candidates (66,67).

Besides their very interesting fundamental properties, these alloys might act much faster than classic SMA-based actuators that are thermally driven. The bandwidth of the latter is limited to a few hertz (for very thin wires) due to cooling restrictions. In magnetoelastic martensitic transformations, bandwidths of some orders larger can be obtained.

FUNCTIONAL PROPERTIES OF SHAPE-MEMORY ALLOYS

Shape-memory alloys have different shape-memory effects and can be used in different ways. These effects and ways of use are described in general terms here. As explained before, binary and ternary Ni–Ti alloys are probably used for more than 90% of new SMA applications. Therefore, quantitative data refer to Ni–Ti alloys, unless otherwise stated.

One-Way Shape-Memory Effect

A shape memory element can be deformed in its martensitic state to almost any “cold shape.” The basic restriction is that the deformations may not exceed a certain limit, typically 8%. These apparent plastic deformations can be recovered completely during heating when the reverse transformation occurs and results in the original “hot shape.” This strain and shape recovery during heating is called the *one-way shape-memory effect* because only the hot shape is memorized (Fig. 1).

The physical basis for this *one-way effect* is a reverse martensitic transformation from a preferentially oriented martensitic phase and shape to the original high-temperature phase and shape, as explained more in detail earlier and in many review papers on shape-memory alloys. The preferential orientation of the martensitic variants originates from the application of stress either below M_f that causes martensitic reorientation, or during the forward transformation that causes preferentially oriented formation of martensite. Thus, the apparent plastic strain is caused by the preferential orientation of martensite.

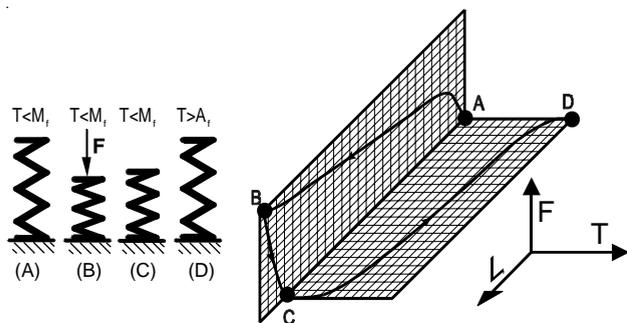


Figure 1. The one-way memory effect. The sample is deformed (A→B) and unloaded (B→C) at a temperature below M_f . The apparent plastic deformation is restored during heating to a temperature above A_f (C→D). Length change, load, and temperature are indicated, respectively, by L , F , and T [from (69)].

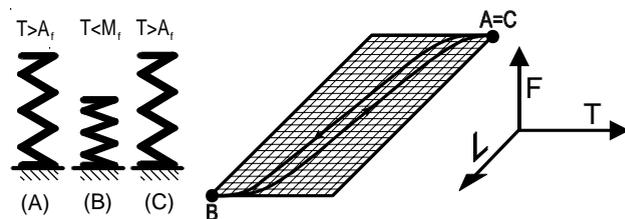


Figure 2. The two-way memory effect. A spontaneous shape change occurs during cooling to a temperature below M_f (A→B). This shape change is recovered during subsequent heating to a temperature above A_f (B→C) [from (69)].

Therefore, the reverse transformation to the parent phase during heating is accompanied by a strain and shape recovery. The one-way shape-memory effect is thus a property *inherent* in the reversible, thermoelastic martensitic transformation.

Thermoelasticity was observed as early as 1938 by Greninger and Mooradian (70). Since then, thermoelasticity and the one-way memory effect have been found in many different alloy systems (71). In 1962, the one-way shape-memory effect was also found in Ni–Ti (72).

Two-Way Shape-Memory Effect and Training

The two-way memory effect involves memorization of two shapes. Figure 2 shows that a cold shape is obtained spontaneously during cooling. Different from the one-way memory effect, no external forces are required to obtain the “memorized” cold shape. During subsequent heating, the original hot shape is restored. The maximum strains are in general substantially smaller than those of the one-way memory effect. A strain limit of about 2% has been mentioned (73), although higher TWME strains have been found in specific cases.

In 1972, Tas et al. proposed the term “two-way memory effect” (abbreviated to TWME) to refer to this spontaneous, reversible shape change between a “hot” shape linked to the parent phase and an acquired “cold” shape linked to the martensitic phase (74). This spontaneous shape change was observed only after particular thermomechanical procedures. Since that time, many papers have been published on the TWME, the thermomechanical procedures and the mechanisms of the TWME, especially for Cu-based SMAs (73,75–83). There have not been many systematic studies of Ni–Ti shape-memory alloys, however, to investigate the TWME and the effects of the thermomechanical procedures (77,78).

The essential difference from the one-way effect is the macro stress-free shape change during the forward transformation or, in physical terms, the spontaneous formation of preferentially oriented martensitic variants. Thus, the TWME requires some sort of asymmetry in the microstructure of the parent phase, such as retained martensite or dislocation structures (76,82). The microstructural asymmetry and the resulting TWME are not inherent characteristics of shape memory alloys, as is the one-way effect, and can be induced only after particular thermomechanical procedures. These thermomechanical procedures are in

general based on the repetition of thermomechanical cycles through the transformation region (73,76,79), that consists of transformation cycles from the parent phase to preferentially oriented martensite. The goal of these repetitive procedures is to acquire the cold shape; therefore, these procedures are referred to as “training.” Some examples of such training procedures are temperature cycling at a constant strain or at a constant stress and superelastic cycling. It can be easily understood that many combinations and variants of these procedures can also be applied. As a result, new procedures have been regularly reported in the literature. In most of these publications, the new aspects of the procedures are emphasized, but almost no attention is paid to the points of similarity to previously described procedures (81).

It is important to note that the training results in concomitant effects, such as changes of the transformation temperatures and heats and residual deformations of the austenitic shape (75,76,81,84). In general, these characteristics become insensitive to cycling as the number of training cycles increases. So, the repetitive procedures to induce the TWME can also be used merely to stabilize the shape-memory behavior. The term training is also generally used to indicate such stabilization treatments, though the TWME can be negligible and should in this case be considered a side effect of the stabilization treatment. It is generally agreed that cyclic training procedures generate some kind of microstructural asymmetry in the parent phase, so that preferential martensitic variants are formed in subsequent thermal cycles, thus causing the TWME (76,82). Basically, three mechanisms for the TWME have been proposed in the literature. It has been observed that training cycling results in generating complex dislocation arrays (76,79,80). Based on this observation, the TWME has been attributed to the residual stress fields of these dislocation arrays (75,76). It was proposed that these residual stress fields favor the nucleation and the beginning of the growth of some preferentially oriented variants, and at the same time, the residual stresses are relaxed by the accompanying shape change. During further cooling, these preferential variants grow without any assistance and result in the TWME.

A second proposed mechanism is based on local stabilization of remnants of preferential martensitic variants that are retained above the original A_f temperature. During cooling, these small martensitic plates would grow and influence the subsequent positioning of other variants, thus causing the TWME (76,79). However, specific experimental observations obtained on Cu-based alloys invalidate these two proposed TWME mechanisms (82). A third mechanism became widely accepted in the past years. Based on a thermodynamic analysis of specific experimental results, it was shown that the defect energy of the complex dislocation arrays generated during training is minimal in the trained variants, that is, in the preferentially oriented variants that have been repeatedly induced during training cycling (82,85). From thermodynamic considerations, it follows that the growth of these trained variants are also favored during subsequent thermal cycling, which explains the TWME. This thermodynamic analysis has also allowed us to explain many other phenomena

related to the TWME (83,86). Because the TWME is closely related to the “trained” dislocation arrays, the TWME can be removed by annealing at moderate temperatures (75).

Next to the cyclic training procedures, the following one-time procedures to induce the TWME have been reported. Remnants of preferentially oriented variants can be stabilized by holding a constrained or stressed sample at temperatures above the nominal A_f for a sufficiently long time. The TWME obtained by this procedure is attributed to the growth of those remnants (76). Aging of a sample at sufficiently high temperatures and stresses can also result in a reversible shape effect (73,76,87,88). Precipitates are formed during the aging. The observed reversible effect is attributed to the residual stress field generated by these precipitates and to the interaction between martensitic formation and the preferentially oriented precipitates. A TWME of small magnitude can also be obtained by a single, sufficiently high plastic deformation of the martensitic phase (73). However, the disadvantages of these one-time procedures are numerous, including large deformations of the hot shape; large shifts of the transformation temperatures; and strong dependence on the stabilization or aging temperatures, stresses, and times (76).

Superelasticity

The shape-memory effects described before require temperature changes. In contrast, the superelastic effect, also called the pseudoelastic effect, is isothermal (89,90). The two-dimensional graph of Fig. 3 shows that a superelastic specimen exhibits normal elastic behavior until a critical stress is reached. Under further stressing, the specimen elongates substantially, as if it were plastically deformed. However when the stress is removed, the specimen contracts to its original dimensions, and the apparent plastic strain is recovered.

Superelasticity can be considered the mechanical analog of the thermal shape-memory effect. Isothermal loading at a temperature above A_f results in a stress-induced martensitic transformation that starts at a critical stress

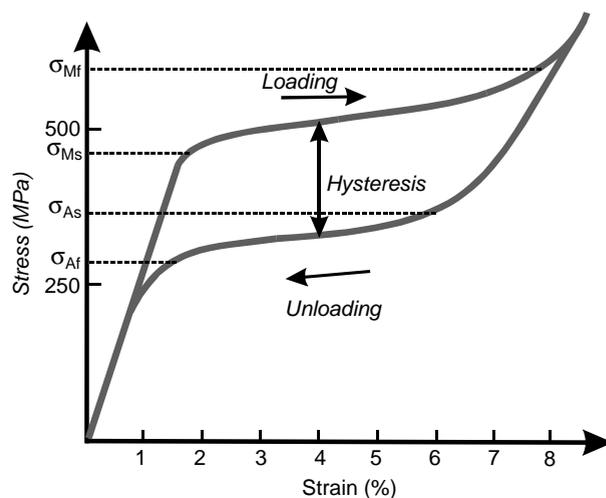


Figure 3. Superelastic behavior at constant temperature due to stress-induced transformation and retransformation.

σ_{M_s} . Further straining occurs at a nearly constant stress level until the transformation finishes at σ_{M_f} . Thus, the apparent plastic strain is caused by the shape strain that accompanies the stress-induced formation of preferentially oriented martensite. During subsequent unloading, the reverse transformation occurs at a lower stress level between σ_{A_s} and σ_{A_f} , and the apparent plastic strains are recovered. Large reversible deformations up to 10% can be obtained, compared to 0.2–0.5% elastic strain in most other metallic alloys. Further straining at stress levels above σ_{M_f} results in elastic straining of the stress-induced martensite, followed at σ_y by plastic yielding of the martensite.

The stress-induced transformation exhibits stress hysteresis that is revealed by the different stress levels for the forward and reverse transformations. This hysteresis is typically 150–300 MPa in Ni–Ti and results in dissipating of energy during superelastic cycling. The energy dissipated per cycle is given by the area enclosed between the upper and lower curves in Fig. 3. Superelasticity also involves the storage of potential energy, given by the area under the unloading curve in Fig. 3. The capacity of this elastic energy storage can be as high as 10 J/g. All of the previously mentioned superelastic characteristics are strongly affected by processing and composition (91,92).

The critical transformation stresses (σ_{M_s} , σ_{M_f} , σ_{A_s} , and σ_{A_f}) increase, in a first approximation, linearly as temperature changes starting from zero at the corresponding transformation temperature, as described by the Clausius–Clapeyron equation. It follows that at a temperature, denoted as M_d , the stress for plastic yielding becomes equal to the stress for martensitic formation. Thus, superelasticity occurs only across a relatively narrow temperature window between the temperatures A_f and M_d . This temperature range of typically 50–100 K is too small for applications in most industrial and consumer fields. The strong temperature dependence of mechanical behavior, described by the Clausius–Clapeyron equation, is a further impediment to the general use of superelasticity (93).

The temperature dependence and small temperature range are no barrier to use in mammalian bodies, where temperature is constant. Moreover, the superelastic effect

results in a unique combination of deduced properties. No other material or technology can offer the combination of high strength, high stiffness, high pliability, and high kink resistance. Ni–Ti has excellent biocompatibility after proper surface treatment. The concept of a metallic material that has extreme elasticity and nearly constant stress levels across a large strain area is something that can be tackled more easily by designers than the concepts related to temperature-activated shape-memory effects (94). Accordingly, the largest commercial successes of SMAs in recent years are linked to using superelasticity in biomedical applications (93–95). Other advantages related to superelasticity and relevant for medical applications have been described in detail by Duerig et al. (93).

It must be mentioned also that Ni–Ti alloys in their as-cold worked state can exhibit nearly linear elastic behavior across an extremely broad strain range. This linear behavior that has a low elastic modulus of typically 30 GPa is called linear superelasticity (96,97). Reversible deformations as high as 4% can be induced that have a very small hysteresis between the loading and unloading curve. Clearly different from the superelastic effect described before, stress-induced martensitic transformation is not the controlling mechanism of linear superelasticity. As a result, temperature and composition have only a minor effect on this behavior.

Generation of Recovery Stresses

If an external constraint prevents an SMA element from returning to its hot shape when heated, high recovery stresses are gradually generated during heating, as illustrated in Fig. 4. Stresses up to 800 MPa can be obtained (99).

Four parameters have to be introduced to describe this shape-memory property (100): the contact strain e_c , the contact temperature T_c , the recovery stress σ_r , and the stress rate $d\sigma_r/dT$. Similar to the one-way and two-way memory effects, the generation of recovery stresses starts from a macroscopic deformation in the martensitic state. During subsequent heating, free recovery occurs until a

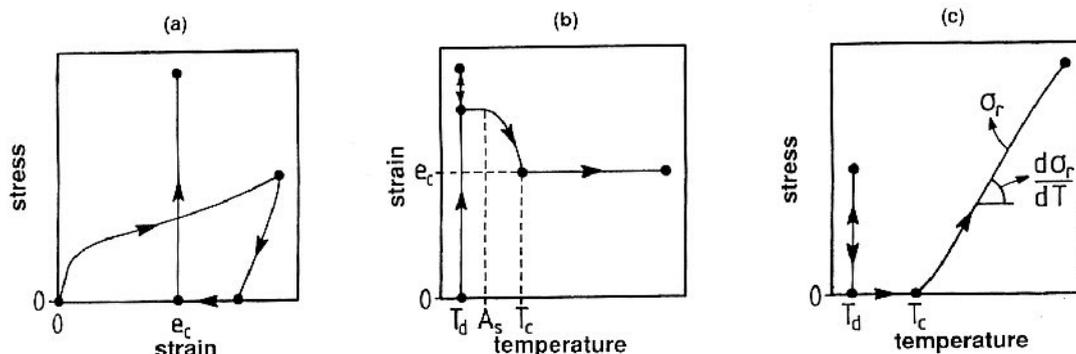


Figure 4. The generation of recovery stresses is shown in three two-dimensional figures: (a) stress–strain, (b) strain–temperature, and (c) stress–temperature. A deformation is imposed at a temperature T_d in the martensitic state. Shape recovery is impeded at a contact strain e_c . From the corresponding contact temperature T_c , recovery stress σ_r are generated at a stress rate $d\sigma_r/dT$ [from (69,98)].

temperature T_c at which an external mechanical obstacle is contacted. The recovery of the remaining deformation, characterized by the contact strain e_c , is impeded by this obstacle. Therefore, from the temperature T_c , recovery stresses are generated at a nearly constant stress rate $d\sigma_r/dT$, often described by a Clausius–Clapeyron equation. This shape-memory property is the basis of some very successful shape-memory applications, especially in earlier years (100,101). However, compared to the number of papers on superelasticity and the one- and two-way memory effects, the number of publications on generating recovery stresses is rather limited (99–114). Moreover, most of these publications focus on engineering applications. Also important to note is that these applications are based mostly on the stresses generated during the first heating after the initial predeformation. Therefore, cyclic generation of recovery stresses was not investigated in earlier papers.

The recent substantial efforts in developing hybrid composites that have embedded shape-memory elements have resulted in an increasing interest in understanding the generation of recovery stresses (99,113–115). In these smart composites, thin SMA wires or other SMA elements are embedded in advanced structural materials without losing the structural integrity of the matrix material. Most functional properties of such adaptive composites are directly related to the following mechanism on the level of the SMA elements. During heating, prestrained martensitic SMA elements operate against the elastic stiffness of the host matrix, and bias the strain recovery and the reverse transformation of the SMA elements. The result is that recovery stresses are generated gradually by the SMA elements in the composite during heating, and the strain recovery and reverse transformation of the SMA elements are delayed. After overcoming temperature hysteresis, the reverse process occurs during cooling. A crucial difference from earlier applications that use the generation of recovery stresses is that these hybrid composites are activated repeatedly, that is, recovery stresses are generated cyclically.

A remarkably common feature of several papers on recovery stress generation is that elementary principles of thermoelastic martensitic transformation are frequently violated. Although review publications can give the impression that the generation of recovery stresses is well understood, careful analysis has shown that the scientific understanding of this property in many papers is still limited (69,86). The limited understanding can be partly explained by the lack of experimental results. In addition, the majority of the published experimental results are questionable because a hard tensile machine was used in most studies. In that case, the thermal dilatations of the crosshead arms can have such a dominating effect that the measured results are unreliable (81,100).

As a result of the lack of data, many papers refer to an old NASA report, published in 1972, for quantitative data on recovery stress generation (102). However, the data in that report refer to the first heating after predeformation in isostrain conditions. This would correspond to a biasing element that has no thermal dilatation and infinite rigidity.

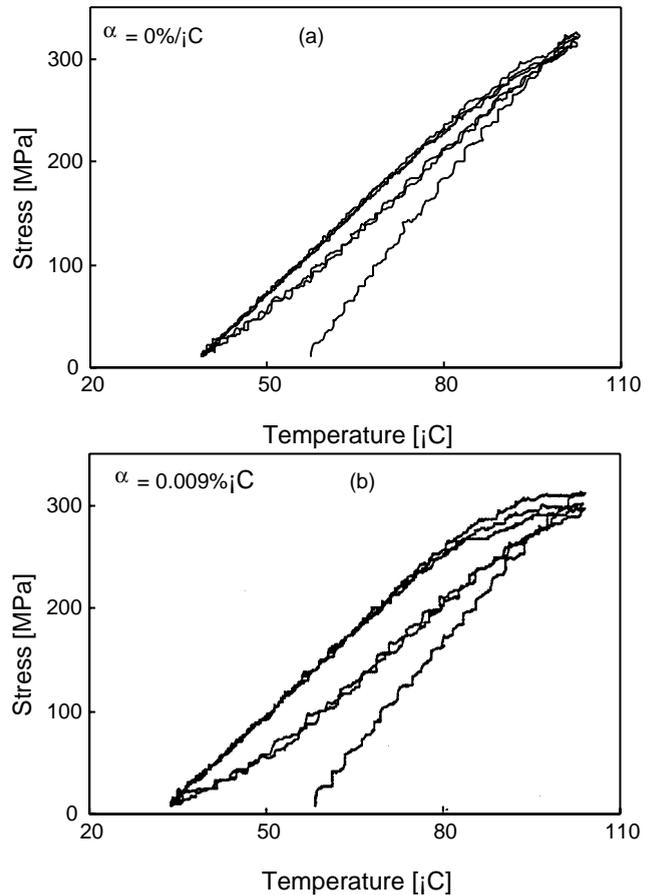


Figure 5. Recovery stress versus temperature for a constant prestrain of 2%. (b) Recovery stress versus temperature for a variable prestrain. The prestrain e is a function of temperature, given by $e = 2\% + \alpha x (T - 55)$, where T is the temperature in $^{\circ}\text{C}$. More details can be found in (114).

Yet, in most practical cases, the conditions are quite different. Only recently, the first results for more complex constraining conditions and for cyclic activation appeared in the literature. An example evidencing some important features is shown in Fig. 5.

The figure shows clearly that the stress–temperature path during the first heating can be quite different from the second heating path. Second and subsequent heating and cooling paths coincide. This difference between the first and second heating path can be caused by a difference in hysteresis paths (99) or by one-time shifts of the reverse transformation temperatures that result from the initial predeformation (116).

It can be also seen that the temperature hysteresis between the heating and cooling paths can be extremely small in isostrain conditions (Fig. 5a) but can increase significantly for nonisostrain conditions (Fig. 5b). This can be simply explained. The fraction of martensite transforming in the isostrain conditions of Fig. 5a is much smaller than the transformed fraction in Fig. 5b. The temperature hysteresis is, in a first approximation, proportional to the transformed fraction, which explains directly the small temperature hysteresis in Fig. 5a (99). An

important conclusion is that for most practical constraining conditions, only a small fraction of martensite is actually transforming during the constrained heating and cooling. Therefore, the observed hysteresis is much smaller than the overall hysteresis, typically below 10 K for Ni–Ti, below 5K for Ni–Ti–Cu, and below 2K for the R-phase transformation (99,114).

As stated before, it is often assumed that stress changes linearly as temperature changes. Moreover, the stress rate $d\sigma_r/dT$ is often considered a material constant directly derived from a Clausius–Clapeyron equation. Such descriptions should be considered very elementary simplifications. The results in Fig. 5 show clearly that the stress rate decreases during constrained heating. It has been also found that the stress is affected by many other parameters, including the thermomechanical history and the prestrain (69,86).

Depending on the magnitude of the prestrain, either a plastic upper limit or an elastic upper limit to σ_r exists. At lower prestrains, the stress increases during heating until the reverse transformation is completed. The upper stress limit in this case is given by the strain divided by Young's modulus. When the prestrain is sufficiently high, the stress increases during heating until plastic yield occurs at a temperature M_d . So, the upper stress limit in this case is the plastic yield stress σ_y . Evidently, in cyclic actuation, the maximum temperature should be kept below M_d .

In all cases discussed, a constraint prevents the SMA element from returning to the hot shape when heated. Thus, a more specific name would be “hot shape” recovery stress. It has been found in trained Cu-based SMA-elements that stresses can also be generated when the TWME is impeded during cooling (69,83). Because the constraint in this case prevents the sample from returning to the cold shape when cooled, the generated stress was called “cold shape” recovery stress to contrast with “hot shape” recovery stress. Practical applications have not been reported so far.

Quantitative comprehension of recovery stress generation presented in the literature is far below the comprehension of the other functional properties of shape-memory alloys. Therefore, recovery stress generation was discussed a bit more extensively than other functional properties. Considering the substantial research efforts in developing hybrid composites that have embedded shape-memory elements, substantial progress in quantitatively understanding recovery stress generation can be expected in the near future.

Work Output

One- and two-way memory effects can be used for free recovery applications in which the single function of the SMA element is to cause motions without any biasing stress. Under constant strains, shape-memory elements can generate substantial recovery stresses. Between these two extremes of free recovery and completely constrained recovery, shape-memory components can yield a wide variety of combinations of strains and stresses. As shown in

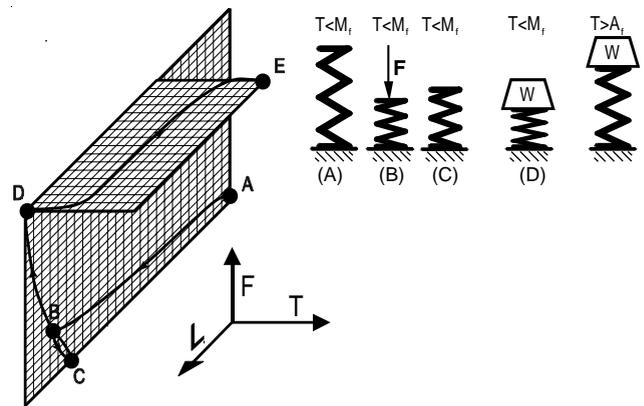


Figure 6. The work output. The sample is deformed at a temperature at a below M_f (A→B), followed by unloading (B→C) and loading again using a bias weight W (C→D). Shape recovery occurs under an opposing force W during heating to a temperature above A_f (D→E). So work is done [from (69)].

Fig. 6, a shape-memory element can be deformed by low force in the martensitic condition or during the forward transformation and can exert a substantially higher force as it reverts to the hot shape when heated. So, work up to 5 J/g is done during heating. This concept can be used in thermal actuators in which the SMA element is activated by an increase in the environmental temperature, or in electrical actuators in which the SMA element is in general activated by direct Joule heating. The work needed to deform the SMA element is much lower than the work that can be obtained during heating. This has been the basis of many prototypes of heat engines that convert heat into useful work [see (117)].

SMA actuators offer distinct advantages compared other types of actuators (118). The main advantage is that by far SMA actuators offer the highest work and power-to-weight ratios of all available actuating technologies at low levels of weight (119). These high-work and high-power densities enable a whole class of applications (e.g., in the field of micro actuation) that are impossible to realize by using other actuating technologies (120–123). SMA actuators can be reduced mostly to a single SMA element without auxiliary parts, resulting in simple compact and reliable devices.

Several important drawbacks that limit the use of SMA actuators to specific niches should also be considered. The conversion of heat into mechanical energy via SMA actuators was studied extensively 15 to 25 years ago. Simple thermodynamic calculations showed that the maximum theoretical efficiency of an SMA actuator is less than 10% (124). In practice, the conversion of heat into mechanical work is less efficient, and the result is that real efficiency is even one order of magnitude smaller than the theoretical value. Another drawback is that the SMA actuator has to be heated and cooled. The low cooling rate, especially limits the use of SMA actuators to relatively low-frequency applications. It was discussed before that stresses in trained Cu-based SMA elements can also be generated when the

TWME is impeded during cooling. Similarly, it has been shown that these trained SMA elements can do a small amount of work during cooling (82,83).

High Damping Capacity

SMA elements have high damping capacity in the austenitic and martensitic conditions. Shape-memory alloys show strong amplitude-dependent internal friction in the martensitic condition. For impact loads, the specific damping capacity can be as high as 90%. Starting from the austenitic condition, energy is dissipated during superelastic cycling as a result of stress hysteresis between superelastic loading and unloading, as explained before. A detailed analysis can be found in (125,126).

BIBLIOGRAPHY

1. L. Delaey, M. Chandrasekaran, M. Andrade, and J. Van Humbeeck, *Proc. Int. Conf. Solid-Solid Phase Transform.* pp. 1429–1453.
2. T. Maki, and T. Tamura, *Proc. ICOMAT-86*, 1986, pp. 963–970.
3. J. Van Humbeeck, M. Chandrasekaran and R. Stalmans, *Proc. ICOMAT-92*, Monterey, CA, 1993, pp. 1015–1025.
4. Q. Gu, J. Van Humbeeck, and L. Delaey, transformation and shape memory effect in Fe-Mn-Si alloys, *J. Phys. IV Col. C3 4*: C3-135–140 (1994).
5. J. Van Humbeeck, *Trans. ASME* **121**: 98–101 (1999).
6. T. Maki, in *Shape Memory Materials*, K. Otsuka, and C.M. Wayman eds., Cambridge University Press, 1998, Chap. 5.
7. A. Sato, K. Sama, E. Chishima, and T. Mori, *J. Phys.* **12**(C4): 797–802 (1982).
8. A. Sato, E. Chishima, K. Soma, and T. Mori, *Acta Metall.* **30**: 1177–1183 (1982).
9. J.S. Robinson and P.G. McGormick, *Mater. Sci. Forum* **56**: 649–653 (1990).
10. A. Sato, *MRS Int. Mtee. Adv. Mater.* 1989, pp. 431–445.
11. M. Murakami, H. Otsuka, H.G. Suzuki, and S. Matsuda *Proc. Int. Conf. Martensitic Transform. (ICOMAT-86)*, 1986, pp. 985–990.
12. D.Z. Liu, W.X. Liu, and F.Y. Gong, *J. Phys. IV, Suppl. J. Phys. III* **5**: (1995), *ICOMAT-95 Proc.*, pp. 1241–1246.
13. J. Van Humbeeck and L. Delaey, *The Martensitic Transformation in Science and Technology*, E. Hornbogen and N. Jost, eds., DGM Verlag, Oberursel, Germany, 1989, pp. 15–25.
14. T. Tadaki, in *Shape Memory Materials*, K. Otsuka, and C.M. Wayman, eds., Cambridge University Press, 1998.
15. M. Ahlers, *Progress in Materials Science*, Vol. 30, J.W. Christian, P. Haasen, and T.B. Massalski, eds., 1986, pp. 135–186.
16. S.K. Wu and H. Ming, in *Engineering Aspects of Shape Memory Alloys*, T. Duerig et al., eds., Butterworth Scientific, U.K., 1990, pp. 69–88.
17. R. Rapacioli and M. Chandrasekaran, *Proc. Int. Conf. Martensitic Transform.* Cambridge, MA, 1979, pp. 596–601.
18. R. Rapacioli and M. Ahlers, *Acta Metall.* **27**: 777–784 (1979).
19. J. Van Humbeeck, L. Delaey and D. Roedolf, *Proc. Int. Conf. Martensitic Transform. (ICOMAT-1986)*, 1986, pp. 862–867.
20. P. Rodriguez and G. Guénil, in *The Martensitic Transformation in Science and Technology*, E. Hornbogen and N. Jost, eds., DGM Verlag, Oberursel, Germany, 1989, pp. 149–156.
21. *Proc. Int. Conf. Martensitic Transform.* L. Delaey, and M. Chandrasekaran, eds., *J. de Phys.* **43** Colloque C4, Suppl. 12 1982; see papers 1. by G. Scarsbrook et al., pp. 703–708; 2. A. Abu Arab et al, pp. 709–714; 3. J. Janssen et al., pp. 715–720.
22. A. Abu Arab, M. Chandrasekaran, and M. Ahlers, *Scripta Metall.* **18**: 709–714 (1984).
23. X. Duan and W. Stobbs, *Scripta Metall.*, **23**: 441–445 (1989).
24. M. Chandrasekaran, E. Cesari, and J. Van Humbeeck, *Proc. Int. Conf. Martensitic Transform ICOMAT-92*, Monterey Inst. of Advanced Studies, 1993, pp. 575–580.
25. J. Van Humbeeck, L. Delaey, and T. Suzuki, *Scripta Metall.* **18**: 893–903 (1984).
26. F.C. Lovey et al., “The influence of γ -phase precipitates on the martensitic transformation in Cu-Zn-Al alloys”, *Proc. Intl. Conf. Martensitic Transform*, Sydney, 1990, **56–58**, p. 493–498.
27. R. Elst, J. Van Humbeeck, and L. Delaey, *Mater. Sci. Technol.* **4**: pp. 644–648 (1988).
28. R. Elst, J. Van Humbeeck, and L. Delaey, *Acta Met.* **36** 1723–1729 (1988).
29. E. Hornbogen, *Acta Metall.* **33**: 595–601 (1985).
30. D.N. Adnyana, *Proc. Int. Conf. Martensitic Transform.*, 1986, pp. 774–779.
31. Z. Nishiyama, *Martensitic Transformation*. Academic Press, 1978.
32. R. Elst, J. Van Humbeeck, and L. Delaey, *Proc. Int. Conf. Martensitic Transform*, 1986, pp. 891–895.
33. W. Jianxin, J. Bohong, and T.Y. Hsu, *Acta Metall.* **36**: 1521–1526 (1988).
34. J.V. Wood and P.H. Shingu, *Met. Trans A* **15A**: 471–480 (1984).
35. K. Mukunthan and L.C. Brown, *Met. Trans A* **19A**: 2921–2929 (1988).
36. A.Q. Khan and L. Delaey, *Zeit. für Met.* **60**: 949–951 (1969).
37. T. Saburi, *MRS Int. Meet. Adv. Mater.* 1989, Vol. 9, pp. 77–91.
38. T. Saburi, in *Shape Memory Materials*, K. Otsuka and C.M. Wayman, eds., Cambridge University Press, 1998, Chap. 3.
39. T. Todoriki and H. Tamura, *Trans. Jpn. Inst. Met.* **28**: 83–94 (1987).
40. D. Treppman, E. Hornbogen, and D. Wurzel, 1995, *Proc. ICOMAT-95*, pp. 569–574.
42. M. Nishida, C.M. Wayman, and T. Honma, *Met. Trans.* **17A**: 1505–1515 (1986).
43. C.M. Wayman, *Proc. Int. Conf. Martensitic Transform.* 1986, pp. 645–652.
44. K.H. Eckelmeyer, *Scripta Met.* **10**: 667–672 (1976).
45. T. Honma, M. Matsumoto, Y. Shugo, and I. Yamazaki, *Proc. ICOMAT-79*, pp. 259–264.
46. P. Huisman-Kleinherenbrink, Ph. D. Thesis, University of Twente (NL), 1991.
47. V.N. Kachin, “Martensitic transformation and shape memory effect in B2 intermetallic compounds of Ti”, *Rev. Phys. Appl.* **24**: 733–739 (1989).

48. V.N. Kachin, V.P. Voronin, V.P. Sivokha, and V.G. Pushin, *ICOMAT-95 Proc. Part I and II*, Lausanne, 1995, pp. 707–172.
49. T. Saburi, T. Takagaki, S. Nenno, and K. Koshino, *MRS Int. Meet. Adv. Mater.* 1989, Vol. 9, pp. 147–152.
50. H. Horikawa and T. Ueki, *Trans. Mater. Res. Soc. Jpn.* **18B**: 1113–1116 (1993).
51. K.N. Melton, J.L. Proft, and T.W. Duerig, *MRS Int. Meet. Adv. Mater.* 1989, Vol. 9, pp. 165–170.
52. J.H. Yang and J.W. Simpson, *ICOMAT-95 Proc. Part I and II*, Lausanne, 1995, pp. 771–776.
53. E.W. Collings, *The Physical Metallurgy of Titanium Alloys*, ASM Series in Metal Processing, 1984.
54. O.M. Ivasishin, N.S. Kosenko, and S.V. Shevskenko, *ICOMAT-95 Proc. Part I and II*, Lausanne, 1995, pp. 1017–1022.
56. H. Sasano and T. Suzuki, *5th Proc. Int. Conf. Titanium*, 1984, Vol. 3, pp. 1667–1674.
57. T. Hamada, T. Sodeoka, and M. Miyagi, *Proc. Int. Conf. Titanium*, 1988, Vol. 2, pp. 877–882.
58. T. Sugimoto, M. Ikeda, S. Komatsu, K. Sugimoto, and K. Kamei, *Proc. Int. Conf. Titanium*, 1988, Vol. 2, pp. 1069–1074.
59. T.W. Duerig, J. Abrecht, D. Richter, and P. Fisher, *Acta Met.* **30**: 2161–2172 (1982).
60. T. Sohmura and H. Kimura, *Proc. ICOMAT-86*, pp. 1065–1070.
61. S.G. Fedotov, T.V. Chelidze, Y.K. Kovernisty, and V.V. Sanadze, *Phys. Met. Metall.* **62**: 109–112 (1986).
62. Europäische Patentanmeldung 0062365, anmeldetag 11, 03, 1982.
63. C.Y. Lei, M.H. Wu, L. McD. Schetky, and C.J. Burstone, *Proc. Wayman Symp.*, May 1996, pp. 413–418.
64. T. Kakeshita and K. Shimizu, *Trans. Mater. Res. Soc. Jpn.* **18B**: 981–984 (1994).
65. T. Kakeshita, T. Saburi, and K. Shimizu, *MRS Symp.*, Boston, 1997, Vol. 459, pp. 269–279.
66. V.A. Chernenko, A. Amengual, E. Cesari, V.V. Kokorin, and I.K. Zasmichuk, *ESOMAT-94 Proc.*, Barcelona, 1994, *J. Phys. IV* Vol. 5; C2, pp. 95–98.
67. V.A. Chernenko, E. Cesari, V.V. Kokorin, and I.N. Vitenko, *Scripta Met. Mater.* **33**: 1239–1244 (1995).
68. M. Cohen, G.B. Olson, and P.C. Clapp, phase transformations, *Proc. ICOMAT-79*, Cambridge, pp. 1–11.
69. R. Stalmans, Doctorate Thesis, Catholic University of Leuven, Department of Materials Science, Heverlee, 1993.
70. A.B. Greninger and V.G. Mooradian, *Trans. Met. Soc.* **128**: *AIME*, 337 (1938).
71. L. Delaey, in *Materials Science and Technology*, Vol. 5: Phase transformations in Materials, P. Haasen, ed., VCH Verlagsgesellschaft, Weinheim, 1991, pp. 339–404.
72. W.J. Buehler, J.W. Gilfrich, and R.C. Wiley, *J. Appl. Phys.* **34**: 1616 (1963).
73. J. Perkins and D. Hodgson, in *Engineering Aspects of Shape Memory Alloys*, T.W. Duerig, K.N. Melton, D. Stöckel, and C.M. Wayman, eds., Butterworth-Heinemann, London, 1990, pp. 195–206.
74. H. Tas, L. Delaey, and A. Deruyttere, *J. Less-Common Met.* **28**: 141–151 (1972).
75. L. Contardo and G. Guénin, *Acta Metall. Mater.* **38**: 1267–1272 (1990).
76. G. Guénin, in *The Martensitic Transformation in Science and Technology*, E. Hornbogen and N. Jost, eds., DGM Informationsgesellschaft, Oberursel, 1989, pp. 39–52.
77. Y. Liu and P.G. McCormick, *Scripta Metall.* **22**: 1327–1330 (1988).
78. P.Y. Manach and D. Favier, *Scripta Metall.* **28**: 1417–1421 (1993).
79. J. Perkins and R.O. Sponholz, *Met. Trans A* **15**: 313–321 (1984).
80. D. Rios-Jara and G. Guénin, *Acta Metall.* **35**: 109–126 (1987).
81. R. Stalmans, J. Van Humbeeck, and L. Delaey, *Acta Metall. Mater.* **40**: 501–511 (1992).
82. R. Stalmans, J. Van Humbeeck, and L. Delaey, *Acta Metall. Mater.* **40**: 2921–2931 (1992).
83. R. Stalmans, J. Van Humbeeck, and L. Delaey, *Mater. Trans. JIM* **33**: 289–293 (1992).
84. R. Stalmans, J. Van Humbeeck, and L. Delaey, *Scripta Metall.* **31**: 1573–1576 (1994).
85. M. Sade, A. Hazarabedian, A. Uribarri, and F.C. Lovey, *Proc. Int. Conf. Solid Phase Transform.*, Cambridge, 1987, pp. 279–281.
86. J. Van Humbeeck and R. Stalmans, in *Shape Memory Materials*, K. Otsuka and C.M. Wayman, eds., Cambridge University Press, Cambridge, 1998, pp. 149–183.
87. T. Saburi, in *Shape Memory Materials*, K. Otsuka and C.M. Wayman, eds., Cambridge University Press, Cambridge, 1998, pp. 49–96.
88. R. Kainuma and M. Matsumoto, *Scripta Metall.* **22**: 475–478 (1988).
89. K. Otsuka and K. Shimizu, *Int. Met. Rev.* **31**: 93–114 (1986).
90. T.W. Duerig and R. Zadno, in *Engineering Aspects of Shape Memory Alloys*, T.W. Duerig, K.N. Melton, D. Stöckel, and C.M. Wayman, eds., Butterworth-Heinemann, London, 1990, pp. 369–393.
91. S.M. Russell, D.E. Hodgson, and F. Basin, *Proc. 2nd Int. Conf. Shape Memory Superelastic Technol.*, SMST, Santa Clara, 1997, pp. 429–436.
92. C.S. Zhang, Y.Q. Wang, J.X. Cheng, and C.L. Zhao, *Proc. 1st Int. Conf. Shape Memory Superelastic Technol.*, Monterey, 1994, pp. 383–388.
93. T.W. Duerig, A.R. Pelton, and D. Stöckel, *Metall* **50**: 569–574 (1996).
94. K.N. Melton, in *Shape Memory Materials '94*, C. Yougi and T. Hailing, eds., International Academic, Beijing, 1994, pp. 523–529.
95. *Proc. 1st Int. Conf. Shape Memory Superelastic Technol.*, MIAS, Monterey, 1994, pp. 341–400.
96. O. Mercier and E. Torok, *J. Phys.* **C4**: 267–272 (1982).
97. G.R. Zadno and T.W. Duerig, in *Engineering Aspects of Shape Memory Alloys*, T.W. Duerig, K.N. Melton, D. Stöckel, and C.M. Wayman, eds., Butterworth-Heinemann, London, 1990, pp. 414–419.
98. K. Otsuka and C.M. Wayman, eds., *Shape Memory Materials*. Cambridge University press, Cambridge, 1998, p. 150.
99. P. Sittner, D. Vokoun, G.N. Dayananda, and R. Stalmans, *Mater. Sci. Eng.* **286**: 298–311 (2000).
100. J.L. Proft and T.W. Duerig, in *Engineering Aspects of Shape Memory Alloys*, T.W. Duerig, K.N. Melton, D. Stöckel and C.M. Wayman, eds., Butterworth-Heinemann London, 1990, pp.115–129.
101. M. Kapgan and K. N. Melton, in *Engineering Aspects of Shape Memory Alloys*, T.W. Duerig, K.N. Melton, D. Stöckel, and

- C.M. Wayman, eds., Butterworth-Heinemann, London, 1990, pp. 137–148.
102. C. Jackson, H. Wagner, and R. Wasilewski, NASA-SP-5110, 1972.
 103. T. Borden, in *Engineering Aspects of Shape Memory Alloys*, T.W. Duerig, K.N. Melton, D. Stöckel eds., and C.M. Wayman, Butterworth-Heinemann, London, 1990, pp. 158–169.
 104. G.R. Edwards, J. Perkins, and J.M. Johnson, *Scripta Metall.* **9**: 1167–1171 (1975).
 105. J. Perkins, G.R. Edwards, C.R. Such, S.M. Johnson, and R.R. Allen, in *Shape Memory Effect in Alloys*, J.Perkins, ed., Plenum Press, NY, 1975, pp. 273–283.
 106. T.W. Duerig and K.N. Melton, in *Shape Memory Alloy '86*, Chu Yougi, T.Y. Hsu, and T. Ko, eds., China Academic, Beijing, 1987, pp. 397–405.
 107. K. Madangopal, S. Banerjee, and M.K. Asundi, in *Shape Memory Alloy '86*, Chu Yougi, T. Y. Hsu, and T. Ko, eds., China Academic, Beijing, 1987, pp. 181–186.
 108. K. Madangopal, R. Ganesh Krishnan, and S. Banerjee, *Scripta Metall.* **22**: 1593–1598 (1988).
 109. Y. Furuya, H. Shimada, Y. Tanahashi, M. Matsumoto, and T. Honma, *Scripta Metall.* **22**: 751–755 (1988).
 110. T.W. Duerig and K.N. Melton, *MRS Int. Meet Adv. Maters.*, 1989, Vol. 9, pp. 581–597.
 111. E. Cydzik, in *Engineering Aspects of Shape Memory Alloys*, T.W. Duerig, K.N. Melton, D. Stöckel, and C.M. Wayman, eds., Butterworth-Heinemann, London, 1990, pp. 149–157.
 112. H. Tobushi, P.H. Lin, K. Tanaka, C. LExcellent, and A. Ikai, *Proc. 1st Int. Conf. Shape Memory Superelastic Technol., MIAS, Monterey*, 1994, pp. 163–168.
 113. R. Stalmans, L. Delaey, and J. Van Humbeeck, *Mater. Res. Soc. Symp. Proc.: Mater. Smart Syst. II*, Pittsburgh, 1987, Vol. 459, pp. 119–130.
 114. D. Vokoun and R. Stalmans, *SPIE* **3667**: 825–835 (1999).
 115. Y. Zheng, L. Cui, D. Zhu, and D. Yang, *Mater. Lett.* **43**: 91–96 (2000).
 116. M. Piao, K. Otsuka, S. Miyazaki, and H. Horikawa, *Mat. Trans. JIM* **14**: 919–929 (1993).
 117. R.M. Banks and M.C. Fleming, *J. Phys. Colloque C-4*: 833–838 (1982).
 118. J. Van Humbeeck, R. Stalmans, and P. A. Besselink, in *Metals as Biomaterials*, J. Helsen and H. Breme, eds., Wiley, 1998, pp. 73–100.
 119. K. Ikuta, *Proc. IEEE Workshop*, 1990, 2156–2161.
 120. Y. Bellouard, T. Lehnert, T. Sidler, R. Gotthardt, and R. Clavel, *MRS Symp., Mater. Smart Syst. III*, Vol. 604, pp. 177–182.
 121. Y. Bellouard, Thèse, Nr. 2308 (2000), Lausanne EPFL.
 122. D. Reynaerts, J. Peirs, and H. Van Brussel, *Proc. 2nd Int. Conf. SMST*, Asilomar, 1997, pp. 533–538.
 123. A.D. Johnson and J.D. Bush, *Proc. 1st Int. Conf. SMST*, Asilomar, 1994, pp. 299–304.
 124. P. Wollants, M. De Bonte, L. Delaey, and J. Roos, *Z. Metallk.* **74**: 146–151 (1979).
 125. J. Van Humbeeck, *Mater. Sci. Eng. A* **273–275**: 134–148 (1999).
 126. J. Van Humbeeck, Chapter 3, pp. 46–60 in “Shape Memory Implants” ed. by L' Hacine Yahia, Springer-Verlag 2000, ISBN 3-540-67229-X.
 127. C. Baber, *Metal Science Journal*, 1971, Vol. 6, pp. 92–100.

SHAPE-MEMORY MATERIALS, MODELING

DAVIDE BERNARDINI

Università di Roma “La Sapienza”
Rome, Italy

THOMAS J. PENCE

Michigan State University
East Lansing, MI

INTRODUCTION

An understanding of shape-memory behavior requires knowledge of various physical processes that operate on different length scales. The crystallographic shifts that are responsible for shape-memory behavior take place in unit cells of atomic dimension. Here, however, these microscopic length scales are not the focus, rather, this article considers the macroscopic scale modeling that allows for the engineering assessment of thermomechanical response (stress–strain–temperature) and energy balances (including damping) for devices such as connectors, actuators, vibration absorbers, and biomedical stents. The macro scale is useful for primary design evaluation such as predicting triggering forces and determining range of motion. Since the shape-memory material is typically incorporated into a larger engineered device or structure, there is also a need for detailed computational simulation of the system as a whole.

BASIC MATERIAL BEHAVIOR AND MODELING ISSUES

The term *shape-memory material* (SMM) is meant to encompass a wide class of metallic alloys with the common feature that they exhibit, at the macroscopic scale, some peculiar and useful functional properties such as pseudoelasticity and shape memory. Nickel-titanium (NiTi) is perhaps the best known and most widely used such material. SMM functional properties derive from transformations between two different solid phases: austenite (**A**) and martensite (**M**). Aspects of the **A** ↔ **M** phase transformation are essential for model development and implementation. In addition, certain intermediate phases may also occur, but these are neglected here because their effect on the macroscopic response is small in comparison.

The **A** ↔ **M** transformation can be induced by a variety of energy inputs (mechanical, thermal, magnetic, ultrasonic, etc.), and it is influenced by grain boundaries, dislocations, inclusions, and other material defects. This article considers the standard thermomechanical setting, namely the **A** ↔ **M** transformation that is induced by temperature T and stress σ . In general, austenite is favored at high temperatures and low stress, whereas martensite is favored at low temperatures and high stress. We will use a boldface σ (and ϵ) to denote a general stress and strain tensors with components σ_{ij} (and ϵ_{ij}). Models involving only a single stress component σ will be developed more than those involving the tensor σ .

Austenite is of higher crystallographic symmetry and so can transform into one or more martensite variants that differ mainly by their orientation relation to the austenite

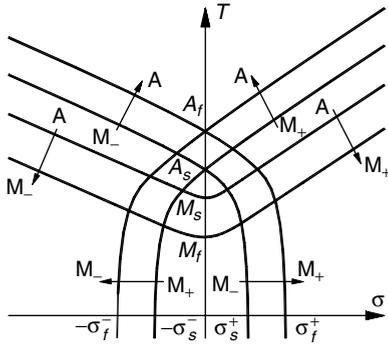


Figure 1. A phase diagram defines the zones of the (σ, T) -plane where the various phase transformations can occur. Each curve represents the (σ, T) -points at which a transformation can either be activated or else completed. Hence there are two curves (start and finish) for each transformation. The figure shows a sketch of a phase diagram that can arise when the SMM is modeled as a mixture of austenite **A** and two martensite variants **M+**, **M-**. Such a diagram can be viewed as an unfolding of a conventional phase diagram triple point so as to include the effects of phase mixing and transformational hysteresis.

parent. By contrast, all martensite variants tend to transform into a single common austenite crystal structure. The transformation from the **A** structure to that of a particular **M** variant is characterized by a crystallographic transformation strain. Typically, an **A** material region transforms into a martensitic microstructure with several variants that combine in complicated twin arrangements and plate morphologies. These microstructures provide a *local transformation strain* γ^* . This in turn gives a *macroscopic transformation strain* ε^* at the engineering scale. This ε^* gives a potentially large strain in stress-induced **A** \rightarrow **M** transformations. It is negligible in cooling-induced **A** \rightarrow **M** transformations because the resulting microstructures involve so-called self-accommodated martensite with local strains γ^* that cancel each other.

Since the phase transformations can be activated under very different conditions to obtain different effects, to have a picture of SMMs behavior, it is necessary to see how stress and temperature differ with respect to **A** \leftrightarrow **M** transformation. Figure 1 shows a stress and temperature phase diagram of a single austenite phase **A** and two families of martensite variants **M+** and **M-**. The curves in this diagram show the relation between stress and temperature levels at which various phase transformations begin and end. This partitions the (σ, T) -plane into three single phase regions, three double phase regions, and a triple phase region.

Purely Thermal Transformation

In the absence of stress, austenite is stable at high temperatures, and martensite is stable at low temperatures. Stress-free cooling of austenite gives **A** \rightarrow **M** conversion beginning and concluding at temperatures M_s and M_f , respectively ($M_f < M_s$). The resulting microstructure is an *unbiased martensite* with a fine-scale arrangement of variant twins with opposing local transformation strains γ^* .

This produces a negligible engineering scale transformation strain $\varepsilon^* = 0$. Similarly, a temperature increase causes **M** \rightarrow **A** conversion, the start and finish temperatures being A_s and A_f , with $A_f > A_s$. The transformation temperatures M_f , M_s , A_s , and A_f are the basic material parameters for purely thermal transformations. They are highly sensitive to the alloy concentrations and to the granular and defect structure as determined by heat treatment and cold work. Once the material is ready for service, these temperatures are easily determined by various means including resistivity and calorimetry testing.

Low-Temperature Martensite Reorientation

Austenite is not present at temperatures below M_f . Nonzero stress at these low temperatures causes certain martensite variants to be relatively more favored. This favoritism correlates with the value of the transformation work $\sigma \cdot \gamma^* = \sigma_{ij} \gamma_{ij}^*$. In the important special case of uniaxial tension/compression, the variants favored in tension are those for which γ^* projects onto the tensile axis as a positive quantity. It is convenient to group all of the variants favored in tension into an **M+** variant family, and all variants favored in compression into an **M-** variant family. Unbiased martensite involves a mixture of both **M+** and **M-**. Sufficiently high tensile loading at temperatures below M_f causes movement of the internal boundaries separating the martensite plates, which can be viewed as a conversion from the **M-** family into the **M+** family. This **M-** \rightarrow **M+** transformation yields *biased martensite*, namely $\varepsilon^* \neq 0$. Unloading does not cause the reverse transformation (**M+** \rightarrow **M-**) so long as the load does not become compressive. Hence the transformation strain ε^* is like a conventional plastic strain upon unloading (Fig. 2). The **M-** \leftrightarrow **M+** transformation is referred to as *reorientation*. As a result, a plasticlike *reorientation plateau* is observed on the isothermal stress-strain curve, with the tensile reorientation beginning and concluding at stresses σ_s^+ and σ_f^+ ($\sigma_f^+ > \sigma_s^+ > 0$). These reorientation stresses are relatively insensitive to temperature changes (there may be a mild increase in σ_s^+ and σ_f^+ with a temperature

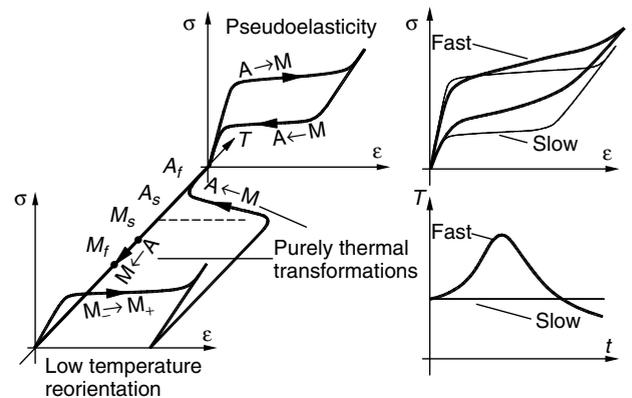


Figure 2. Sketch of the macroscopic effects of the various phase transformations in the stress-strain-temperature $(\sigma-\varepsilon-T)$ space (left). Effect of the loading rate on the pseudoelastic behavior: faster loads give rise to greater hardening and temperature variations (right, where t denotes time).

decrease). Similar reorientation processes occur in compressive loading due to the $\mathbf{M}^+ \rightarrow \mathbf{M}^-$ transformation, with the start and finish of the reorientation plateau described by stresses $-\sigma_s^-$ and $-\sigma_f^- (-\sigma_f^- < -\sigma_s^- < 0)$.

Pseudoelasticity

Austenite can be present at temperatures above M_f . Such austenite is converted to biased martensite upon application of sufficiently high stress, either tensile (activating $\mathbf{A} \rightarrow \mathbf{M}^+$) or compressive (activating $\mathbf{A} \rightarrow \mathbf{M}^-$).

Consider the tensile case. A stress increase gives the $\mathbf{A} \rightarrow \mathbf{M}^+$ transformation, which again generates a plateau on the stress-strain diagram. This is referred to as a *pseudoelastic* plateau so as to distinguish it from the reorientation plateau observed at the lower temperatures. For $T > A_f$, the start and finish stresses for pseudoelasticity are greater than the $\mathbf{M}^- \rightarrow \mathbf{M}^+$ reorientation stresses σ_s^+ and σ_f^+ . They are also highly temperature sensitive, increasing with temperature at an approximately constant rate. However, if the temperature is close to M_f , then little distinction can be made between pseudoelasticity and reorientation because the loading plateau stresses match the reorientation values σ_s^+ and σ_f^+ . Unloading activates the $\mathbf{M}^+ \rightarrow \mathbf{A}$ transformation if $T > A_s$, resulting in an unloading plateau below the $\mathbf{A} \rightarrow \mathbf{M}^+$ loading plateau. The unloading plateau rejoins the loading curve if $\mathbf{M}^+ \rightarrow \mathbf{A}$ goes to completion ($T > A_f$), and so defines a hysteresis loop (Fig. 2). At temperatures $A_s < T < A_f$, the $\mathbf{M}^+ \rightarrow \mathbf{A}$ unloading conversion does not go to completion and the unloading plateau intersects the strain axis before reaching the origin of the stress-strain diagram. If $T < A_s$, then the $\mathbf{M}^+ \rightarrow \mathbf{A}$ transformation is not even activated upon unloading. Thus, in all cases where $T < A_f$, there is some residual strain due to the presence of \mathbf{M}^+ martensite when unloading is complete.

Shape-Memory Effect

At all temperatures, where $T < A_f$, after sufficiently high load causing either $\mathbf{A} \rightarrow \mathbf{M}^+$ or $\mathbf{M}^- \rightarrow \mathbf{M}^+$ transformation, residual strain is present after unloading due to the presence of biased martensite. Unlike conventional plastic flow in metals (generated by dislocations) the SMM plasticlike residual strain is recovered by heating above A_f , because this converts martensite to austenite. Since this austenite converts to unbiased martensite upon any later stress-free cooling, the residual plastic strain does not return (unless there is further loading/unloading). This heating/cooling elimination of an apparently “plastic” strain due to previous loading/unloading is the *shape-memory effect*.

While the preceding discussion has covered the basic aspects of the material behavior that macroscopic models should reproduce, SMM is often employed in situations involving further effects that are important objectives for useful modeling. The most important of these are briefly described next.

Response to Complex Loading Paths

At constant temperature, loading reversals that interrupt $\mathbf{A} \rightarrow \mathbf{M}$ and $\mathbf{M} \rightarrow \mathbf{A}$ before completion lead to internal

subloops within the major stress-strain hysteresis loop associated with complete transformation. Load paths involving simultaneous change in T and σ generally augment or diminish transformation that would occur under either T or σ alone. This is critical for modeling the rate effect that is described next.

Rate Dependency due to Transformational Heating

If mechanical loads producing phase transformation are not applied in a quasi-static way, then temperature variation occurs in the sample and a rate-dependent response is observed. This is due to the exothermic and endothermic nature of the $\mathbf{A} \rightarrow \mathbf{M}$ and $\mathbf{M} \rightarrow \mathbf{A}$ transformations, respectively. During $\mathbf{A} \rightarrow \mathbf{M}$ the material self-heats and the temperature rise works against the transformation (conversely during $\mathbf{M} \rightarrow \mathbf{A}$ the sample self-cools). This might involve a number of consequences: different onset stresses for transformation plateaus, plateau steepening, and variation in the shape of internal loops (Fig. 2). The extent of this effect is governed by the heat exchange with the environment: high rates of loading can cause significant departure from isothermal behavior. High rates of loading can occur in both actuator and damping shape-memory devices.

Tension/Compression Asymmetries

In uniaxial loading, significant differences in the stress-strain behavior have been observed between tension and compression. This is due to the different microstructures that are formed. In particular, the behavior of the \mathbf{M}^+ variant family in tension is not the symmetric image of the behavior of the \mathbf{M}^- variant family in compression. Such a phenomenon is modeled by a phase diagram that is asymmetric with respects to σ .

Three-Dimensional States

In the three-dimensional case involving tensor σ rather than scalar σ , the experimental behavior is less well understood, and complex multivariant structures are to be expected in most cases. A key issue in three-dimensional modeling is the proper constitutive description of an appropriate local transformation strain that transcends the lack of information about the actual multivariant microstructure. In view of the correlation of variant favoritism with the transformation work $\sigma \cdot \gamma^* = \sigma_{ij} \gamma_{ij}^*$, some sort of coaxiality relations between stress and transformation strain are conjectured at modeling scales appropriate to a multivariant microstructure. These difficulties are compounded under nonproportional loading, since the transformation strain then evolves as a consequence of both pseudoelastic $\mathbf{A} \leftrightarrow \mathbf{M}$ processes and $\mathbf{M} \leftrightarrow \mathbf{M}$ reorientation of existing variants.

STATE OF THE ART AND HISTORICAL DEVELOPMENTS

Modeling of the macroscopic behavior of SMM has been the subject of much activity since the beginning of the 1980s, attracting the interest of engineers, applied mathematicians, and materials scientists. This section surveys the state of the art on the basis of the huge literature

available on the subject. The survey is restricted to models that directly relate to macroscopic modeling and does not delve into the voluminous literature on metal physics and purely microstructural development, even though much of this literature provides enormous insight. While an effort has been made to be rather comprehensive, the survey is still far from complete. We have attempted to give careful bibliographic references by selecting one representative paper for each approach. Each of these models then typically gives rise to refinements, generalizations, verification studies, and implementation strategies. For the sake of conciseness, complete bibliographic references cannot be given for all of these modeling extensions.

The discussion of the previous section makes clear that the behavior of SMM observable at the macroscopic scale is the effect of several complex microstructural phenomena. This section is organized, as in the list below, with respect to contributions that include an explicit model for such microstructural phenomena and others that do not. Although certain models that will be discussed can be viewed as spanning more than one such approach, the classification given below aids in organizing the numerous modeling approaches that have been proposed.

Approaches modeling one or more microstructural phenomena	<i>Lattice cell mechanics</i> <i>Interface nucleation and propagation</i> <i>Micromechanics</i>
Approaches modeling directly the macroscopic behavior	<i>Models without internal variables</i> <i>Hysteresis models</i> <i>Models with internal variables</i>

Approaches Modeling One or More Microstructural Phenomena

Models included in this group are grounded in theories that analyze the material at a scale in which the multiphase nature of the material is rendered explicit and one or more effects of phase transformations can be described by some direct microstructural model. Macroscopic behavior is then recovered by some kind of averaging procedure. In a continuum setting, this implies that each point belongs to one phase and the first spatial derivatives of the displacement and temperature fields can be discontinuous.

Lattice Cell Mechanics. In this approach, the macroscopic response of the material is determined by studying the behavior of a collection of lattice cells that can be in a particular phase or phase variant. In response to loads and temperature changes at the system boundary, cell transitions between different phases can take place. Two approaches for the transition kinetics can be broadly identified: *statistical mechanics* and *strain energy minimization*.

The statistical mechanics approach has roots in Muller and Wilmanski (1) and has been further developed by Achenbach (2). The cellular array is grouped as a stack of layered aggregates of cells that can be found in one of the

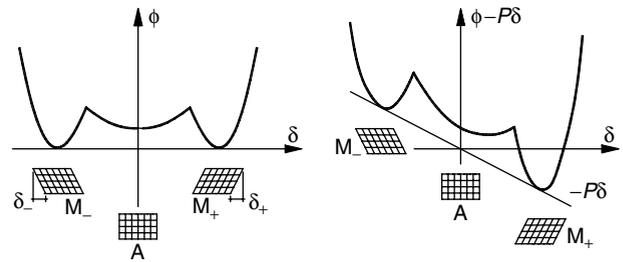


Figure 3. The three-well energy function of the Muller-Achenbach model without (*left*) or with (*right*) mechanical load. Each minimum corresponds to a phase with different structure (austenite **A** and two martensite variants **M₊**, **M₋**). The differences among the ordinates of the various minima represent the energy barriers that cells have to overcome to undergo phase transformations. The right graph shows the effect of a mechanical load P that lowers the right minimum and causes the **M₊** phase to be the energetical favorite.

three phases: austenite **A** and two martensite variants **M₊**, **M₋**; each characterized by a different cell length. Cells are in random thermal motion, and thermal fluctuations permit them to transform from one phase to another. Such transitions lead to variations in the stack composition that are monitored by the phase fractions ξ_A , ξ_+ , and ξ_- . A three-well potential energy ϕ whose minima are each associated with one phase is the basic constitutive ingredient from which all material parameters are derived by statistical arguments (Fig. 3). Macroscopic strain and temperature are obtained respectively from the normalized length of the whole layer aggregate and from a measure of the thermal fluctuation. The phase fraction evolution is governed by a system of ordinary differential equations expressing the transition rate balance between layers on the basis of the probability of overcoming the energy barriers that separate the minima of ϕ . This finally provides a complete model for uniaxial stress pseudoelasticity and reorientation.

A second approach for the description of the transitions between lattice cells is based on strain energy minimization. A model developed by Morris and his collaborators (3) involves a multidimensional lattice of cells with corresponding multidimensional transformation strains for each cell. The total free energy is the sum of a temperature-dependent chemical free energy and a strain-dependent elastic strain energy. The strain energy contribution is highly dependent on cell location and choice of transformation variant, due to the constraint of surrounding cells. The computation is based on isotropic elasticity, and the underlying mathematical technique developed by Khachaturyan requires equality of elastic constants in all phases. For a given change in temperature, the overall transformation process is simulated by a stepwise energy minimization. At each step, the particular cell that transforms is selected as the one that most lowers the energy, and such transformations continue so long as the overall energy is lowered. Complex microstructures and internal stress states occur. These complex states are the main focus of such modeling, as opposed to providing a macroscopic model for overall stress-strain-temperature behavior.

Interface Nucleation and Propagation. Large deformation theories of continuum mechanics have also been used to treat martensitic phase transformation. These treatments stem from analysis by Ericksen of a uniaxial continuum with a nonconvex elastic strain energy density (4). These energies give rise to nonmonotone stress–strain curves. Fundamental thermodynamic arguments harking back to Gibbs and Maxwell indicate that strains on the descending branches of these stress–strain curves are unstable in that they cannot be part of a deformation field that minimizes energy. Energy minimization naturally gives rise to distinct material regions, each involving continuous strain, that are separated from each other by interfaces across which the strain is discontinuous so as to avoid unstable branches. The connection to stress-induced phase transformation follows by placing the distinct stable branches of the stress–strain curve into correspondence with distinct material phases. Avoidance of the unstable branches is then formally similar to spinodal decomposition. One outgrowth of this work has focused on putting the crystallographic theory of martensite on a rigorous mathematical foundation so as to predict microstructure without invoking the approximations inherent in the linear theory of elasticity (5–7).

The fundamental nature of much of the work cited immediately above renders it outside the scope of this modeling survey. However, under suitable interpretation, certain treatments of this type do provide a macroscopic model for the thermomechanical behavior of shape-memory materials. In particular, SMM uniaxial response follows from a thermoelastic free-energy density with either two or three minima that each define a distinct phase. Boundary value problems give solutions in which the continuum is subdivided via phase boundary interfaces into different phase regions. Phase transformation proceeds from the nucleation of new interfaces or from the propagation of the existing ones (8). The jump in the Gibbs free energy across the interface follows from the Eshelby energy-momentum tensor in the form of a generally nonzero driving traction. Phase boundary movement gives either energy dissipation or energy accumulation as determined by the direction of interface motion. Standard boundary value problems do not have a unique solution when phase boundaries are present unless the constitutive theory is augmented with both interface nucleation criteria and interface kinetic motion criteria. These typically depend on the driving traction, and they can be formulated so that the overall load–displacement–temperature relation reproduces pseudoelastic and reorientation behavior. Kinetic criteria can be derived, for example, similarly to that of Achenbach and Muller, so as to involve a probability of overcoming the energy barriers between phases on the basis of thermal fluctuation. Quasi-static and fully dynamic treatments follow for isothermal, adiabatic, and heat conducting cases. The kinetic motion criteria can be extracted from more refined theories in which the phase boundaries are regarded as transition zones exhibiting additional physical effects (9).

Micromechanics. In the models of this group each macroscopic point is put in correspondence with a representative volume element (RVE) of a multiphase material in

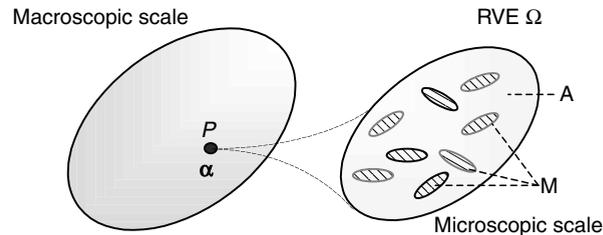


Figure 4. Macroscopic versus microscopic scale modeling. Each macroscopic point P corresponds to a microscopic region Ω where the multiphase nature of the alloy (A , M) can be appreciated explicitly. At the macroscopic scale the features of Ω are synthetically described by the internal variables α .

which some regions are subjected to local transformation strains due to the different crystal structure of the phases (Fig. 4).

Under proper boundary conditions, a boundary value problem on the RVE that models the effects of the phase transformations is obtained. Local quantities are volume averaged over the RVE, giving macroscopic quantities that generally retain a dependence on the microstructural features through some overall descriptor α (usually the phase fractions ξ). The resulting equations for the macroscopic behavior fit into the framework of internal variable models, as described later in this section. Constitutive ingredients are a macroscopic free-energy function Φ and a set of kinetic rate equations for the microstructural descriptors α . The free energy provides, via partial differentiation with respect to stress (or strain), equations for strain (or stress) while derivatives with respect to α give generalized forces that drive the phase transformations. The free-energy functions are structured as the sum of an elastic strain energy and a chemical free energy. While the chemical contribution is specified mainly by standard thermodynamic expressions, the elastic term varies considerably among such models as it follows from different micromechanics assumptions on the accommodation process due to phase interaction (10). Kinetic equations are derived from phase transformation criteria stating that the transformations occur when the generalized driving forces meet experimentally determined threshold values.

Patoor, Eberhardt, and Berveiller initiated such an approach in 1987 by combining ideas from transformation plasticity, continuum micromechanics, and crystallographic theories of martensitic transformation (11). In its recent formulation, their model considers, at the single-crystal level, a linear elastic RVE consisting of an austenite matrix with inclusions of 24 martensite variants, each exhibiting a local transformation strain computable from the lattice parameters. Each variant is assumed to grow, mixed with austenite, in a well-defined cluster. The interaction energy describing the accommodation between pairs of variants is then computed using the interfacial operator method of Hill. The minimization of this interaction energy determines the cluster orientation and the overall free energy finally depends only on the variant fractions. At the polycrystalline level, a second RVE consisting of nontextured assemblies of spherical grains is considered and a self-consistent approach is used to derive the

final macroscopic constitutive equations. Large numerical simulations involving the representation of each grain are required. Comparisons with experimental data are given, mainly for uniaxial stress, and show good agreement even in the prediction of tension–compression asymmetry. Recent developments also include the consideration of non-isothermal behavior. Promising applications of this model have been proposed by Gall and Sehitoglu (1999) who used experimentally determined grain orientation distribution functions to simulate the effect of texture.

In 1993, Sun and Hwang proposed to treat the problem focusing directly on the polycrystals with an RVE consisting of grains that are each wholly in the austenite or martensite phase. The phase interaction energy is computed using Mori-Tanaka theory by considering martensite grains as randomly dispersed spherical inclusions within the matrix of austenite grains. While martensite is considered as a single phase without explicitly accounting for the different variant orientations, neglect of the multivariant structure is overcome by proposing a direct relation between the local transformation strain and the average stress in the matrix so as to simulate the biasing effect of stress in the variant selection process. The local transformation strain is therefore not strictly crystallographic, and the resulting description is in terms of an equivalent transformation strain. Reorientation effects are taken into account by introducing a second martensite fraction. Issues related to nonproportional loading are also discussed (12).

Starting in the early 1990s, Levitas developed models from a somewhat different viewpoint (13). The SMM is modeled as a dissipative material already at the microscale where, due to the phase transformations, relevant field quantities vary between two values reflecting an underlying two-phase model for the microstructure. A first-averaging procedure is performed over an internal time scale representative of the transformation duration in order to obtain an average dissipation rate and driving force. Different energetic transformation criteria are given: an overall nucleation criterion results after integration over the RVE while a criterion for interface propagation is given after integration over the propagating interface. An extremum principle with respect to the variation of the RVE boundary conditions is invoked to determine the evolution of the microstructural parameters.

Goo and Lexcellent (14) proposed a model for single crystals based on a free-energy function and a dissipation rate function. The free-energy function is derived by a self-consistent evaluation of internal stresses among the phases. The model allows for nonisothermal behavior, reorientation, and two-way shape-memory effect. The influence of the interaction energy on the macroscopic modeling is examined, and the comparisons with experimental data under uniaxial stress show good agreement with the modeling prediction.

The analysis of Lu and Weng (15) treats each grain as a mixture of austenite and a single martensite variant whose local transformation strain is computed from lattice parameters. The particular variant is selected in analogy with the Patel-Cohen criterion on maximum transformation work. Polycrystals are then modeled by an assembly of

nontextured spherical grains, and a self-consistent method is used to compute the macroscopic response. As with the Patoor-Eberhardt-Berveiller model, this requires large numerical simulations involving the representation of each grain.

Huang and Brinson (16) propose a different microstructural model at the single-crystal level. An austenite matrix with martensite inclusions made of groups of self-accommodating variants, each exhibiting the local crystallographic transformation strain, is arranged in a way reminiscent of the experimentally observed wedgelike microstructure. Free energy is then computed by assuming a random distribution of such inclusions that are taken to be of spherical shape. This idealization is shown to be useful in modeling thermally activated transformations and low-temperature reorientation. The model captures also the tension/compression asymmetry and the different response observed experimentally when the loading direction varies with respect to crystal axes. The model has been extended to cover penny-shaped inclusions and polycrystalline behavior by studying an assembly of nontextured spherical grains homogenized by a self-consistent method.

Summarizing, micromechanic approaches incorporate several features into the modeling, including the effect of a multiple-variant microstructure and the effect of its polycrystalline texture. This permits explanations for most macroscopically observed behaviors, though certain detailed issues remain under discussion. Such issues include the determination of the number of variants and the modeling of their arrangement (17), as well as the modeling of nonproportional multi-axial loading; for recent experimental studies, see (18,19).

Approaches Modeling Directly the Macroscopic Behavior

Direct modeling will be understood as including theories where each point of the material, instead of being in an identifiably distinct phase, is representative of a phase mixture whose microstructural features are described by one or more descriptive variables. In a continuum setting, the associated strain and temperature gradient fields are continuous.

Models without Internal Variables. In such models the material behavior is described by strain, stress, temperature, and entropy without the introduction of quantities representing the phase mixture. Constitutive information is provided by a free-energy function whose partial derivatives provide constitutive equations for strain (or stress) and entropy.

In 1980, Falk proposed a Landau-Devonshire type of free-energy function based on the analogy between SMM uniaxial stress–strain curves and the electric field–magnetization curves of ferromagnetic materials (20). Nonmonotone stress–strain curves are obtained, and the unstable negative slope part is interpreted as the occurrence of the phase transition. The actual pattern followed during transformation is assumed to proceed at constant stress. The particular form of the Landau-Devonshire free-energy accounts for the temperature dependence of the

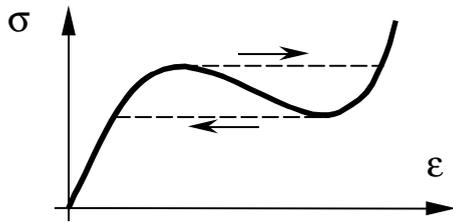


Figure 5. Nonmonotone stress–strain curves. The negative slope part is unstable and the dashed lines represent an assumed transformation path.

isothermal stress–strain behavior. Hysteresis arises as consequence of the different stress levels of the extremal points of the unstable region, as indicated in Fig. 5.

Under suitable interpretation, many aspects of this model correspond to aspects of Ericksen’s treatment and its subsequent extensions. In particular, although the time evolution of a phase transformation is not treated directly, such information can be inferred by associating the extent of the transformation with the strain distance on the constant stress transitions.

Hysteresis Models. Hysteresis models seek to reproduce experimentally observed curves that involve high nonlinearity and complex looping. They have been widely used in several fields, with that of magnetic materials being most developed. In this approach, constitutive equations are proposed directly on the basis of their mathematical properties, often without explicit focus on their link with the physical phenomena of interest. Reliability and the robustness of the model are favorably matched to experiments, and the resulting algorithm allows for the treatment of arbitrarily complex driving input.

Two main algorithm classes have received special attention in the context of SMM phase transformation. The first one is based on tracking subdomain conversion/reversion and lead to integral based algorithms. The most common of these is known as the Preisach algorithm and it has been used to describe uniaxial isothermal pseudoelastic stress–strain SMM response (21,22).

The second algorithm class involves differential equations with separate forms for driving input increase and driving input decrease. Differential equations of Duhem-Madelung form have been used to model SMM phase fraction evolution during thermally induced transformation (23,24). This gives phase fraction subloops for temperature histories obeying $M_f < T(t) < A_f$. Under sustained thermal cycling, these subloops collapse onto a final limiting subloop, with the resulting shakedown behavior registering the fading influence of the initial phase-fraction state. By being formulated so as to link the internal variable of phase fraction to the driving force variable of temperature, such algorithms lend themselves to a wider internal variable framework as is described next.

Models with Internal Variables. The key feature of this approach is to introduce one or more *internal variables*

(*order parameters*) α describing the internal structure of the material (see again Fig. 4). A general thermodynamical treatment then proceeds by providing equations for the evolution of these internal variables (23,26). The first application of such an approach to SMM seems to be due to Tanaka and Nagaki (27) where internal variables are employed to describe the development of the underlying phase mixture. The internal variables α , along with a set of mechanical and thermal *control variables*, then define a collection of *state variables* \mathbf{s} . Typical mechanical control variables are either strain ε or stress σ . Typical thermal control variables are either temperature T or entropy η . The internal variables α typically include one or more phase fractions ξ and/or macroscopic transformation strains. For example,

$$\mathbf{s} = \{\sigma, T, \alpha\}. \quad (1)$$

The temperature gradient ∇T must also be incorporated into \mathbf{s} if heat conduction is considered. The theory is composed of the *physical laws*, the *constitutive equations* that characterize the features typical of each material, and *material behavior requirements* that ensure thermodynamical process restrictions.

Constitutive information is specified by two kinds of relations:

1. *State equations* for the entities that are conjugate to the control variables. These can be formulated directly or else obtained as partial derivatives of a suitable free energy function after enforcing the Clausius-Duhem inequality for every process. The Gibbs free-energy G is appropriate for Eq. (1) and gives

$$\begin{cases} \varepsilon = \widehat{\varepsilon}(\mathbf{s}) \\ \eta = \widehat{\eta}(\mathbf{s}) \end{cases} \quad \text{with} \quad \begin{cases} \varepsilon = -\frac{\partial G(\mathbf{s})}{\partial \sigma} \\ \eta = -\frac{\partial G(\mathbf{s})}{\partial T} \end{cases}. \quad (2)$$

If heat conduction is to be included, then a constitutive equation relating temperature gradient ∇T and heat flux \mathbf{q} (usually the Fourier equation) is also required. The relation between microscale phenomena and the structure of the macroscale free-energy functions is discussed in (10).

2. A set of *kinetic equations* for the internal variables α . In view of phase transformation hysteresis, these equations generally depend on the past history of the material. Standard practice in most internal variable models is to specify this dependence through equations relating the rates of the internal variables to the state and its time derivatives. The internal state then follows from the solution of differential equations in time:

$$\widehat{\mathbf{F}}(\mathbf{s}, \dot{\mathbf{s}}) = \mathbf{0} \quad \text{typically giving } \dot{\alpha} = \widehat{\mathbf{f}}(\mathbf{s}, \dot{\sigma}, \dot{T}), \quad (3)$$

and often linear in $\dot{\sigma}$ and \dot{T} as well. The superposed dot in Eq. (3) denotes time differentiation.

The full thermomechanical behavior of a system involving SMM is then described by a complete “initial value problem”: *given an initial state, an initial time, and an arbitrary loading history, predict the state reached by the system at subsequent times.* This initial value problem is governed by the preceding constitutive equations together with appropriate initial and boundary conditions and with the physical laws of:

Energy conservation (first law of thermodynamics)

$$\dot{e} = \sigma \cdot \dot{\varepsilon} + \dot{Q}, \quad (4)$$

Conservation of linear momentum (equation of motion)

$$\text{div } \sigma = \rho \ddot{\mathbf{u}} - \mathbf{b}, \quad (5)$$

Entropy balance and Clausius-Duhem inequality (second law of thermodynamics)

$$\Gamma = \dot{\eta} - \frac{\dot{Q}}{T} \quad \text{and} \quad \Gamma \geq 0, \quad (6)$$

where e is the internal energy, ρ the mass density, \dot{Q} the rate of heat exchange with the environment (positive if absorbed by the SMM), \mathbf{u} the displacement, \mathbf{b} the body force density, and Γ the internal entropy density production rate. The Gibbs free energy is $G = e - T\eta - \sigma \cdot \varepsilon$.

Although sometimes employing formalisms that are quite different, several models fitting into this basic framework have been proposed to describe SMM behavior. Irrespective of their derivation, they involve a constitutive description prescribed via state equations and kinetic equations. Differences involve the choice and interpretation of the internal variables α and the form of the kinetic equations. The following survey proceeds in chronological order, using a common notation that may depart from that of the original work.

One of the first explicit macroscopic models for SMM has been given by Tanaka and his coworkers for uniaxial isothermal pseudoelasticity (28). This model for $\mathbf{A} \leftrightarrow \mathbf{M}$ transformation employs one scalar internal variable, the martensite fraction ξ_M , together with strain and temperature as control variables so that $\alpha = \{\xi_M\}$ and $\mathbf{s} = \{\varepsilon, T, \xi_M\}$. The constitutive equation for stress is given as $\sigma = L(\varepsilon - \varepsilon^*)$, where the overall transformation strain is $\varepsilon^* = \gamma^* \xi_M$. Here L is the elastic modulus and γ^* the local transformation strain, both of which are regarded as material parameters. The kinetic equation for the martensite fraction ξ_M is derived from a dissipation potential resulting in a form analogous to the exponential Koistinen-Marburger relation used in metallurgy. The kinetic equation is especially simple and so enables closed form integration to give

$$\xi_M = \begin{cases} 1 - e^{\beta_F} & (\mathbf{A} \rightarrow \mathbf{M}) \\ e^{\beta_R} & (\mathbf{M} \rightarrow \mathbf{A}) \end{cases} \quad \text{with} \quad \begin{cases} \beta_F = a_M(M_s - T) + b_M \sigma, \\ \beta_R = a_A(A_s - T) + b_A \sigma, \end{cases}$$

where a_M, b_M, a_A and b_A are material constants.

In 1987 Frémond proposed a three-dimensional model based on the following state description:

$$\mathbf{s} = \{\varepsilon, T, \xi_A, \xi_+, \xi_-, \nabla \xi_A, \nabla \xi_+, \nabla \xi_-\},$$

where ξ_A, ξ_+, ξ_- are the fractions of austenite and two martensite variants and ∇ denotes gradient (29). The free energy is the sum of the pure phase energies weighted by the respective fractions plus a term aimed to enforce the constraint $\xi_A + \xi_+ + \xi_- = 1$. The evolution equations are derived from a dissipation potential. The balance equations governing the model are derived on the basis of the principle of virtual power taking into account explicitly the contribution of the internal variables. The model performances are discussed within simplified constitutive assumptions, and show the ability of the model to capture the main qualitative features of pseudoelasticity and low temperature reorientation.

In 1988 Bondaryev and Wayman developed a multi-surface plasticity theory for SMM to account for three-dimensional isothermal pseudoelasticity and reorientation (30). This yields a three-dimensional framework that would allow for the generalization of many of the uniaxial stress models that are the major focus of this article. The macroscopic transformation strain tensor ε^* is treated as an internal variable on its own so that $\mathbf{s} = \{\sigma, T, \varepsilon^*\}$. Free energies are given for the austenite phase and for a general martensite phase in which the transformation strain is of arbitrary orientation. The free-energy difference among the phases initiates the transformation activity when resistive the thresholds are met. This defines temperature-dependent threshold surfaces in stress space, which are analogous to traditional yield surfaces. These surfaces, govern transformation strain accumulation ($\mathbf{A} \rightarrow \mathbf{M}$), reduction ($\mathbf{M} \rightarrow \mathbf{A}$), and reorientation ($\mathbf{M} \rightarrow \mathbf{M}$). The change in transformation strain proceeds according to a normality condition with respect to these threshold surfaces and so determines the orientation of a transformation strain increment. This gives:

$\mathbf{A} \rightarrow \mathbf{M}$: Transformation strain increment $d\varepsilon^*$ coaxial with the stress deviator

$\mathbf{M} \rightarrow \mathbf{A}$: Transformation strain annihilated without regard to stress orientation

$\mathbf{M} \rightarrow \mathbf{M}$: $d\varepsilon^*$ oriented according to the difference between the current stress deviator orientation and the current ε^* orientation

The magnitude of the transformation strain increment follows by analogy to traditional yield surface plasticity. Under hardening, continued plastic straining requires a sustained increase in stress. A consistency condition for remaining on the evolving yield surface then determines the magnitude of the plastic strain increment. In the absence of hardening, the yield surface is fixed so that continued plastic straining can be sustained under constant stress. The magnitude of the plastic strain increment is then determined from boundary conditions. Bondaryev-Wayman initially present $\mathbf{A} \rightarrow \mathbf{M}$ and $\mathbf{M} \rightarrow \mathbf{A}$ thresholds that do not harden (corresponding to $M_s = M_f$ for $\mathbf{A} \rightarrow \mathbf{M}$ and $A_s = A_f$

for $\mathbf{M} \rightarrow \mathbf{A}$). In contrast, the $\mathbf{M} \rightarrow \mathbf{M}$ reorientation threshold is naturally dependent on the current ε^* and so exhibits a changing form analogous to hardening. At any instant of time there may be (1) no transformation, (2) a single transformation from among $\mathbf{A} \rightarrow \mathbf{M}$, $\mathbf{M} \rightarrow \mathbf{A}$, $\mathbf{M} \rightarrow \mathbf{M}$, or (3) a multiple transformation consisting of $\mathbf{M} \rightarrow \mathbf{M}$ in concert with one of $\mathbf{M} \rightarrow \mathbf{A}$, $\mathbf{A} \rightarrow \mathbf{M}$. Case 3 requires the determination of a transformation strain incremental multiplier for each simultaneous process. Modifications are then presented for $\mathbf{A} \rightarrow \mathbf{M}$ and $\mathbf{M} \rightarrow \mathbf{A}$ threshold hardening ($M_s > M_f$ and $A_s < A_f$). The few model predictions that are presented confirm the correct qualitative features of this approach, but a detailed discussion of the full range of model predictions is not given.

Muller and his coworkers (31) proposed one of the first models for stress-strain curve sublooping when phase transformations do not go to completion. The main constitutive ingredient of the theory is an overall Helmholtz free energy of the form

$$\Phi(\varepsilon_A, \varepsilon_M, T, \xi_M) = (1 - \xi_M) \phi_A(\varepsilon_A, T) + \xi_M \phi_M(\varepsilon_M, T) + \Omega \xi_M (1 - \xi_M),$$

where ϕ_A , ϕ_M , ε_A , ε_M are the free energies and strains of the pure phases and ξ_M is the martensite fraction internal variable. The last term is a phase interaction energy, with Ω being a material parameter. The overall strain is $\varepsilon = (1 - \xi_M)\varepsilon_A + \xi_M\varepsilon_M$. Values of ξ_M , ε_A , and ε_M are determined by minimizing Φ under strain constraint. As in the models of Ericksen (4) and Falk (20), the resulting uniaxial stress-strain response is nonmonotone with the negative slope part being unstable. Constant stress lines from the maxima and minima define a stress-strain outer envelope loop associated with complete transformations. The unstable negative slope portion of the stress-strain response then provides a triggering threshold for phase transformations within the interior of the stress-strain envelope (Fig. 6). The associated model for internal sublooping determines the evolution of ξ_M and so, roughly speaking, plays a similar role to the kinetic evolution equations in other internal variable models.

Ortin and planes have developed a detailed thermodynamic framework for SMM materials (32) that provides a basis for thermomechanical modeling and the experimental determination of material characterization parameters. They develop a model uniaxial stress, describing the state as $\mathbf{s} = \{\sigma, T, \xi_M\}$. Energy balance during phase transformation generates a transformation kinetic in which the free-energy differential during transformation is balanced by the sum of an elastic energy storage rate differential and an energy dissipation rate differential. The evolution of σ and T then determines the evolution of ξ_M once the dissipation rate is given a constitutive prescription. Dissipation functions can be constructed so as to ensure known qualitative aspects of phase transformation hysteresis, including fine sublooping features. Experimental data fitting with reference to purely thermal transformation allows for explicit functional forms. Full stress-strain-temperature dependence for uniaxial $\mathbf{A} \leftrightarrow \mathbf{M}$ transformation then follows.

In 1990, Liang and Rogers proposed a modification of the Tanaka $\mathbf{A} \leftrightarrow \mathbf{M}$ transformation model so as to account

for the effect of phase fraction values at the beginning of the transformation in the event of an initial phase mixture (33). This allows for the treatment of internal subloops. They also replaced the exponential Koistinen-Marburger kinetic equation with a trigonometric expression, giving

$$\xi_M = \begin{cases} \frac{1 - \xi_0}{2} \cos \beta_F + \frac{1 + \xi_0}{2}, & (\mathbf{A} \rightarrow \mathbf{M}), \\ \frac{\xi_0}{2} \cos \beta_R + 1, & (\mathbf{M} \rightarrow \mathbf{A}), \end{cases}$$

where ξ_0 is the value of ξ_M when the transformation process is first activated and β_F, β_R are as given previously.

In 1992, Raniecki, Tanaka, and Lexcellent proposed a model for three-dimensional pseudoelasticity based on the martensite fraction as internal variable so that $\mathbf{s} = \{\varepsilon, T, \xi_M\}$ (34). They proposed a Helmholtz free-energy function in the form

$$\Phi(\mathbf{s}) = \frac{1}{2}(\varepsilon - \varepsilon^*) \cdot \mathbf{L}(\varepsilon - \varepsilon^*) - (T - T_0) \mathbf{m} \cdot \mathbf{L}(\varepsilon - \varepsilon^*) + E^{ch}(T) + \Omega(T)\xi_M(1 - \xi_M),$$

where \mathbf{m} provides thermal expansion coefficients, E^{ch} is the chemical free energy and the last term is the phase interaction energy. The overall transformation strain tensor is given as $\varepsilon^* = \gamma^* \xi_M$. The local transformation strain tensor γ^* is assumed to be traceless and coaxial with the strain deviator. In this context the phase equilibrium corresponds to the vanishing of the quantity $\Pi = \partial \Phi / \partial \xi_M$, which is identified as the driving force for $\mathbf{A} \leftrightarrow \mathbf{M}$ phase transformation. The condition $\Pi = 0$ gives rise to a stress-strain curve with a descending branch which, as in Muller's model, triggers the activation of certain transformations (Fig. 6). Here, however, the transformation evolution is described by a kinetic equation for ξ_M . A generalized expression for such a kinetic equation is proposed and includes the exponential form of Tanaka as a special case. Subsequent development of the model includes the incorporation of micromechanical considerations into the derivation of the free energy and the proposal of a modified relation between γ^* and the state

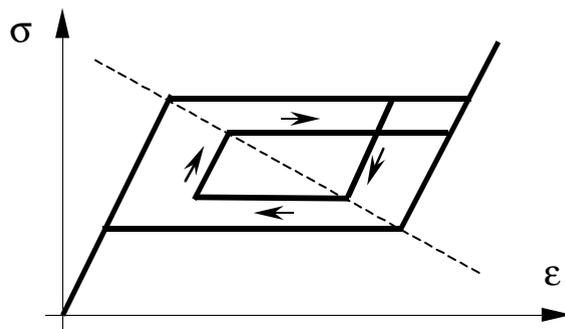


Figure 6. Modeling of subloops in a model of Muller (31). The upper and lower constant stress lines arise as in Fig. 5 and bound an internal region in stress-strain space where subloops can occur. If a transformation associated with these bounding lines is interrupted before completion due to load reversal, then the stress-strain path enters the internal region. The reverse transformation is only activated if the internal path encounters the negatively sloped line associated with unstable stress-strain response.

variables so as to capture an optimal variant arrangement on the assumption that the actual value of γ^* minimizes the free energy.

In 1993, Brinson proposed a further extension of the Tanaka-Liang-Rogers model in order to distinguish between stress-induced (oriented) martensite and thermally induced (unbiased) martensite (35). Accordingly, $\xi_M = \xi_s + \xi_T$ where ξ_s is the fraction of stress-induced martensite resulting in a local transformation strain γ^* and ξ_T is the fraction of thermally induced self-accommodated martensite resulting in a zero transformation strain. Hence $\alpha = \{\xi_s, \xi_T\}$, and the resulting model gives a true shape-memory capability, formally absent in the earlier Tanaka-Liang-Rogers model, in that transformation strain at zero stress is annihilated on heating without reappearing on cooling. Kinetic equations for ξ_s and ξ_T then model both $\mathbf{A} \leftrightarrow \mathbf{M}$ pseudoelasticity and low-temperature conversion of thermally induced martensite to oriented martensite under isothermal uniaxial stress. Subsequent development explicitly correlates the kinetic equations to the (σ, T) -phase diagram, with the resulting model permitting both analytical and numerical treatment of initial value problems for uniaxial response (36).

In 1994, Ivshin and Pence proposed a model for uniaxial pseudoelasticity based on $\mathbf{s} = \{\sigma, T, \xi_A\}$, where $\xi_A = 1 - \xi_M$ is the austenite phase fraction (24,37). State equations are given for strain and entropy by assuming that the pure phases are subject to a common overall stress. Kinetic equations are given in terms of the Duhem-Madelung hysteresis algorithm. Both stress and temperature-induced $\mathbf{A} \leftrightarrow \mathbf{M}$ transformation are treated in a unifying thermodynamic framework by the identification of the proper driving input via the Clausius-Clapeyron equation. This allows for a straightforward treatment of arbitrary thermomechanical loading paths in (σT) -space under uniaxial tension. Load-cycling behavior and the resulting shakedown response are then easily determined. Unlike earlier treatments, rate effects due to heating and cooling intrinsic in the $\mathbf{A} \leftrightarrow \mathbf{M}$ transformation are systematically investigated. To treat uniaxial compression in the same framework, Wu and Pence (38) decompose the martensite phase into two martensite variant families \mathbf{M}_+ and \mathbf{M}_- each characterized by its own transformation strain and, respectively, favored under tension (\mathbf{M}_+) or compression (\mathbf{M}_-). Accordingly, $\xi_M = \xi_+ + \xi_-$, leading to $\mathbf{s} = \{\sigma, T, \xi_+, \xi_-\}$. In this setting, unbiased (thermally induced) martensite is the specific mixture of the two variant families giving zero overall transformation strain. While maintaining all of the features of the Ivshin-Pence formulation, the Wu-Pence model provides a complete uniaxial description for SMM not only at temperatures near A_f but also at temperatures well below M_f where reorientation applies.

In Lubliner and Auricchio (39), a three-dimensional model for isothermal pseudoelasticity is proposed on the basis of $\mathbf{s} = \{\sigma, T, \xi_M, \mathbf{v}\}$, where \mathbf{v} defines the orientation of ε^* via $\varepsilon^* = \gamma^* \mathbf{v}$ and γ^* is a material parameter. Kinetic equations for \mathbf{v} and ξ_M are given by normality conditions to proper loading functions specified for each type of transformation. A Drucker-Prager form is taken for \mathbf{v} so as to treat the dependence of phase transformation on hydrostatic stress. Pseudoelasticity and high-temperature

reorientation as induced by nonproportional loads are considered. In the uniaxial setting, the Brinson decomposition $\xi_M = \xi_s + \xi_T$ is introduced to account for low-temperature reorientation. The extension to finite kinematics is developed. Numerical implementation is given in the context of finite elements.

Boyd and Lagoudas (40) present a three-dimensional model for SMM behavior based on $\mathbf{s} = \{\sigma, T, \xi_M, \varepsilon^*\}$. A free energy with a structure similar to that of Muller is generalized so that the phase interaction term is specified by a series expansion whose coefficients are left to experimental identification. The macroscopic transformation strain rate is decomposed into the sum of a pseudoelastic and a reorientation contribution, $\dot{\varepsilon}^* = \dot{\varepsilon}^{pe} + \dot{\varepsilon}^{re}$. Similar to Tanaka et al. (28), dissipation potentials are used to derive kinetic equations for $\dot{\xi}_M, \dot{\varepsilon}^{re}$, while $\dot{\varepsilon}^{pe}$ is related to the martensite fraction rate via $\dot{\varepsilon}^{pe} = \Lambda \dot{\xi}_M$. The orientation tensor Λ is assumed as coaxial with a modified stress deviator for $\mathbf{A} \rightarrow \mathbf{M}$ transformation, whereas it is aligned with the ε^* deviator for $\mathbf{M} \rightarrow \mathbf{A}$ transformation. Various aspects of the model, such as sublooping, connection with micromechanics, and numerical implementation, have been developed in subsequent papers.

Rajagopal and Srinivasa (41) apply the concept of materials with multiple natural configurations to uniaxial pseudoelasticity with the martensite fraction ξ_M acting as an internal variable. The Green-Naghdi approach to balance and constitutive equations is used together with a principle of maximum dissipation. This elegant treatment accounts for nonisothermal behavior and finite deformations within a rigorous and innovative framework of continuum thermomechanics.

Sittner, Stalmans, and Tokuda (18) have developed a hysteretic model for uniaxial pseudoelasticity based on a martensite fraction that is governed by an evolution equation with a driving force that involves the concept of an effective equilibrium temperature. This effective equilibrium temperature, which generally differs from the actual material temperature, is formally dependent on the martensite volume fraction. Unlike other models, the macroscopic transformation strain is assumed in the form $\varepsilon^* = \chi \gamma^* \xi_M$, where $\chi = \chi(\sigma)$ is a nonlinear stress-dependent function. The model is able to capture a wide variety of sublooping behavior including a notion of return point memory.

Within the setting of finite kinematics Govindjee and Hall (42) present a model based on two martensite variants \mathbf{M}_+ and \mathbf{M}_- . A phase diagram approach is used, and the transformation kinetics are derived by arguments reminiscent to those of Muller-Achenbach and Abeyaratne-Knowles. The resulting model allows for both pseudoelasticity and reorientation. Algorithmic issues specific to finite element implementation are carefully considered, and civil engineering scale applications are presented.

A COMPREHENSIVE MODEL FOR UNIAXIAL STRESS

As seen from the previous section, SMM uniaxial behavior is well understood. This covers a wide range of applications including several kinds of actuators, vibration absorbers,

and applications exploiting the material in the form of wires. Models available for this setting treat reorientation, shape memory, and pseudoelasticity under arbitrary temperature-stress paths so as to reproduce nonisothermal behavior, transformational heating/cooling, and internal subloops. Accordingly, this section summarizes a practical and complete one stress-component material model. It synthesizes aspects of the previous section's discussion with a focus on the complementary roles of state equations and kinetic equations in generating a well-posed and complete model. The development is framed in the context of tensile and compressive loading, although it also applies to other choices of stress component, such as a particular shear stress component. The equations that are presented in the final implementation follow those of Ivshin and Pence (24,37) and Wu and Pence (38), although the discussion is framed so as to permit alternative implementations.

The SMM is treated as subject to *mechanical loads* specified by histories of prescribed stress $\sigma(t)$ or strain $\varepsilon(t)$ or a combination of both (e.g., representing bias springs in actuators among elastic restraints) and *thermal loads* specified by histories of prescribed temperature or heat rate. Alternatively, under mechanical loads, the temperature can be determined as a consequence of the heat exchange with a known environmental temperature T_E .

The state in Eq. (1) is given by $\mathbf{s} = \{\sigma, T, \alpha\}$ with $\alpha = \{\xi_A, \xi_+, \xi_-\}$ subject to

$$\xi_A + \xi_+ + \xi_- = 1, \quad 0 \leq \xi_A \leq 1, \quad 0 \leq \xi_+ \leq 1, \quad 0 \leq \xi_- \leq 1. \quad (7)$$

The \mathbf{M}_+ family has an associated scalar transformation strain $\gamma_+^* > 0$, representing the maximum positive macroscopic transformation strain when an \mathbf{M}_+ microstructure is maximally oriented with the tensile stress. Similarly, the \mathbf{M}_- family has an associated scalar transformation strain $-\gamma_-^* < 0$. The positive constants γ_+^*, γ_-^* are basic material parameters.

State Equations for Strain and Entropy

The first group of constitutive equations are obtained from a Gibbs free energy that is taken as

$$G(s) = \xi_A g_A(\sigma_A, T) + \xi_+ g_+(\sigma_+, T) + \xi_- g_-(\sigma_-, T), \quad (8)$$

where g_A, g_+, g_- and $\sigma_A, \sigma_+, \sigma_-$ are the free energies and the stresses relative to the pure phase. The phase stresses depend on the microstructural phase arrangement. This arrangement is henceforth regarded as giving

$$\sigma_A = \sigma_+ = \sigma_- = \sigma. \quad (9)$$

In conditions different from Eq. (9), an additional term expressing the interaction energy between the phases would otherwise arise (10). In the present model, hysteresis properties that would be influenced by such an interaction energy are instead modeled with the aid of envelope functions introduced below. This permits easy specification of desired thermal hysteresis properties.

Standard forms for the pure phase free energies are

$$\begin{aligned} g_A &= -\frac{1}{2} \frac{\sigma^2}{E_A} - m_A \sigma (T - T^*) + C_A \left(T - T^* + T \ln \frac{T}{T^*} \right) \\ &\quad - \eta_{A0} (T - T^*) + g_{A0}, \\ g_+ &= -\frac{1}{2} \frac{\sigma^2}{E_+} - \gamma_+^* \sigma - m_+ \sigma (T - T^*) + C_+ \left(T - T^* + T \ln \frac{T}{T^*} \right) \\ &\quad - \eta_{+0} (T - T^*) + g_{+0}, \\ g_- &= -\frac{1}{2} \frac{\sigma^2}{E_-} + \gamma_-^* \sigma - m_- \sigma (T - T^*) + C_- \left(T - T^* + T \ln \frac{T}{T^*} \right) \\ &\quad - \eta_{-0} (T - T^*) + g_{-0}. \end{aligned} \quad (10)$$

Here E_A, E_+, E_- are the elastic moduli; m_A, m_+, m_- are the coefficients of thermal expansion, and C_A, C_+, C_- are the specific heats at constant stress of the various phases. The expressions (10) make use of a stress-temperature reference state $(\sigma, T) = (0, T^*)$, which is regarded as at the center of the multiphase region of Fig. 1 by taking T^* as the average of the four transformation temperatures. The constants $\eta_{A0}, \eta_{+0}, \eta_{-0}$ and g_{A0}, g_{+0}, g_{-0} are the single-phase entropies and single-phase free energies at this reference state. Additional simplification follows by assuming the following:

- A common specific heat C in martensite and austenite.
- A common reference state entropy η_{M0} in all martensite variants.
- Negligible thermal expansion.
- Negligible slip plasticity.

None of these simplifications is essential, and the associated generalizations are easily made. According to the general formulation, the macroscopic strain and entropy follow from Eq. (2) as

$$\varepsilon = -\frac{\partial G}{\partial \sigma} = \bar{D}\sigma + \varepsilon^*, \quad \eta = -\frac{\partial G}{\partial T} = C \ln \frac{T}{T^*} + \bar{\eta}_0, \quad (11)$$

where

$$\begin{aligned} \bar{D}(\xi) &= \frac{\xi_-}{E_-} + \frac{\xi_A}{E_A} + \frac{\xi_+}{E_+}, \quad \varepsilon^*(\xi) = \xi_+ \gamma_+^* - \xi_- \gamma_-^* \\ \bar{\eta}_0(\xi) &= (\xi_+ + \xi_-) \eta_{M0} + \xi_A \eta_{A0}. \end{aligned} \quad (12)$$

Here $\varepsilon^*(\xi)$ is again the transformation strain, and $\bar{D}(\xi)$ is the Reuss estimate of the effective compliance. Introduce $\Delta \eta_0 = \eta_{A0} - \eta_{M0}$, whereupon $\bar{\eta}_0(\xi) = \eta_{M0} + \xi_A \Delta \eta_0$. Other than the baseline value of η in Eq. (11), the model dependence on η_{M0} and η_{A0} is only via the basic material parameter $\Delta \eta_0 > 0$. It is given by $\Delta \eta_0 = \Delta H / T^*$ where ΔH is the latent heat of the $\mathbf{M} \rightarrow \mathbf{A}$ transformation as measured from calorimetry.

These considerations account for the following essential material parameters: $M_f, M_s, A_s, A_f, \Delta \eta_0$ (or ΔH), $E_A, E_+, E_-, \gamma_+^*, \gamma_-^*, C$. These eleven parameters are sufficient for modeling the basic SMM behavior with the exception of $\mathbf{M} \leftrightarrow \mathbf{M}_+$ reorientation. As will be explained in what follows, reorientation is modeled by including four

additional material parameters $\sigma_s^+, \sigma_f^+, \sigma_s^-, \sigma_f^-$. All 15 material parameters are taken to be positive by definition.

Phase Transformation Kinetics

The second group of constitutive equations give the phase transformation kinetics. The model derives such equations from two constitutive ingredients that are, in principle, both experimentally measurable:

- The *phase diagram* that defines the loci of the points in the stress-temperature plane in which the various phase transformations can be activated (an example of which is reported in Fig. 1).
- The *envelope functions* $\zeta_{M \rightarrow A}(T)$ and $\zeta_{A \rightarrow M}(T)$ that, as explained below, determine the equations for the internal variables thus describing how the phase transformations evolve once activated.

Here, $\zeta_{A \rightarrow M}(T)$ is the value of ξ_A associated with $\sigma = 0$ and T decreasing from above A_f to below M_f . Similarly, $\zeta_{M \rightarrow A}(T)$ is the value of ξ_A associated with $\sigma = 0$ and T increasing from below M_f to above A_f . Each function has a graph that monotonically increases from zero to one as T increases over an appropriate interval: $M_f < T < M_s$ for $\zeta_{A \rightarrow M}(T)$; $A_s < T < A_f$ for $\zeta_{M \rightarrow A}(T)$. In the absence of detailed experimental data, these functions can be approximated on their transition interval. The simplest approximation is a linear function of T , but this gives a slope change at the associated start and finish temperatures. Such slope changes give rise to sharp corners on the model stress-strain curves at the beginning and end of the transformation plateaus. Smoother forms, which eliminate such corners, involve hyperbolic or trigonometric functions. A simple and useful representation is the Liang-Rogers trigonometric form:

$$\begin{aligned} \zeta_{M \rightarrow A}(T) &= \frac{1}{2} \left\{ 1 - \cos \left(\left(\frac{T - A_s}{A_f - A_s} \right) \pi \right) \right\}, \\ \zeta_{A \rightarrow M}(T) &= \frac{1}{2} \left\{ 1 - \cos \left(\left(\frac{T - M_f}{M_s - M_f} \right) \pi \right) \right\} \end{aligned} \quad (13)$$

on the transition intervals with

$$\zeta_{M \rightarrow A} = \begin{cases} 0 & \text{if } T \leq A_s, \\ 1 & \text{if } T \geq A_f, \end{cases} \quad \zeta_{A \rightarrow M} = \begin{cases} 0 & \text{if } T \leq M_f, \\ 1 & \text{if } T \geq M_s. \end{cases}$$

For the sake of clarity of exposition, it is useful to discuss the features of the phase transformation kinetics with reference first to purely thermal transformation and then to combined stress and thermally induced transformation.

Purely Thermal Transformation. In this case the driving input is the temperature history $T(t)$. Times t of temperature reversal are *switching instants*. The martensite is then unbiased in that it involves \mathbf{M}_+ and \mathbf{M}_- in the ratio γ_+^* to γ_-^* so as to give $\varepsilon^* = 0$. Thus $\xi_+ = \lambda_+ \xi_M$ and $\xi_- = \lambda_- \xi_M$, where the constants λ_+ and λ_- are given by

$$\lambda_+ = \frac{\gamma_-^*}{\gamma_+^* + \gamma_-^*}, \quad \lambda_- = \frac{\gamma_+^*}{\gamma_+^* + \gamma_-^*}, \quad (14)$$

and $\xi_M = \xi_+ + \xi_- = 1 - \xi_A$ is the overall martensite fraction. The transformation kinetic reduces to the determination of ξ_A as T changes. The graph of ξ_A versus T will generally involve complicated sublooping if there are numerous switching instants obeying $M_f < T < A_f$ that prevent the $\mathbf{M} \rightarrow \mathbf{A}$ and $\mathbf{A} \rightarrow \mathbf{M}$ transformations from going to completion. However, no matter how complicated, this graph will be contained between the two envelope functions so that $\zeta_{M \rightarrow A}(T) \leq \xi_A \leq \zeta_{A \rightarrow M}(T)$.

The phase transformation kinetic is now proposed by means of a Duhem-Madelung (D-M) differential equation of the form:

$$\begin{aligned} \dot{\xi}_M &= H_{A \rightarrow M}(\xi_A, T) \dot{T} \quad \text{if } \dot{T} \leq 0 \quad (\mathbf{A} \rightarrow \mathbf{M}), \\ \dot{\xi}_M &= H_{M \rightarrow A}(\xi_M, T) \dot{T} \quad \text{if } \dot{T} \geq 0 \quad (\mathbf{M} \rightarrow \mathbf{A}). \end{aligned} \quad (15)$$

These equations, in conjunction with the conservation condition $\xi_A + \xi_M = 1$, determines ξ_A, ξ_M from the prevailing conditions at the last switching instant. Each D-M equation in (15) can be given a time independent form, such as $d\xi_M/dT = H_{A \rightarrow M}(\xi_A, T)$ for $\mathbf{A} \rightarrow \mathbf{M}$. The centerpieces of Eq. (15) are the *governing functions* $H_{M \rightarrow A}$ and $H_{A \rightarrow M}$, which must be formulated so that the phase fraction obeys the envelope containment condition $\zeta_{M \rightarrow A}(T) \leq \xi_A \leq \zeta_{A \rightarrow M}(T)$. Envelope coincidence must occur for transformations that begin from a pure \mathbf{M} or a pure \mathbf{A} state so long as temperature reversal is avoided. The following general form for the governing functions ensures these properties:

$$\begin{aligned} H_{A \rightarrow M}(\xi_A, T) &= - \left(\frac{\xi_A}{\zeta_{A \rightarrow M}(T)} \right)^n \frac{d\zeta_{A \rightarrow M}(T)}{dT}, \\ H_{M \rightarrow A}(\xi_M, T) &= - \left(\frac{\xi_M}{1 - \zeta_{M \rightarrow A}(T)} \right)^n \frac{d\zeta_{M \rightarrow A}(T)}{dT}, \end{aligned} \quad (16)$$

where $n > 0$ is a constant. Larger values of n give flatter internal subloops. The choice $n = 1$ gives a standard D-M model for shape-memory materials (24). Experimental magnetic susceptibility measurements (43) indicate that enhanced experimental correlation is obtained for n near $n = 3$. Further refinements and modifications can be invoked, and certain sublooping situations that are difficult to describe with a D-M equation have been noted (44). Some examples of (T, ξ_A) -trajectories within the bounding envelope functions are shown in Fig. 7.

The purely thermal process involves simultaneous $\mathbf{A} \rightarrow \mathbf{M}_+$ and $\mathbf{A} \rightarrow \mathbf{M}_-$ transformation for temperature decrease, and simultaneous $\mathbf{M}_+ \rightarrow \mathbf{A}$ and $\mathbf{M}_- \rightarrow \mathbf{A}$ transformation for temperature increase, always with $\xi_+ = \lambda_+ \xi_M$ and $\xi_- = \lambda_- \xi_M$. It follows that ξ_M can be eliminated from Eq. (15) by rewriting them as

$$\begin{aligned} \dot{\xi}_+ &= \lambda_+ H_{A \rightarrow M}(\xi_A, T) \dot{T} \quad \text{if } \dot{T} \leq 0 \quad (\mathbf{A} \rightarrow \mathbf{M}_+), \\ \dot{\xi}_+ &= \lambda_+ H_{M \rightarrow A} \left(\frac{\xi_+}{\lambda_+}, T \right) \dot{T} \quad \text{if } \dot{T} \geq 0 \quad (\mathbf{M}_+ \rightarrow \mathbf{A}), \end{aligned} \quad (17)$$

and

$$\begin{aligned} \dot{\xi}_- &= \lambda_- H_{A \rightarrow M}(\xi_A, T) \dot{T} \quad \text{if } \dot{T} \leq 0 \quad (\mathbf{A} \rightarrow \mathbf{M}_-), \\ \dot{\xi}_- &= \lambda_- H_{M \rightarrow A} \left(\frac{\xi_-}{\lambda_-}, T \right) \dot{T} \quad \text{if } \dot{T} \geq 0 \quad (\mathbf{M}_- \rightarrow \mathbf{A}). \end{aligned} \quad (18)$$

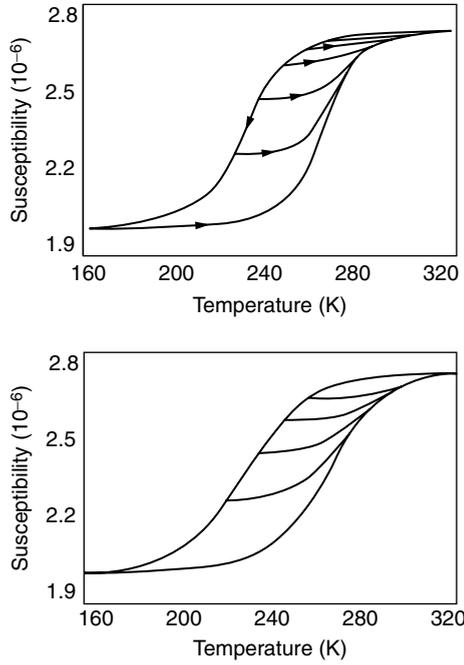


Figure 7. Magnetic susceptibility versus T under purely thermal loading for a NiTi thin film as determined by measurement (*above*). Magnetic susceptibility correlates directly with austenite phase fraction. Both the outer hysteresis loops for cooling (upper envelope) and heating (lower envelope) are shown. The internal curves leave the upper envelope if T is increased before $\mathbf{A} \rightarrow \mathbf{M}$ transformation is complete. This behavior is modeled (*below*) on the basis of Eq. (15), with a hyperbolic tangent envelope form and $H_{M \rightarrow A}$ governing function with $n = 3$.

These equations, in conjunction with Eq. (7), determine the evolution of the phase fraction variables ξ_A, ξ_+, ξ_- for purely thermal transformation. Although unwieldy compared to Eq. (15), the alternative formulations (17) and (18) recast the phase transformation kinetic in terms of the internal variables $\alpha = \{\xi_A, \xi_+, \xi_-\}$.

Pseudoelasticity. While temperature is the driving input in the purely thermal case, under simultaneous change in temperature and stress, the functions

$$\begin{aligned} \Pi_{+A}(\sigma, T) &:= \frac{\partial G}{\partial \xi_+} - \frac{\partial G}{\partial \xi_A} = -\gamma_+^* \sigma + \Delta \eta_0 (T - T^*) \\ &\quad - \frac{1}{2} \left(\frac{1}{E_+} - \frac{1}{E_A} \right) \sigma^2, \\ \Pi_{-+}(\sigma, T) &:= \frac{\partial G}{\partial \xi_-} - \frac{\partial G}{\partial \xi_+} = (\gamma_+^* + \gamma_-^*) \sigma - \frac{1}{2} \left(\frac{1}{E_-} - \frac{1}{E_+} \right) \sigma^2, \\ \Pi_{-A}(\sigma, T) &:= \frac{\partial G}{\partial \xi_-} - \frac{\partial G}{\partial \xi_A} = \gamma_-^* \sigma + \Delta \eta_0 (T - T^*) \\ &\quad - \frac{1}{2} \left(\frac{1}{E_-} - \frac{1}{E_A} \right) \sigma^2, \end{aligned} \quad (19)$$

are the more generalized driving forces for the respective transformations $\mathbf{M}_+ \leftrightarrow \mathbf{A}$, $\mathbf{M}_- \leftrightarrow \mathbf{M}_+$, and $\mathbf{M}_- \leftrightarrow \mathbf{A}$. Changes in (σ, T) that cause Π_{+A} , Π_{-+} , and Π_{-A} to increase favor $\mathbf{M}_+ \rightarrow \mathbf{A}$, $\mathbf{M}_- \rightarrow \mathbf{M}_+$, and $\mathbf{M}_- \rightarrow \mathbf{A}$, respectively.

Conversely, the decrease favors transformation in the opposite direction. The associated transformation is activated only if the current value of (σ, T) is also within the corresponding transformation zone of the (σ, T) -phase diagram (e.g., see Fig. 1).

The $\mathbf{M}_- \leftrightarrow \mathbf{M}_+$ transformation zone does not extend into $T \geq A_f$, in which case only transformations $\mathbf{M}_+ \leftrightarrow \mathbf{A}$ and $\mathbf{M}_- \leftrightarrow \mathbf{A}$ can occur. Accordingly, for $T \geq A_f$ attention is focused on Π_{+A} and Π_{-A} which are renormalized as follows:

$$\begin{aligned} \beta^+(\sigma, T) &:= \frac{\Pi_{+A}(\sigma, T)}{\Delta \eta_0} + T^* = T - k_1^+ \sigma - k_2^+ \sigma^2, \\ \beta^-(\sigma, T) &:= \frac{\Pi_{-A}(\sigma, T)}{\Delta \eta_0} + T^* = T + k_1^- \sigma - k_2^- \sigma^2, \end{aligned} \quad (20)$$

where $k_1^+ = \gamma_+^* / \Delta \eta_0 > 0$, $k_1^- = \gamma_-^* / \Delta \eta_0 > 0$, $k_2^+ = (E_A - E_+) / (2E_A E_+ \Delta \eta_0) \geq 0$, and $k_2^- = (E_A - E_-) / (2E_A E_- \Delta \eta_0) \geq 0$. The inequalities follow from $\Delta \eta_0 > 0$, $\gamma_+^* > 0$, $\gamma_-^* > 0$, $E_A \geq E_+$, and $E_A \geq E_-$. Under this renormalization $\beta^+(0, T) = T$. For $\sigma \neq 0$, the function $\beta^+(\sigma, T)$ behaves like a “stress-adjusted temperature” that governs the $\mathbf{M}_+ \leftrightarrow \mathbf{A}$ transformation for general changes in (σ, T) . The function $\beta^-(\sigma, T)$ plays a corresponding role with respect to $\mathbf{M}_- \leftrightarrow \mathbf{A}$ transformation.

Extending the purely thermal algorithm with driving input $T(t)$ to the case where the driving input is given by $\sigma(t)$ in conjunction with $T(t) \geq A_f$ amounts to replacing $T(t)$ in Eq. (17) with $\beta^+(\sigma(t), T(t))$ and to replacing $T(t)$ in Eq. (18) with $\beta^-(\sigma(t), T(t))$. This gives

$$\begin{aligned} \dot{\xi}_+ &= \lambda_+ H_{A \rightarrow M}(\xi_A, \beta^+(\sigma, T)) \dot{\beta}^+ \quad \text{if } \dot{\beta}^+ \leq 0, \\ \dot{\xi}_+ &= \lambda_+ H_{M \rightarrow A} \left(\frac{\xi_+}{\lambda_+}, \beta^+(\sigma, T) \right) \dot{\beta}^+ \quad \text{if } \dot{\beta}^+ \geq 0, \end{aligned} \quad (21)$$

and

$$\begin{aligned} \dot{\xi}_- &= \lambda_- H_{A \rightarrow M}(\xi_A, \beta^-(\sigma, T)) \dot{\beta}^- \quad \text{if } \dot{\beta}^- \leq 0, \\ \dot{\xi}_- &= \lambda_- H_{M \rightarrow A} \left(\frac{\xi_-}{\lambda_-}, \beta^-(\sigma, T) \right) \dot{\beta}^- \quad \text{if } \dot{\beta}^- \geq 0. \end{aligned} \quad (22)$$

The switching instants are now times when either $\dot{\beta}^+$ or $\dot{\beta}^-$ changes sign. Note, for example, that

$$\dot{\beta}^+ = \frac{\partial \beta^+}{\partial \sigma} \dot{\sigma} + \frac{\partial \beta^+}{\partial T} \dot{T} = (-k_1^+ - 2k_2^+ \sigma) \dot{\sigma} + \dot{T}.$$

Kinetic Eqs. (21), (22), with the balance condition Eq. (7), determine the evolution of the phase fraction variables ξ_A, ξ_+, ξ_- . According to the structure assumed for $H_{A \rightarrow M}$, $H_{M \rightarrow A}$ and $\zeta_{A \rightarrow M}$, $\zeta_{M \rightarrow A}$, it follows that the $\mathbf{A} \rightarrow \mathbf{M}_+$ and $\mathbf{M}_+ \rightarrow \mathbf{A}$ transformations are activated and completed for those values of (σ, T) such that $\beta^+ = M_s$, $\beta^+ = M_f$, $\beta^+ = A_s$, $\beta^+ = A_f$, respectively. Along with Eq. (20), they define the boundaries of the relevant zones of the (σ, T) phase diagram. A similar remark applies to the transformations $\mathbf{A} \rightarrow \mathbf{M}_-$ and $\mathbf{M}_- \rightarrow \mathbf{A}$ vis-à-vis the function $\beta^-(\sigma, T)$. The strain and entropy follow from Eqs. (11) and (12). The resulting model for $T \geq A_f$ faithfully predicts general

tension/compression asymmetry, pseudoelastic transformation, and pseudoelastic sublooping.

Low-Temperature Reorientation. If the temperature is below A_f , then two additional phenomena have to be taken into account: The $\mathbf{M}_- \leftrightarrow \mathbf{M}_+$ reorientation transformation and the possibility of multiple transformations. An example of the latter is provided in the purely thermal case when $\mathbf{A} \leftrightarrow \mathbf{M}_-$ and $\mathbf{A} \leftrightarrow \mathbf{M}_+$ take place simultaneously. The model can be extended to cover such situations provided a phase diagram is available that describes the activation and the completion of all possible transformations. Equations (21), (22) can then model the additional transformation possibilities provided that the functions $\beta^+(\sigma, T)$ and $\beta^-(\sigma, T)$ are modified from the specification (20) so as to account for the change in zonal geometry of the (σ, T) -phase diagram when $T < A_f$. A modification that accomplishes this purpose and so establishes such a phase diagram is given by

$$\begin{aligned} \beta^+(\sigma, T) &= T - k_1^+ \sigma - k_2^+ \sigma^2 \quad \text{if } \sigma \geq 0, \\ \beta^+(\sigma, T) &= \frac{(A_f - A_s)T - (A_f k_s^+ - A_s k_f^+) \sigma}{A_f - A_s - (k_s^+ - k_f^+) \sigma} \quad \text{if both } \sigma \leq 0 \\ &\text{and } T \geq M_f - k_1^- \sigma + k_2^- \sigma^2, \\ \beta^+(\sigma, T) &= \frac{(A_f - A_s)(M_f - k_1^- \sigma + k_2^- \sigma^2) - (A_f k_s^+ - A_s k_f^+) \sigma}{A_f - A_s - (k_s^+ - k_f^+) \sigma} \\ &\text{if both } \sigma \leq 0 \text{ and } T \leq M_f - k_1^- \sigma + k_2^- \sigma^2, \quad (23) \end{aligned}$$

and

$$\begin{aligned} \beta^-(\sigma, T) &= T + k_1^- \sigma - k_2^- \sigma^2 \quad \text{if } \sigma \leq 0, \\ \beta^-(\sigma, T) &= \frac{(A_f - A_s)T + (A_f k_s^- - A_s k_f^-) \sigma}{A_f - A_s + (k_s^- - k_f^-) \sigma} \quad \text{if both } \sigma \geq 0 \\ &\text{and } T \geq M_f + k_1^+ \sigma + k_2^+ \sigma^2, \\ \beta^-(\sigma, T) &= \frac{(A_f - A_s)(M_f + k_1^+ \sigma + k_2^+ \sigma^2) + (A_f k_s^- - A_s k_f^-) \sigma}{A_f - A_s + (k_s^- - k_f^-) \sigma} \\ &\text{if both } \sigma \geq 0 \text{ and } T \leq M_f + k_1^+ \sigma + k_2^+ \sigma^2. \quad (24) \end{aligned}$$

Again $k_1^+, k_2^+, k_1^-, k_2^-$ are defined after Eq. (20). The new constants: $k_s^+, k_f^+, k_s^-, k_f^-$ are

$$\begin{aligned} k_s^+ &= \frac{A_s - M_f - k_1^- \sigma_s^- - k_2^- (\sigma_s^-)^2}{\sigma_s^-}, \\ k_f^+ &= \frac{A_f - M_f - k_1^- \sigma_f^- - k_2^- (\sigma_f^-)^2}{\sigma_f^-}, \\ k_s^- &= \frac{A_s - M_f - k_1^+ \sigma_s^+ - k_2^+ (\sigma_s^+)^2}{\sigma_s^+}, \\ k_f^- &= \frac{A_f - M_f - k_1^+ \sigma_f^+ - k_2^+ (\sigma_f^+)^2}{\sigma_f^+}. \end{aligned} \quad (25)$$

Contours of constant β^+ and constant β^- give continuous curves on the phase diagram, although each such curve will have up to two sharp corners by virtue of the abrupt

formula changes in Eqs. (23) and (24). Smooth contours more resembling those in Fig. 1 can be obtained by a more complicated redefinition of β^+, β^- .

Note that the phase diagram of Fig. 1 is defined by six continuous curves. The present treatment provides similar such curves. The four curves that continue into $T < M_f$ as approximately constant stress curves parallel to the T -axis are defined by $\beta^+ = A_s$, $\beta^+ = A_f$, $\beta^- = A_s$, and $\beta^- = A_f$. The zone $A_s < \beta^+ < A_f$ bounded by two of these curves, $\beta^+ = A_s$ and $\beta^+ = A_f$, is associated with transformations that deplete \mathbf{M}_+ . This depletion gives $\mathbf{M}_+ \rightarrow \mathbf{A}$ for $\sigma > 0$. For $\sigma < 0$, this depletion gives $\mathbf{M}_+ \rightarrow \mathbf{M}_-$ reorientation provided that T is sufficiently low. The low-temperature threshold for such pure reorientation is given by $T = M_f - k_1^- \sigma + k_2^- \sigma^2$ as specified in Eq. (23). In particular, this $\mathbf{M}_+ \rightarrow \mathbf{M}_-$ reorientation is activated in compression, beginning at $\sigma = -\sigma_s^-$ and finishing at $\sigma = -\sigma_f^-$. The portion of the zone $A_s < \beta^+ < A_f$ that is near the central region of the phase diagram is characterized by $\sigma < 0$ and $T > M_f - k_1^- \sigma + k_2^- \sigma^2$. Here \mathbf{M}_+ transforms into a mixture of \mathbf{A} and \mathbf{M}_- . Similar comments hold with respect to the zone $A_s < \beta^- < A_f$ provided that the roles of \mathbf{M}_+ and \mathbf{M}_- are interchanged. Note that neither zone is specifically associated with depletion of \mathbf{A} . The zone of \mathbf{A} depletion is bounded by the remaining two curves on the phase diagram. These two curves pass through $T = M_s$ and $T = M_f$ on the T -axis and are given by $\beta^+ = M_s$ and $\beta^+ = M_f$ on $\sigma > 0$ and by $\beta^- = M_s$ and $\beta^- = M_f$ on $\sigma < 0$. The austenite depletion is via $\mathbf{A} \rightarrow \mathbf{M}_+$ if σ is sufficiently tensile, and is via $\mathbf{A} \rightarrow \mathbf{M}_-$ if σ is sufficiently compressive. For σ near zero, the depletion of \mathbf{A} is into a mixture of \mathbf{M}_+ and \mathbf{M}_- , as in the case of thermally induced martensite.

In summary, Eqs. (7), (21), and (22) with (23) and (24) now describe general $\mathbf{A} \leftrightarrow \mathbf{M}_+$, $\mathbf{A} \leftrightarrow \mathbf{M}_-$, $\mathbf{M}_- \leftrightarrow \mathbf{M}_+$ transformation throughout the full extent of the (σ, T) -phase diagram. A more detailed discussion is given in Wu and Pence (38) for the special case of symmetric tension/compression behavior. Equations (23), (24) also give the proper tension/compression asymmetry if any of the equalities $\gamma_+^* = \gamma_-^*$, $\sigma_s^+ = \sigma_s^-$, $\sigma_f^+ = \sigma_f^-$, or $E_+ = E_-$ do not hold. As discussed above, this model gives multiple concurrent transformations in the core region of the phase diagram. Low-temperature reorientation involves passing through an unexpressed unstable austenite intermediary (45). Algorithm (15) is recovered under purely thermal transformation. At temperatures $T < A_f$, isothermal tension followed by compression generates open stress-strain curves that are sometimes referred to as ferroelastic response.

Operative Equations for Various Driving Conditions

The set of equations required for the actual computation of SMM response are summarized in the following. This set may be different depending on the conditions in which the material is used and the way that the loading input specified.

If T and σ in the material are prescribed, then the determination of ξ_A, ξ_+, ξ_- proceeds directly on the basis of Eqs. (7), (21), and (22). In this case the strain ε as provided by Eq. (11) decouples from the kinetics.

If T in the material is prescribed, but σ in the material is not, then the first of Eqs. (11) must be solved in conjunction with Eqs. (7), (21), (22). This includes cases where ε is prescribed (various hard constraint situations) or when σ is related to ε by external constraint (e.g., a bias spring).

If T in the material is not prescribed directly, then the complete thermomechanical behavior requires consideration of the material thermal balance and heat exchange with the environment. The heating rate \dot{Q} can be written

$$\dot{Q} = \dot{Q}_{\text{rev}} + \dot{Q}_{\text{irrev}} \quad (26)$$

where, in view of Eq. (6), the reversible contribution is

$$\dot{Q}_{\text{rev}} = T\dot{\eta}. \quad (27)$$

The irreversible contribution $\dot{Q}_{\text{irrev}} = -T\Gamma$ includes dissipative effects that are intrinsic with the phase transformation, and by Eqs. (2), (4), (8), (26), and (27), it can be expressed as

$$\dot{Q}_{\text{irrev}} = \frac{\partial G}{\partial \xi_A} \dot{\xi}_A + \frac{\partial G}{\partial \xi_+} \dot{\xi}_+ + \frac{\partial G}{\partial \xi_-} \dot{\xi}_-. \quad (28)$$

According to Ortin and Planes (46), \dot{Q}_{irrev} can be neglected in a first approximation. More generally,

$$C\dot{T} + T\Delta\eta_0\dot{\xi}_A + \left(\frac{\partial G}{\partial \xi_A} \dot{\xi}_A + \frac{\partial G}{\partial \xi_+} \dot{\xi}_+ + \frac{\partial G}{\partial \xi_-} \dot{\xi}_- \right) = \dot{Q}. \quad (29)$$

Here \dot{Q} provides the heat exchange with the environment and requires additional description to this effect. Important situations in which T is not prescribed include adiabatic conditions and conditions of convective heat transfer to an ambient temperature T_E . The adiabatic case is $\dot{Q} = 0$, and involves no heat exchange with the environment. It is an appropriate model under sustained and rapid mechanical loading. The case of convective heat transfer may often be described by

$$\dot{Q} = \kappa(T_E - T), \quad (30)$$

where $\kappa > 0$ is a known parameter. Here $\kappa \rightarrow 0$ gives the adiabatic limit, while $\kappa \rightarrow \infty$ enforces $T(t) = T_E(t)$. If T is in fact prescribed, then Eqs. (29) and (30) determine either the heat exchange \dot{Q} or the fluid/atmospheric media temperature T_E that is necessary to sustain the prescription. Table 1 summarizes the equations one uses for the determination of ξ_A , ξ_+ , ξ_- .

Table 1. Modeling Equation Summary

Prescribed Quantities		Equations
Stress σ	Temperature T	(7), (21), (22)
Strain ε	Temperature T	(7), (11) ₁ , (21), (22)
Stress σ	Heat rate \dot{Q}	(7), (21), (22), (29)
Strain ε	Heat rate \dot{Q}	(7), (11) ₁ , (21), (22), (29)
Stress σ	Media temperature T_E	(7), (21), (22), (29), (30)
Strain ε	Media temperature T_E	(7), (11) ₁ , (21), (22), (29), (30)

Table 2. Representative Values for Material Constants

M_f [K]	M_s [K]	A_s [K]	A_f [K]	σ_s [Mpa]	σ_f [Mpa]
235	260	295	310	120	210
γ^*	ΔH [J/m ³]	E_A [Mpa]	E_M [Mpa]	C [J/(m ³ K)]	
0.06	115*10 ⁶	40*10 ³	25*10 ³	5*10 ⁶	

Discussion

Some features of the model can be shown explicitly by the presentation of a few representative numerical simulations and experimental validations for a NiTi alloy. The results are given in this section with reference to governing functions (16), with $n = 1$ and tension/compression symmetry, which implies that $\gamma_+^* = \gamma_-^* = \gamma^*$, $\sigma_s^+ = \sigma_s^- = \sigma_s$, $\sigma_f^+ = \sigma_f^- = \sigma_f$, and $E_+ = E_- = E_M$. Under these circumstances, typical values of the full set of material parameters needed by the model are given in Table 2.

Subloops. The subloop model provides internal stress-strain subloops within the fuller stress-strain curve that is associated with complete transformation. Repeated cycling between either fixed stresses or fixed strains causes the subloops to converge onto a final limiting response. This allows for the prediction of shakedown behavior associated with either repeated stress cycling or repeated strain cycling (Fig. 8).

Pseudoelasticity and Reorientation. For $T > A_f$, combined tension/compression loading gives transformation behavior that alternates: $\mathbf{A} \rightarrow \mathbf{M}_+ \rightarrow \mathbf{A} \rightarrow \mathbf{M}_- \rightarrow \mathbf{A} \rightarrow \mathbf{M}_+ \rightarrow \dots$. As the temperature is lowered, the model gives isothermal stress-strain behavior with plateau stresses that decrease with temperature in the correct way. Once the temperature is lowered below A_f , the isothermal stress-strain behavior under tension/compression begins to involve $\mathbf{A} \leftrightarrow \mathbf{M}$ transformation in conjunction with direct $\mathbf{M}_- \leftrightarrow \mathbf{M}_+$ reorientation. The model Eqs. (7), (21), (22) track this multiple transformation activity. The model also retrieves pure $\mathbf{M}_- \leftrightarrow \mathbf{M}_+$ reorientation with temperature-independent plateau stresses when $T < M_f$. For $T < A_f$, isothermal tension/compression loading excursions

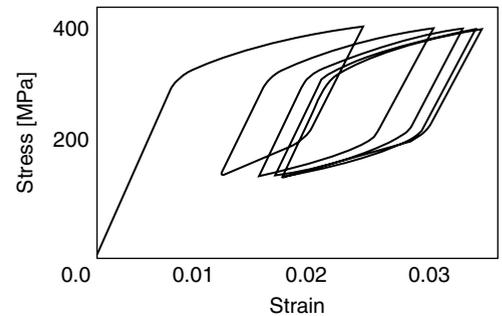


Figure 8. Stress cycling between 150 MPa and 400 MPa is modeled at $T = 315 \text{ K} > A_f$ using the trigonometric envelope functions (13) and D-M governing functions (16) with $n = 1$. Each new loop is richer in \mathbf{M}_+ and leaner in \mathbf{A} than the previous loop. After about five transient loops, the response has converged to a repeating loop that stabilizes the cycling between \mathbf{A} and \mathbf{M}_+ .

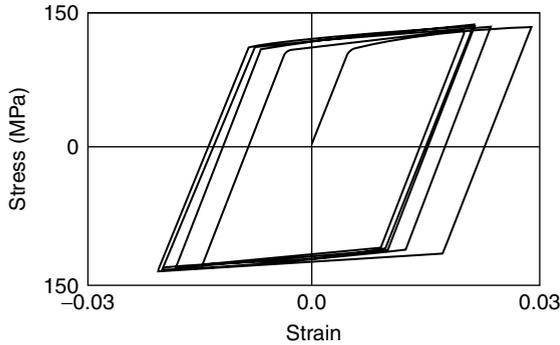


Figure 9. Stress cycling between -150 MPa and 150 MPa is modeled at $T = 235$ K $> M_f$ using the same envelope and governing functions as in Fig. 8. Here $\sigma_s^+ = -\sigma_s^- = 120$ MPa and $\sigma_f^+ = -\sigma_f^- = 210$ MPa. Since $120 < 150 < 210$ MPa, the cycling generates a sequence of transformations: $\mathbf{M}_- \rightarrow \mathbf{M}_+ \rightarrow \mathbf{M}_- \rightarrow \mathbf{M}_+ \rightarrow \dots$, all of which are incomplete. The convergent stable loop is symmetric, because the particular material parameter choice represents a tension/compression symmetric material.

generate open stress–strain curves (ferroelasticity). As is the case for high-temperature behavior, loop convergence takes place under repeated cycling. The convergence is immediate after the first cycle if the stress or strain magnitude is sufficient to complete all of the transformations. If, however, the cycling magnitudes do not cause complete transformation, then the curves again shake down to their limiting response (Fig. 9).

Nonisothermal Pseudoelasticity. Rate dependency follows from this model under the common condition of convective heat transfer as described by Eqs. (29), (30). For fixed κ , faster loading gives less phase transformation because convection inefficiency gives heat retention that works against the $\mathbf{A} \leftrightarrow \mathbf{M}$ transformation. The adiabatic limit is approached under very rapid loading. The opposite limit of isothermal transformation occurs under very slow loading. Figure 10 shows the connection among isothermal, convective, and adiabatic loading as predicted by this modeling.

Experimental Validation. As an example of validation of the model, we report a comparison of the model predictions with experimental data. Figure 11 shows the results of mechanical loading tests with temperature measurements performed on 1 mm near equiatomic commercial grade NiTi wires (47) and the corresponding prediction of the model after straightforward material parameter identification. The comparison gives excellent agreement in predicting sublooping behavior and temperature change under the combined conditions of cyclic loading, transformational heating/cooling, and convective heat transfer.

Further features of the model that also follow but are not obvious from the previous figures include temperature-dependent stress–strain curves due to the change in pseudoelastic yield stress with temperature, stress-free strain upon unloading due to the presence of stress-induced martensite, and the shape-memory effect.

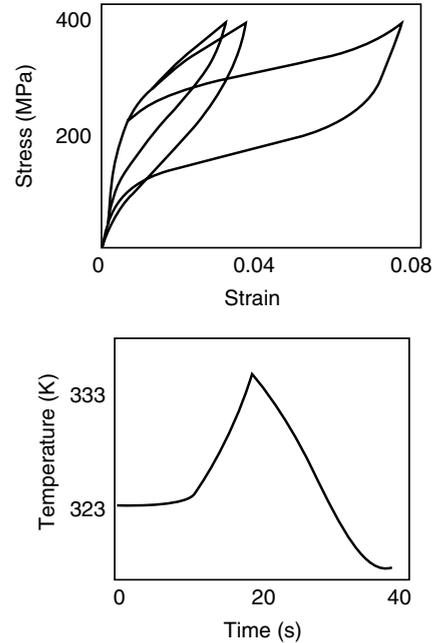


Figure 10. Three stress–strain simulations beginning at $T = 323$ K $> A_f$ (above). This simulation uses a hyperbolic tangent envelope function and slightly changed material parameters suggestive of a different NiTi material microstructure (e.g., $A_f = 315$ K, $E_M = 20000$ MPa). The extended stress–strain curve is the isothermal response. The curve on the far left is the adiabatic response, and the middle curve models heat transfer to $T_E = 323$ K over 20 s loading and unloading intervals. Nonisothermal loading gives heating since $\mathbf{A} \rightarrow \mathbf{M}_+$ is exothermic. Nonisothermal unloading gives cooling since $\mathbf{M}_+ \rightarrow \mathbf{A}$ is endothermic. Under adiabatic conditions, the cooling returns the material temperature back to $T = 323$ K at the conclusion of unloading. Under convective conditions, there is temperature undershoot on unloading (below) because of net heat transfer to the generally cooler ambient.

SUMMARY AND CONCLUSIONS

The thermomechanical stress–strain–temperature behavior of SMM can be modeled so as to predict shape memory, pseudoelasticity, and martensite reorientation. In order to achieve broad engineering utility, it is also necessary to predict sublooping, shakedown, and nonisothermal behavior. Current models can reliably capture these effects under

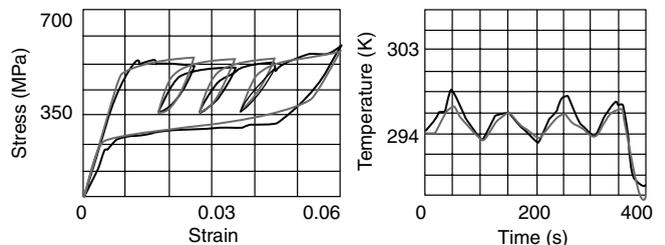


Figure 11. Comparison of this modeling (red) to measured response (black) in a NiTi wire at a loading rate for which convection is important.

simple states of stress such as uniaxial tension or simple shear. These settings cover a rather wide range of applications such as actuators, vibrations absorbers and other devices. Models for multi-axial loads are not yet as thoroughly validated with experiment, especially with regard to reorientation under nonproportional loading. As was indicated by the literature surveyed above, these models differ principally in how refined a description of the microstructure is implicit in their framework. Some models explicitly treat interactions between microstructural entities, whereas others treat the effect of microstructural evolution in a thermodynamic framework with consolidated internal variables.

A brief description of a generally useful internal variable model for uniaxial stress and a variety of heating effects was also presented. The internal variables describe the microstructure in terms of local phase fractions for austenite \mathbf{A} and two variants of martensite \mathbf{M}_+ and \mathbf{M}_- with transformation strains of opposite orientation. The internal variables evolve with temperature and stress according to kinetic equations (7), (21), (22). Strain and entropy follow from Eqs. (11) in a way that is consistent with their status as thermodynamic conjugates to stress and temperature. Equations (29), (30) treat the effect of transformational heating/cooling and heat transfer, which naturally provides a rate effect to the stress-strain behavior. The examples presented here have not addressed the effect of inhomogeneous deformation or nonuniform temperature within the SMM material. These issues can be treated for this and similar models with appropriate field equations, such as Eqs. (4) to (6), and Fourier's law. This approach in turn allows for computational simulation of SMM components within larger systems.

BIBLIOGRAPHY

- I. Muller and K. Wilmanski, *Il Nuovo Cimento* **B57**: 283–318 (1980).
- M. Achenbach, *Int. J. Plast.* **5**: 371–398 (1989).
- P. Xu and J.W. Morris, *Metall. Trans.* **A24**: 1281–1294 (1993).
- J.L. Ericksen, *J. Elast.* **3-4**: 191–201 (1975).
- J.G. Ball and R.D. James, *Arch. Rational Mech. Anal.* **100**: 13–52 (1987).
- K. Bhattacharya, *Continuum Mech. Thermodyn.* **5**: 205–242 (1993).
- K. Bhattacharya and R.V. Kohn, *Arch. Rational Mech. Anal.* **139**: 99–180 (1997).
- R. Abeyaratne, S.J. Kim, and J.K. Knowles, *Int. J. Sol. Struct.* **31**: 2229–2249 (1993).
- L. Truskinovsky. In P.M. Duxbury and T.J. Pence, eds., *Dynamics of Crystal Surfaces and Interfaces*, Plenum, New York, 1997, pp. 185–197.
- D. Bernardini, *J. Mech. Phys. Sol.* **49**: 813–837 (2001).
- E. Patoor, A. Eberhardt, and M. Berveiller, *Arch. Mech.* **40**: 755–794 (1988).
- Q.P. Sun and K.C. Hwang. (1994). In J. Hutchinson and T.W. Wu, eds., *Advances in Applied Mechanics*, Vol. 31, Academic Press, San Diego, CA, 1994, pp. 249–298.
- V.I. Levitas, *Int. J. Sol. Struct.* **35**: 889–940 (1998).
- B.C. Goo and C. LExcellent, *Acta Mater.* **45**: 727–737 (1997).
- Z.K. Lu and G.J. Weng, *Acta Mater.* **46**: 5423–5433 (1998).
- M. Huang and L.C. Brinson, *J. Mech. Phys. Sol.* **46**: 1379–1409 (1998).
- X. Gao, M. Huang, and L.C. Brinson, *Int. J. Plast.* **16**: 1345–1369 (2000).
- P. Sittner, R. Stalmans, and M. Tokuda, *Smart Mater. Struct.* **9**: 452–465 (2000).
- T.J. Lim and D.L. McDowell, *ASME J. Eng. Mater. Tech.* **121**: 9–18 (1999).
- F. Falk, *Acta Metall.* **28**: 1773–1780 (1980).
- Y. Huo, *Continuum Mech. Thermodyn.* **1**: 283–303 (1989).
- J. Ortin, *J. Appl. Phys.* **71**: 1454–1461 (1992).
- A.A. Likhacev and Y.N. Koval, *Scripta Metall. Mater.* **27**: 223–227 (1992).
- Y. Ivshin and T.J. Pence, *Int. J. Eng. Sci.* **32**: 681–704 (1994).
- B.D. Coleman and M.E. Gurtin, *J. Chem. Phys.* **47**: 597–613 (1967).
- J.R. Rice, *J. Mech. Phys. Sol.* **19**: 433–455 (1971).
- K. Tanaka and S. Nagaki, *Ingenieur Archiv* **51**: 287–299 (1982).
- K. Tanaka, S. Kobayashi, and Y. Sato, *Int. J. Plast.* **2**: 59–72 (1986).
- M. Fremond, *Shape Memory Alloys*. Springer Vienna, 1996, pp. 1–68.
- E.N. Bondaryev and C.M. Wayman, *Metall. Trans.* **A19**: 2407–2413 (1988).
- I. Muller, *Continuum Mech. Thermodyn.* **1**: 125–142 (1989).
- J. Ortin and A. Planes, *Acta Metall. Mater.* **37**: 1433–1441 (1989).
- C. Liang and C.A. Rogers, *J. Intel. Mater. Syst. Struct.* **1**: 207–234 (1990).
- B. Raniecki, C. LExcellent, and K. Tanaka, *Arch. Mech.* **44**: 261–288 (1992).
- L.C. Brinson, *J. Intell. Mater. Syst. Struct.* **4**: 229–242 (1993).
- A. Bekker and L.C. Brinson, *J. Mech. Phys. Sol.* **45**: 949–988 (1997).
- Y. Ivshin and T.J. Pence, *J. Intel. Mater. Syst. Struct.* **5**: 455–473 (1994).
- X. Wu and T.J. Pence, *J. Intel. Mater. Syst. Struct.* **9**: 335–354 (1998).
- J. Lubliner and F. Auricchio, *Int. J. Sol. Struct.* **33**: 991–1003 (1996).
- J.G. Boyd and D.C. Lagoudas, *Int. J. Plast.* **12**: 805–842 (1996).
- K.R. Rajagopal and A.R. Srinivasa, *Z. Angew. Math. Phys.* **50**: 459–496 (1999).
- S. Govindjee and G.J. Hall, *Int. J. Sol. Struct.* **37**: 735–760 (2000).
- R. Loloee, T.J. Pence, and D.S. Grummon, *J. Physique IV*, **C8**: 545–550 (1995).
- A. Amengual, A.A. Likhacev, and E. Cesari, *Scripta Mater.* **34**: 1549–1554 (1992).
- R. Wasilewski, *Metall. Trans.* **2**: 2973–2981 (1971).
- J. Ortin and A. Planes, *J. Physique IV*, **C4**: C4–C13 (1991).
- D. Bernardini and F. Brancaloni, In *Proc. Manside Workshop*, January 28–29 1999, Rome, Italy, part II, pp. 73–84.
- K. Gall and H. Sehitoglu, *Int. J. Plast.* **15**: 69–92 (1999).
- P. Sittner, M. Takakura, and M. Tokuda, *Mater. Sci. Eng. A*, **234-236**: 216–219 (1997).

SHIP HEALTH MONITORING

JASON S. KIDDY
 PETER C. CHEN
 PAUL ROSS
 Alexandria, VA

INTRODUCTION

Although ships have been sailing for hundreds of years, the field of ship health monitoring is relatively new. Two main forces have spurred the rise in the popularity and importance of ship health monitoring systems. First, reduced budgets and reduced manpower have resulted in smaller crew sizes and less routine maintenance. This reduced maintenance has brought into question the health of many ships, their components, and the ability to detect a problem quickly. Clearly, a failure aboard any vessel will bring a loss of revenue and productivity, if not worse. The second driver behind the growth of ship health monitoring systems is the decrease in the expense of computers and sensors coupled with an increase in their capabilities and the associated data processing algorithms. Until recently, an automated health monitoring system was either impossible to achieve or prohibitively expensive to install and maintain. Today, ship health monitoring systems are becoming increasingly more common as their capabilities are further demonstrated and understood by the maritime community. Nevertheless, many issues remain to be settled as technology improves, resulting in continuously expanding requirements.

OVERVIEW OF SHIP HEALTH MONITORING

Modern ships, both commercial and military, are extremely complicated machines, so that the aspects of the ship to which a ship health monitoring system can be applied are numerous. To date, efforts have focused primarily on the global hull response and diagnostic monitoring of the propulsion system. Although these two applications have the greatest potential for financial and safety improvements, other areas, such as local hull stresses and cargo tanks, can benefit from monitoring (1). In military applications, many specialized monitoring applications can be envisioned, such as weapon systems or ordinance monitoring and battle damage estimates (2).

Potential Benefits

Ship health monitoring systems can be employed for many reasons. In almost all cases (except research vessels), the primary reason is to reduce the overall cost of ship operations. In isolated cases, other reasons have dictated the use and development of monitoring systems. Another readily apparent benefit of an installed system is failure prevention. Although this is rarely the primary factor for ship monitoring, the safety benefit gained from avoiding a catastrophic failure is a strong motivator for using a system. Finally, because of the advanced technology and recent emergence of ship health monitoring systems, many

systems have been installed for research. The research has focused on the capabilities, benefits, and logistics of long-term ship health monitoring system use.

Financial. Ultimately, for the field of ship health monitoring to be viable, installed systems must decrease the overall costs of operating a ship and/or provide a substantial performance benefit. In both cases, a financial impetus will exist for installing and using a ship monitoring system. In general, cost savings will come from three sources: a reduction in maintenance labor-hours, reduced equipment costs, and increased readiness and uptime. Unfortunately, of these three, only the cost savings from the reduction in maintenance labor-hours can be easily calculated. If a ship health monitoring system is operating properly, many equipment failures are likely to be detected earlier before they become disastrous. Early detection will in turn help prevent one failure from affecting other components and lower the costs to fix a failure, resulting in a reduction in equipment costs that is hard to quantify. Furthermore, unscheduled maintenance will be reduced, resulting in an increase in ship availability. The increased availability will then create higher revenues for the entire system.

Safety. An additional benefit of ship health monitoring is the inherent safety provided by such a system. If the physical health of critical areas is constantly monitored, catastrophic failures are less likely to occur. Most critical regions of the ship can be monitored for stress overloads. Should a stress overload occur, the monitoring system quickly informs the bridge and provides pertinent information to the crew to lessen the amount of damage caused by the overload. If a failure has occurred, this information will also be passed to the bridge. In this case, although the system was unable to prevent a major failure, it will give the crew additional time and information to deal with the problem. For military vessels, the safety of the crew can be protected by the provision of real-time, accurate battle damage estimates. Again, this information will allow for a rapid damage assessment that could easily save lives onboard a crippled vessel.

Potential Dangers to Ships

Ocean and seagoing ships are subject to many potential hazards. The following sections describe some of the operational hazards in detail. In general, these hazards are caused by wind and waves, ice, and material (cargo) handling/storage. In addition to the hazards faced by commercial vessels, military vessels face the obvious threats from enemy actions. However, this section discusses only hazards that are encountered during typical ship operations.

Wind and Waves. Waves, both large and small, are an ever-present hazard to shipping. Waves stress a ship through several different phenomena that are adapted from (3) and described here.

Quasi-Static Hull Girder Stress (Global Stress). Hull girder shears and moments are caused by the cyclic buoyancy of a wave that is superimposed on the ship's geometry in a quasi-static balance with the ship accelerations. The

moment values depend more on the projected wave length superimposed on the hull (wave length/cosine of the heading angle) than on the encounter frequency. However, the pitch and heave resonance (a function of the encounter frequency versus the motion natural frequency) can increase the hull girder moments.

Hull Girder Whipping (Global Stress). When a structure is impacted, much of the impact energy is absorbed by the structure as vibrational energy. This vibration generally forms as motion of the structure at its first natural frequency. When a ship is impacted, such as during a slam, the ship hull vibrates in its fundamental bending modes (vertical and lateral). This is termed hull girder whipping. Slams can occur on the bottom and on the flare of the ship's bow. Bottom slams occur when the forefoot of the ship is lifted clear of the sea by severe ship and wave motions during rough seas. A slam occurs as the ship reenters the sea and the bow impacts the water. Flare slamming may occur as the result of relative motion between the vessel and the sea, even without bow emergence, but can also occur when there is little motion between the vessel and the sea, if the wave is steep enough. Bottom slams tend to have shorter lengths and higher loading frequencies than flare slams. The dominant slam type depends on the ship type. High-speed containerships that have finer forward lines and a flaring bow experience greater stresses from a flare slam, but the opposite is true for full-form tankers that have little flare. Measurements on an aircraft carrier have shown that the whipping moments are of the same magnitude as the quasi-static moments during flare slam (1). Although the whipping vibrations and energy dissipation mechanisms are not well understood for large complex vessels, they are generally less severe in flexible (i.e., high L/D ratio) ships (4). The whipping moment components are usually small compared to the quasi-static moment, but the whipping moments occur at higher frequencies. Recent investigations suggest that whipping may increase fatigue damage by 20% to 30% (5).

Springing (Global Stress). When a ship impacts waves at a frequency that is at or close to the primary hull resonant frequency (the two-noded, vertical bending mode), springing may occur. Springing is steady-state resonance of the ship at its natural frequency. Although this is also a low-frequency event, springing frequencies are of an order of magnitude higher than quasi-static hull girder stresses. The resulting moment, especially when superimposed on the quasi-static stresses, may be significant in long-term fatigue damage. Ships can experience springing in small and moderate seas, as long as the encounter frequency approaches the ship's natural frequency.

Wave Refraction (Local Stress). Hull girder stresses are generally caused by larger waves of the order of the ship's length. But the lower stresses created by smaller waves impacting on the sides of the ship can cause localized long-term fatigue damage and may lead to cracks and crack propagation. This effect is intensified by wave reflection in beam seas. Localized fatigue cracking has been a problem on some Trans Alaska Pipeline System (TAPS) trade tankers where the local waves generally strike the starboard side during the southern voyages and the port side on the northerly routes.

Slamming (Local Stress). The damage from forefoot slamming has been mentioned previously as a cause of whipping. In addition, the locally high stresses can cause damage to the bow structure. Bottom slamming in ships often results in dishing of the bottom shell plate, and flare slamming results in dishing of the side shell and possibly the loss of the flare strake.

Ice. In addition to the stresses caused by sea loading, ice represents a significant danger to shipping. The danger from ice comes primarily from localized impact loading on the hull. However, global stresses can also become a factor due to hull girder whipping.

Ice Transit (Local Stress). Because ship speed, heading, and ice conditions can vary greatly, local ice loads on the ship's structure are complex. The danger from ice depends more on local stresses than on the global hull forces. Shipboard measurements have shown that typical hull girder stresses induced by ice transit are less than those induced by open-water waves (3). The pressures and forces encountered during ship-ice impacts are random and follow log-normal type probability distributions (6). Ice loads are nonuniform, and high loads are applied to small areas of the hull (e.g., 0.5 m²). In addition, these loads occur at many locations along the hull, predominately on the bow. The instantaneous area of the hull that is most highly loaded depends on the type of operation (ramming, turning, etc.) and the geometry and strength of the hull structure. Ice loads are more difficult to measure than slamming loads because of their high frequencies and random distributions. Studies indicate that strain rates for ice loading in the local structure are similar to those for the global response and that neither of these is significantly different from those experienced from sea loading (3).

Hull Girder Whipping (Global Stress). Hull girder whipping caused by slamming is described in the previous section. Because whipping is the vibration of the ship's primary bending modes due to impulsive loading, ice ramming can also induce hull girder whipping. This effect is usually felt during initial ice impact and less during steady ice transit.

Cargo. As for ice, cargo and material handling can also add additional stresses to a ship's hull. Cargo loads can be categorized into two forms.

Quasi-Static Hull Girder Stress (Global Stress). Quasi-static hull girder stress was mentioned previously when caused by long length waves superimposed on the geometry of the ship, causing high moment and stress values. For cargo, quasi-static girder stresses are caused by differences in the loading distribution curve and the ship's buoyancy curve along the length of the ship. Care must be taken during cargo loading and unloading so that the maximum allowable stress values are not exceeded.

Cargo Loads (Local Stress). Cargo loading anomalies can often result in localized regions of high stress. Two examples of loading anomalies are uneven loading in bulk ships (this, it has been hypothesized, is the cause for a number of bulk ship failures) and unequal hydrostatic pressure heads

across tank boundaries. The American Bureau of Shipping (ABS) SafeHull code specifically considers a checkerboard loading pattern in cargo and ballast tanks as a worst case scenario. Hence, the loading sequence can result in excessively high global and local stresses.

Different Ship Types and Requirements

Ship Types. Specific ship designs are susceptible to various hull responses. Table 1, taken from (3), shows the typical monitoring requirements for several common ship types.

Operating Environment. Ships are designed for many functions and operate in a wide variety of conditions. The size, type, and operating environment of a ship greatly influence the requirements placed on a health monitoring system. For example, because container ships that operate in calm, smooth waters in the tropical zones are rarely subjected to icy conditions or severe storms, whipping and other wave/ice-induced fatigue stresses are not critical factors. Therefore, a proper health monitoring system should concentrate on cargo-induced loads and maximizing operational efficiency.

On the other hand, TAPS trade tankers are constantly subjected to severe storms, high waves, and very directional sea states (7). As cargo runs are made to the south, waves primarily strike the starboard side of the ship. On the return trip to the north, waves strike the port side primarily. This pattern has resulted in localized fatigue problems and requires a high density of local stress sensors to detect the onset of cracking. In the North Sea, hull girder bending, slamming, and green water are present because of

the very steep waves and require sensors similar to those for the TAPS trade tankers to detect any potential problems. Bulk ships operating on the Great Lakes are susceptible to springing because of their high L/D ratios and require global hull bending sensors. Ice breakers are subject to high localized stresses, especially around the bow, but the global bending stresses are typically lower than those experienced by other oceangoing vessels.

These basic generalizations apply only to small groups of ships that continuously operate in a given environment. Many ships are not easy to classify; their health monitoring considerations must include all of the potential operating conditions for a specific vessel. Most military vessels are prime examples of ships that are commonly operated with no fixed route and that experience many types of sea states throughout a voyage.

Additional Capabilities

Industry has realized that ship operators expect a ship health monitoring system to do more than simply measure the state of stress throughout the ship and monitor machinery health. As mentioned previously, ship health monitoring systems must be financially attractive to be commercially acceptable. It is clear that preventing the loss of a ship or increasing the useful life of a ship by structural monitoring are financial benefits. However, additional benefits can be gained by incorporating external (non-ship-based) sensor readings into a comprehensive monitoring system. These external readings are currently composed of two primary types: weather avoidance/planning and route monitoring.

Weather Reports. Current weather reports are available to ships from a variety of environmental sensor platforms, including fixed land sensors, ocean buoys, other ships, aircraft, and satellites. Collectively, these platforms can provide the necessary information to a ship to help prepare an optimal route to increase the efficiency of the route and to lessen any damage that may be caused by ice or poor weather.

Fixed Land Sensors. Fixed land sensors are primarily used to measure basic meteorological conditions such as wind speed and direction, temperature, and precipitation.

Ocean Buoys. The National Data Buoy Center (NDBC) operates the Ocean Data Acquisition System (ODAS), which is a network of more than 60 buoys that are anchored in deep ocean areas off of North America. These buoys send satellite transmissions to the National Weather Service (NWS) that provide weather and oceanographic data from their stations in the Atlantic, Pacific, Gulf of Mexico, and the Great Lakes. The wind speed and direction data, once processed by NDBC, is reportedly accurate to within $\pm 10^\circ$ and ± 1 m/s.

Ships. NWS receives weather reports every 3 hours from ships that participate in the U.S. and World Meteorological Organization (WMO) Voluntary Observing Ship Program (VOS). These reports include basic meteorological conditions and best estimates of the current sea state, ice, and visibility. These programs include 49 participating countries and approximately 7000 ships that provide about

Table 1. Monitoring Requirements by Ship Type

Ship Type	Typical Hull Response
Passenger ship	Ship motion (roll) Bow flare slam
Tanker/products carrier	Midship hull girder stress Bow/amidships shell stiffeners Forefoot slam Explosive environment
Bulk ships	Cargo loading hull girder stresses Cargo hold frame stresses Stress concentrations at hatch corners Forefoot slam
Container ships	Stress concentrations at hatch corners Hull girder torsion Bow flare slam Green water Whipping/cargo accelerations
LNG/internal tank	Forefoot slam Temperature/explosive environment Sloshing
Barges/platforms	Towline/mooring tension Motions and inertial forces Lateral motion
Naval combatant	Bow flare slam Fire control plane deflections

1000 reports a day. The U.S. program has existed as a descendant of the U.S. Coast Guard Ocean Weather Station ships for several decades. The data provided by the VOS is commonly used for weather forecasting and is distributed by the National Ocean Weather Service through the Global Telecommunications System.

Aircraft. Aircraft are most commonly used to track hurricanes in the Atlantic Ocean. But they have also been experimentally used to track ice conditions in the polar regions (8) and have been used to provide Synthetic Aperture Radar (SAR) readings to estimate local sea states.

Satellites. In general, satellites provide a great deal of information to weather forecasters. Due to the advent of modern satellite imaging technology, forecasters can accurately predict the weather many days in advance. In addition to the basic weather forecasting functions, several satellite instruments have been used to aid in ship navigation.

Advanced High-Resolution Radiation (AVHRR) sensors are used to sense ocean temperatures and map sea currents. Flown by NOAA since 1978, AVHRR sensors detect infrared radiation to measure the sea surface temperature. This data is critical to helping oceanographers track ocean currents. Generally, two AVHRR satellites are in polar orbit on 24-hour cycles phased 12 hours apart to give both day and night readings. Although clouds interfere with AVHRR readings, this interference has not significantly affected their usefulness.

Radar altimetry has been used to measure the distance between a satellite and the ocean waves to estimate the sea state. Although the technology was first demonstrated onboard the GEOS-3 in the 1970s, suitable accuracy was not obtained until the recent launch of the TOPEX/Poseidon in 1992. However, until additional satellites become operational, this technology will be used primarily for research.

A final sensor technology is scatterometry. Scatterometry measures the scatter within a return pulse from a radar altimeter to determine the roughness of the seas. Calm seas give a clear, concise radar reflection, whereas rough seas tend to distort the return. The sea state is then related to the local wind speed through an empirical correlation. Again, scatterometry is a new technology that is currently in the development stages and that will hopefully be available to health monitoring systems in the future.

Route Monitoring/Planning. Adding the capability of monitoring weather conditions to a ship health monitoring system can improve the crew's ability to plan an optimal course through the weather. Ideally, though, the monitored weather conditions would be integrated with a route planning system to provide an optimal route to the crew automatically. By integrating the local health monitoring system with a real-time routing system, the ship's handling can also be adjusted to minimize danger to the ship. Using this type of system, it is possible to plan the best route to reach a given destination while reducing time and fuel consumption from unwanted ship responses. Several technologies needed for such a system are already in existence, including weather forecasting, the predicted ship responses, and the local sensors to determine the actual ship response. However, an intelligent software product for

performing the optimization at real-time speeds has not been fully realized.

The benefits of route planning were demonstrated in 1993 by ARCO Marine (9). Two sister TAPS trade tankers traveled from San Francisco, California to Valdez, Alaska with the same ballast condition. One ship contained a voyage planning system based on the predicted wave heights and directions; the other did not. The ships remained within a narrow corridor and varied only the timing and speed. The ship that used route planning arrived approximately 18 hours earlier than the sister ship, and the sister ship suffered \$400,000 in wave-induced damage.

ENVIRONMENTAL ISSUES

For a ship health monitoring unit to be acceptable, it should not adversely affect the operation of the ship or its crew. Furthermore, the system must be reliable and require very low maintenance. A primary driving benefit of a ship health monitoring system is a reduction in crew workload. If the system is constantly in need of repairs, this benefit does not materialize. Unfortunately, the maritime environment is very harsh and unforgiving. Every component of a health monitoring system must be considered for reliability and maintenance. The primary factor that reduces the life of ship components is the highly corrosive marine environment. Many other sensor location-specific factors may also contribute to component failures. These include explosive environments, inadvertent physical damage by the crew, and operational overloads.

Corrosive Marine Environment

The marine environment is tremendously harsh and corrosive. This environment quickly affects most exposed metals and many other materials. Therefore, a ship's health monitoring system components must either be protected from this environment or constructed of materials that are not subject to marine corrosion. Although any material will eventually corrode, corrosion in the marine environment is more severe for several reasons. First, most metals corrode slowly at ambient temperatures and low humidity. Increasing the humidity provides water, which is necessary as an electrolyte for charge transfer. In addition to the ever-present water, the saline environment of ocean water further increases corrosion by speeding up the localized breakdown of oxide films. The chloride also increases the conductivity of seawater compared to freshwater, again increasing the corrosion of many metals. A final contributor to corrosion is the low acidity of seawater. The pH of seawater is usually 8.1 to 8.3.

Other

In addition to the highly corrosive marine environment, ship health monitoring sensors are placed in other types of extreme environments. Many modern ships are designed for and dedicated to transporting large quantities of oil or natural gas. For these ships, it is desirable to measure the health of the fuel storage tanks. This environment places additional safety requirements on the sensor

and any associated electronics due to the explosive nature of the cargo. A few sensors are intrinsically safe in an explosive environment, but most sensors must be encapsulated or encased in an explosion-proof container, which can substantially increase the cost and complexity of the sensor.

Another concern is to protect the health monitoring system's components from physical harm. It is often necessary to place sensors or wiring in locations that are susceptible to physical damage. Examples include deck-mounted components that can be damaged by the crew's activities or forefoot-mounted pressure sensors that must be designed to handle the high forces experienced during slams or ice-breaking duties. In some early systems, pressure sensors failed most frequently of all equipment (3). Without the proper protection, these sensors fail quickly.

SENSOR TECHNOLOGY

The sensor network is crucial to providing real-time, accurate information from the ship health monitoring unit to the ship's crew. It is the distributed sensors that directly measure the motions and health of the ship. To provide a detailed picture of the entire ship, these sensors must comprise many different types and must be placed throughout the vessel. Therefore, current and future health monitoring systems will incorporate a wide variety of sensor types that measure a diverse number of physical parameters. Figure 1 illustrates a possible ship health monitoring sensor arrangement. In addition to choosing the physical parameters that must be measured, where they are to be measured, and the type of sensor to be used, one must make decisions concerning the cost, reliability, and safety of the specific sensor. Current state-of-the-art sensors can meet these objectives, but novel sensor designs are continuously being developed and must be considered as possible improvements over existing designs.

Measurands and Potential Sensors

As mentioned, a wide variety of parameters must be measured to obtain an adequate picture of a ship's motions and health. The following sections describe many of these measurands and the potential sensors that are currently available for the measurement.

Pressure. Pressure gauges are most often used in a ship health monitoring system to measure slamming pressures, forefoot emergence, and in-tank hydrostatic pressures. Multiple pressure gauges are often located longitudinally along the forefoot to detect the extent of emergence and to determine the magnitude and extent of bottom impact pressures (10). To facilitate maintenance, it is important to ensure that any sensors, especially forefoot pressure sensors, can be replaced or repaired without entering dry-dock.

There are two primary types of pressure transducers: diaphragm and piezoelectric types. Both types are commonly available commercially at about the same cost, but each has certain advantages and disadvantages.

Diaphragm-Type Pressure Transducers. There are two types of diaphragm style pressure transducers: one has a clamped circular plate, and the other employs a hollow cylinder. However, the clamped plate design is more suited for the pressure ranges of ship health monitoring, and only they are discussed in this section. The strain distribution on a clamped circular plate of constant thickness has been solved analytically and has been experimentally validated. Based on these results, a special purpose diaphragm strain gauge has been designed to take advantage of this strain distribution. Using this type of strain gauge arrangement, one finds that the pressure is proportional to the measured strain. Typical strain gauge instrumentation can be used to sample the data.

Piezoelectric-Type Pressure Transducer. This type of pressure transducer uses a piezoelectric crystal as both the diaphragm and sensor. In general, the piezoelectric crystal

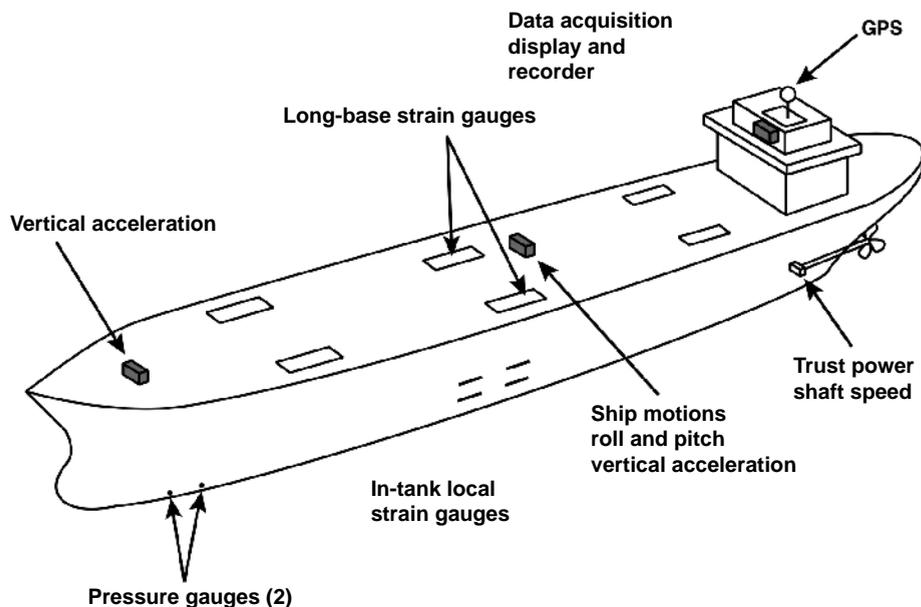


Figure 1. Potential ship health monitoring sensor arrangement.

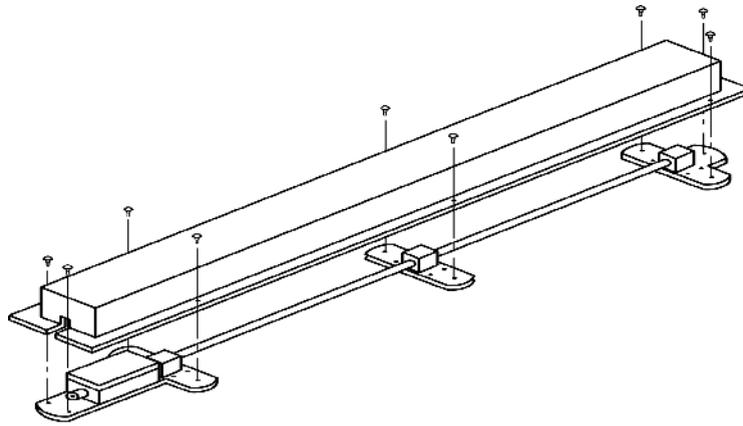


Figure 2. Long-base strain gauge.

(most commonly quartz) is placed inside a hollow cylinder. Because of the piezoelectric effect, an applied pressure generates an electrostatic charge that is proportional to the pressure. The piezoelectric crystal has a high output impedance. Therefore, a charge amplifier is commonly employed to convert the charge into an amplified voltage that is read by a standard voltage recorder. The low-frequency response of the transducer depends on the time constant of the amplifying circuit but can be designed to use frequencies that are low enough for ship health monitoring. The primary disadvantage of this type of transducer is that the charge amplifier electronics make the system less intrinsically safe than diaphragm-type pressure transducers.

Global Strain. Global strain is most commonly measured using long-baseline strain gauges. Long-baseline strain gauges normally consist of a long rod (approximately 2 meters long) rigidly attached at one end to the hull. The second end is allowed to move freely through a set of guides that ensure only axial movement. The extent of movement of the free end, as measured from a set “zero” strain locations, divided by the length of the rod gives the average strain in the deck across a region of the hull that is the length of the rod. Figure 2 is an illustrative example of a long-base strain gauge.

Although the basic mechanism is simple and robust, these devices are commonly placed on the deck of a ship and must be appropriately protected from both physical damage and the environment. Three common techniques are currently available for measuring the displacement of the free end.

Linear Potentiometer. Linear potentiometers are simple resistors that have a varying resistance that is proportional to the displacement. The displacement is measured by sending a low voltage and current through the potentiometer and measuring the resistance. This method is very inexpensive, and recent advances have made precision linear potentiometers as accurate and repeatable as other technologies. One disadvantage is that the resistor’s life is limited and it must be protected from the environment.

Linear Variable Differential Transformers (LVDT). LVDTs are the most popular variable-inductance sensor used for displacement measurements. In an LVDT, a magnetic core

moves through an insulated bobbin without physical contact. Three symmetrically spaced coils are wound around the bobbin. The position of the magnetic core controls the mutual inductance between the center primary coil and the two outer secondary coils. When a voltage is applied to the primary coil, a voltage is set up in the two secondary coils. The secondary coils are wired 180° out of phase with each other. Therefore, when the core is centered within the bobbin, the secondary voltages are of equal magnitude and cancel out. However, a small movement of the core results in a larger voltage in one of the secondary coils and hence a sensor reading. Because there is no contact between surfaces, LVDTs are free from friction and have very long lives. The response is also free from hysteresis. The resolution of an LVDT is partially determined by the voltage recorder, which give LVDTs superb resolution and accuracy. The two primary disadvantages of LVDTs are their higher cost compared to linear potentiometers and the relatively high voltage (5 to 15 volts) that must be supplied to the primary coil.

Linear Displacement Transducer. A linear displacement transducer is a magnetostrictive sensor that measures the time between an interrogating magnetic pulse and a return pulse that is generated by a magnet connected to the free end of the rod. As for an LVDT, this type of device has no contacts and a long service life. It is also intrinsically safe for hazardous environments but is much higher in cost than either of the other two options.

Local Strain. Local strain measurements are often used in ship health monitoring systems. They are applicable to detecting a wide variety of potential hazards. They are most commonly used along the bow, in cargo and ballast tanks, and along any other critical internal structures. Local strain gauges are also ideal for detecting cracking around certain areas such as hatch corners and highly stressed welds. In general, the number of local strain gauges is limited by the cost of installation and data acquisition as opposed to the cost of the gauges or the desire to measure more parameters. Ideally, almost any failure can be detected by a local strain gauge if it is installed in the right location.

By far the most common method of measuring local strains is to use a resistance foil strain gauge. Details of

the performance and use of resistance strain gauges can be found in many texts. The strain gauge operates by detecting a change in resistance of metal wire as it is stretched. The change in resistance is caused by a decrease in the conductive cross-sectional area due to Poisson's ratio. To amplify the change in resistance, one employs thin metal-foil grids to maximize the amount of conductor within a given region. Because the change in resistance is very slight, a Wheatstone bridge is usually used to convert the changing resistance to a variable voltage. Resistance strain gauges can be purchased in a wide variety of sizes, sensitivities, and geometries. It is also possible to purchase strain gauge rosettes that permit measuring all three surface stress components simultaneously. Because strain gauge readings depend on measuring a changing resistance, they can be made safe by using very low voltages.

An important consideration is the temperature sensitivity of strain gauges. As a material is heated or cooled, the material expands or contracts, depending on the material's coefficient of thermal expansion. If the strain gauge does not similarly expand or contract, an apparent strain will be seen by the strain gauge. Two methods exist for eliminating this apparent strain. The first method is ensuring that the coefficient of expansion is identical for both the structure and gauge. Therefore, both expand at the same rate, and no apparent strain is seen. For this reason, commercial vendors sell strain gauges made from a wide variety of materials. The second approach is not as attractive for practical applications. In this approach, a second gauge/material combination is bonded together but is placed in a region that is completely free of stress. Through this technique, the amount of apparent temperature-induced strain can be calculated from the second gauge and can be subtracted from the active gauge to provide the final mechanical strain reading. However, finding a suitable location for the dummy gauge is not generally possible.

The installation of strain gauges is also a well-documented field that has a myriad of options. The most common techniques include adhesive bonding and welding, but other methods are also available for packaged sensors. Generally, it is preferable to protect the strain gauge after installation. Again, this is done by using a variety of protective coatings and encapsulation techniques. Many strain gauges sold have the gauge itself prepackaged into an encapsulated enclosure for protection.

Motion. Ships' motions are possible in all six degrees of freedom. The three translational degrees of freedom are surge (longitudinal), sway (lateral), and heave (vertical). Roll, pitch, and yaw are the three rotational degrees of freedom. Roll, pitch, and heave are generally considered the three most important degrees of freedom for the following reasons:

Roll. Roll generally affects the crew and cargo loads. In addition to crew and passenger discomfort, rolling creates lateral cargo loads that must be resisted by horizontal restraints. In fluid tanks, rolling may increase the hydrostatic pressure head or induce sloshing. A ship's master usually turns the ship into waves to reduce rolling, but this increases hull girder stresses.

Pitch. Pitching has the same effect on fluid-filled cargo tanks as rolling and results in an increased hydrostatic pressure head and sloshing. In addition, the length of the ship increases the distance between the ends of the ships and the pitch axis, resulting in high pitch accelerations that may in turn result in slamming.

Heave. Heave is more pronounced than surge or sway because of wave motions and the coupling with pitch motions. Heave causes effects similar to roll and pitch in fluid-filled tanks. In addition, the vertical accelerations require vertical cargo restraints.

Ship motions are commonly measured by a variety of sensor types. The most common devices are accelerometers and gyros, although there are many variations of both. Table 2 lists the current state of the art in ship motion sensors (11).

Shaft Speed, Power, and Thrust. The ship's performance parameters (shaft speed, power, and thrust) can provide measurements of the propulsive efficiency relative to the current environmental conditions. These measurements can also provide an indication of the health of the propulsion equipment. Most ships are currently configured with equipment to measure these parameters directly from the propeller shaft. In general, this information is already passed into the engine control room and can be further routed to the ship health monitoring system.

Global Positioning. Similar to the propulsion system measurements, ships usually already have an installed Global Positioning Systems (GPS). As a rule, GPS units output a serial signal dictating the ship's position (latitude and longitude), heading, and speed on a regular basis. Therefore, the ship health monitoring unit only needs to read this information from the preexisting unit.

Advanced Sensors

Fiber-Optic Strain Gauges. One promising new technology is fiber optic strain gauges. There are two primary types of fiber-optic strain gauges, Fabry-Perot and fiber Bragg grating sensors (12). Both types offer a number of advantages over traditional resistance strain gauges. Because fiber-optic strain gauges use light as the sensing and transmitting element, they are intrinsically safe and pose no fire or explosive hazards. Furthermore, fiber optics are very resistant to corrosive elements because they have no metallic components and are covered by protective hermetically sealed coatings. A final common advantage is that fiber optics are immune to electromagnetic interference. Therefore, they can be placed in regions of high electric or magnetic fields without any degradation of performance. Both types of fiber-optic strain gauges are also capable of strain resolutions equal to or greater than that of resistance strain gauges.

Fabry-Perot Fiber-Optic Strain Gauges. Fabry-Perot strain gauges are manufactured by placing a small air gap (or an internal mirror) within a fiber, followed by a reflective surface, which can be either a micromirror or more optical fiber. A broadband light wave is transmitted down the length of the fiber. At the first junction between the

Table 2. Ship Motion Sensor Technology

Motion	Sensor Type	Advantages	Potential Problems
Roll and pitch	Vertical gyro	Reliable, may be able to use existing ship unit	Drift, cost, power
	Magnetometer	Moderate cost	Calibration on steel ships
	Solid-state gyro	Low cost and power, units packaged with integral rates and displacements	Sensitive to external vibrations
	“Watson meter”	Reliable, accurate for pendulum-based design	Moderate cost
Yaw (heading)	Bubble/simple pendulum	Low cost	Inaccurate when not at the center of motion
	Gyrocompass	Current state of the art	Expensive, frequently needs service
	Solid-state gyro (KVH)	Low-cost combination of rate gyro and flux gate compass	Unproven, unknown life and reliability
	Solid-state gyro (fiber optic)	New laser ring technology, no moving parts	Expensive, not yet commercialized for ship use
	Flux gate compass	Good for small ships once compensated for, low cost	Difficult to use effectively unless able to swing the ship for compensation
	Magnetometer	Moderate cost	Calibration on steel ships
	Piezoelectric accelerometer	Good for machinery vibrational measurements	Unsuitable for ship response frequencies
Surge Sway Heave	Piezoresistive accelerometer	Low cost, good for short-term ship motions	Subject to temperature cross-axis errors
	Servo-accelerometer	Excellent stability, accuracy, and reliability	Expensive
	Capacitive accelerometer	Moderate cost, performance nearing that of a servo-accelerometer	Cross-axis sensitivity higher than that of servoaccelerometers

fiber and the air gap (or mirror), some of the light is reflected back to the source, and some of the light is transmitted into the gap. At the mirror or second air/fiber interface, light is again reflected and transmitted. Now, two separate light sources are reflecting light back along the length of the fiber. The length of the air gap dictates the phase difference between the two waves. Allowing this air gap to expand or contract, based on the local strain, creates a strain sensor. Hence, a measurement of the phase offset can be correlated to a strain measurement. There are extrinsic Fabry–Perot sensors, intrinsic Fabry–Perot sensors, and in-line fiber etalon (ILFE) sensors. All three are based on the same principle, the differences lie in the choice of the reflective medium. Although intrinsic Fabry–Perot sensors are sensitive to strain and temperature, extrinsic Fabry–Perot sensors and ILFEs have very low thermal sensitivity. One disadvantage of Fabry–Perot sensors, compared to fiber Bragg grating sensors, is the difficulty in multiplexing many sensors along a single fiber.

Fiber Bragg Grating Strain Gauges. Fiber Bragg gratings are based on the photorefractive effect. Bare fiber is exposed to a hydrogen environment and then imprinted using an ultraviolet laser. The imprinting is done by one of several methods, and it leaves a series of equally spaced lines along a region of the fiber. This series of lines is called a Bragg grating; the lines are actually very small regions that have a slightly different index of refraction. Bragg gratings can be fabricated through an interferometric (holographic) method or by using phase masks. In a fiber that has a Bragg grating, transmitted broadband light is

reflected back toward the source at a specific frequency that corresponds to the grating wavelength. All other frequencies of light pass unaffected through the Bragg grating.

Because the frequency of the reflected light is proportional to the spacing of the Bragg grating, a change in the spacing will result in a change in the reflected wavelength. Hence, a strain gauge can be made by bonding a Bragg grating of a specific wavelength to a structure. As the structure is strained, the Bragg grating will expand or contract, thereby changing the wavelength of the reflected light. By measuring the wavelength of the reflected light, one can deduce the strain at the location of the grating (13).

One major advantage of using fiber Bragg gratings as local strain sensors is the capability of using wavelength and/or time division multiplexing to place many Bragg gratings (strain sensors) along a single optical fiber (14). When a broadband light source is used with a Bragg grating, every wavelength, except the wavelength corresponding to the grating, is transmitted through the grating. Therefore, a second Bragg grating, at a different wavelength, may be placed further along the fiber. This second Bragg grating will reflect a different wavelength back to the source. Now, two separate strain readings can be taken by monitoring the two reflected wavelengths. This process can be repeated many times along the length of the fiber, which allows making many distributed local strain readings within a single fiber-optic cable that also transmits all of the data back to the control computer. Similarly, time division multiplexing can be achieved by monitoring the time of return of the Bragg grating wavelengths, enabling interrogation of

multiple sensors along a fiber. One disadvantage of Bragg grating strain sensors is its strong thermal sensitivity. Numerous methods have been proposed to compensate for this thermal sensitivity, but none have yet been commercially successful.

Other Fiber-Optic Gauges. In addition to fiber-optic strain gauges, Fabry–Perot and Bragg grating strain sensors have been incorporated into other designs to enable detecting pressure, temperature, or even chemical content. In general, these sensors have many of the same advantages due to the nature of fiber optics as opposed to electrical components. But again, very few of these hybrid sensors are currently available commercially. These sensors will probably become more available and less expensive as the technology matures. Using Bragg grating-type sensors, it will also be possible to construct a series of varying sensors that are connected to the same fiber-optic transmission cable. For example, the fiber-optic cable from several forefoot pressure sensors could be run up to the cargo tanks where several Bragg grating strain sensors were located. Each of these sensors could be multiplexed together so that all of them are interrogated by a single fiber-optic cable running back to the control computer.

Sensor Power

Another consideration for ship health monitoring sensor systems is the source of the required power. Most common sensors, including strain gages and accelerometers, require a constant electrical input to operate. This power is usually provided by the control computer and is sent through installed wiring to the individual sensors. The disadvantage is that this approach often leads to additional bulky cabling. An alternative is to use the ship's existing power distribution network. This approach is, however, complicated because the existing power, especially near the bow, is limited and of poor quality, and high voltage spikes are common. The control computer and critical sensors must also be connected to an uninterruptable power supply (UPS) to maintain operation in the event of a power failure.

DATA

The wealth of information obtained from the remote and on-board sensors of a health monitoring system must be transmitted and processed into a form that is both useful and concise. It is widely accepted that a useful health monitoring system must have a bridge terminal to display all of the pertinent information to the ship's crew. This first step is to transmit the data to a central location, either on the bridge or nearby. After the data has been transmitted to a central location, it is input into a computer system that analyzes and formats it into easy-to-read displays. In addition to displaying the real-time data to the crew during the voyage, it is often desirable to store this data for future analysis. Each step in this process is an involved function, and each is described in the following sections.

Transmission

The first requirement is to transmit the data obtained from the various sensors that are located throughout the ship to a central location. This task is made difficult by several factors. First, the corrosive marine environment is extremely harsh to any exposed wiring. The distances onboard modern vessels also require that many signals be transmitted hundreds of feet before being amplified or received by the central computer. In general, there have been three main methods of transmitting the data: hard wiring, radio links, and fiber optics (3).

Hard Wiring. Hard wiring is the most common form of data transmission in currently installed health monitoring systems. In ships that have protected longitudinal passageways, shielded and grounded cables offer the route that has the lowest installation expense. However, many ships, including tankers and product carriers, do not have these passageways and require more extensive cable routing and cost. The use of armored cable is also recommended for any external cable routing to protect it from physical damage. In explosive environments, it is extremely important to ground all cables to reduce the risk of sparking; in fact this procedure is generally preferred in all applications to reduce noise.

Radio Links. Radio links between sensors and the control computer offer the advantage of simplicity of installation because no cables need to be run through the ship's hull. Nevertheless radio links have increased costs because of the transmitter and antenna. An additional advantage for explosive environments is that radio links eliminate the spark hazard found in hard wiring. This type of signal transmission becomes less attractive for large numbers of localized sensors. When wiring, one can lay a multitude of sensor cables at one time. For radio transmission, multiple transmitters are required for additional sensors. Although radio transmission does not degrade over the length of the ship, signal interference is possible and can corrupt the data with spurious signals.

Fiber-Optic Network. Fiber-optic networks represent a good alternative to hard wiring and many advantages but higher cost. As proved by the telecommunications industry, fiber optics can easily transmit many signals hundreds of feet without any signal degradation. Fiber optics are also inherently safe for explosive environments and do not suffer from electromagnetic interference. For many of these reasons, current naval vessels are being outfitted with wide area distributed fiber-optic networks (15). Similar to hard wiring, fiber-optic networks suffer the disadvantage of needing a fiber-optic cable from the sensors to the control computer. In addition, most standard analog electrical sensors require expensive signal converters and decoders to convert the data into a corresponding light signal and back to electrical signals at the control computer. However, for a ship that has a preexisting fiber optic network, this form of transmission is extremely attractive and will be attractive for other systems as the cost of fiber-optic components continues to drop.

Analysis

The analysis and display of the sensor data are the primary objectives of the standard ship health monitoring unit. It is the responsibility of the control computer to perform data acquisition from the individual sensors and to process this data to determine whether the ship has been damaged or faces any immediate danger. The most common method of analyzing the information is to monitor each sensor in terms of the absolute magnitude of the sensor reading. Although the algorithm is relatively simple, the crew should determine whether the ship has experienced a reading that approaches or exceeds the maximum allowable level. If an overload has occurred, the crew must know immediately, so that appropriate action may be taken. This may dictate a change in the ship's heading, an adjusted cargo loading pattern, or possibly a visual inspection of the sensor location.

More complicated analyses are also performed to determine the overall fatigue experienced by the ship and to locate any general trends in the data that might indicate a potential failure. These analyses include average sensor levels, standard deviations, and peak values. More complicated signal processing techniques have been developed for machinery health monitoring, but such techniques have not yet been used for ship health monitoring systems. Limited attempts have been made at this point to use ship health monitoring information to predict the remaining fatigue life of ship components.

Display

A key component of the ship health monitoring system is the bridge display. The information displayed on at the bridge is the primary interface between the monitoring system and the ship's crew. To be easily used by the crew, the information must meet many different requirements.

Information. The information given by the display is the most important function of the entire system. The crew requires simple displays that can quickly inform them of any potential dangers and the effect of various maneuvers on the state of the ship. Complicating the information display is the crew members' needs for different types and amounts of data. The information that is required during cargo loading is very different from the information needed when traveling through rough seas. Support personnel are interested in different types of information as they post-process the data. To meet these many demands simultaneously, it is common practice for the bridge display to consist of numerous (more than five) different screens.

Experience has also shown which data formats and types of information are the most helpful to bridge personnel. For example, the ABS requires displaying hull girder stresses over a relatively short period of time so that the effect of speed or heading changes on the measured stress levels can be evaluated. Experience has also shown that bridge crews want the stress information to be displayed as a percentage of the maximum allowable stress as opposed to actual stress or strain readings. The actual values are

important and are generally saved for later use, but the immediate needs do not require this information.

Alarms. Both audible and visual alarms are standard in all ship health monitoring systems. The alarms are needed to inform the crew quickly of any potential dangers. However, it is very important to set the alarm sensitivity levels high enough so that the crew does not become frustrated by the alarm system. Excessively sensitive alarms have resulted in disconnection of the alarm system in previous applications, especially because of ice-induced local stresses (3).

Color/Lighting. The graphical display of the ship health monitoring system must be capable of operating in two modes. During the daytime, it is desirable to have a bright screen with obvious color clues to inform the crew quickly of the ship's status. Standard danger colors such as red and yellow should be used to highlight high sensor levels or overloads. Similarly, cool colors should be used to indicate that the ship is operating normally. However, the display must not interfere with the crew's night vision during evening hours. Hence, the display must be able to switch to a second mode where lower intensity schemes can be used to maintain night vision. Intensity variations can then be used to signify danger as opposed to a color change.

Storage

After the information has been displayed to the crew, the sensor information must be stored for later retrieval and analysis. Using modern storage media, it is not difficult to store vast amounts of data in relatively little space. Nonetheless, a continuously operating ship health monitoring system can generate huge amounts of data. Several options exist for storing data, including magnetic disks and tapes and optical disks. The primary considerations for the storage medium are the cost and capacity of the device versus the frequency with which the medium needs to be changed by the ship's crew during a voyage.

COMMERCIAL SYSTEMS

To date, there are a few commercially available comprehensive ship health monitoring systems. In addition, several manufacturers commercially produce Hull Response Monitoring Systems (HRMS). Although these systems are not comprehensive health monitoring systems that encompass the entire vessel, they provide detailed and adequate monitoring of the vessel's hull structure and include many other associated monitoring functions. These systems will surely form the basis for a comprehensive ship health monitoring system. To regulate commercial HRMS, the ABS published, in 1995, classifications for Hull Condition Monitoring Systems. Although these guidelines are general, they do provide a minimum compliance level for all HRMS. These requirements are listed in Table 3.

Table 3. ABS HRMS Requirements

Measurement Device	Parameter	Sensitivity
Accelerometer	Range	None
	Accuracy	± 0.01 g's
	Frequency	$3 \times$ required response
Strain Gage	Range	Yield stress
	Accuracy	$\pm 5 \mu\epsilon$
	Frequency	5 Hz
Fatigue	Bin size	$50 \mu\epsilon$

Current Systems

Currently, there are approximately 10 commercial manufacturers of HRM systems. Although many of the systems focus on recording the same physical parameters, no two systems are identical and almost every manufacturer will specially design a system to the end user's needs. Nonetheless, it is helpful to give approximate capabilities of these individual systems. The values in Table 4 are given only as a reference; individual or all values may be higher or lower for any given manufacturer and system.

Future Enhancements

The future of HRMS and comprehensive ship health monitoring systems is both exciting and dynamic. As the cost of new vessels continues to rise, there is an increasing demand to maintain and extend the operational life of new and existing vessels. A ship health monitoring unit is uniquely capable of extending the life of a vessel by providing the optimal course to the crew to avoid severe storms and to limit the damage incurred by the ship. The health monitoring system may also allow aging ships to be kept in service for longer periods of time by accurately identifying any failures before they become catastrophic.

Almost every aspect of future systems is likely to be enhanced over the current state of the art within the next decade. These advancements include the following:

1. *Improved Sensors.* Almost every type of sensor will have more capabilities, reduced cost, and improved safety and reliability.
2. *Increased Sensor Density.* As the cost of individual sensors and data acquisition hardware decreases, the number of sensors installed in a typical system will increase. This will bring a greater density of sensors to a given region and will allow for more detailed measurements of local stresses.

Table 4. Typical HRMS Features

Parameter	Value
Input channels	16–64
Sample rate	10–50 Hz
Data storage capacity	1 GB
Real-time display	1 min
Average display length	5 min
No. of display screens	5–10

3. *Additional Monitoring Functions:* In addition to improving the ability to detect and locate any potential damage to the ship's hull, advancements will allow monitoring the health of additional regions of the ship or ship components. For example, researchers in related fields have demonstrated the capability of accurately detecting transmission faults before they become catastrophic. Faults such as cracked or broken gear teeth, damaged bearing raceways, or misaligned shafts have all been successfully detected before failure. These systems generally use accelerometers to measure the vibrations close to the gears, bearings, and shafts in many transmission systems. The measured vibrational signals can be analyzed by a wide variety of methods of varying complexity and computing power. Although the global parameters of power, speed, and torque are measured in current systems, they can only inform the crew of a transmission fault after it has occurred and has begun to affect the ship's operation. Therefore, this type of transmission health monitoring system is a logical addition to current ship health monitoring systems.
4. *Increased Computing Power:* The continued push for faster computers will enable health monitoring systems to perform more detailed analysis in real time. Such advances will lead to more sophisticated and sensitive algorithms that can inform the crew of a potential problem before it occurs. This is especially helpful in machinery diagnostics.
5. *Condition-Based Maintenance:* Currently, ship components are repaired or replaced based on one of two factors. Either the part is replaced on a time-dictated schedule, to prevent the part statistically from ever failing, or it is repaired/replaced after failing. Neither of the current methods is optimal. In the first case, healthy components are discarded only because they have been used for a set amount of time. In the latter case, the ship is potentially unavailable for service because of the required repairs. An alternative approach, which may be provided by an advanced ship health monitoring system, is condition-based maintenance (16–18). By monitoring the health of a structure, one can also monitor the remaining life of that component. For example, a crack may be detected, but the monitoring system may show that the part has an additional 6 months of life before failure. This monitoring capability will ensure that the ship's availability is not lost because of repairs or waiting for a component to become available, and healthy parts will not be inadvertently wasted.
6. *Increased Data Storage:* It is well known that the cost of data storage is continuously dropping. This trend will be very beneficial to ship health monitoring systems as increased amounts of raw data will be stored for possible retrieval and postprocessing.
7. *Improved Weather Forecasting and Route Planning:* Although weather forecasting is not a direct part of a ship health monitoring unit, the advances in weather forecasting will give the system greater confidence in

the predicted weather and will improve the effectiveness of route planning activities. This will be coupled to improved route planning algorithms that will be developed, as the relationships between the ship's response and weather conditions are more fully understood.

BIBLIOGRAPHY

1. J.A. Kuny, R.R. Lewis, and M.D. Dianora, *Am. Soc. Nav. Eng., Int. Ship Symp. II*, Philadelphia, Nov. 1996.
2. S. Phoa, *Am. Soc. Nav. Eng., Int. Ship Symp. II*, Philadelphia, Nov. 1996.
3. S.B. Slaughter, M.C. Cheung, D. Sucharski, and B. Cowper, Ship Structure Committee, NTIS #PB98-100431, 1997.
4. E.V. Lewis, *Structural Dynamics of Ships*. Royal Institute of Architects, 1974.
5. P.B. Lacey and H. Chen, *SNAME Los Angeles Metropolitan Sect. Paper*, 1993.
6. J.W. St. John et al., *SNAME Ictech '94*, Mar. 1994.
7. D.J. Witmer and J.W. Lewis, *SNAME Trans.* **102**: 501–533 (1994).
8. E.D. Leavitt and G. McAvoy, *MTS J.* **21**: 29–36 (1987).
9. R. Lovdahl, P. Lacey, and H. Chen, *SNAME 1995 Calif. Joint Sect. Meet.*, Apr. 2, 1995.
10. D.J. Witmer and J.W. Lewis, *SNAME 1994 Los Angeles Metropolitan Sect. Meet.*, Jan. 13, 1994.
11. F.H. Ashcroft, R.D. Goebel, and W.F. Hennessy, *SNAME 1995 Joint Calif. Sect. Meet.*, Apr. 22, 1995.
12. G.A. Johnson, S.T. Vohra, and S. Mastro, *Am. Soc. of Nav. Eng., Int. Ship Symp III*, Philadelphia, June 1999.
13. P. Ross, P. Chen, R. Wagreich, S. Chen, J. Sirkis, J. Kuny, and R. Lewis, *Am. Soc. Nav. Eng., Int. Ship Symp. III*, Philadelphia, June 1999.
14. K. Pran, G. Johnson, A.E. Jensen, K.A. Hegstad, G. Sagvolden, Y. Farsund, C.C. Change, L. Malsawma, and G.W. Wang, *Proc. SPIE 7th Int. Symp. Smart Struct. Mater.*, March 2000.
15. D.J. Coyle and R.J. Patterson, *Am. Soc. Nav. Eng., Int. Ship Symp. II*, Philadelphia, Nov. 1996.
16. D.K. Hoth, *Am. Soc. Nav. Eng., Int. Ship Symp. II*, Philadelphia, Nov. 1996.
17. E. Rerisi and J. Hutter, *Am. Soc. Nav. Eng., Int. Ship Symp. III*, Philadelphia, June 1999.
18. J.D. Keenan and W.H. Sims, *Am. Soc. Nav. Eng., Int. Ship Symp. III*, Philadelphia, June 1999.

SMART PEROVSKITES

ZHONG L. WANG
Georgia Institute of Technology
Atlanta, GA

Perovskite and perovskite-related structures are a class of smart materials (1). Perovskite-structured materials have important applications in ferroelectricity, piezoelectricity, ferromagnetism, magnetoresistance, superconductivity, ionic conductivity, and dielectricity. Typical perovskite

materials of technological importance are piezoelectric $\text{Pb}(\text{Zr},\text{Ti})\text{O}_3$, electrostrictive $\text{Pb}(\text{Mg},\text{Nb})\text{O}_3$, magnetoresistant $(\text{La},\text{Ca})\text{MnO}_3$, and superconductive $\text{YBa}_2\text{Cu}_3\text{O}_7$. Perovskite-related materials are versatile matrices for generating transition- and rare-earth metal oxides that exhibit a broad spectrum of properties and functions (2) that are related to the following characteristics: (1) Nearly innumerable combinations of metal cations can be accommodated within perovskite-related structural systems. (2) By reduction / reoxidation processes, nonstoichiometry (i.e., controlled amounts of ordered oxygen vacancies) can be introduced into the structure. In turn, high oxygen ion mobility or modified electronic and magnetic features can be implemented, and (3) the design of composite structural systems containing perovskite building units (perovskite slabs of different thicknesses) allows fine-tuning electronic and magnetic properties.

From the viewpoint of crystal structure, the ABO_3 type structure, in which the cation A usually has valence 2+ and the cation B has valence 4+, is the fundamental perovskite. The perovskite family is created by doping other types of cations into the stoichiometry and/or introducing anion deficiency. Understanding the structures and the relationships among the abundant structures in the perovskites may lead to some insights into the intrinsic connection between structure and properties. This article focuses on the structure and structural evolution of perovskites and explores the intrinsic linkages among the members of the perovskite family. First, we introduce the "smart" properties of perovskites. Then, the intrinsic connection among the perovskites is explored. Finally, the analysis of mixed valences and oxygen deficiency is addressed.

THE FAMILY OF PEROVSKITE-STRUCTURED MATERIALS

Examples of Perovskite Structures

The most typical perovskite structure is BaTiO_3 (Fig. 1a). The Ba atoms appear at the corners of the unit cell and oxygen atoms at the face centers. Both the Ba and O make up a face-centered lattice structure. The octahedrally coordinated titanium ion is located at the center of the unit cell. This structure can be generically written as ABO_3 , which is the fundamental structural configuration of perovskites.

Materials that have perovskite-like structures are numerous. The most typical are ceramic high-temperature superconductors such as $\text{YBa}_2\text{Cu}_3\text{O}_7$ (Fig. 1b). The unit cell can be considered a stack of three perovskite units along the c-axis direction, where the cation lattice preserves that of the perovskite, and oxygen vacancies are introduced. The distortion in the oxygen lattice sites is due mainly to the vacancies. The long periodicity of the c axis (e.g., superstructure) is the result of alternate distribution of the Y and Ba cations and the ordered structure of oxygen vacancies.

A comparison of BaTiO_3 with $\text{YBa}_2\text{Cu}_3\text{O}_7$ indicates the following. In BaTiO_3 , the cations are screened by anions so that two cations are not directly face-to-face. In $\text{YBa}_2\text{Cu}_3\text{O}_7$, although the cation distribution is the same as that in BaTiO_3 , the Cu ions at the top layer of the

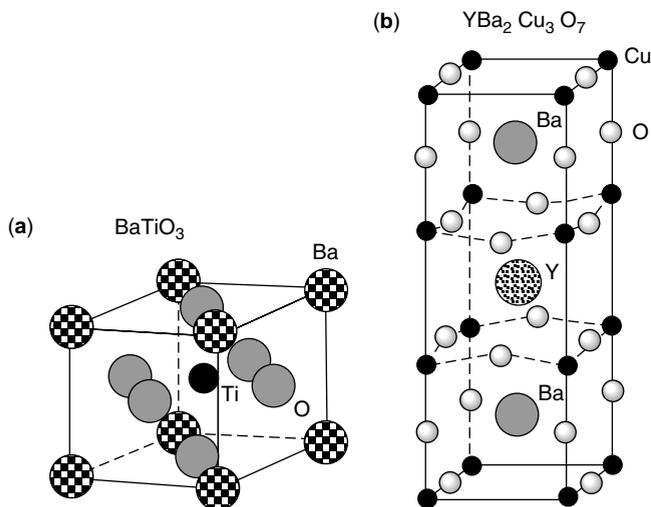


Figure 1. Atomic structural models of (a) BaTiO_3 and (b) $\text{YBa}_2\text{Cu}_3\text{O}_7$.

unit cell are face-to-face without the screening of anions, whereas the Cu cations next to Y are well coordinated. This structural configuration is possible only if Cu ions have different valence states at the two types of lattice sites. Therefore, perovskites have three major structural characteristics: cation substitution, ordered oxygen vacancies, and mixed valences of cations.

Perovskite-Like Structures

In the ABO_3 structure, the valences of the A (12-coordinated) and B (6-coordinated) cations are usually $2+$ and $4+$, respectively. The valence variation at the A cation position can cause distortion or displacement of the oxygen anion array, possibly resulting in distortion in the B-cation-centered octahedron. The B cation must have the flexibility to tolerate this effect, and the transition-metal elements are candidates for filling the B-cation position because of their multivalences and their special 3d and 4d electronic configurations. This is the reason that transition-metal oxides have perovskite-type structures (3). Perovskite-like structures can be sorted by the valence combination of the A and B cations as follows (4):

1. $\text{A}^{1+}\text{B}^{5+}\text{O}_3$ type, such as KNbO_3 , NaNbO_3 , LiNbO_3 and KTaO_3 .
2. $\text{A}^{2+}\text{B}^{4+}\text{O}_3$ type, in which the A^{2+} cations are alkaline-earth ions such as cadmium or lead, and the B^{4+} ions can be Ce, Fe, Pr, Pu, Sn, Th, Hf, Ti, Zr, Mo, and U. BaTiO_3 and PbTiO_3 are typical examples. These two compounds are well known for their remarkable ferroelectric properties (see later section).
3. $\text{A}^{3+}\text{B}^{3+}\text{O}_3$ type, such as GdFeO_3 , YAlO_3 , PrVO_3 , PrCrO_3 , NdGaO_3 and YScO_3 .
4. $\text{A}^{2+}(\text{B}^{3+}_{0.67}\text{B}^{6+}_{0.33})\text{O}_3$ type, such as $\text{Ba}(\text{Sc}_{0.67}\text{W}_{0.33})\text{O}_3$ and $\text{Sr}(\text{Cr}_{0.67}\text{Re}_{0.33})\text{O}_3$.
5. $\text{A}^{2+}(\text{B}^{2+}_{0.33}\text{B}^{5+}_{0.67})\text{O}_3$ type, for example $\text{Ba}(\text{Sr}_{0.33}\text{Ta}_{0.67})\text{O}_3$, and $\text{Pb}(\text{Mg}_{0.33}\text{Nb}_{0.67})\text{O}_3$.
6. $\text{A}^{2+}(\text{B}^{3+}_{0.5}\text{B}^{7+}_{0.5})\text{O}_3$, $\text{A}^{2+}(\text{B}^{2+}_{0.5}\text{B}^{6+}_{0.5})\text{O}_3$, $\text{A}^{2+}(\text{B}^{1+}_{0.5}\text{B}^{7+}_{0.5})\text{O}_3$, and $\text{A}^{3+}(\text{B}^{2+}_{0.5}\text{B}^{4+}_{0.5})\text{O}_3$ types, such as $\text{Ba}(\text{Sr}_{0.5}\text{W}_{0.5})\text{O}_3$, $\text{Pb}(\text{Sc}_{0.5}\text{Ta}_{0.5})\text{O}_3$ and $\text{Pb}(\text{Sc}_{0.5}\text{Nb}_{0.5})\text{O}_3$. The compounds, $\text{Pb}(\text{Mg}_{0.33}\text{Nb}_{0.67})\text{O}_3$, $\text{Pb}(\text{Sc}_{0.5}\text{Ta}_{0.5})\text{O}_3$ and $\text{Pb}(\text{Sc}_{0.5}\text{Nb}_{0.5})\text{O}_3$, are very important ferroelectric materials, and they are usually called “relaxors.”
7. $\text{A}^{2+}(\text{B}^{1+}_{0.25}\text{B}^{7+}_{0.75})\text{O}_3$ type, such as $\text{Ba}(\text{Na}_{0.25}\text{Ta}_{0.75})\text{O}_3$ and $\text{Sr}(\text{Na}_{0.25}\text{Ta}_{0.75})\text{O}_3$.
8. $\text{A}^{2+}(\text{B}^{2+}_{0.5}\text{B}^{5+}_{0.5})\text{O}_{2.75}$ and $\text{A}^{2+}(\text{B}^{3+}_{0.5}\text{B}^{4+}_{0.5})\text{O}_{2.75}$ that are anion deficient, such as $\text{Sr}(\text{Sr}_{0.5}\text{Ta}_{0.5})\text{O}_{2.75}$ and $\text{Ba}(\text{Fe}_{0.5}\text{Mo}_{0.5})\text{O}_{2.75}$.
9. $\text{A}^{2+}(\text{B}^{3+}_{0.5}\text{B}^{2+}_{0.5})\text{O}_{2.25}$.

It is apparent that the perovskite structures cover a large group of materials. Three questions are particularly interesting: What are the special properties of perovskites as far as smart materials are concerned? What is the relationship between these structures, for example, the structural evolution in perovskite, and what is the relationship between the cation valence and its coordination? The following analysis explores the answers to these questions.

STRUCTURES AND PROPERTIES

Ferroelectricity

Ferroelectric materials are candidates for robust non-volatile memories (5). Figure 2 gives a high-resolution transmission electron microscopy (TEM) image of BaTiO_3 oriented along [100] (or [001]), where the cations are in dark contrast and the contrast is directly related to the

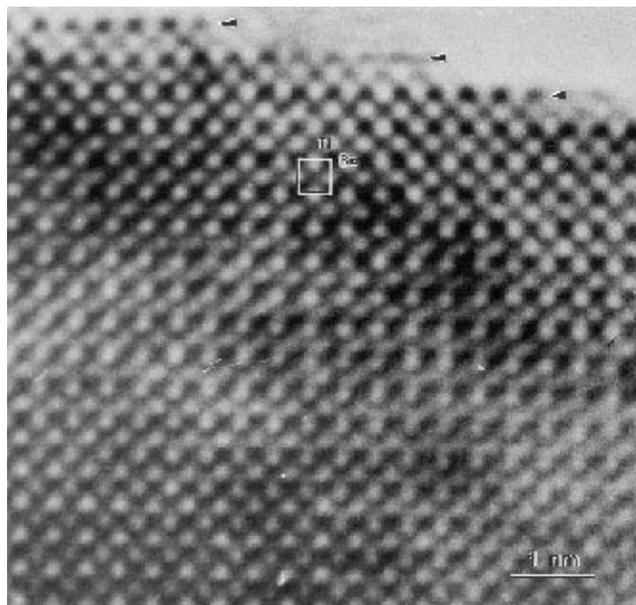


Figure 2. High-resolution transmission electron microscopy image of BaTiO_3 oriented along [100], showing the cation (in dark contrast) distribution in the crystal. The atom types can be clearly identified. At the top of the film, surface steps of one unit cell height are seen, and the termination layer is Ba–O.

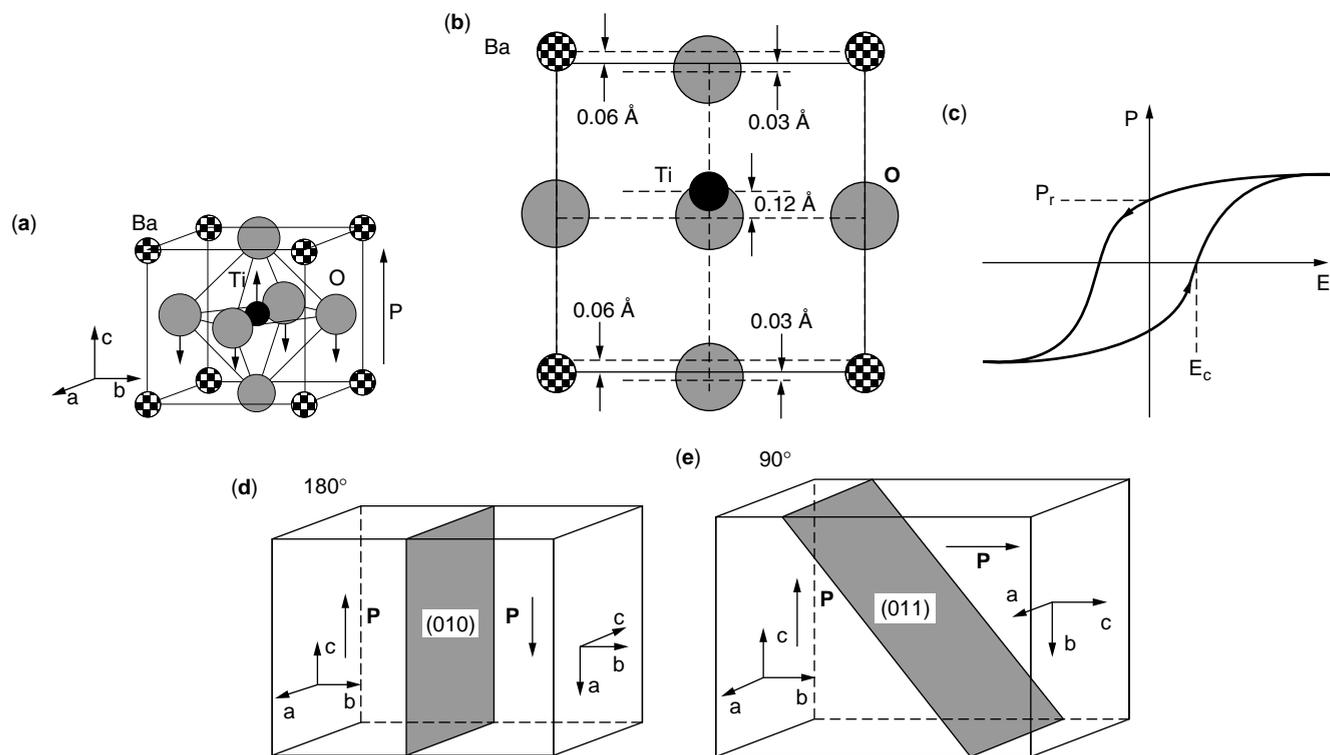


Figure 3. (a) Intrinsic spontaneous polarization in tetragonal BaTiO₃. (b) Displacements of the cations and anions in the tetragonal structure. (c) Schematic of a hysteretic loop for a ferroelectric material induced by an external electric field E , where P and E_c are the polarization and coercive fields, respectively. (d, e) 180° and 90° domain boundaries in BaTiO₃.

atomic number. The oxygen anions are not clearly resolved in the image because of its weak scattering power. At the top of the film, the last ending layer is the Ba–O layer, clearly indicating that the Ti atom strongly demands a complete octahedral coordination even at the boundary of the crystal. The octahedral coordination of Ti is at the root of ferroelectricity.

The Ti ion is surrounded by six oxygen ions in an octahedral configuration (Fig. 1a). BaTiO₃ has a cubic structure at $T > 120^\circ\text{C}$. For $5 < T < 120^\circ\text{C}$, it is tetragonal. In the low-temperature range of $-90 < T < 5^\circ\text{C}$, it has an orthorhombic structure, and for $T < -90^\circ\text{C}$, it is rhombohedral. Therefore, the structural transformation from centrosymmetric to noncentrosymmetric occurs at 120°C , and ferroelectricity occurs at $T < 120^\circ\text{C}$. Below the 120°C transition temperature, the oxygen and titanium ions are displaced to new positions (Fig. 3a,b), forming a tetragonal structure where $c/a = 1.01$ (3). A unilateral displacement of the Ti⁴⁺ ion against O²⁻ results in a dipole moment. When all of the dipoles of different domains point in the same direction, the material is *ferroelectric*. If the dipoles have equal strength but are aligned in an antiparallel configuration so that they cancel each other and the material does not exhibit a macroscopic dipole, it is *antiferroelectric*. If these dipoles cannot completely cancel each other, the residual dipoles add up, forming a macroscopic dipole, which is *ferroelectricity*.

The spontaneous alignment of dipoles that occurs at the onset of ferroelectricity is often associated with a

crystallographic phase change from a centrosymmetric, nonpolar lattice to a noncentrosymmetric polar lattice. If an external electric field is applied to the crystal, the population of the domains whose polarizations are parallel to the field increases, and those whose polarizations are antiparallel and not parallel to the field decrease. If the external electric field is removed, the domains cannot spontaneously compensate for each other again, and a remanent polarization P_r remains. To remove the remanent polarization, an oppositely oriented electric field whose field strength is E_c , called the *coercive field*, has to be applied to the crystal. The polarization hysteretic loop (Fig. 3c) is the basis of electric data storage using ferroelectric materials.

An increasing number of materials have been found that demonstrate spontaneous polarization. Lead titanate (PbTiO₃), which has the same perovskite structure as BaTiO₃, is ferroelectric. Other examples include Rochelle salt (potassium sodium tartrate tetrahydrate), KH₂PO₄, KH₂AsO₄; perovskites NaCbO₃, KCbO₃, NaTaO₃, and KTaO₃; ilmenite structures, LiTaO₃ and LiCbO₃; and tungsten oxide, WO₃.

Domains and domain boundaries can be formed in ferroelectric materials. The spontaneous polarization of the Ti and oxygen ions creates an electrostatic polarization P along the c axis. This anisotropic structural configuration can form 90 and 180° domain boundaries defined with reference to the orientations of the c axes or the P vectors that belong to the two crystal domains (Fig. 3d,e). The 90° domain boundary is just a (101) [or (011)] twin boundary of

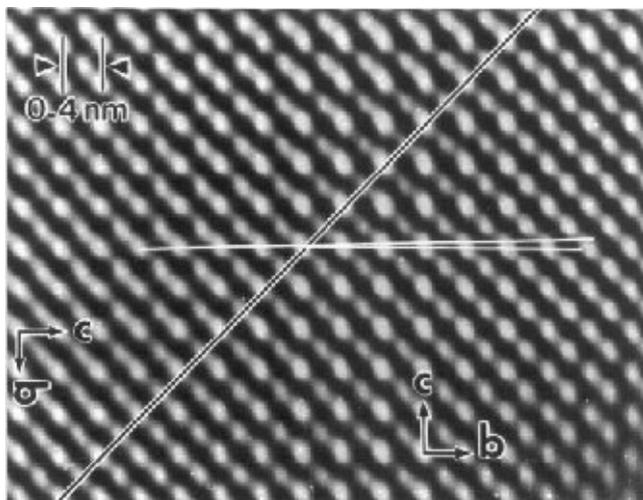


Figure 4. High-resolution transmission electron microscopy image of BaTiO_3 oriented along $[100]$, showing the presence of a (011) 90° domain boundary.

the tetragonal crystal (6). The \mathbf{P} polarization or c axis turns its direction by an angle $\alpha = \delta - 90^\circ$ across the twin or domain boundary, where δ is determined by the axial ratio of the tetragonal unit cell: $\delta = 2 \arctan(c/a)$; $\alpha = 0.57^\circ$ for $c/a = 1.01$.

Figure 4 is a high-resolution transmission electron microscopy image of a 90° domain boundary in BaTiO_3 . The domain boundary is parallel to the electron beam direction. A relative rotation α of the crystal lattice across the domain boundary (dark line) can be seen along the white lines traced along the rows of atoms. Domain structure, such as its sharpness, thickness, and size, is a key to many ferroelectric properties.

Another candidate for nonvolatile memories is $\text{SrBi}_2\text{Ta}_2\text{O}_9$ (Fig. 5). The structure is a c -axis stacking of perovskite units accommodated by bismuth oxide layers. The noncentr. symmetric structure produces the relative displacements of the cations and anions, for example, ferroelectricity.

Piezoelectricity

Piezoelectric Effect. *Piezoelectricity* is induced by a linear relationship between an electric field and mechanical strain. To illustrate piezoelectricity, consider a positively charged atom that is surrounded tetrahedrally by anions (Fig. 6). The center of gravity of the negative charges is at the center of the tetrahedron. By exerting pressure on the crystal along the cornering direction of the tetrahedron, the tetrahedron experiences distortion, the center of gravity of the negative charges and the position of the positive central atom no longer coincide, and an electric dipole is generated. If all of the tetrahedra in the crystal have the same orientation or some other mutual orientation that does not allow for cancellation among the dipoles, the crystal has a macroscopic dipole. The two opposite faces of the crystal have opposite electric charges. This is an electromechanical transducer that converts a mechanical force into an electric signal or vice versa.

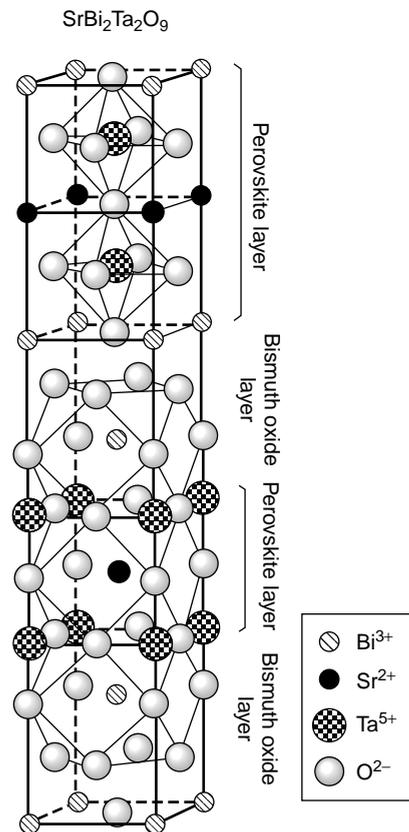


Figure 5. Structure of ferroelectric $\text{SrBi}_2\text{Ta}_2\text{O}_9$.

$\text{Pb}(\text{Zr}_x\text{Ti}_{1-x})\text{O}_3$ (PZT) is the most typical piezoelectric material (Fig. 7) (7). PZT has the perovskite structure. When cooling from high temperature, the crystal structure of PZT undergoes a displacive phase transformation and atomic displacements of $\sim 0.1 \text{ \AA}$. For titanium-rich compositions, the point symmetry changes from $m\bar{3}m$ to tetragonal $4mm$ at the Curie temperature. To maximize the polarization of the PZT phase, compositions near a second phase transition are chosen. At the Curie point (T_c), PZT converts from a paraelectric state that has the ideal cubic

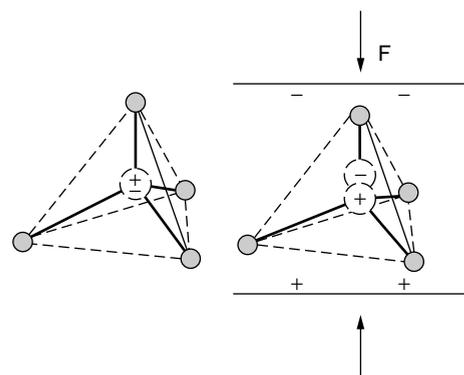


Figure 6. Mechanism of the piezoelectric effect. External pressure causes the deformation of a coordination tetrahedron that results in a shift of the gravity centers of the electric charges and creates a local polarization dipole.

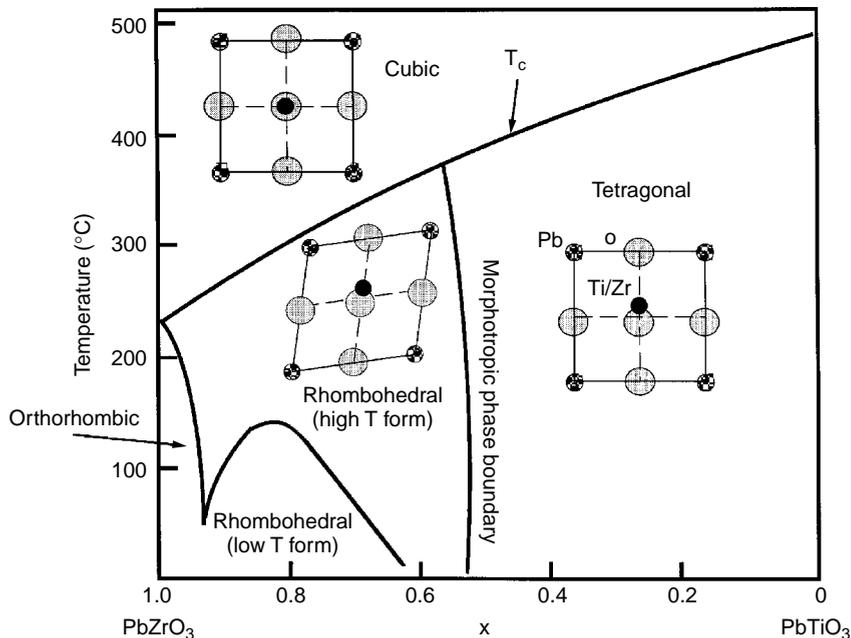


Figure 7. (a) Structural model of $\text{Pb}(\text{Zr}_x\text{Ti}_{1-x})\text{O}_3$ (PZT). (b) Phase diagram $\text{Pb}(\text{Zr}_x\text{Ti}_{1-x})\text{O}_3$ and the associated structural changes at the Curie temperature (T_c) and the morphotropic phase boundary (MPB). Compositions near MPB have 14 possible poling directions (six perpendicular to the planes and eight along the corners).

perovskite structure to a ferroelectric phase located near a morphotropic phase boundary (MPB) between the tetragonal and rhombohedral states. Very large piezoelectric coupling between electric and mechanical variables is attained near the MPB.

Parameters for Piezoelectric Ceramics. The proportional constants between the strain and the applied electric field are related by $\mathbf{x} = \mathbf{d}\cdot\mathbf{E}$, where \mathbf{d} is the *piezoelectric coefficient* tensor (8). The state of strain in the volume is fully described by a second-rank tensor, a “strain tensor,” and the state of stress by a stress tensor, again of second rank. Therefore, the relationships between the stress and strain tensors, that is, Young’s modulus or compliance, are fourth-rank tensors. The relationship between the electric field and electric displacement, the permittivity, is a second-rank tensor. In general, a vector (formally regarded as a first-rank tensor) has three components, a second-rank tensor has nine components, a third-rank tensor has 27 components, and a fourth-rank tensor has 81 components. However, not all of the tensor components are independent. The crystal symmetry and the choice of reference axes reduce the number of components. A poled ceramic has ∞ -fold symmetry in a plane normal to the poling direction. Therefore, the symmetry of a poled ceramic is described as ∞mm .

The convention is to define the poling direction as the 3 axis. The shear planes are indicated by the subscripts 4, 5, and 6 and are perpendicular to directions 1, 2, and 3, respectively (see Fig. 8). For example, d_{31} is the coefficient that relates the field along the polar axis to the strain perpendicular to it, and d_{33} is the corresponding coefficient for both strain and field along the polar axis. Shear can occur only when a field is applied at right angles to the polar axis so that there is only one coefficient, d_{15} . There are also piezoelectric coefficients corresponding to hydrostatic stress, for example, $d_h = d_{33} + 2d_{31}$.

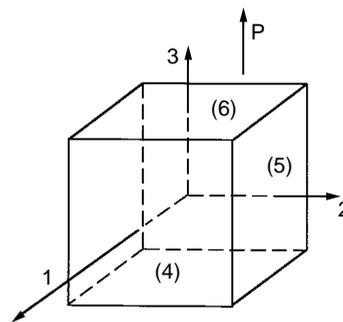


Figure 8. Labeling of reference axes and planes for piezoceramics (note that the polarization direction is upward).

Both intrinsic and extrinsic factors contribute to piezoelectric coefficients. The intrinsic effects from the distortions of the crystal structure under mechanical stress appear in Fig. 9. Under mechanical stress parallel to the dipole moment, an enhancement of the spontaneous polarization P along x_3 is induced. When stress is applied perpendicularly to that dipole moment, electric charges develop transversely. These are the d_{33} and d_{31} effects, respectively. When the dipole is tilted by shear stress, charges appear on the side faces, the d_{15} coefficient. The extrinsic contribution to the piezoelectric coefficient comes mainly from domain-wall motion.

Piezoelectric properties are described in terms of these parameters: dielectric displacement (\mathbf{D}), electric field (\mathbf{E}), stress (\mathbf{X}), and strain (\mathbf{x}). The direct piezoelectric effect relates polarization to stress and is used in sensors. The converse effect relates strain to electric field and is used in actuators. If the piezoelectric coefficient \mathbf{d} is assumed constant, the direct and converse effect can be written as

$$\mathbf{D} = \mathbf{d}\cdot\mathbf{X} + \epsilon^x\cdot\mathbf{E}, \quad (1)$$

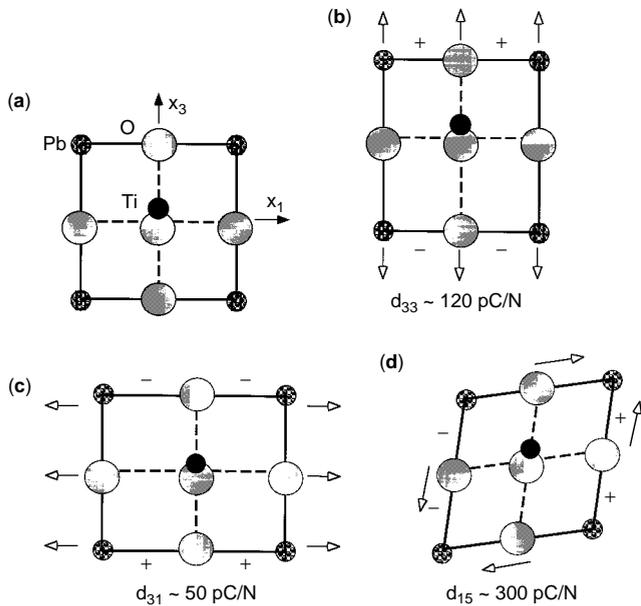


Figure 9. Schematics showing the relationship between the intrinsic piezoelectric effect and the piezoelectric coefficients in polar lead titanate.

and

$$\mathbf{x} = \mathbf{s}^E \cdot \mathbf{X} + \mathbf{d} \cdot \mathbf{E}, \quad (2)$$

where \mathbf{s} is elastic compliance and the superscripts denote the parameters held constant. The crystal structure and its distortion determine the piezoelectric coefficients (Fig. 9). Table 1 gives typical values of the parameters for some piezoelectric materials. Piezoelectric ceramics are probably the most important actuating materials (9).

Electrostrictive Perovskites

Electrostriction describes a quadratic relationship between mechanical strain and the square of the electric polarization. $\text{Pb}(\text{Mg},\text{Nb})\text{O}_3$ (PMN) is the most typical electrostrictive perovskite (10). PMN is cubic at high temperatures, and when annealed, transforms to a partially

ordered state. When cooled further, PMN passes through a diffuse phase transformation at room temperature where it exhibits very large dielectric and electrostrictive coefficients. Just below room temperature, it transforms to a ferroelectric rhombohedral phase. The piezoelectric d_{33} coefficient is the slope of the strain–electric field curve when strain is measured in the same direction as the applied field. Its value for $\text{Pb}(\text{Mg}_{0.3}\text{Nb}_{0.6}\text{Ti}_{0.1})\text{O}_3$ is zero at zero field, but it increases to a maximum of 1300 pC/N, three times larger than that of PZT, under a bias field of 3.7 kV/cm. This means that the electromechanical coupling coefficient can be tuned across a wide range to convert the transducer from inactivity to extreme activity. The dielectric constant of PMN depends on the bias field. Polarization saturates at a high field and causes a decrease in the dielectric constant. Therefore, the electrical impedance can also be controlled. This tunable property has its roots in so-called compositional microdomains. Because the ionic radii of Sc^{2+} , Ta^{6+} , Mg^{2+} , and Nb^{6+} in perovskites are close to each other, the long-range order is not favorable, but the shorter range order is preferable. The ordered phase has a double size in each axis of the perovskite unit cell and a face-centered-cubic structure. They are reexpressed as the $\text{Pb}_2\text{MgNbO}_6$ -order domain and the PbNbO_3 -disorder domain if the original compound is $\text{Pb}(\text{Mg}_{0.33}\text{Nb}_{0.67})\text{O}_3$ in which the valence states of Nb, Mg, and Pb are 5+, 2+ and 2+, respectively. High-resolution TEM reveals that the domains are about 3 nm in diameter and the ordered regions are small islands separated by narrow walls of niobium-rich PMN (11). The ordered regions, $\text{Pb}_2\text{MgNbO}_6$, have a negative charge, but the disordered areas, PbNbO_3 , have a positive charge. The ordered regions have to be very small to minimize the coulombic energy, but the distortion of perovskite is so small that the interfacial energy is negligible. The compositions of the ordered and disordered regions are different; therefore, the domains may be called compositional domains. The negatively charged ordered islands are enclosed by positively charged disordered layers to make the system neutral on a local scale and minimize the electrostatic energy.

Electrostriction is a 6×6 matrix, that relates strain to the square of the electric polarization. For a cubic crystal, Q_{11} , Q_{12} , and Q_{44} are the coefficients normally used.

Table 1. Typical Properties of Some Piezoelectric Materials^a

Material	T_c (°C)	ϵ_{33}^X	ϵ_{11}^X	d_{31}	d_{33} pC/N	d_{15}	s_{11}^E	s_{12}^E	s_{13}^E $\mu\text{m}^2/\text{N}$	s_{33}^E	s_{44}^E
α -Quartz ^{ab}	—	4.6	—	—	—	—	12.8	−1.8	−1.2	9.6	20.0
BaTiO ₃	130	1900	1600	−79	190	270	8.6	−2.6	−2.9	9.1	23
PZT A ^c	315	1200	1130	−119	268	335	12.2	−4.1	−5.8	14.6	32
PZT B ^c	220	2800	—	−234	480	—	14.5	−5.0	−6.7	17.8	—
PbNb ₂ O ₆	560	225	—	−11	80	—	29	—	−5 to −8	25	—
Na _{0.5} K _{0.5} NbO ₃	420	400	600	−50	160	—	9.6	—	—	10	—
LiNbO ₃ ^b	1210	29	85	−0.85	6	69	5.8	−1.2	−1.42	5.0	17.1
LiTaO ₃ ^b	665	43	53	−3.0	5.7	26	4.9	−0.52	−1.28	4.3	10.5
PbTiO ₃	494	203	—	−7.4	47	—	11	—	—	11	—

^aFrom Ref. 8.

^bSingle crystals.

^cPZT A and PZT B are two typical PZT materials that illustrate, in particular, the wide range of tunable properties.

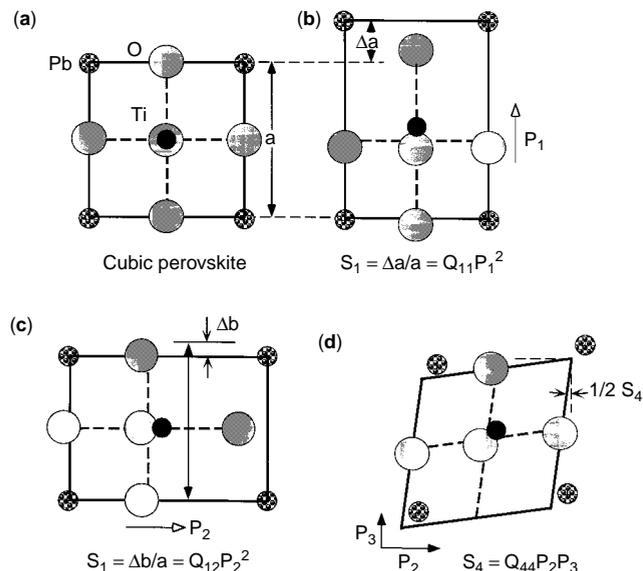


Figure 10. Electrostriction in cubic perovskite, showing the physical origin of electrostrictive coefficients Q_{11} , Q_{12} , and Q_{44} . The cross represents the lattice site of the atom.

Electrostrictive transducers rely on microregions that fluctuate in polarization. On average, the atoms reside in the ideal cubic sites but are continually shifting from these positions. Atomic scale descriptions of Q_{11} , Q_{12} , and Q_{44} are given in Fig. 10 (12).

Colossal Magnetoresistivity

Magnetoresistance is the change in electrical resistance R due to an external magnetic field. Positive or negative magnetoresistance refers to the increase or decrease of resistivity as the applied magnetic field changes. Two kinds of magnetoresistance are defined by the two possible orientations of the magnetic field H and the current J applied to the samples measured. $\Delta R_{//} = R_{//}(H) - R_{//}(0)$ when H is parallel to J , and $\Delta R_{\perp} = R_{\perp}(H) - R_{\perp}(0)$ when H is perpendicular to J . In ferromagnetic metals, magnetoresistance is anisotropic where $\Delta R_{//} > \Delta R_{\perp}$, positive $\Delta R_{//}$ and negative ΔR_{\perp} . They are saturated under a modest field (10 Oe for PermalloyTM), and the magnitude depends on the direction of the spontaneous magnetization. These materials are often referred to as anisotropic magnetoresistant materials. A typical value of $(R_{//} - R_{\perp})/R(0)$ for Permalloy TM is about 2% at room temperature.

Colossal magnetoresistance (CMR) has been observed in a perovskite-structured $(La,A)MnO_3$ ($A = Ca, Sr,$ or Ba) (13,14). A $(La,Ca)MnO_3$ epitaxial film prepared by laser ablation reportedly exhibited a CMR peak value of 127,000% at 77 K and 1300% at 260 K (13). Figure 11 gives a magnetoresistance curve of a new material, $La_{0.5}Ca_{0.23}MnO_{3-\delta}$, that displays CMR at room temperature (15). The resistance drops as the applied magnetic field increases, and it increases as the temperature increases. This property is very useful in sensor applications.

CMR magnetic oxides have a perovskite type crystal structure that has ferromagnetic ordering in the a,b plane

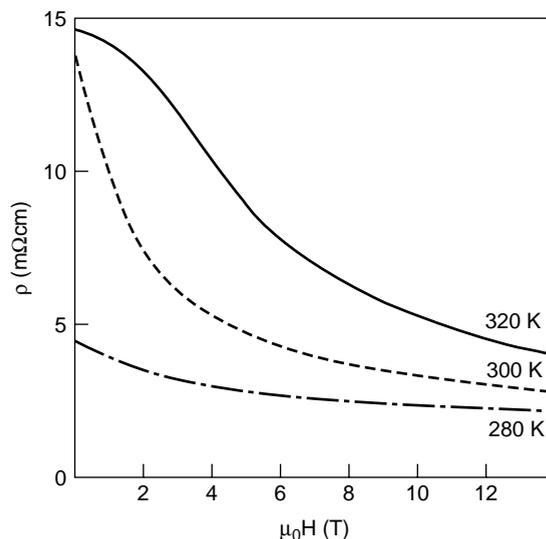


Figure 11. Magnetoresistance observed in an A site deficient $La_{0.5}Ca_{0.23}MnO_{3-\delta}$ around room temperature. The field was applied perpendicularly to the film.

and antiferromagnetic ordering along the c axis (Fig. 12b). The ferromagnetically ordered Mn–O layers of the a,b planes are isolated by a nonmagnetic La(A)–O monolayer. This spin-coupling structure is intrinsic. $La_{1-x}A_xMnO_3$ and $La_{1-x}A_xCoO_3$ compounds that have the extreme values $x = 0, 1$ are neither ferromagnetic nor good electrical conductors; they are semiconductors. Only compounds that have intermediate values of x are ferromagnetic, and the strongest ferromagnetism occurs in the range of $0.2 < x < 0.4$. Within this same range, the materials are electrically conductive.

Cation Doping-Induced Insulator-to-Conductor Transition

The electric conductivity of a perovskite, such as $LaMnO_3$, can be tuned by cation doping. We limit our discussion here to the case in which the dopants are distributed in a long-range order rather than as random impurities, as in conventional ceramics. The substitution of trivalent La^{3+} in $LaMnO_3$ by divalent A^{2+} is balanced by the conversion of Mn valence states between Mn^{3+} and Mn^{4+} and creation of oxygen vacancies as well. The following ionic structure of $La_{1-x}A_xMnO_3$ is proposed (16):



Whenever Mn^{3+} and Mn^{4+} are on neighboring Mn sites, the possibility exists of conductivity from electrons hopping from Mn^{3+} to the Mn^{4+} assisted by the oxygen anion. That this hopping current should be spin polarized is required for a process of two simultaneous electron hops (from Mn^{3+} onto O^{2-} and from O^{2-} onto Mn^{4+} , thus interchanging Mn^{3+} and Mn^{4+} ; this is called *double exchange* (17) (Fig. 13). The electron hopping away from the Mn^{3+} remembers the spin state that it has on the ion, as determined by *Hund's rule*, and the electron hopping onto Mn^{4+} must have the same spin state. This is only possible,

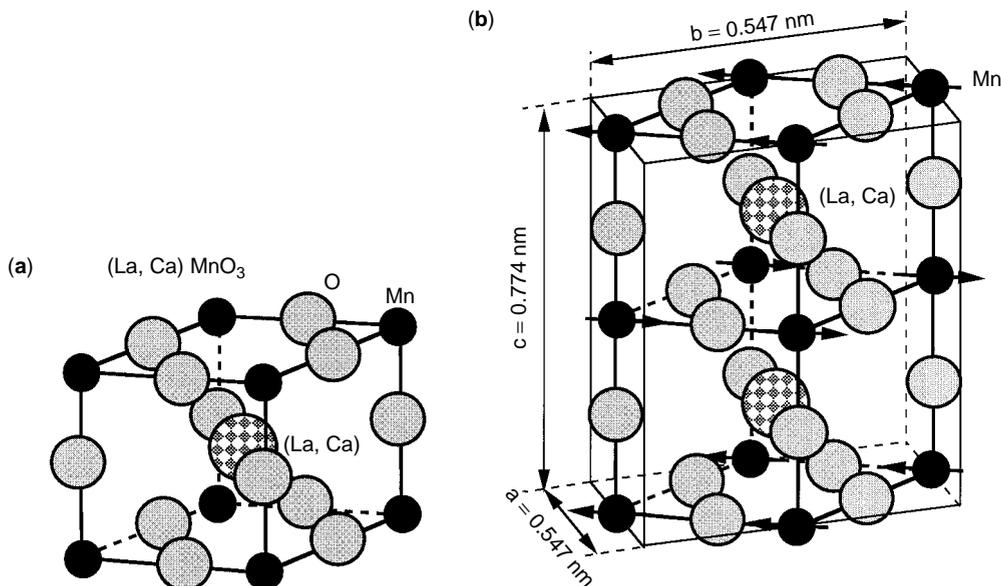


Figure 12. (a) Perovskite unit cell of $(\text{La,Ca})\text{MnO}_3$. (b) The newly introduced unit cell considering that the Jahn–Teller distortion results from magnetic ordering, so that the structure is orthorhombic and has space group $Pnma$. (c) High-resolution transmission electron microscopy image of $\text{La}_{0.67}\text{Ca}_{0.33}\text{MnO}_3$ oriented along $[100]$.

without violating Hund's rule on Mn^{4+} , if the net ion spins of the neighboring Mn^{3+} and Mn^{4+} are in the same spin direction. Hopping is impossible if the spin directions of the two states are antiparallel. In fact, the likelihood of electron hopping between two magnetic ions turns out to depend on the spin or transformation (18,19), which is characterized by a transfer integral (analogous to the transfer probability)

$$t_{ij} = b_{ij} \cos(\theta_{ij}/2), \quad (4)$$

where b_{ij} is a constant depending on the isolation between the ions and θ_{ij} is the angle between the direction of the spin ions. Thus, the electric resistance of the material is a function of its intrinsic magnetic order.

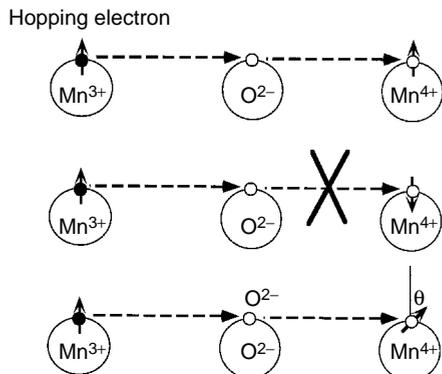


Figure 13. Schematic illustration of the double-exchange effect for parallel, antiparallel, and general cases. This effect is responsible for the electric conductivity of a doped perovskite.

Cation doping in PZT usually preserves the standard ABO_3 structure, in which the B-atom site can be occupied by either Zr or Ti. Cation doping, in some cases, can create superstructures. Figure 14a shows a high-resolution TEM image of $\text{La}_{0.33}\text{Sr}_{0.67}\text{CoO}_3$ (20), where the cation positions and types can be identified from the image. The oxygen atoms are not seen in the image due to the limited resolution power of the TEM and its weak scattering power. This structure is the result of cation substitution in a perovskite, and its structural model of fully coordinated anions is given in Fig. 14b. If the structural model is projected along $[100]$, the cation positions match the positions of the white dots observed in the image.

From Eq. (3), it can be seen that charge ordering of Mn^{3+} and Mn^{4+} can also occur in CMR materials. The ordered arrays of Mn^{3+} and Mn^{4+} can produce long, periodic, structural modulations, and this structure is important for coordinating the intrinsic electron hopping effect.

Ionic conductivity

Compounds that have perovskite or perovskite-related structures usually can be used as oxygen ion conductors, such as LaMnO_3 , and fuel cells, such as $\text{Fe}_{1-x}\text{Sr}_x\text{CoO}_3$. The sizes of A cations should be similar to that of oxygen anions, for example, Ba^{2+} (0.135 nm), Pb^{2+} (0.119 nm), Sr^{2+} (0.118 nm), Ca^{2+} (0.1 nm), and O^{2-} (0.14 nm). In LaMnO_3 , the La cations are smaller than oxygen anions (about 24%). Thus, oxygen migration is possible but limited. For faster diffusion of oxygen anions, it is best to create oxygen vacancies. Increasing temperature, of course, could create oxygen vacancies, but for lower temperature diffusion, the substitution of La^{3+} by divalent cations, such as Sr^{2+} and Ca^{2+} , can create oxygen vacancies [see Eq. (3)]. In

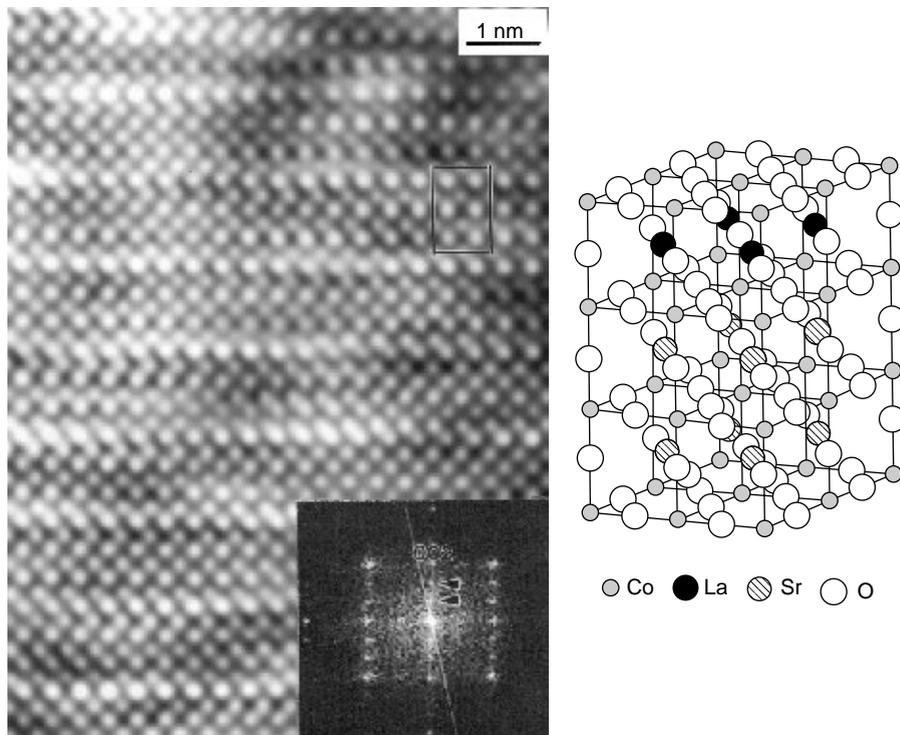


Figure 14. High-resolution TEM image of $\text{La}_{0.33}\text{Sr}_{0.67}\text{CoO}_3$, viewed along [100], showing the superstructure induced by cation substitution. The inset is a Fourier transform of the image. The structural model of $\text{La}_{0.33}\text{Sr}_{0.67}\text{CoO}_3$ is also given.

$(\text{LaO}_3)^{3-}$ (111) stacking layers (as displayed in Fig. 17a), the oxygen anions are closely packed around the A cation; a missing oxygen anion (due to creation of a vacancy) will cause strong coulombic interaction between the A cations (Fig. 15). This repulsive force makes the distance between the two adjacent A cations larger, and the A cations attract the neighboring oxygen anions to move in. If this distortion is too large, it may introduce buckling of the $(\text{LaO}_3)^{3-}$ layers and change the Bravais

cell. In other words, oxygen diffusion may induce phase transition.

Aurivillius phases have the general formula $\text{Bi}_2\text{O}_2(\text{A}_{n-1}\text{B}_n\text{O}_{3n+1})$ and can be described as layered structures that consist of perovskite blocks sandwiched between fluorite-like $(\text{Bi}_2\text{O}_2)^{2+}$ sheets (see Fig. 5). $\text{Bi}_2\text{NaNb}_2\text{O}_{8.5}$ is an $n = 2$ defect Aurivillius phase that exhibits good oxide ion conductivity at high temperature because of a structural phase transition (21). CaTiO_3 , SrCeO_3 , and

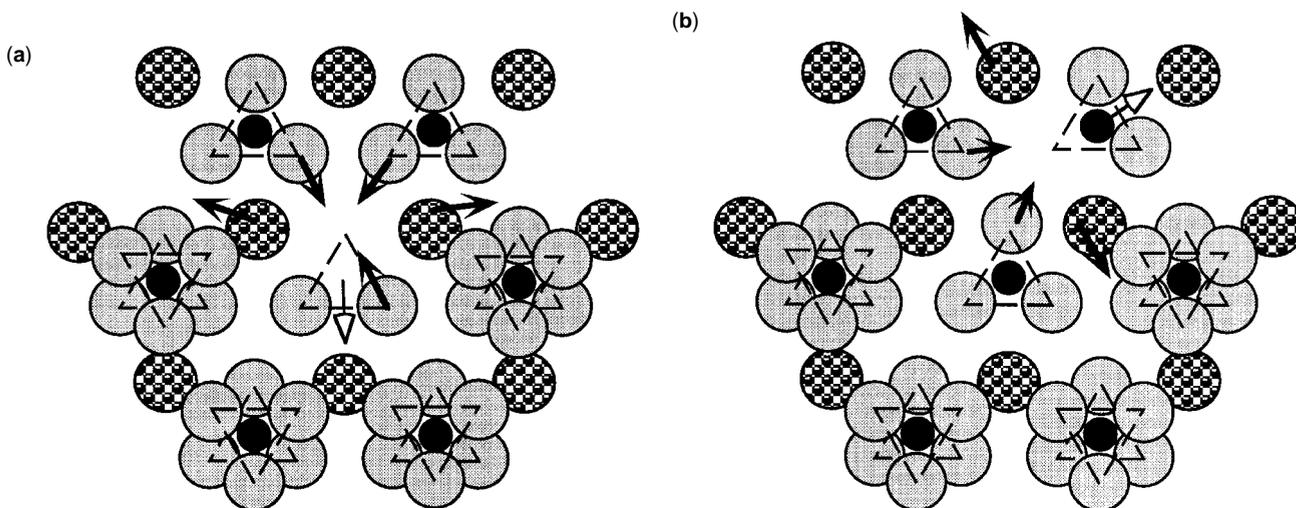


Figure 15. Migration of oxygen vacancies in perovskite and an ionic conductor. (a) Creation of one oxygen vacancy may lead to the diffusion of another oxygen anion toward the vacancy site, thus (b) the vacancy site is shifted, resulting in some distortion near the new vacancy site in the (AO_3) (111) stacking layer.

BaCeO₃ can be good conductors of ions and electrons or protons by simply changing the valence of the B cations to mixed valency.

Superconductivity

Superconductivity is the phenomenon of vanishing electrical resistance below the superconducting transition temperature (T_c). Superconductivity has important applications in power transmission, nuclear magnetic resonance, extremely strong magnets, and high-speed computing. High-temperature superconductors, such as YBa₂Cu₃O₇, Bi₂Sr₂CaCu₂O_{8+d} and Tl₂Ba₂Ca₂Cu₃O_x, are the focus of today's research (22). Most ceramic superconductors are based on the perovskite structure. The structure of YBa₂Cu₃O₇ will be used as an example to show the construction of its structure from anion-deficient perovskites (see later section).

A superconductor is characterized by three physical quantities. The first is the *critical transition temperature* T_c , below which superconductivity appears. The second is the *critical magnetic field* H_c , below which a superconducting body exhibits perfect diamagnetism and excludes a magnetic field. If the applied magnetic field is higher than H_c , the materials revert to the normal state. H_c is temperature-dependent. The third is the *critical current density* J_c , above which superconductivity is destroyed and the superconductor reverts to the normal state. Polycrystalline superconducting materials are limited mainly by low J_c due to the weak link between grain boundaries. Improving J_c in the field is the essential task of current research in superconductivity.

THE FUNDAMENTAL STRUCTURAL CHARACTERISTICS OF ABO₃ PEROVSKITE

Perovskite and related structures cover a large portion of smart materials, and their crystal structures can vary to a large extent. The key questions are do the perovskite-type structures have smart properties and are there any intrinsic connections among the structures? The answers may be found in the following areas: (1) nonstoichiometry

of the cation and/or the anions; (2) distortion of the cation configuration; and (3) the mixed valence and the valence mixture electronic structure. From the viewpoint of crystal structure, each of these features can be introduced by doping a third type of cation into the stoichiometric phase of a base structure. It is important to understand how oxygen stoichiometry and lattice distortion are introduced as a result of doping another type of cation that has different valence states. First, we explore the fundamental perovskite structure (2).

Vertex Sharing of Oxygen Octahedra

In the ABO₃ type structure, the cation B whose valence is 4+ is usually a transition-metal element that prefers to form a six-coordinated octahedron with its neighboring oxygen anions, and itself is located at the center. The octahedron is the basic unit of these structures. The geometric configuration of the arrangement of the octahedra that has the lowest interactive energy is a linear 180° vertex-sharing connection. If the octahedra are connected to each other at every vertex, they form a 3-D network (Fig. 16a). Because the oxygen at every vertex is shared by adjacent octahedra, the composition of this configuration is BO₃, and the unit cell is a simple cubic, as given in Fig. 16a. This structure, however, cannot exist unless B has a valence of 6+ because the valence charges are not balanced. Thus, a cation of valence 2+ must be introduced into the structure to balance the local excess negative charge. A vertex-sharing octahedral network, on the other hand, has a large cavity in the center of the unit cell. A cation of valence 2+ can occupy this cavity. Then, the unit cell still preserves the simple cubic structure, and the composition is ABO₃, simply the basic model of the perovskite structure (Fig. 16b). It is clear, therefore, that six-coordinated octahedra are essential structural and compositional building blocks, and sharing of all of the vertexes is required for the stoichiometry and the structure of the perovskite. Based on this ideal, we can outline some characteristics of the perovskite structure:

1. Any cation, that prefers to have six coordination could occupy the B position even if its valence is

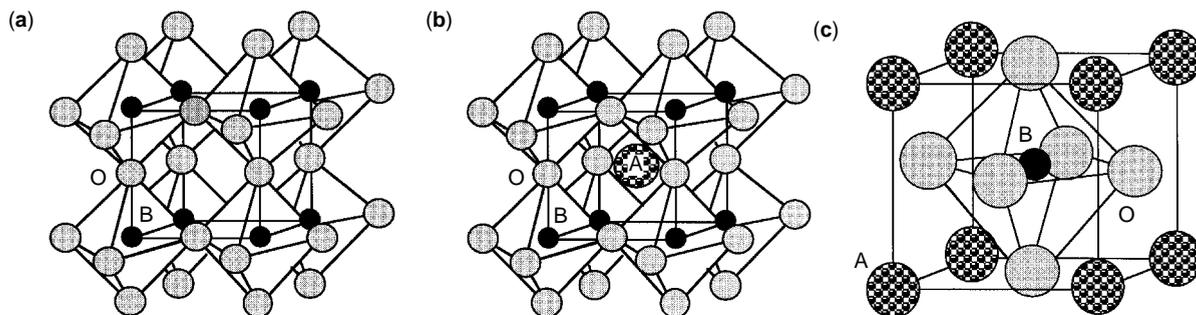


Figure 16. ABO₃ perovskite structure formed by corner-sharing octahedron chains. The corners of the Bravais cell are the B cations. (a) A postulated structure where the A cation is absent at the center, and (b) a structure that has the A cation. The anions shadowed by the octahedra are not shown for clarity. (c) ABO₃ perovskite structure drawn by setting the A cations as the corners of the Bravais cell, showing the BO₆ octahedron located in the A-cation cube.

- different from $4+$, but to balance the valence charges, the average valence at the B site usually equals $4+$.
- The A cation of valence $2+$ is expected to have a larger radius because, as a general rule, the valence increases as the radius of the cation decreases and vice versa. The coordination number of the A cation is 12. Elements that have a coordination number of more than six are likely to bond ionically. This is possible because the A cation in the perovskite structure is usually an alkaline-earth metal element. From this viewpoint, we may say that any cation, of higher ionicity or less polarity could occupy the position of the A cation, although its valence is different from $2+$. But to balance the charge, the average valence at this site is likely to be $2+$.
 - From 1 and 2, we can see that B cations have some degree of covalency, or in other words, they are easier polarize than A cations, and bonding between the B cation and the adjacent oxygen is stronger than that of the A cation. The octahedron of oxygen is the basic unit of the perovskite structure although it may be distorted, as is discussed later. Oxygen anions have significant influence on the cation valence if it can be changed. For example, a vacancy of oxygen can result in modifying the valence state of the B cation. The A cation may not be easily modified because it bonds ionically to the adjacent oxygen. However, if the A cation changes valence, the surrounding oxygen anions must be affected to balance the valence charges by creating vacancies. Because the oxygen anion cannot have a valence other than -2 , an oxygen deficiency must form to balance the local charge.

The roles of the A and B cations in the structural evolution of ABO_3 type perovskite can be understood as follows. *The A cation plays a key role in oxygen deficiency because of its stronger ionic interaction with oxygen anions.* We will see later that the A cation is closely packed together with oxygen. The B cation prefers six-coordination (i.e., an octahedron) although its valence can vary. The flexibility in the valences of the B cation makes the oxygen anion deficiency acceptable, and in a reverse process, the oxygen deficiency is autocontrolled by adjusting the valence of the B cation. This process makes the perovskite structure the most fascinating structural configuration for smart materials.

If the A cation changes its valence, for example, from $+2$ to $+3$, the oxygen anion must modify its occupancy to match the valence variation of the A cation and to balance the local charge, resulting in oxygen deficiency. But this change will feed back to the B cation, leading to disproportionation of the local valence state. The disproportionation of the B cation's oxidation state usually changes the electronic band structure of the perovskite compound, resulting in a transition from an insulator to a semiconductor, conductor, or even superconductor. CaMnO_3 is an insulator, but when the Ca (i.e., A^{2+} cation) is totally replaced by La($+3$), LaMnO_3 could be a conductor if the monovalence Mn^{4+} were replaced by mixed valent states of Mn^{3+} and Mn^{4+} . From the viewpoint of charge balance, if the valence of the Mn cations is $+3$ and $+4$, the relative content of the oxygen anions should be more than 3, in other words, its formula

should be LaMnO_{3+x} . Based on the crystallography of the perovskite structure illuminated discussed before, the excess oxygen anions have no place to locate because the existing oxygens have already been closely-packed in the structure. Therefore, if Mn^{3+} and Mn^{4+} must coexist in the system, the only choice is to change the A cation's valence state from La^{3+} to a mixture of La^{3+} and A^{2+} , a divalent cation. For example, if a small portion of La^{3+} is replaced by Ca^{2+} , the La^{3+} stoichiometry will change from 1 to $1-x$. Then we can have a perovskite type structure that has $+3$ and $+4$ mixed Mn cations at the B site.

Unit Cell by Taking the A Cation as the Origin

The perovskite structure has a simple cubic Bravais cell, in which the octahedra share corners and the origin of the Bravais cell is at the B cation, as shown in Fig. 16a. Alternatively, we can also take the A cation as the origin, and the unit cell is transformed into the form given in Fig. 16c. In this configuration, the A cations locate at the cubic vertices, and the oxygens occupy the face centers of the six faces to form an octahedron where the B cation is at the center. This geometric arrangement makes the oxygen form linear $\text{O}^{2-}-\text{B}^{4+}-\text{O}^{2-}$ triples parallel to the x , y , and z axes, and the oxygens are located at the centers of the squares formed by the A^{2+} cations. If an electric field is applied parallel to the z axis, for example, the $\text{O}^{2-}-\text{B}^{4+}-\text{O}^{2-}$ chain parallel to the z axis is polarized, but the $\text{O}^{2-}-\text{B}^{4+}-\text{O}^{2-}$ chains parallel to the x and y axes may not be disturbed, resulting in polarization of the crystal parallel to the z axis (ferroelectricity). Moreover, the displacement of the oxygen anions could cause a distortion in the configuration of the A cations, resulting in a change in the shape of the unit cell (the piezoelectric effect).

Oxygen Cubic Close Packing

Now, we reexamine the ABO_3 perovskite structure from a different viewpoint. We use the structural model of perovskite formed by the corner-sharing chains of octahedra, for example, the unit cell that has B cations as the origin, as shown in Figure 16b, in which the octahedron is sketched for clarity. The relationship between the oxygen anions and A cations in the $\{111\}$ stacking layers are revealed in Fig. 17a, where an A cation is surrounded by six oxygen anions and forms a closely packed structure. Note that three oxygen atoms form a triangular unit. Naturally, the A cations also form a hexagonal array, but are separated by the triangularly packed oxygen anions. The chemical composition of this layer is AO_3 . Because the charge of A is $2+$, the valence charge of the layer is $(\text{AO}_3)^{4-}$. To balance the local negative charge, a layer that has a charge of $4+$ must be introduced. Therefore, a B^{4+} cation layer (without oxygen anions) should be the next stacking layer (Fig. 17b). The A cation has three stacking positions, indicated by α , β , and γ , respectively, and the B cation has similar stacking. An A cation hexagon contains six apex-sharing oxygen triangles. This characteristic of the fundamental stacking layer $(\text{AO}_3)^{4-}$ determines the features of the perovskite structure. The $(\text{AO}_3)^{4-}$ and the B^{4+} layers are stacked together to form a combined layer, and the perovskite structure is obtained by stacking the new layers following a sequence of $\alpha \beta \gamma$ by translating the cation

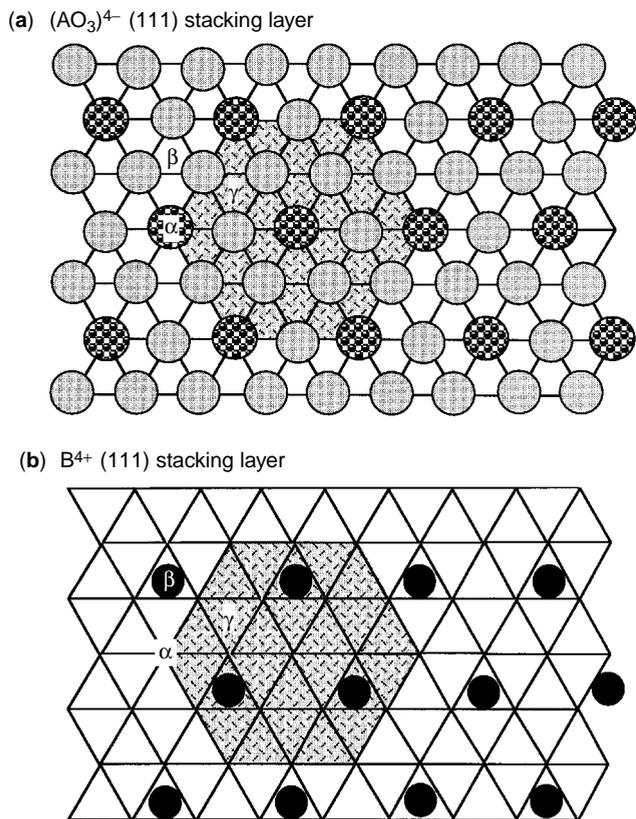


Figure 17. (a) The fundamental $(\text{AO}_3)^{4-}$ (111) close-packing layer, and (b) the B^{4+} (111) close-packing layer in the perovskite structure. The shadowed area represents the hexagonal unit from stacking the two layers into one. α , β , and γ represent the three cation stacking positions for forming the three-dimensional perovskite structure.

sites (similar to the stacking to form a face-centered-cubic lattice).

Anion Close Packing and Formation of Tetrahedra and Octahedra

The introduction of oxygen vacancies in the unit cell can create a wide range of perovskite structures. If one of the oxygen sublattices is vacant, the layer composition is $(\text{AO}_2)^{2-}$ (see Fig. 18b). If two of the oxygen sublattices are vacant, the layer composition is AO ; thus, the layers are neutral, provided the valence of cation A is 2+. There is no electrostatic force attracting the B^{4+} cation layers, and thus, the perovskite structure cannot be formed. If oxygen vacancies are partially formed in the $(\text{AO}_3)^{4-}$ layer, the close-packing layer should be $(\text{AO}_{3-x})^{(4-2x)-}$ (Fig. 18c,d where $x = 0.67$ and $x = 0.33$, respectively). Therefore, there are two choices to have oxygen vacancies in the three oxygen sublattices. If one oxygen vacancy is formed in the oxygen sublattice, it can locate in different layers and/or different sublattices of oxygen. The symmetry of oxygen triangles makes vacancies possible at different corners of the triangles that belong to alternate stacking layers. This creates the five- and four-coordinated oxygen polyhedra, shown in Fig. 19, where the structure of the B cations does not change, but their valence states may change to balance the local charge. The A cations in the

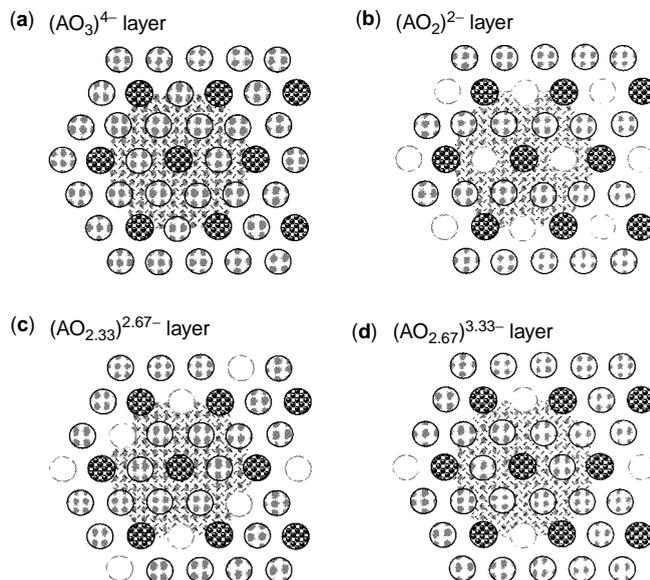


Figure 18. Two-dimensional oxygen sublattices in the $(\text{AO}_3)^{4-}$ layers (a) no vacancy $[(\text{AO}_3)^{4-}]$, (b) one vacancy $[(\text{AO}_2)^{2-}]$, and (c, d) partial vacancies.

$(\text{AO}_3)^{4-}$ layer may change their valence states and/or their total number in the layer (e.g., stoichiometry). Therefore, the substitution of the A cations by an element of different valence may be the optimum choice to induce oxygen deficiency and/or the mixed valence states of the B cations. This is the tailoring characteristic of the perovskite structure.

In the ABO_3 structure, the A cations are closely packed with the oxygen anions and are dominated by ionic bonding. The B cations may be regarded as the dependent subordinations that rely on the structure of the $(\text{AO}_3)^{4-}$ layer. Bonding of B cations with the surrounding oxygens may be ionic, covalent, or partially covalent. If the $(\text{AO}_3)^{4-}$ layer has defects, such as vacancies, stacking these layers will form distorted or deficient oxygen octahedra. Thus, the B cation must have the flexibility of modifying its valence state to balance the local charge. In other words, if we intend to modify the valence of the B cation, changing the structure surrounding the A cations via doping is recommended. This is an important principle for modifying the structure and properties of perovskites.

ANION-DEFICIENT PEROVSKITE STRUCTURAL UNITS—THE FUNDAMENTAL BUILDING BLOCKS FOR NEW STRUCTURES

Perovskite types of structures are the basis of many oxides, in which oxygen-deficient perovskites are a key group of materials that possess functionality. Anion-deficiency can change the coordination number of the B cation octahedra and the size and type of the Bravais unit cell of the compounds. It is important to elucidate the possible types of anion-deficient perovskite structures to enhance insight into fabricating new materials.

As shown in Fig. 18, there are three equivalent stacking positions for the layers parallel to the (111) plane. Figure 19a shows an $\alpha\beta$ stacking of the $(\text{AO}_3)^{4-}$ layers (the

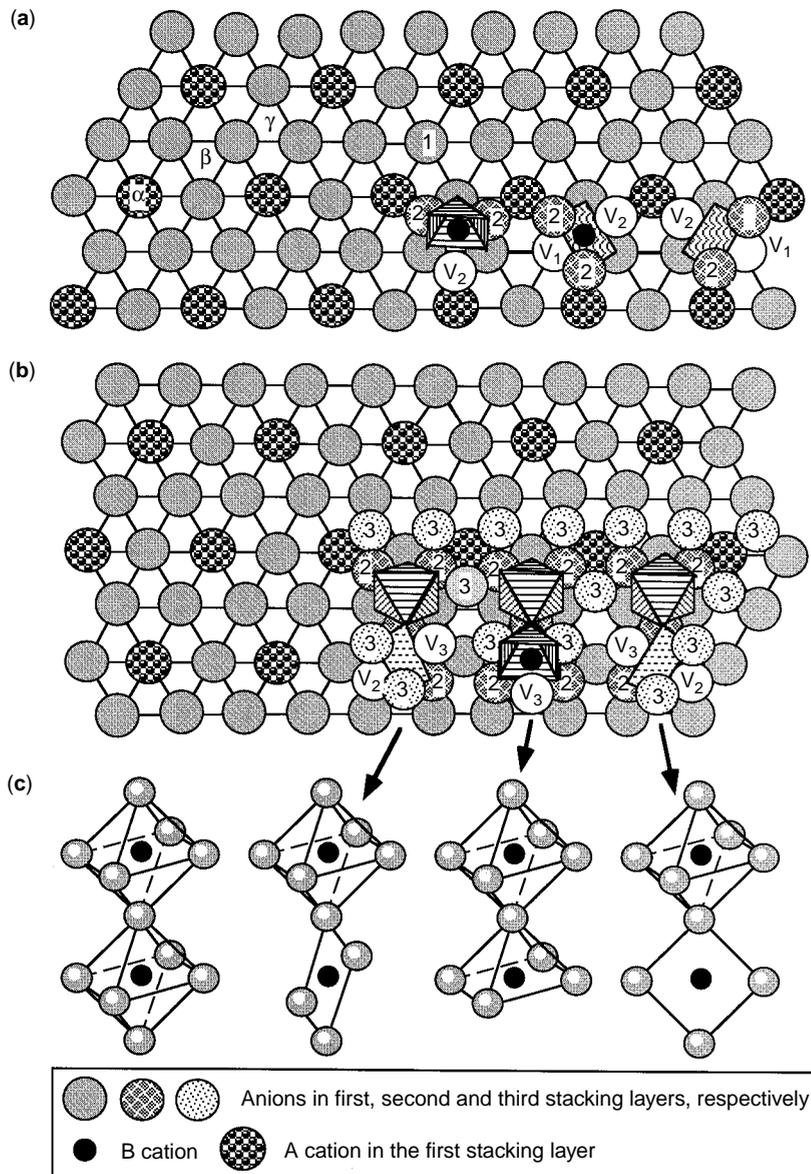


Figure 19. The local oxygen vacancy arrangements and the stacking of the $(\text{AO}_3)^{4-}$ type layers (with vacancies). (a) Single and double oxygen vacancies in BO_6 octahedron, where 1, 2, and 3 represent the anions in the α , β , and γ layers, respectively. (b, c) Connections of the octahedra without, with one, and with two oxygen vacancies.

α , β , and γ positions are indicated in the left part of this figure). The octahedra are formed by superimposed oxygen triangles that belong to the α and β layers (this is most easily seen in Fig. 18). If an oxygen vacancy is created in the β layer, as indicated by V_2 , a five-coordinated polyhedron is formed. By the same token, oxygen vacancies can be created in the octahedra formed by the superposition of the β and γ layers, as shown in Fig. 19a,b, where the sequence numbers 1, 2, and 3 represent the anions in the α , β , and γ layers, respectively. As discussed before, any $(\text{AO}_3)^{4-}$ stacking layer can have oxygen vacancies. A five-coordinated polyhedron is formed if only one oxygen vacancy is created in the β layer, and the α layer is perfect. If two oxygen vacancies (V_1 and V_2) are created in the two adjacent layers within the same octahedron, a four-coordinated square is formed, as shown in Fig. 19a. It is impossible to have more than two vacancies in one octahedron because four apexes are the minimum to form a 3-D connection for the

perovskite type of frame. The geometric assemblies of the four-, five-, and six-coordinated A cations are given in Fig. 19c.

Figure 20a,b shows the BO_5 and BO_4 structures, respectively, in an octahedral sheet. The lack of one oxygen at a vertex forms a five-coordinated unit (Fig. 20a), and the lack of two oxygens located at two opposite vertices results in a four-coordinated square (Fig. 20b). The lack of three oxygens in one octahedron forces the unit to break its connections to other octahedra. Figure 20c shows a cluster constructed on the basis of the symmetrical distribution of two oxygen octahedra, resulting in four 5-coordinated polyhedra and two 4-coordinated squares. It has the composition AB_8O_{30} , and it demonstrates how the different coordinated polyhedra may be connected to each other in 3-D.

Figure 20d–n shows eight possible interconnecting configurations of the five- and four-coordinated polyhedra (2). Any of these types of anion-deficient perovskite-like

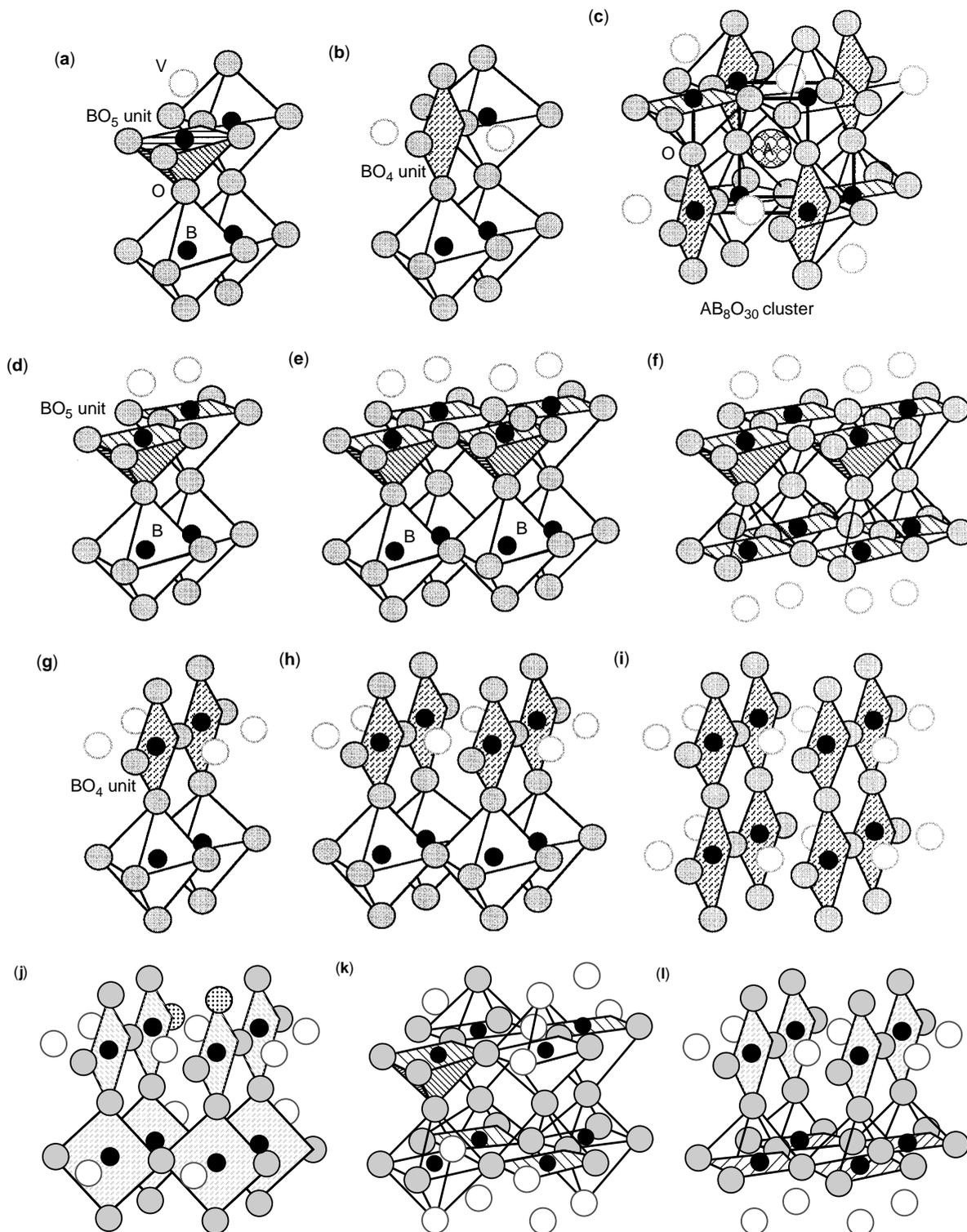


Figure 20. A total of 14 perovskite-type structural modules that have oxygen vacancies, where the oxygen anions are represented by different patterns to distinguish their stacking layers and V stands for vacancy. The combination of these modules can reproduce the crystal structures of many compounds (see text).

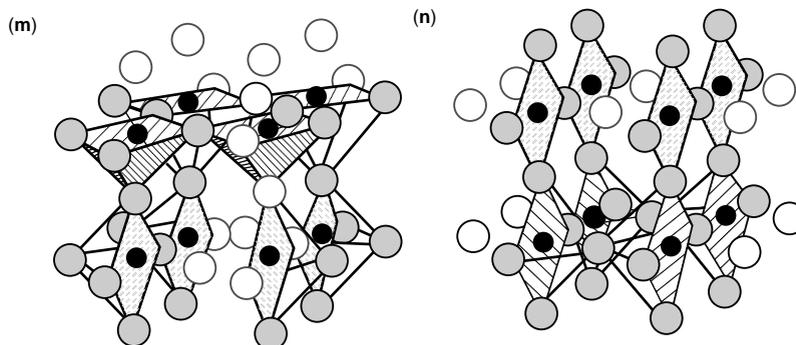


Figure 20. (Continued)

structural modules may be connected to the ideal perovskite structure's unit cells to form a new structure. The compound YBaCuFeO_5 (23), for example, has a structure that contains CuO_5 and FeO_5 units, corresponding to configuration (f) shown in Fig. 20. $\text{LaBa}_2\text{Cu}_2\text{TaO}_8$ (24) and $\text{Ba}_2\text{La}_2\text{Cu}_2\text{Sn}_2\text{O}_{11}$ (25) belong to the module shown in Fig. 20(l). The idealized structure of this compound contains TaO_6 or SnO_6 octahedra and five-coordinated CuO_5 . The structure of $\text{YBa}_2\text{Cu}_3\text{O}_7$ (26) is built by the unit in Fig. 20(l), and it contains five-coordinated CuO_5 and four-coordinated CuO_4 squares. The compound $\text{La}_2\text{Ni}_2\text{O}_5$ is constructed by the (h) type in Fig. 20, and it contains NiO_6 octahedra and NiO_4 four-coordinated squares. These types of anion-deficient perovskite-like units can be combined with tetrahedra, octahedra, or others including themselves to form a variety of different structures. The compound $\text{Ca}_2\text{Mn}_2\text{O}_5$ (27), for example, contains distorted MnO_5 units similar to the (m) type unit shown in Figure 20. The modules shown here are based the squares, tetrahedra, square-based pyramids (half-octahedra) and octahedra. These are the most fundamental "bricks" for constructing anion-deficient perovskite structures.

It must be pointed out that although the structural modules proposed in Fig. 20 assume that the atom sites are the same as in perfect perovskite structures, in practice, lattice relaxation/distortion is possible due to unbalanced anion coordination and oxygen vacancies. Fine-tuning of the structure must rely on quantitative fitting of X-ray or neutron diffraction data.

STRUCTURAL EVOLUTION IN THE FAMILY OF PEROVSKITES

High-Temperature Superconductors

In high T_c superconductors, the B cation is Cu that has +1, +2, and +3 valence states and coordination numbers 2, 3, and 4 for Cu(I) and 4 and 6 for Cu(II) and Cu(III). If the A and B cations in ABO_{3-x} have valences +3 and +2, respectively, the $(\text{AO}_3)^{4-}$ layers must be replaced by $(\text{AO}_{3-x})^{(3-2x)-}$. Because the Cu cation (the B cation) may have valence +1, +2, and +3, the perovskite-type cell that contains oxygen vacancies gives the Cu cation the possibility of disproportionating its valence from

+2 to +1 plus +3. Figure 21a gives the module shown previously in Fig. 20(l). The B cation in this module has two types of coordinations: BO_5 and BO_4 . If we flip this module over, as shown in Fig. 21b, and combine these two by superimposing the BO_4 units of the two, we obtain the new module given in Fig. 21c. This new module is the building block for the Y–Ba–Cu–O system of high T_c superconductors and introduces Ba cations between the modules. The structure of $\text{YBa}_2\text{Cu}_3\text{O}_7$ is given in Fig. 21d. Combining BO_4 units and the $\text{YBa}_2\text{Cu}_3\text{O}_7$ module (Fig. 21d) produces the structure of $\text{YBa}_2\text{Cu}_4\text{O}_8$ (Fig. 21f). Stacking $\text{YBa}_2\text{Cu}_3\text{O}_7$ and the $\text{YBa}_2\text{Cu}_4\text{O}_8$ modules creates the structure of $\text{Y}_2\text{Ba}_4\text{Cu}_7\text{O}_{15}$ (Fig. 21g).

Anion Deficiency-Induced Brownmillerite Structure

As we discussed before, the ABO_3 type perovskite has a fundamental stacking layer $(\text{AO}_3)^{4-}$ in which the A cations play a key role in creating oxygen vacancies. In general, the A cations have the valence +2 and an ionic character, which means that their coordination numbers are higher than six, for instance, 12. If the A cations have the valence +3 and an ionic character, the oxygen in the $(\text{AO}_3)^{4-}$ layers must be modified by creating oxygen vacancies. The charge of the layer is $(\text{AO}_3)^{3-}$; thus, the B cation layers must be modified to balance the local excess positive charge. The B cations usually have polarization or partial covalence, and they have favorable coordination numbers for reducing energy. If the valence of partial A cations is changed from +3 to +2, the negative charge of (AO_3) layers increases from -3 to $-(3+x)$, where x depends on the relative population of A^{2+} . Simultaneously, part of the B cations should modify their valence state from +3 to +4 to balance the local charge. The percentage of B cations whose valence is modified is related to the x value, as $\text{La}_{1-x}\text{Sr}_x\text{MnO}_3$. Mn^{3+} and Mn^{4+} cations are both favorable for six-coordinated octahedra, but the situation is different for Co cations. Both Co^{3+} and Co^{4+} can have six and four coordinations. Four-coordinated Co^{4+} requires the presence of oxygen vacancies; thus, the (AO_3) layer is replaced by $(\text{A}_{1-x}^{3+}\text{A}_x^{2+}\text{O}_{3-y})^{(3+x-2y)-}$.

On the other hand, if some B cations are reduced and their valence state is decreased from +3 to +2, the $(\text{AO}_3)^{3-}$ layers are required to compensate for the excess negative charge, resulting in oxygen vacancies and the change in

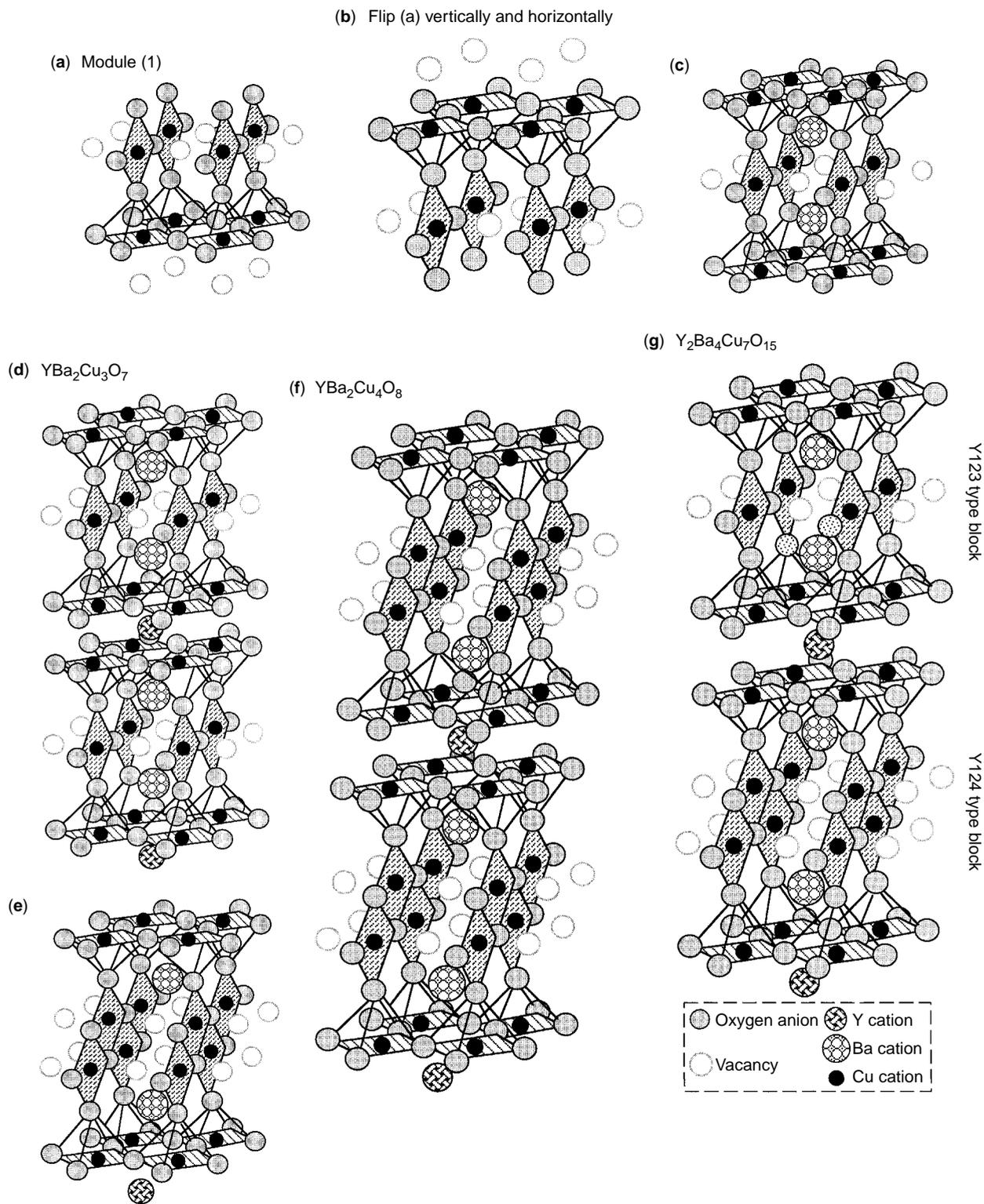


Figure 21. Evolution of the oxygen-deficient perovskite modules into the crystal structures of the Y–Ba–Cu–O system. (a) The module given in Fig. 20(l); (b) the module is flipped over vertically and horizontally. (c) The modules in (a) and (b) are combined by superimposing the BO_4 units to form a new module, which is the structural building block for $\text{YBa}_2\text{Cu}_3\text{O}_7$. (d) The structure of $\text{YBa}_2\text{Cu}_3\text{O}_7$. (e) A new module created by combining the two modules in (a) and (b) to share the edges of the BO_4 units. This new module is the building block of the structure of $\text{YBa}_2\text{Cu}_4\text{O}_8$. (f) The structure of $\text{YBa}_2\text{Cu}_4\text{O}_8$. (g) The structure of $\text{Y}_2\text{Ba}_4\text{Cu}_7\text{O}_{15}$ is a combination of the modules in (c) and (e).

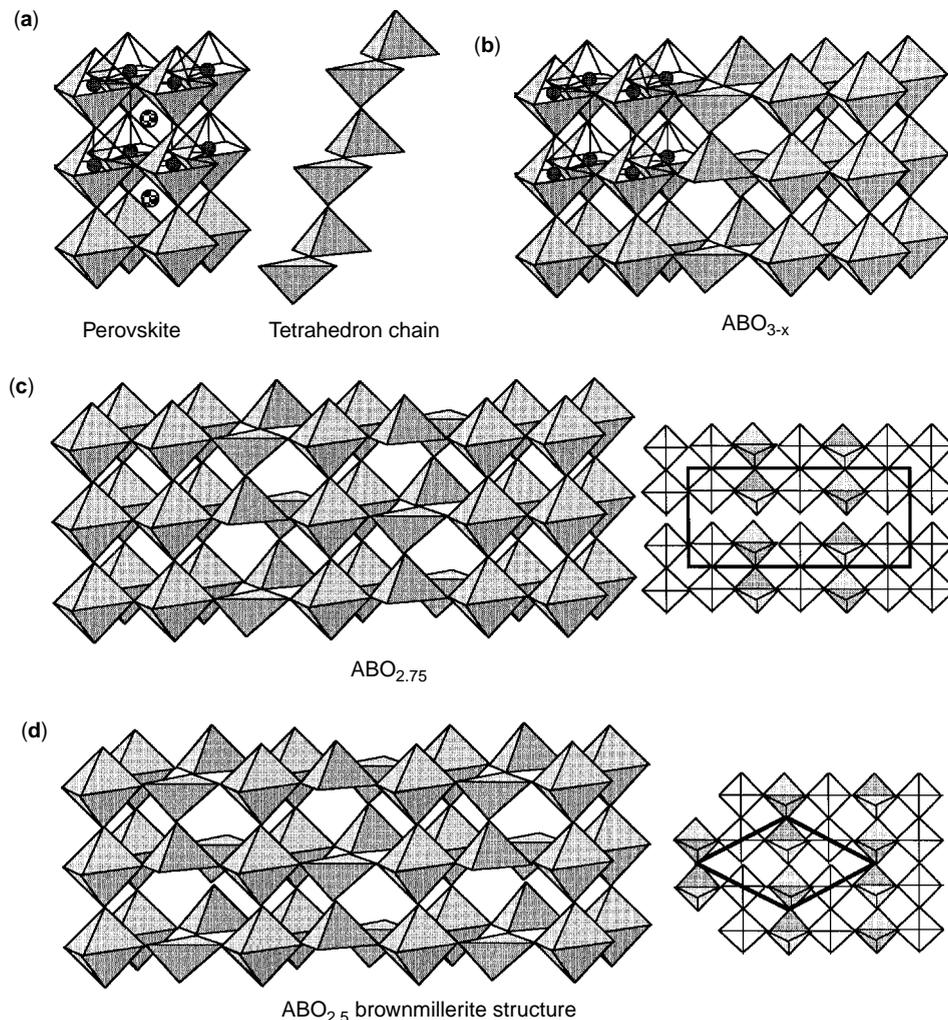


Figure 22. Structural evolution from perovskite and tetrahedron chains to form ABO_{3-x} , $ABO_{2.75}$, and $ABO_{2.5}$ (brownmillerite) structures owing to ordered anion deficiency. In all of these structures, the A and B cation lattice remains the same as that in standard perovskite, although for clarity they are not shown in the models; the only change is the introduction of anion vacancies.

the coordination number of the B cations. If the percentage of reduced B cations is small, the perovskite structure still holds. If the percentage of reduced B cations reaches an upper limit, the perovskite structure has to be changed to another structure that might be related to perovskite. The structural evolution from perovskite into the brownmillerite structure is an example.

$A\text{CoO}_3$ ($A = \text{La, Pr, Nd, Gd}$) perovskite is a typical example. Methane gas can be oxidized by LaCoO_3 above 1000°C . This means that LaCoO_3 can release lattice oxygen (28,29). Co^{2+} cations can have coordination numbers of six, five, and four. During the reduction process, Co^{3+} can be reduced to Co^{2+} and Co^0 . The coordination is changed from octahedral to tetrahedral. During the reduction process, the anion framework should hold, but it can have vacant sites. The perovskite unit and the possible corner-sharing tetrahedron chain are shown in Fig. 22a. When these chains are connected to neighboring octahedra, the remaining two corners of each tetrahedron are shared. If we randomly

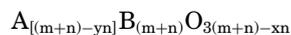
insert these tetrahedra chains into the perovskite structure (Fig. 22b), the compound is ABO_{3-x} . As the number of tetrahedron chains increases and reaches a number at which two octahedron slabs (that have the thickness of the perovskite unit cell) and one corner-sharing octahedron slab are separated by a “slab” of the tetrahedron chains, as shown in Fig. 22c, the $ABO_{2.75}$ structure is formed. Ordered structures of $A^{2+}(B_{0.5}^{3+} B_{0.5}^{4+})O_{2.75}$ and $A^{2+}(B_{0.5}^{2+} B_{0.5}^{5+})O_{2.75}$ can be formed. If the octahedron slab and the slab of the tetrahedron chains are stacked alternately via corner-sharing, the brownmillerite structure $ABO_{2.5}$ is constructed (Fig. 22d). As the relative number of the tetrahedron chains increases, the interaction between A cations also increases. If the number of the tetrahedron slabs is more than that of the octahedron slabs, the system will be unstable. Then, LaCoO_{3-y} is likely to be reduced to two phases: La_2O_3 and CoO . In other words, all of the Co^{2+} will have tetrahedral coordination, and there is no octahedral-coordinated Co^{2+} .

Tailoring Perovskite-Related Structures

In general, compounds that have perovskite-like structures can be represented by the chemical formula $A_mB_mO_{3m-x}$. In this formula the $(AO_3)^{4-}$ and the B^{4+} layers are the basic stacking layers, and they are stacked alternately following the sequence $\alpha\beta\gamma$, respectively, as defined by the three positions shown in Fig. 17. If the compound contains m layers of $(AO_3)^{4-}$ and m layers of B^{4+} and they are stacked together in the sequence $\alpha\beta\gamma$, respectively, alternately, we can have $A_mB_mO_{3m}$, such as ABO_3 when $m = 1$, the perovskite structure. If oxygen vacancies are introduced, an $(AO_3)^{4-}$ layer may have one oxygen sublattice possessing a vacancy, and the layer $(AO_3)^{4-}$ is transformed into $(AO_2)^{2-}$. If these types of anion-deficient layers are mixed with the ideal $(AO_3)^{4-}$ layers, the new compounds should have the formula $A_mB_mO_{3m-n}$ (where $n \leq m$), in which n represents the number of the $(AO_2)^{2-}$ layers that contain sublattices that have oxygen vacancies. $Ca_2Mn_2O_5$ ($m = 2$ and $n = 1$), $YBa_2Cu_3O_6$ or $YBa_2Cu_3O_8$ ($m = 3$ and $n = 3$), $YBa_2Cu_3O_7$ ($m = 3$ and $n = 2$), $LaBa_2Cu_2TaO_8$ ($m = 3$, $n = 1$), and $Ba_2La_2Cu_2Sn_2O_{11}$ ($m = 4$ and $n = 1$) are typical examples.

Furthermore, if the $(AO_3)^{4-}$ layer lack only a portion of the oxygens, the composition and the charge of the layer becomes $(AO_{3-x})^{(4-2x)-}$. If n and m are the numbers of the $(AO_{3-x})^{(4-2x)-}$ type and the $(AO_3)^{4-}$ type layers, respectively, the compounds formed by stacking these layers and a total of $(m+n)$ B cation layers alternately, should be $A_{m+n}B_{m+n}O_{3(m+n)-xn}$. For example, when $n = m = 1$, it is $A_2B_2O_{6-x}$ ($x < 1$). These two types of oxygen-deficient perovskite compounds have been found. Compounds that have $(AO_{3-x})^{(4-2x)-}$ layers may be more stable than those that have $(AO_2)^{2-}$ layers. The high T_c superconductor is an example (21).

Compounds that have some $(AO_2)^{2-}$ layers may have an integral number of oxygens, but the compounds that have $(AO_{3-x})^{(4-2x)-}$ layers may have a nonintegral number of oxygens. As discussed before, the A cation can be substituted by cations that have different valences or by partial cation vacancies to satisfy the local electron orbital and charge balance requirements. These results may help us to understand the structure of oxide functional materials. Most of the useful compounds that have functionality contain anion-deficient perovskite-like structural units, especially those that contain $(AO_{3-x})^{(4-2x)-}$ layers. The 14 structural configurations shown in Fig. 10 are the fundamental building blocks for building these structures. The perovskite-related compounds that have A cation and/or oxygen anion deficiency may have the general formula



where $0 \leq x < 1$, $0 \leq y < 1$, $0 \leq n < m$, and n and m are integral numbers.

Although A or B cations can be partially substituted by another element, the average valences are usually close to +2 and +4 (see later), respectively. The A cation, for example, may be replaced by (A^+, A^{3+}) or (V^+, A^{3+}) , where V stands for a positively charged vacancy site, and B cation

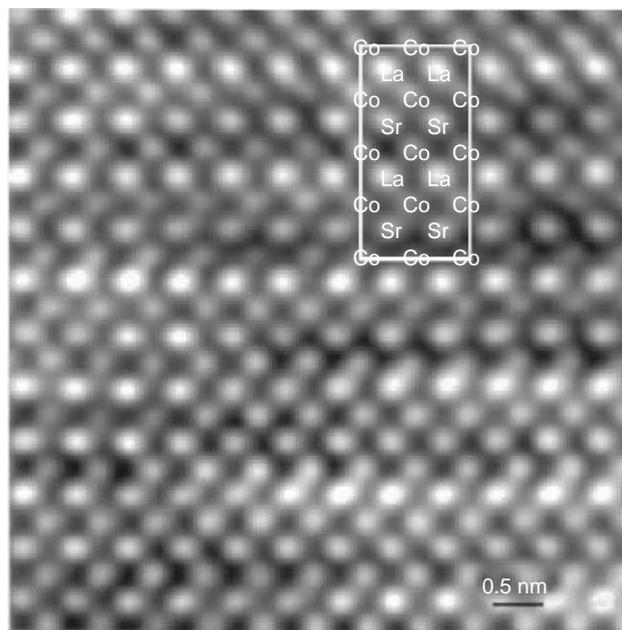


Figure 23. High-magnification TEM images of $La_{0.5}Sr_{0.5}CoO_{2.25}$ recorded along [100], where the white spots correspond to the projected atom columns.

can be (B^{3+}, B^{5+}) or (B^{2+}, B^{6+}) . Substitution of these combinations can give a series of compounds that belong to the perovskite family. Cations of different valences usually have different sizes and electron configurations, and they tend to have a strong influence on the oxygen close packing, especially in the $(AO_3)^{4-}$ layers. A change in the valence of the A cation should have a much stronger effect because it is in the fundamental stacking layer.

Figure 23 shows a cross section of a TEM image of $La_{0.5}Sr_{0.5}CoO_{2.25}$ viewed along [100]. The La and Sr atoms are distributed in different (001) atomic planes and exhibit La–Co–Sr–Co–La–Co–Sr–Co–(001) layered structure (31). The crystal structure is based on a fundamental perovskite module of $LaSrCo_2O_6$ (or $La_{0.5}Sr_{0.5}CoO_3$) without anion deficiency, as shown in Fig. 24a, which is a combination of two perovskite unit cells of $LaCoO_3$ and $SrCoO_3$. The structure of $La_{0.5}Sr_{0.5}CoO_{2.25}$ is composed of eight of this type module that has ordered anion vacancies in each. Two anion-deficient modules of $LaSrCo_2O_{4.5}$ are derived from this stoichiometric module, denoted by M_1 and M_2 (Fig. 24b). A divalent Co is likely to be coordinated by one oxygen, on average, in the top and bottom layers. These modules are the building blocks for constructing the full unit cell of $La_{0.5}Sr_{0.5}CoO_{2.25}$ (Fig. 24c). The unit cell is orthorhombic. The 3-D atomic distribution in the full unit cell is given in Fig. 24d. The coordination numbers of La and Sr are nine, respectively, and those of Co are five and four (Fig. 24e); thus, the structure is chemically stable. The three-oxygen-coordinated cation is a tetrahedron, the four-oxygen-coordinated cation is a square sheet, and the five-oxygen-coordinated cation forms a half-octahedron (e.g., a square-based pyramid).

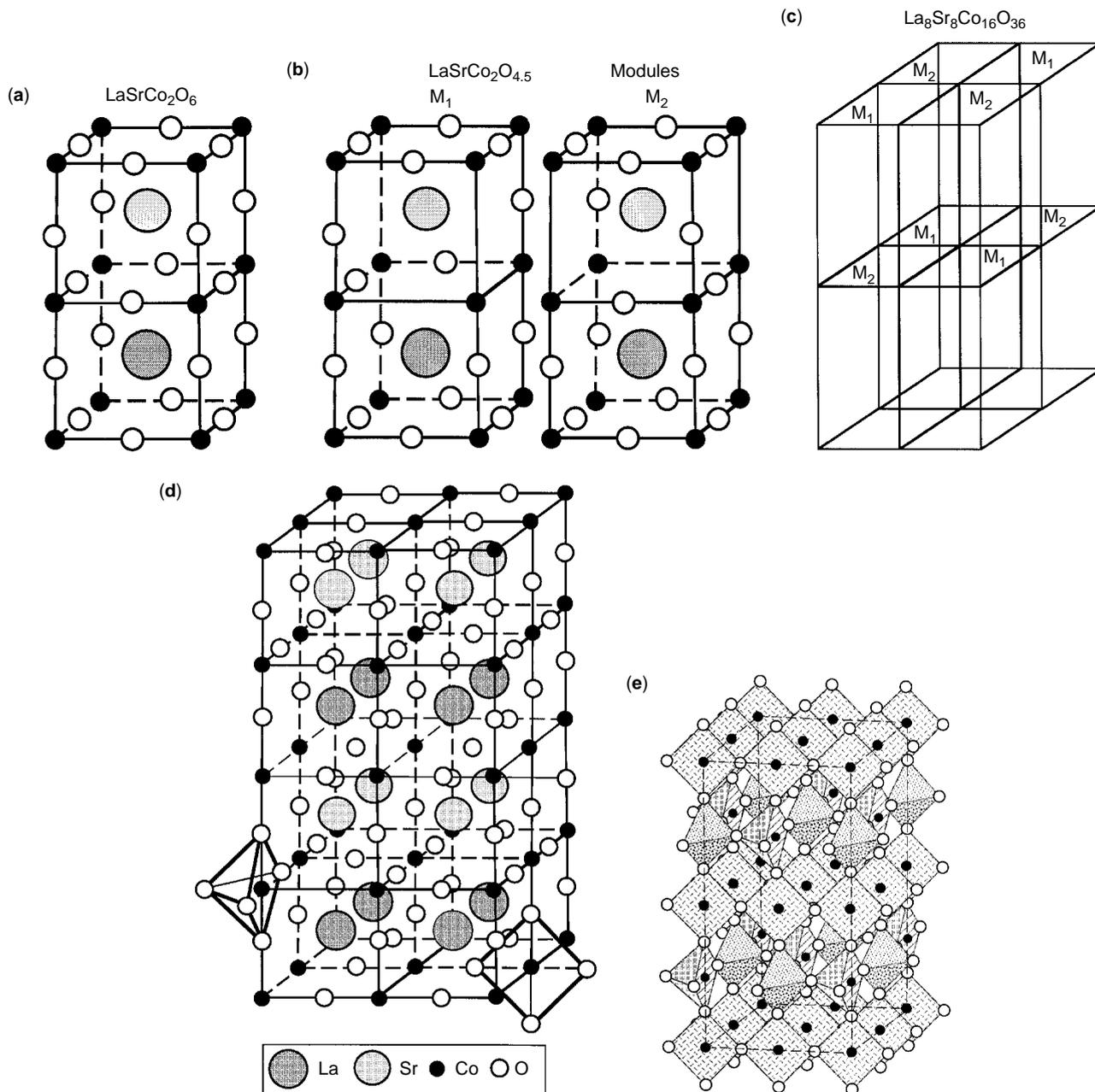


Figure 24. (a) Structural model of the $\text{LaSrCo}_2\text{O}_6$ module. (b) Two anion-deficient modules of $\text{LaSrCo}_2\text{O}_{4.5}$, and (c) the corresponding stacking to form a complete unit cell of $\text{La}_8\text{Sr}_8\text{Co}_{16}\text{O}_{36}$ (or $\text{La}_{0.5}\text{Sr}_{0.5}\text{CoO}_{2.25}$). (d) The 3-D model of the structure proposed for $\text{La}_8\text{Sr}_8\text{Co}_{16}\text{O}_{36}$. (e) A replot of (d) without showing the La and Sr cations for clarity, displaying the five and four coordinations of the Co atoms.

QUANTIFICATION OF MIXED VALENCES BY EELS

Transition- and rare-earth metal oxides are the fundamental ingredients of *mixed valences* in the structural unit (2). The valence states of metal cations in such materials can be chemically determined by using the redox titration, but it is inapplicable to nanophase or nanostructured materials such as thin films. The wet chemistry approaches usually do not provide any spatial resolution. X-ray photoelectron spectroscopy (XPS) can provide information on the average

distribution of cation valences for nanostructured materials that have certain spatial resolution, but the spatial resolution is nowhere near the desired nanometer scale, and the information provided is limited to a surface layer 2–5 nm thick.

Electron energy-loss spectroscopy (EELS), a powerful technique for materials characterization at nanometer spatial resolution, has been widely used in chemical microanalysis and studies of solid-state effects (32). In EELS, the L ionization edges of transition-metal and

rare-earth elements usually display sharp peaks at the near edge region, which are known as *white lines*. In transition metals that have unoccupied 3d states, the transition of an electron from the 2p state to the 3d levels leads to the formation of white lines. The L_3 and L_2 lines are the transitions from $2p^{3/2}$ to $3d^{3/2}3d^{5/2}$ and from $2p^{1/2}$ to $3d^{3/2}$, respectively, and their intensities are related to the unoccupied states in the 3d bands (33).

Numerous EELS experiments have shown that a change in the valence state of cations introduces a dramatic change in the ratio of the white lines, leading to the possibility of identifying the occupation number of the 3d orbital using EELS. The 3d and 4d occupations of transition-metal and rare-earth elements have been studied systematically (34,35). The oxidation states of Ce and Pr have been determined in an orthophosphate, in which the constituents of Ce and Pr are of the order of 100 ppm (36). It has been applied to quantifying the valence transition in Mn and Co oxides (37), determining the concentration of oxygen vacancies (38), refining the crystal structure of an anion-deficient perovskite (31), identifying the crystal structure of nanoparticles (CoO and Co_3O_4) (39), and determining magnetic ordering in spinel (40). The principle of this analysis is illustrated following.

Figure 25 shows an EELS spectrum of Co oxide. The EELS data must also be processed first to remove the gain variation introduced by the detector channels and to

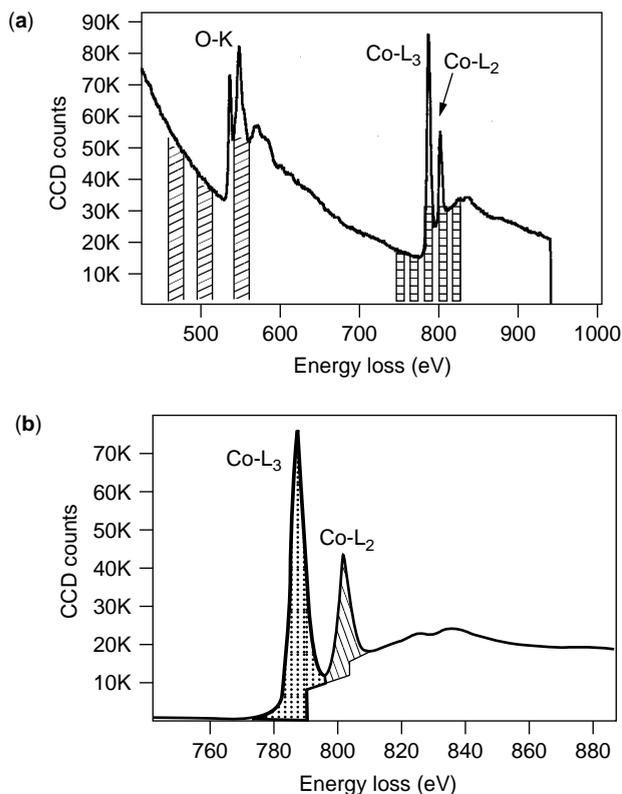


Figure 25. An EELS spectrum acquired from a Co oxide that shows the O-K and Co-L ionization edges. The fine structures arise from the atomic and solid structure of the specimen. (b) The technique used to extract the intensities of the white lines.

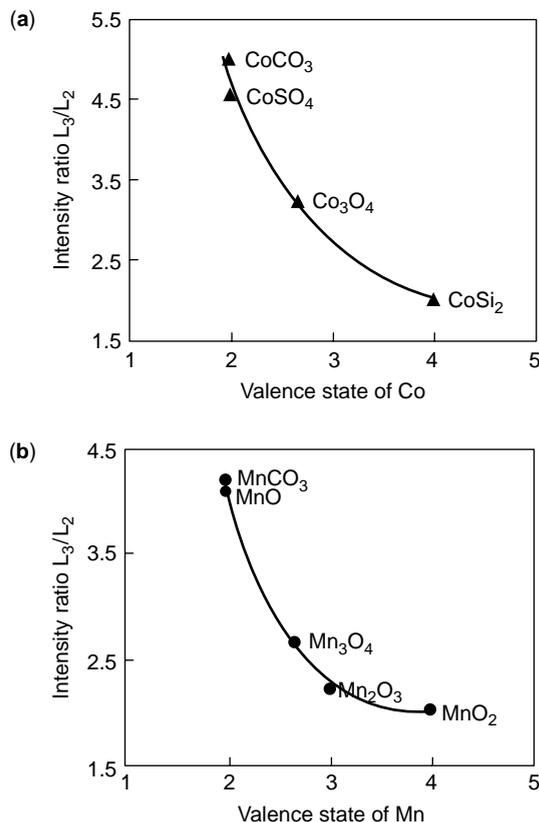


Figure 26. Plots of the intensity ratios L_3/L_2 calculated from the spectra acquired from (a) Co compounds and (b) Mn compounds as a function of the cation valence. A nominal fit of the experimental data is shown by the solid curve.

remove the multiple-inelastic-scattering effect by a deconvolution technique. Consequently, the data presented here are the result of single inelastic scattering. The ratio of white line intensities is likely to be a reliable and sensitive approach. The calculated result for L_3/L_2 is rather stable and is not sensitive to the specimen thickness or the noise level in the spectrum.

EELS analysis of valence states is carried out by reference to the spectra acquired from standard specimens that have known cation valence states. Because the intensity ratio of L_3/L_2 is sensitive to the valence state of the corresponding element, if a series of EELS spectra is acquired from several standard specimens that have known valence states, an empirical plot of these data serves as the reference for determining the valence state of the element in a new compound. The L_3/L_2 ratios for a few standard Co compounds are plotted in Fig. 26a. EELS spectra of Co- $L_{2,3}$ ionization edges were acquired from CoSi_2 (Co^{4+}), Co_3O_4 ($\text{Co}^{2.67+}$), CoCO_3 (Co^{2+}), and CoSO_4 (Co^{2+}). Figure 26b is a plot of the experimentally measured intensity ratios of white lines L_3/L_2 for Mn. The curves clearly show that the ratio L_3/L_2 is very sensitive to the valence state of Co and Mn. This is the basis of our experimental approach for measuring the valence states of Co and Mn in a new material.

To demonstrate the sensitivity and reliability of using white line intensity for determining valence states in

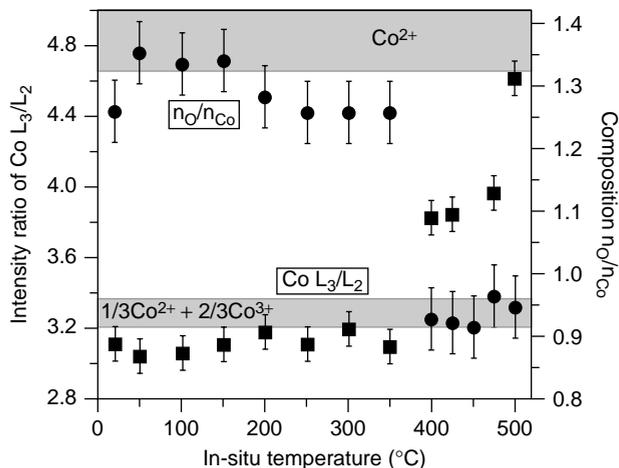


Figure 27. An overlapped plot of the white line intensity ratio of $\text{Co L}_3/\text{L}_2$ and the corresponding chemical composition of $n_{\text{O}}/n_{\text{Co}}$ as a function of the *in situ* temperature of the Co_3O_4 specimen that shows the abrupt change in valence state and oxygen composition at 400°C . The error bars are determined from the errors introduced in background subtraction and data fluctuation among spectra.

mixed valence compounds (37), the *in situ* reduction of Co_3O_4 is examined first. Figure 27 shows the $\text{Co L}_3/\text{L}_2$ ratio and the relative composition of $n_{\text{O}}/n_{\text{Co}}$ for the same crystal, as the specimen temperature was increased. The specimen composition was determined from the integrated intensities of the O–K and Co– $\text{L}_{2,3}$ ionization edges by using ionization cross sections calculated from the SIGMAK and SIGMAL programs (32). The L_3/L_2 ratios corresponding to Co^{2+} determined from the EELS spectra of CoSO_4 and CoCO_3 at room temperature and $\text{Co}^{2.67+}$ obtained from Co_3O_4 are marked by shadowed bands, whose widths represent experimental error and the variation among different compounds. The $\text{Co L}_3/\text{L}_2$ ratio and the composition, $n_{\text{O}}/n_{\text{Co}}$, simultaneously experience a sharp change at $T = 400^\circ\text{C}$. The chemical composition changes from $\text{O}:\text{Co} = 1.33 \pm 0.5$ to $\text{O}:\text{Co} = 0.95 \pm 0.5$ that accompanies the change of the average valence state of Co from $+2.67$ to $+2$ when the temperature is above 400°C . Electron diffraction has also confirmed the reduction process observed.

HIGH-SPATIAL-RESOLUTION MAPPING OF VALENCE STATES

Energy-filtered transmission electron microscopy (EF-TEM) (41) is a rapidly developing field for high spatial-resolution chemical imaging. Images (or diffraction patterns) formed by electrons that have specific energy losses can be obtained by using an energy filter. If the energy-selected electron image can be formed using the white lines described earlier (6), one can map the spatial distribution of the valence states (42). According to Fig. 25, an energy window 12 eV wide is required to isolate the L_3 from the L_2 white lines. A five-window technique is introduced

(see Fig. 25a): two images are acquired at the energy losses before the L ionization edges, and they are to be used to subtract the background for the characteristic L edge signals; two images are acquired from the L_3 and L_2 white lines, respectively, and the fifth image is recorded using the electrons right after the L_2 line that will be used to subtract the continuous background underneath the L_3 and L_2 lines. Then, a L_3/L_2 ratio image will be obtained, which reflects the distribution of valence state across the specimen. It must be pointed out that the thickness effect has been removed in the L_3/L_2 image.

A partially oxidized CoO specimen that contains a CoO and Co_3O_4 grain structure was chosen for this study (44). The CoO and Co_3O_4 phases are separated by clear boundaries, and it is an ideal specimen for testing the optimum resolution. Figure 28 shows a group of energy-filtered TEM images from a triple point in the CoO– Co_3O_4 specimen. The energy-filtered images using the L_2 and L_3 lines and the post- L_2 line region (Fig. 28b–d) show distinctly different contrast distributions due to differences in the relative white line intensities. From these three images, the L_3/L_2 ratio is calculated after subtracting the contribution from the continuous energy-loss region that is due to single atom scattering, the image clearly displays the distribution of cobalt oxides that have different valence states (Fig. 28e), where the diffraction contrast disappears. The region of lower oxidation state (Co^{2+}) shows a stronger contrast, and those that have high oxidation states show darker contrast (see the L_3/L_2 ratio in Fig. 26a). The O/Co image (Fig. 28f) was calculated from the images recorded from the O–K edge and the $\text{L}_3 + \text{L}_2$ white lines for an energy window width of $\Delta = 24$ eV. Although the energy-filtered O–K edge image exhibits some diffraction contrast and the thickness effect, the O/Co compositional ratio image greatly reduces the effect. The high-intensity region in the O/Co image indicates a relatively high local concentration of oxygen (e.g., higher Co oxidation states); the low intensity region contains relatively less oxygen (e.g., a lower Co valence state), entirely consistent with the information provided by the L_3/L_2 image. A line scan across the valence state map clearly illustrates that a spatial resolution of 2 nm can be achieved (42). This is remarkable compared to any existing techniques.

SUMMARY

Perovskite is probably the most important structural type of smart materials. The properties of perovskites depend strongly on their structures. In this article, the characteristics of ABO_3 perovskites are analyzed to reveal the intrinsic connection among the A, B, and O elements and the roles of tetrahedrons and octahedrons in the structure. The (111) alternate stacking of the close-packed $(\text{AO}_3)^{4-}$ and B^{4+} layers is responsible for the cation substitution and the creation of anion vacancies. Careful analysis of the possibilities of creating oxygen vacancies results in a total of 14 fundamental structural units, which are the building blocks for constructing the unit cells of complex functional materials, such as high T_c superconductors.

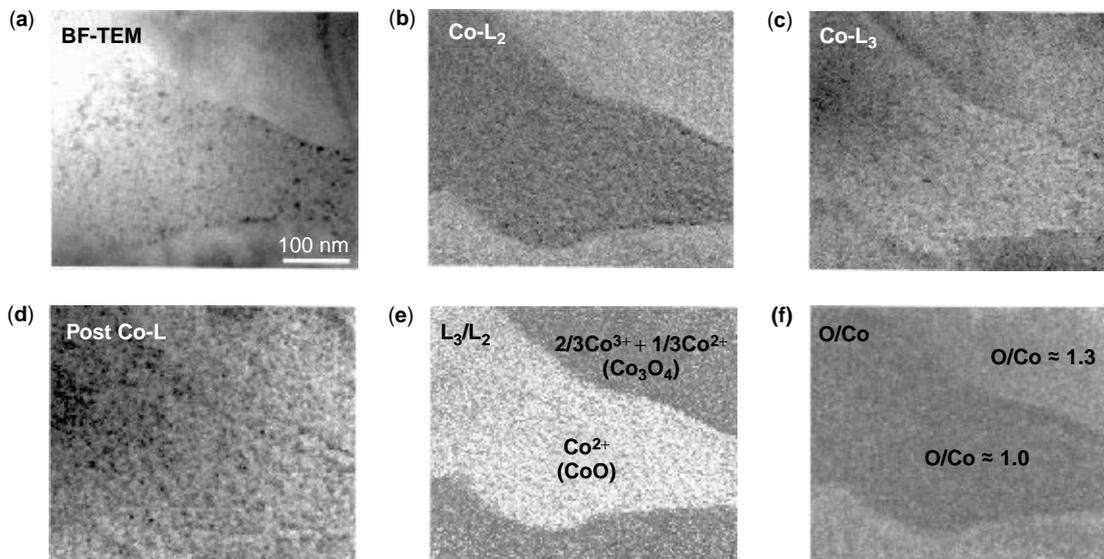


Figure 28. A group of images recorded from the same specimen region using signals of (a) electrons without energy loss, (b) the Co-L₂ edge, (c) the Co-L₃ edge, and (d) the post-Co-L₂ line. (e) The processed L₃/L₂ image that displays the distribution of valence states. (f) The atomic concentration ratio image of O/Co. The O/Co image is normalized with reference to the standard composition of CoO for the low portion of the image to eliminate the strong influence of the white lines on the ionization cross section. Each raw image was acquired using an energy window width of $\Delta = 12$ eV except for O-K at $\Delta = 24$ eV. (g) and (h) are line scan profiles from (e) and (f), respectively, proving the achievement of 2-nm spatial resolution.

ACKNOWLEDGMENT

Thanks to Dr. Z.C. Kang for many stimulating discussions.

BIBLIOGRAPHY

- R.E. Newnham and G.R. Ruschau, *J. Am. Ceram. Soc.* **74**(3): 463–480 (1991).
- Z.L. Wang and Z.C. Kang, *Functional and Smart Materials—Structure Analysis and Structural Evolution*. Plenum Press, NY, 1998, Chap. 3.
- C.N.R. Rao and B. Raveau, *Transition Metal Oxides*, VCH, 1995.
- F.S. Galasso, *Perovskites and High T_c Superconductors*, Gordon and Breach, 1990.
- O. Auciello, J.F. Scott, and R. Ramesh, *Phys. Today*, pp. 22–27 (July 1998).
- M. Tanaka and G. Honjo, *J. Phys. Soc. Jpn.* **19**: 954–970 (1964).
- B. Jaffe, W.B. Cook, and H. Jaffe, *Piezoelectric Ceramics*. Academic Press, London, 1971.
- A.J. Moulson and J.M. Herbert, *Electroceramics—Materials, Properties, Applications*. Chapman & Hall, London, 1990.
- K. Uchino, *Piezoelectric Actuators and Ultrasonic Motors*. Kluwer Academic, Boston, 1997.
- S. Nomura, M. Mizuno, J. Kuwata, M. Abe, and J.C. Burfoot, *Ferroelectrics* **23**: 183–186 (1980).
- Z.C. Kang, C. Caranoni, I. Siny, G. Nihoul, and C. Boulesteix, *J. Solid State Chem.* **87**: 308–320 (1990).
- R.E. Newnham, *MRS Bull.* pp. 20–34 (July 1997).
- S. Jin, T.H. Tiefel, M. McCormack, R.A. Fastnacht, R. Ramech, and L.H. Chen, *Science* **264**: 413–415 (1994).
- R. von Helmolt, J. Wecker, B. Holzapfel, L. Schultz, and K. Samwer, *Phys. Rev. Lett.* **71**: 2331–2334 (1994).
- D.B. Studebaker, M. Todd, T.H. Baum, Y. Berta, and Z.L. Wang, unpublished.
- G.H. Jonker and J.H. Van Santen, *Physica* **16**: 337–349 (1950).
- C. Zener, *Phys. Rev.* **82**: 403–405 (1951).
- J.B. Goodenough, Met., in *Progress in Solid State Chemistry*, H. Reiss, ed., Pergamon Press, 1971, Vol. 5, pp. 145–399.
- P.G. de Gennes, *Phys. Rev.* **118**: 141–154 (1960).
- Z.L. Wang and J. Zhang, *Phys. Rev. B* **54**: 1153–1158 (1996).
- A.Q. Pham, M. Puri, J.F. Dicarolo, and A.J. Jacobson, *Solid State Ionics* **72**: 309–313 (1994).
- B. Reveua, C. Michel, M. Hervieu, and D. Groult, *Crystal Chemistry of High T_c Superconducting Copper Oxides*. Springer-Verlag, Berlin, 1991.
- L. Er-Rakho, C. Michel, P. LaCotte, and B. Raveau, *J. Solid State Chem.* **73**: 531–535 (1988).
- M. Rey, P. Dehaudt, J. Joubert, and A.H. Hewat, *Physics C* **167**: 162–167 (1990).
- M.T. Anderson, J.P. Zhang, K.P. Poeppelmeier, and L.D. Marks, *Chem. Mater.* **5**: 151–165 (1993).
- M.A. Beno, D.W. Soderholm, D.W. Capone, J.D. Jorgensen, K.I. Schuller, C.U. Serge, K. Zhang, and J.D. Grace, *Appl. Phys. Lett.* **51**: 57–59 (1987).
- K.R. Poeppelmeier, M.E. Leonowicz, and J.M. Longo, *J. Solid State Chem.* **44**: 89–98 (1982).
- K.R. Barnard, K. Foger, T.W. Turney, and R.D. Williams, *J. Catal.* **125**: 265–275 (1990).

29. A. Baiker, P.E. Marti, P. Keusch, and A. Reller, *J. Catal.* **146**: 268–276 (1994).
30. H.L. Yakel, *Acta Crystallogr.* **8**: 394–398 (1955).
31. Z.L. Wang and J.S. Yin, *Philos. Mag. B* **77**: 49–65 (1998).
32. R.F. Egerton, *Electron Energy-Loss Spectroscopy in the Electron Microscope*. 2e, Plenum Press, NY, 1996.
33. F.M.F. de Groot, M. Grioni, and J.C. Fuggle, *Phys. Rev. B* **40**: 5715–5723 (1989).
34. H. Kurata and C. Colliex, *Phys. Rev. B* **48**: 2102–2108 (1993).
35. D.H. Pearson, C.C. Ahn, and B. Fultz, *Phys. Rev. B* **47**: 8471–8478 (1993).
36. J.A. Fortner and E.C. Buck, *Appl. Phys. Lett.* **68**: 3817–3819 (1996).
37. Z.L. Wang, J.S. Yin, J.Z. Zhang, and W.D. Mo, *J. Phys. Chem. B* **101**: 6793–6798 (1997).
38. Z.L. Wang, J.S. Yin, Y.D. Jiang, and J. Zhang, *Appl. Phys. Lett.* **70**: 3362–3364 (1997).
39. J.S. Yin and Z.L. Wang, *Phys. Rev. Lett.* **79**: 2570–2573 (1997).
40. Z.J. Zhang, Z.L. Wang, B.C. Chakoumakos, and J.S. Yin, *J. Am. Chem. Soc.* **120**: 1800–1804 (1998).
41. L. Reime, ed., *Energy-Filtering Transmission Electron Microscopy*. Springer Series in Optical Sciences, Vol. 71. Springer, Berlin, 1995.
42. Z.L. Wang, J. Bentley, and N.D. Evans, *J. Phys. Chem. B* **103**: 751–753 (1999).
43. J. Bentley, S. McKernan, C.B. Carter, and A. Revcolevschi, *Microbeam Anal.* **2** (Suppl.): S286–287 (1993).

SOIL-CERAMICS (EARTH), SELF-ADJUSTMENT OF HUMIDITY AND TEMPERATURE

EMILE H. ISHIDA
INAX Corporation
Minatomachi, Tokoname, Aichi, Japan

INTRODUCTION

Earth ceramics are hydrothermally solidified soil bodies that have been developed mainly for use as interior and exterior materials for housing. Generally, manufacture of industrial materials requires the use of refined raw materials. In the case of earth ceramics, a wide range of soils can be used for manufacture without extensive refining. Besides, since solidification is carried out at 150 to 200°C under saturated steam pressure, the energy required for solidification is relatively small. Therefore, the load imposed on the earth by the manufacturing of this material is kept very low. Earth ceramics possess excellent thermal insulation and self humidity control properties because the original nano-pores of the soil remain in the bodies after solidification. Because of these properties, earth ceramics are effective in controlling the interior climate in dwellings, particularly in monsoon regions like Japan. It has been reported that in dwellings where earth ceramics were actually used as the floor material, the humidity control properties of the material allows comfortable and healthy living with almost no necessity for the use of air conditioners throughout the year and that the amount of energy consumed for living is reduced significantly.

The necessity of a new concept of manufacturing goods based on due consideration for humans and the earth, and the thinking behind the development of earth ceramics based on investigations of this concept, are described below. Examples of actual use of earth ceramics are also outlined.

A NEW DEFINITION ON MATERIALS WITH CONSIDERATION FOR HUMANS AND THE EARTH

The Inevitability of a Recirculation-Based Society

The global environment has now begun to have great effects on our lives. Environmental issues such as global warming, desertization, depletion of the ozone layer, and acid rain stem mainly from global scale expansion of the economic activities of the industrially advanced nations and population expansion in the developing countries. These two problems are compelling enough for us to consider radical reforms of the global social structure.

In their book *Beyond the Limits* published in February 1992, Meadows and his co-authors (1) warned that unless countermeasures are taken, the world economy which may still expand until around the year 2020, will cease its growth after that because of various limiting factors, and that the world's social structure will have disintegrated completely by the year 2100. A number of similar reports supplementing these warnings were published (2,3) and created a great stir among material scientists.

The United Nations Conference on Environment and Development (Earth Summit) held in Brazil in June of the same year issued the Rio Declaration on Environment and Development, adopting Agenda 21 as the action plan of the summit. Out of all the activities carried out by humans, those that impose the largest load on the earth are the economic and industrial activities. These activities depend on a flow of materials and energy. To counter the danger of extinction of humankind and build a sustainable society, it is essential that we institute a recirculation-based society. A recirculation-based society is a rejection of the consumption-based economic structure of the industrialized countries. The idea is to construct a type of society not ever experienced in the past. What will the manufacture of goods be like in a recirculation-based society? Here, we need to redefine the basic notion of manufacturing.

Manufacturing of Goods with Awareness of a Recirculation-Based Society

A recirculation-based society can be expressed by a simple diagram as shown in Fig. 1. Since the dawn of civilization the human being has demarcated a border (system boundary) between the natural ecological system and the human life system (human ecological system).

Humans thought that they could not sustain their livelihood without exploiting nature. The amount of intake from nature and the resulting amount of release into nature he increased rapidly after the Industrial Revolution. These now far exceed what nature can offer without serious consequences for the preservation of humankind. A recirculation-based society is one that reduces much of the

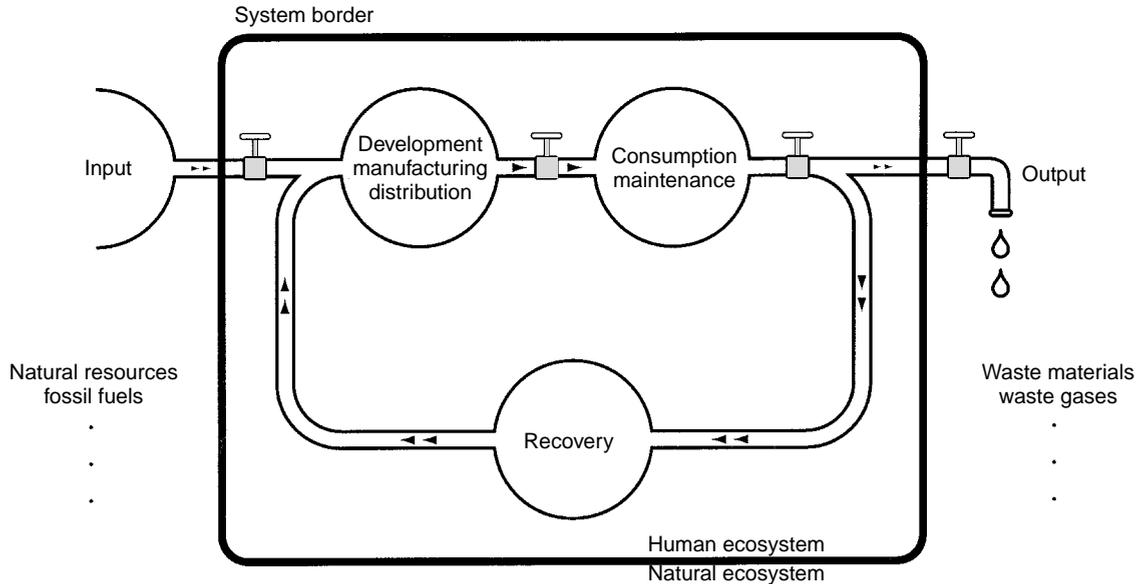


Figure 1. Re-circulation based society.

input from nature (fossil energy, raw materials) as well as its output into nature's ecosystem (exhaust gases, exhaust heat, waste materials), and makes efforts to recirculate and regenerate the input from nature within the human ecosystem. In other words, when manufacturing goods for a recirculation-based society, it is very important to think about the total energy balance and the material balance. In practice, it is necessary to reduce the output by reducing the input, and in addition, synthesize new materials from materials and energy that have the possibility of producing fresh output. We must, therefore, change our attitude from recycling to recirculation.

Manufacturing of Goods with Consideration for the Earth

Although manufacturing of goods with an awareness of a recirculation-based society is unavoidable, it is doubtful whether goods manufactured based on such a concept would be accepted by the world. Let us consider an extreme example. An electric refrigerator not only consumes electricity but also the refrigerating agent, which is Freon gas and a burden to the earth's environment. In order to lower the input and output, should we then stop using electric refrigerators and return to the earlier era of ice-boxes for refrigeration? Most people would answer no. It is not easy to abandon a convenience once it has been experienced. Furthermore, denial of the existence of electric refrigerators means also denial of large businesses that sell refrigerated foods like department stores, supermarkets, and convenience stores. The invention of the electric refrigerator brought great changes to the social system. As long as irreversibility of life values exists, we cannot easily return to the "good old age." If we reach an ultimate state where there is no alternative but to return to the past such a reversal would probably be at the expense of unprecedented patience and great pain. Conversely, if it were easy to return to a former way of living, environmental problems

would not occur, there might not be the need to develop new materials. Our inability to return to a former, energy-efficient way of living is why we need to address the serious issue of environmental pollution and set guidelines for new materials:

Most important, the manufactures need to be made more conscious of the earth. The goods they develop should be useful, convenient, or improvements of existing goods that are necessary to people. Of course, as environmental problems become more severe, the balance between people and the earth will shift, and without doubt, more emphasis will start to be placed on earth friendly materials and manufacturing. Materials and goods produced without consideration of their value or usefulness to people would cease to exist. Taking this argument from a different perspective, no one could argue that the manufacturing of goods does not exploit something (input) from the earth and then discharge some waste to the earth (output) which is produced as the input from the earth is converted into goods with functions that have some value to people (Fig. 2). We could express this relation of the development of goods as,

$$\text{Value} = \frac{P}{I + O},$$

where P is the performance of the material, $I + O$ are the input and output of the material when manufactured, used, and scrapped.

In general, if the value for people is less than one, there is no merit to developing the material or goods. Development should aim to raise this value to five, ten, or even the hundreds. Attain entirely new approaches, might be necessary, disregarding any current concepts. Now, let us think specifically about materials that could be developed using the $P/(I + O)$ valuation concept.

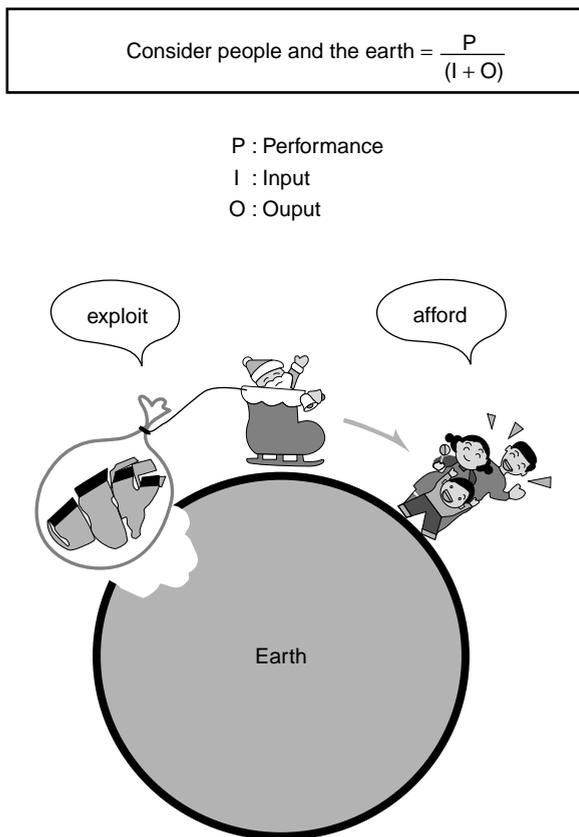


Figure 2. A new value for the manufacturing.

SMART MATERIALS FOR THE LIVING ENVIRONMENT

Foremost among materials for maintaining a comfortable living space, while reducing the burden on the earth's environment, is the development of high heat insulation houses. Of course, this is not a consideration in hot and humid regions with monsoon climates. Primarily, there is the case of Japanese dwellings.

Japan, which is located in a monsoon zone, has a relatively distinct climate compared to other developed nations. In this unique hot and humid climate, much damage is caused by the high humidity. Although the humidity itself is lower than that of say, London, Paris, or San Francisco, Japan is at the top with regard to humidity-related damage to houses. This is due to the fact that fungi and bacilli that affect human health and cause damage to houses are able to breed rapidly under the warm and humid Japanese climate. To counter these adverse influences, elevated houses emerged in Japan as far back as the ninth century. Elevated floors allow for underfloor ventilation. The materials used for construction were paper, wood, and soil. Some 50 years ago airtight houses were introduced in Japan following examples in Europe and America. A new apartment house building trend then took root in Japan. However, it has turned out that the Western airtight dwellings were unsuited to a humid climate, and thus uncomfortable for living. The search for comfort led to the introduction of various indoor implements—electric fans, followed by coolers and air conditioners. Then the first oil

shock in 1973 forced Japanese to make significant reductions in their energy consumption, and the idea of airtight and heat-insulated housing became national policy. In the adoption of such dwelling, the technology to control indoor temperature has improved considerably, although the humidity has proved to be difficult to control. Because the rooms in Japanese houses are small, the interiors of these airtight dwellings tend to be inferior compared to houses in Europe and America (4, 5). In fact the interior environment in Japan has degenerated to a clutter of such appliances humidifiers, dehumidifiers, and air cleaners (Fig. 3). Yet, despite these measures, the sick house syndrome and allergic diseases are continuing to proliferate. Compared to 1973, when the industrial energy consumption in Japan in 1997 was 104%, today dwellings consume energy as high as 217%, and this figure has been rapidly climbing (6). The energy increases have occurred even though much progress has been made in the development of low-energy consumption type of appliances.

The technological solution to this problem would be to have a self-monitoring and self-regulating the indoor climate (humidity, in particular) by materials of the house—the floor, walls, and ceiling materials—that have high P value. At the same time, it is necessary to examine the methods of synthesizing these materials from the perspective of not using even more energy and natural resources for the synthesis—that is, to maintain a low $(I + O)$ value. If such materials could be developed, indoor climate control could be made effective even in the most airtight and heat-insulated homes. Such materials would possess high value as those to society with due consideration to humans and the earth.

CLIMATE CONTROL BY POROUS BODIES

The humidity range in which a person feels comfortable is said to be 40% to 70%. It has been reported that by maintaining humidity within this range, allergy sources such as mites, as well as the breeding of wood-eating bacteria, molds, and the like, that cause degradation of wood in wooden houses, could be restricted (7). This humidity range is also thought to be effective in curbing the spread of viruses and even the accumulation of static electricity. Chemical and physical methods of humidity regulation are available, but here, let us think about a safe method, which is humidity regulation by porous bodies.

The target material would be one that does not absorb the water vapor when the humidity is less than 40%, but if the humidity rises higher, it should work to lower the humidity by rapidly absorbing the water vapor from the atmosphere. Then, as the humidity starts to fall, the material should act to increase the humidity rapidly to the preferred 40% to 70%. In other words, the water vapor absorption isotherm of the material should be steep in the 40% to 70% humidity range (Fig. 4).

So these humidity-regulating porous materials should be capable of making the water vapor in the atmosphere to condense within the capillary pores that exist on their surface when the humidity is high. Conversely, when the humidity is low, they should function to vaporize the

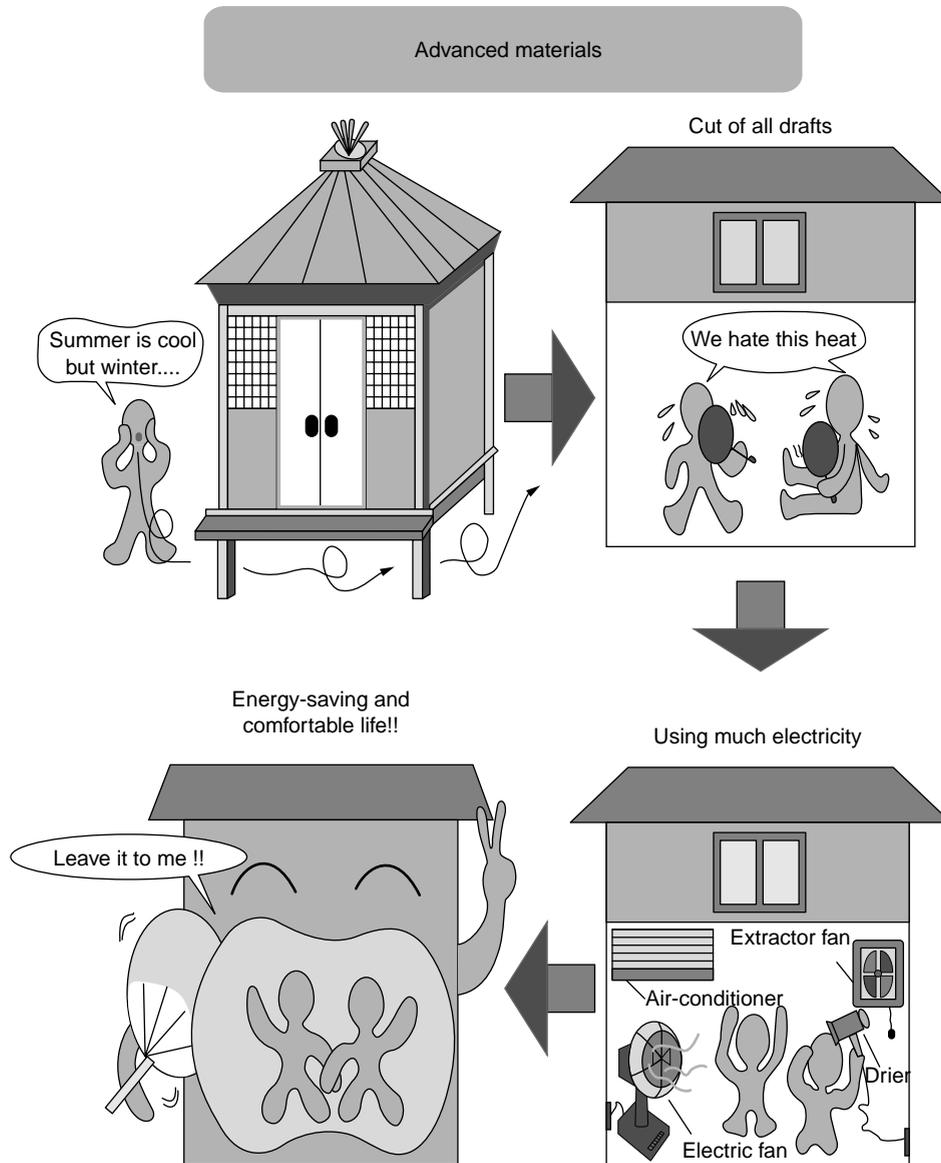


Figure 3. Earth and people conscious materials for the living environment.

condensed water. The relation between the vapor pressure, P/P_0 , required for capillary condensation and pore size, with curvature radii r_1 , r_2 , is expressed by Kelvin's equation of capillary condensation:

$$\ln\left(\frac{P}{P_0}\right) = -\left(\frac{\gamma V}{\cos\theta RT}\right)\left(\frac{1}{r_1} + \frac{1}{r_2}\right).$$

Here, r_1 and r_2 are the radii of curvature of the pores in two perpendicular directions, γ is the surface tension of the condensate, V is the molecular volume of the condensate, and θ the contact angle of the condensate within the pore.

Calculations based on this equation, corrected for the preexistence of a certain thickness of the adsorbed layer prior to capillary condensation (8), yield pore radii values of 3.2 nm for 40% relative humidity and 7.4 nm for 70% humidity. High-humidity regulating performance can be

expected from materials synthesized with their pore radii being controlled to be within this range.

USING THE GREATNESS OF NATURE WISELY

Utilizing Soil

There are many possibilities of synthesizing porous materials with humidity-regulating properties. For example, taking petroleum as the starting material, the synthesis could be done by chemical polymerization or biomimetic methods. However, the use of these methods makes it difficult to lower the input and output of the synthesized material. For this reason, we should select "soil," as the starting material. Natural soil is a material containing incipient micropores that can be effective in imparting humidity regulating performance. Even after its use in the

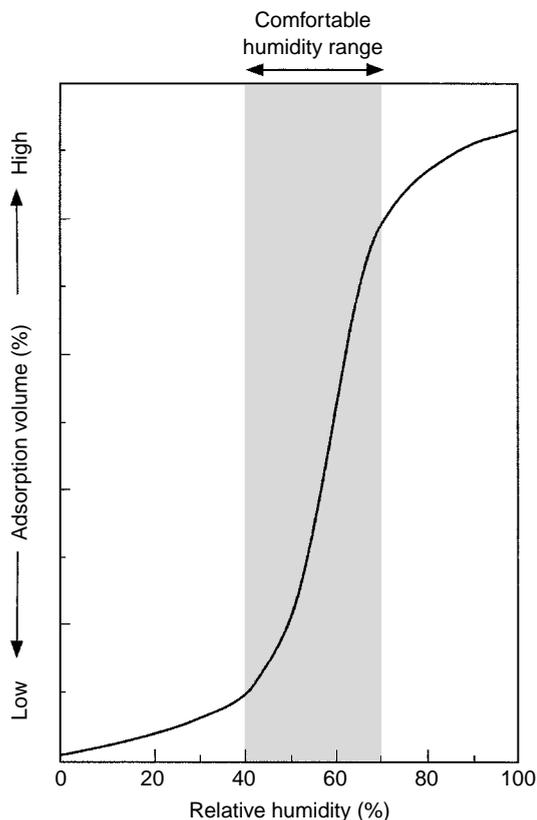


Figure 4. Property for the humidity-regulating porous materials.

human ecosystem is over, soil will not inflict a large burden on the natural ecosystem.

Soil appeared some 400 million years ago at a time when plenty of oxygen was supplied by the atmosphere. The composition of soil is almost the same today. The excess oxygen was decomposed in the stratosphere to form the ozone layer. The ozone layer has prevented strong ultraviolet rays from pouring onto the surface of the earth, and allowed the movement of animals and plants from the sea to land to begin. With the help of this organic matter, the land of stone and sand turned into a land of green, and soil appeared for from the weathering and decomposition of rocks (9). Were it not for soil, the perfect recirculation performance of the current natural ecosystem would not exist. So one cannot ignore the benefits obtained from soil for the existence of humankind. Humankind is, of course, indebted to soil with regard to food crops, but more so, it was through the aid of earthen dwellings that humankind was able to survive the glacial era without running out of seed.

Soil contains numerous pores, both large and small. These pores collect air, water, and many nutrients, allowing soil to carry out its functions. For example, a survey of the virgin forests of the Shiga plateau (Japan) has shown that in a 1 m by 1 m by 15 cm volume of soil, there are 360 living creatures such as centipedes and earthworms ranging about 2 cm in size, then some 2 mm in size thread earthworms, beach fleas amounting to 2.3 million in count, and finally any number of protozoan, mold, and bacteria

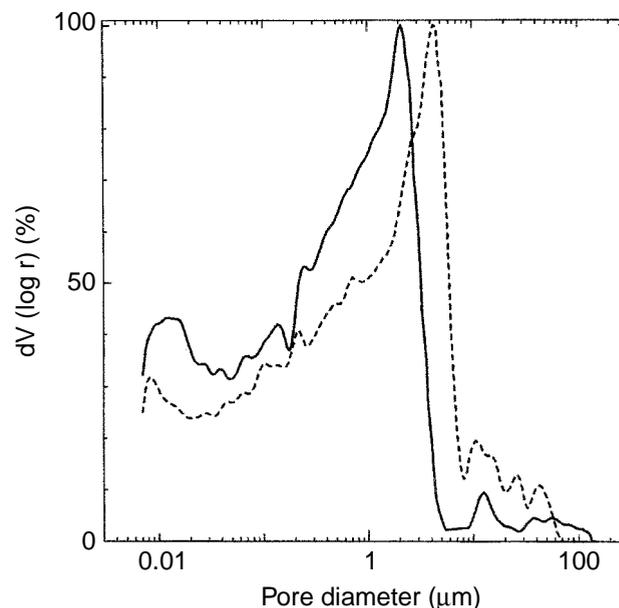


Figure 5. Examples for the pore size distribution of the general soil after dry-pressed under 30 MPa (solid line) and 20 MPa (broken line).

of more than 10 trillion in count (10). Figure 5 shows the pore size distribution of common earth (soil). It is clear that the 10 nm (0.01 μm) pores considered suitable for humidity regulation are incipient in the material. It is also clear that the cohesive structure in the neighborhood of 10 nm does not collapse easily even under pressure. However, even any kinds of soil if lost for some reason or the other, cannot be regenerated for an extremely long time. Furthermore, because of its complex structure, soil is a material that has not been successfully synthesized artificially up to now.

Humankind has utilized earth very cleverly as a construction material in many ways. In its natural state, it has been used for cave-type and pit-type dwellings. Soil has been processed and used to obtain various types of construction materials such as sun-dried bricks. In Japan, it was used as *Tataki*, or earthen walls. The humidity regulation and thermal insulation properties arising from the innumerable pores in soil were utilized in Japan's Edo period to build earthen storehouses that protected stock from wind, fire, and water. The technology of soil utilization was even raised to an artistic level, as can be seen from the *Nurikabe* walls of the Edo era. Once these soil materials have served their purpose in the human ecosystem, they can be returned to the natural ecosystem. Therefore, soil is an extremely rare material that can cross the system boundary zone freely. Unfortunately, the use of earth as soil, is usually not the possible in current construction practice. If, for example, soil is used in its natural state for the flooring of today's airtight and heat-insulated dwellings, the house would become dusty and the health of the inhabitants may be affected adversely. In addition, there are problems regarding strength, durability, and workability.

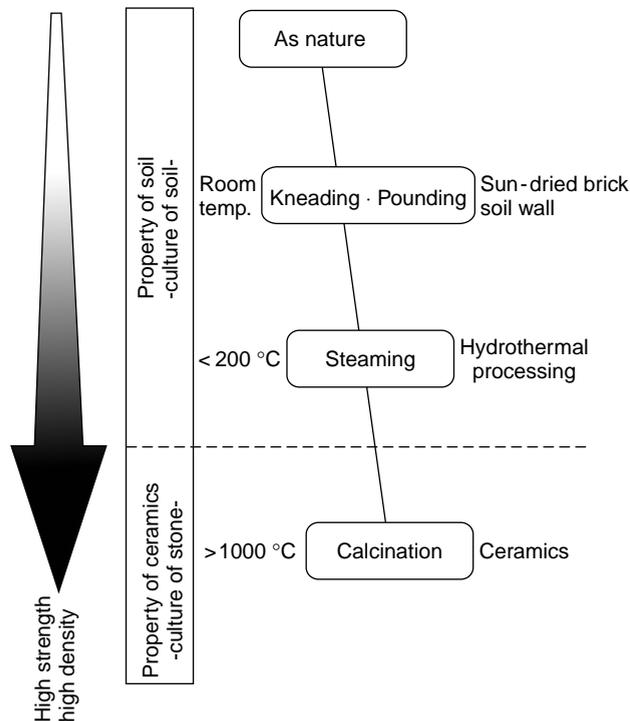


Figure 6. A culture of soil in human history.

Solidified ceramics such as bricks, blocks, and tiles made their appearance early in history in order to solve these problems. However, ceramics, which could be considered to have been developed in the Stone Age, are produced through high-temperature reactions. For this reason, it is difficult to say that ceramics maintain the inherent properties and performance of soil. In order to sustain the inherent properties and performance of soil (pore size distribution), the temperature of manufacture of earthen products must be lower than 500°C. In the case of earth/organic material composites, an even lower temperature is desirable.

The new technology of solidifying soil by hydrothermal treatment is a low-temperature process developed to obtain a material with properties and performance between those of soil and ceramics (6).

Hydrothermal Processing

The largest application of hydrothermal processing is in the field of building materials manufacture. It was developed in Europe and has over 100 years of history as represented by sand lime bricks (11,12). Usually, the process involves mixing quartz (SiO_2) with about 10% lime ($\text{Ca}(\text{OH})_2$) and exposing the mixture to saturated steam at about 200°C. This results in the formation of acicular calcium silicate hydrates (Fig. 7) (13). It is believed that strength development is obtained because these hydrates (tobermorite, xonotolite, etc.) are entwined with each other in the solidified bodies. From the point of view of energy consumption, this method of synthesis is an extremely high efficiency process. Numerous studies have been made regarding the reactions involved during synthesis by hydrothermal treatment and regarding the behavior of

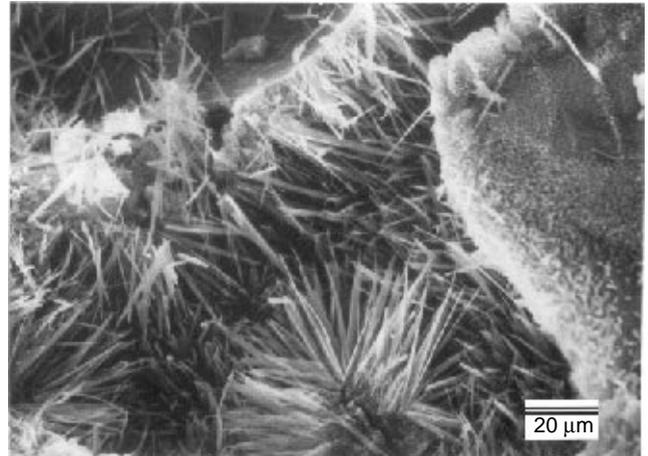
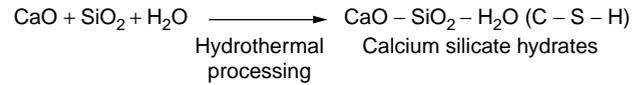


Figure 7. Calcium silicate hydrates, the most important bonding materials in the hydrothermally solidified building material (photo: tobermorite in autoclaved sand lime brick).

calcium silicate hydrate materials, mainly in the $\text{CaO}-\text{SiO}_2-\text{H}_2\text{O}$ system (14–16). Although many of the reports deal with the purity of the starting materials, it is generally accepted that it is desirable to use SiO_2 with high purity (12). Almost no studies seem to have been made with significantly altered SiO_2 compositions. Mitsuda et al. (17,18) have studied the effects of Al_2O_3 source additions to the $\text{CaO}-\text{SiO}_2-\text{H}_2\text{O}$ system and the replacement of Si by Al in tobermorite ($\text{Ca}_5(\text{Si}_6\text{O}_{18}\text{H}_2) \cdot 4\text{H}_2\text{O}$). However, the addition of Al_2O_3 in this study was up to $\text{Al}/(\text{Al} + \text{Si}) = 0.16$, which corresponds to the solid solution limit of Al in tobermorite (19,20). An investigation by Kalousek has reported that when kaolinite is used as the Al source, hydrogarnet forms in the range $\text{Al}/(\text{Al} + \text{Si}) = 0.12 - 0.5$ (20). Generally, this hydrogarnet has a cubic structure and contains $\text{Ca}_3\text{Al}_2\text{Si}_3\text{O}_{12}-\text{Ca}_3\text{Al}_2\text{H}_{12}\text{O}_{12}$ in solid solution (21,22). In a recent study of metakaolin–quartz–lime series slurries, the effect of amount of metakaolin additions and of quartz grain size on hydrogarnet formation have been clarified (23–25). These reports indicate that with increasing $\text{Al}/(\text{Al} + \text{Si})$ ratio in the $\text{CaO}-\text{SiO}_2-\text{Al}_2\text{O}_3-\text{H}_2\text{O}$ system, hydrogarnet forms in addition to calcium silicate hydrates such as C–S–H (calcium silicate hydrate gels) and tobermorite, and that the hydrogarnet becomes the main phase for $\text{Al}/(\text{Al} + \text{Si})$ ratios higher than about 0.2. However, in the composition range over which hydrogarnet is the main phase formed, not much study seems to have been carried out on the strength development of hydrothermally synthesized bodies. Investigations about the relationship between strength development and the microstructural changes that accompany the formation of hydrogarnet cannot be found either. This is probably because of the fact that it was thought that formation of hydrogarnet causes strength deterioration in calcium silicate materials (26). Because of this belief, avoiding the formation of hydrogarnet has been an important direction of research up to now.

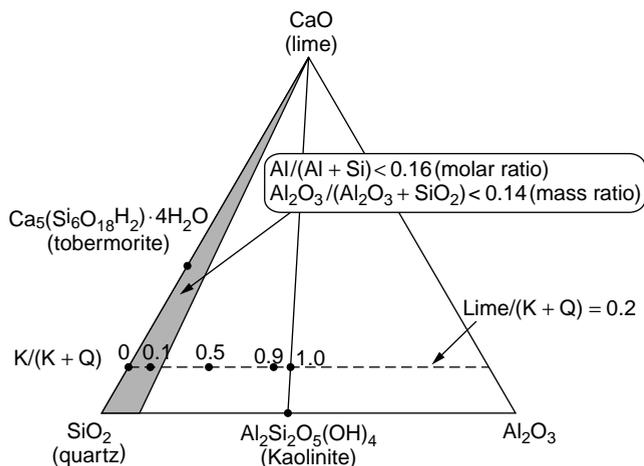


Figure 8. Experimental compositions (mass ratio) in the CaO–SiO₂–Al₂O₃ system. Hatched area was mainly discussed in the past.

An approach completely different from current directions of research is required here. Soil, which usually consists of minerals such as quartz, feldspar, and kaolinite (Al₂(Si₂O₅)(OH)₄), contains besides SiO₂ a considerable amount of Al₂O₃ (about 30%). For this reason, it is unsuitable as the raw material for calcium silicate hydrate formation. In order to produce solidified materials with soil as the starting material, sufficient strength development would be required in solidified bodies that contain hydrogarnet as the main phase.

Hydrothermal Solidification of Kaolinite

An example of a solidification experiment using nearly pure quartz (Indian quartz), kaolinite (Georgia kaolin—a clay mineral without micropores), and lime (CaO) is outlined here. In this experiment, the mass ratio of quartz, kaolinite, and lime were varied (Fig. 8, Table 1). So that

$$\frac{\text{kaolinite}}{\text{quartz} + \text{kaolinite}} : \frac{K}{(Q + K)} = 0, 0.1, 0.5, 0.9, 1.0,$$

and

$$\frac{\text{lime}}{\text{quartz} + \text{kaolinite}} = 0.21.$$

After weighing the materials, the lime was slaked and a further 10% of water was added and mixed to allow for easy forming. The specimens were obtained by uniaxial

Table 1. Compositions of the Experimental Mixtures

$K/(Q + K)$ (mass ratio)	Al/(Si + Al) (atomic ratio)	Ca/Al (atomic ratio)	Ca/(Si + Al) (atomic ratio)
0.0	0.00	0.23	0.23
0.1	0.05	0.24	0.23
0.5	0.24	0.31	0.24
0.9	0.45	0.45	0.24
1.0	0.50	0.50	0.25

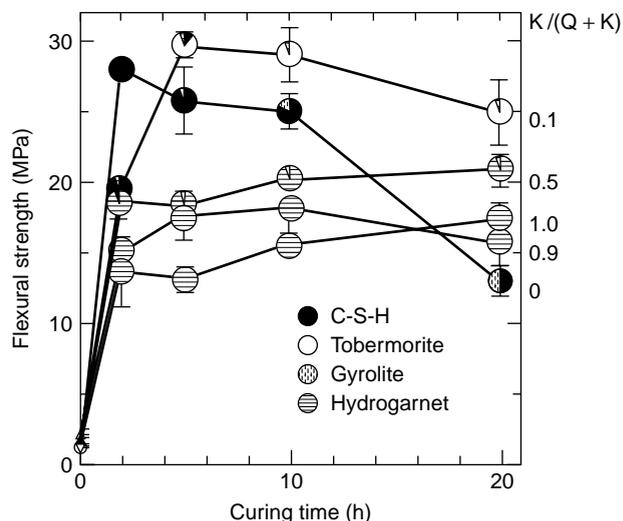


Figure 9. Generation of the flexural strength and the phases along with the hydrothermal processing.

press forming of this mixture at 30 MPa. The press-formed specimens were cured under saturated steam pressure at 200°C for 2 to 20 hours. (The compositional range of the specimens was Ca/(Si + Al) = 0.23 to 0.25, Al/(Si + Al) = 0.0 to 0.50.

The variation in the phases formed with treatment time is shown in Fig. 9, together with the strength development characteristics. It is recognized that in all specimens, hydrothermal treatment results in increased flexural strengths. It is also clear that hydrothermal solidification is possible even in those specimens with Al/(Si + Al) = 0.24 to 0.50 where hydrogarnet is the main phase formed. In the case of specimens with Al/(Si + Al) = 0 and 0.05 that correspond to compositions investigated frequently in the past, mainly calcium silicate hydrates are formed, and the maximum flexural strength of about 30 MPa is attained in 2 hours of treatment. Longer treatment times lead to decreased strength, particularly for the specimens with Al/(Si + Al) = 0.

In kaolinite-rich specimens with Al/(Si + Al) = 0.24 – 0.50, flexural strength reaches approximately 15 to 20 MPa in 2 hours along with the formation of hydrogarnet. Longer treatment times lead to only slight strength increases. The flexural strength is maximum for specimens with Al/(Si + Al) = 0.05, and becomes lower for larger Al/(Si + Al) values. However, the rate of decrease in strength becomes small with increasing Al/(Si + Al) value. In the specimens with Al/(Si + Al) = 0.24 to 0.50, the decrease in maximum flexural strength with increase in Al content is clearly smaller than in those with Al/(Si + Al) < 0.24.

These results are extremely significant:

1. By making hydrogarnet as the main phase, sufficient strength allowing the synthesized bodies to be utilized as building materials is obtained, although the strength may be somewhat lower than those of conventional calcium silicate hydrate materials.
2. Although the strength decreases with increasing amounts of kaolinite, it is possible to limit the

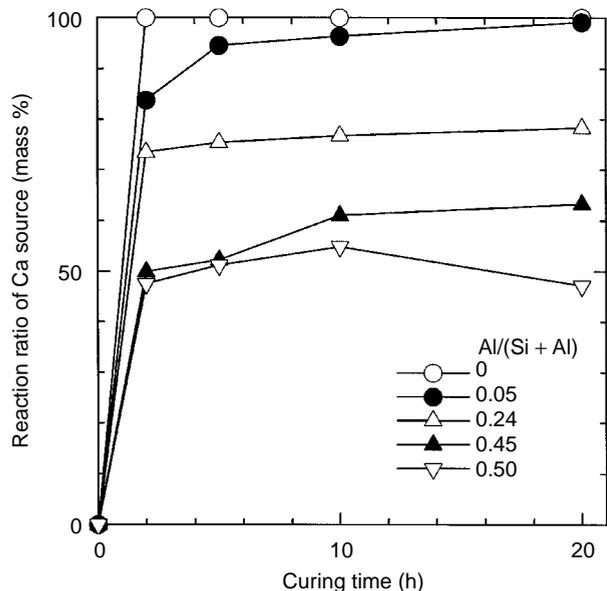


Figure 10. Influence of the starting materials on the reaction rates of Ca source with processing time.

strength decrease to acceptable levels by making hydrogarnet as the main phase. As a result, in actual practice, variations in composition of natural soil will not significantly affect the final physical properties of the synthesized material. In other words, this means that the range of starting materials that can be used is wide.

Unfortunately, the mechanism of strength development by hydrogarnet formation still remains unclarified although there is a possibility that this is related to the reaction scenario of Ca as explained below. In the case of specimens with $Al/(Si + Al)$ ratios lower than 0.24, almost 100% of the calcium is consumed within the first two hours of reaction (Fig. 10). For specimens with $Al/(Si + Al) = 0.24$ to 0.50, calcium is consumed rapidly during the first two hours, but since little or no further calcium consumption occurs, after that, the reaction ratio stagnates in the range 50% to 75%. This low reaction ratio may be one of the factors influencing strength development by hydrogarnet formation.

Some interesting results are provided by microstructural evaluation (Fig. 11). For $Al/(Si + Al) = 0$ and 0.05, acicular and platelike reaction products (identified by XRD to be C-S-H, tobermorite) can be observed to fill up the interparticle spaces after two hours of treatment. In contrast, for $Al/(Si + Al) = 0.45$ and 0.50, only platelike kaolinite particles are recognizable, and hydrogarnet cannot be observed despite the fact that its existence has been confirmed by XRD. Hydrogarnet crystals in sizes ranging from a few μm to some tens of μm have been reported to have been observed in hydrothermally synthesized bodies using slag as the starting material (27) and in slurries (23–25). In the current specimens, however, such hydrogarnet crystals could not be recognized, even through TEM observation. This indicates the possibility

that the hydrogarnet formed may be existing as ultra-fine particles.

There is a clear difference in the pore size distribution between the specimens that contain mainly calcium silicate hydrate phases and those that contain hydrogarnet as the main phase formed (Fig. 12). The pore size distribution of specimens with $Al/(Si + Al) = 0$ and 0.05 shows a peak at about $1 \mu\text{m}$ in the press-formed state. This peak location shifts to smaller values and the total pore volume decreases as hydrothermal treatment progresses. This indicates that the $1 \mu\text{m}$ pores in the press-formed state are filled by the C-S-H and tobermorite formed by hydrothermal treatment, and this densification is accompanied by strength increase (28,29). When the treatment time is increased to five hours or more in $Al/(Si + Al) = 0$, the peak in the neighborhood of $0.01 \mu\text{m}$ formed by two hours of hydrothermal treatment shifts toward coarser sizes. This pore-coarsening phenomenon is due to the formation of gyrolite (28), and it appears to be connected to the observed strength reduction.

The behavior of specimens with $Al/(Si + Al) = 0.45$ and 0.50 differs greatly from that of those with $Al/(Si + Al) = 0$ and 0.05. For $Al/(Si + Al) = 0.45$ and 0.50, the pore size distribution in the press-formed state does not shift with hydrothermal treatment time, but the number of pores decrease as the treatment time increases. As a result, pores around $0.04 \mu\text{m}$ are formed. Another characteristic of these compositions is that the pore volume hardly changes during this time.

From the preceding results, it is clear that the changes in pore size distribution of specimens with $Al/(Si + Al) = 0.45$ and 0.50, where hydrogarnet is the main phase formed, are different from those with $Al/(Si + Al) = 0$ and 0.05 in which pores are filled up by the calcium silicate hydrates formed. The formation of hydrogarnet and strength development without alteration of the pore sizes existing at the time of press-forming suggests that the solidified structure may be one that corresponds to that shown in Fig. 13. It is envisaged that ultra-fine hydrogarnet particles grow densely and in situ from the surface of kaolinite particles inward (forming $0.04 \mu\text{m}$ pores). Through this solidification mechanism, bonding of the kaolinite particles occurs with almost no alteration to the pore sizes existing at the time of press forming. The reason for the amount of hydrogarnet formed reaching a near-saturation level after two hours of hydrothermal treatment may be that the reaction has become diffusion controlled because of the kaolinite particles being covered by the hydrogarnet layer.

In similar experiments using silica sand containing low-purity clay minerals as the starting material, it has been found that although the type of phases formed during hydrothermal treatment change greatly with the treatment temperature and time, the strength of the solidified material is strongly influenced not by the type but by the amount of formed phases (Fig. 14) (30). On the other hand, it has been clearly shown that the strength of bodies formed by extrusion and casting processes, which require the addition of large amounts of water (binder) to the raw material, is much lower than that of bodies formed by uniaxial pressing (dry pressing). Although the detailed mechanism of strength development by hydrothermal treatment

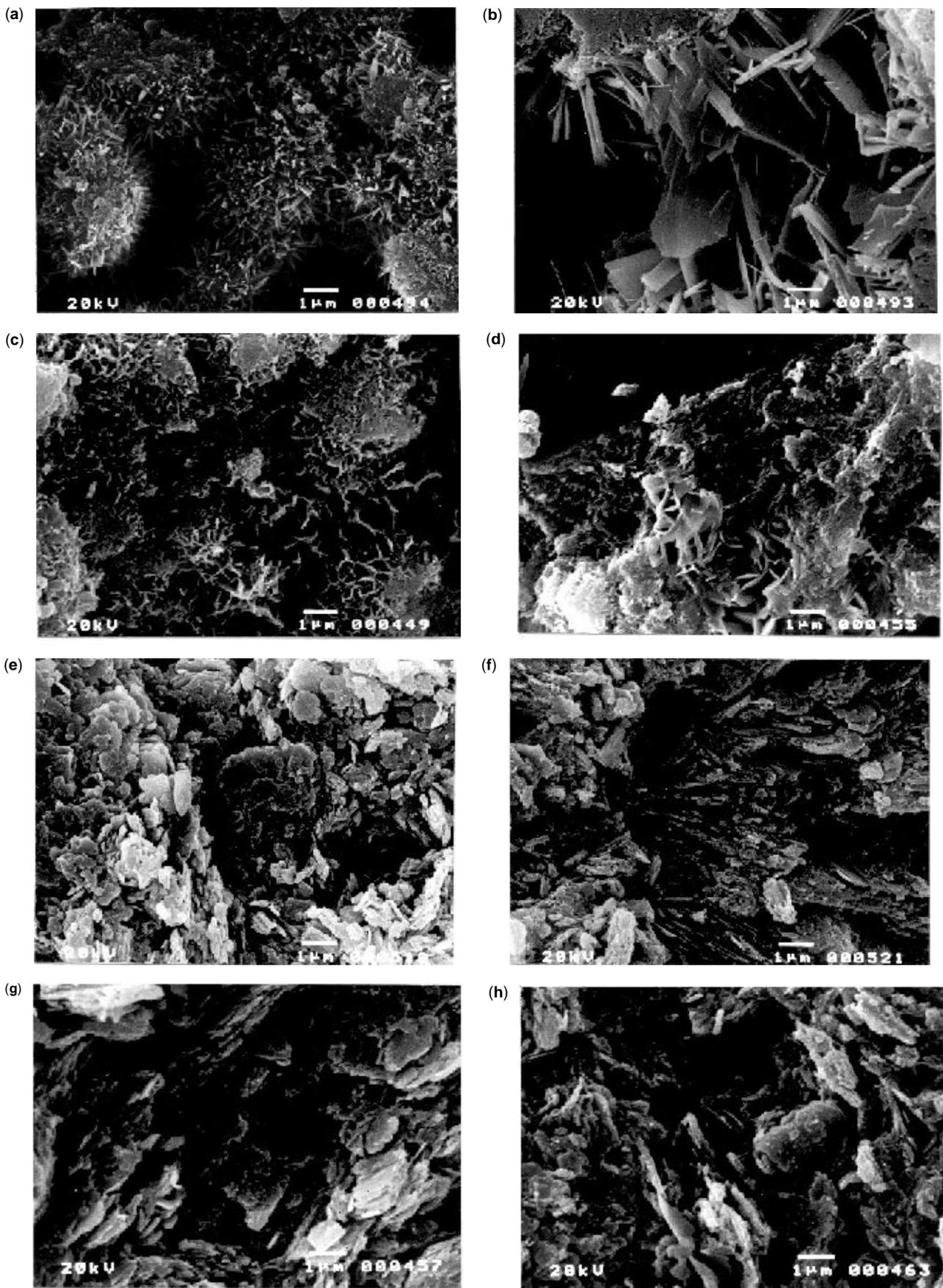


Figure 11. SEM photographs of the fracture surface of the specimens (Al/(Al + Si) = 0 (a, b), 0.05 (c, d), 0.45 (e, f), and 0.50 (g, h) cured for 2 hours (a, c, e, g) and 20 hours (b, d, f, h).

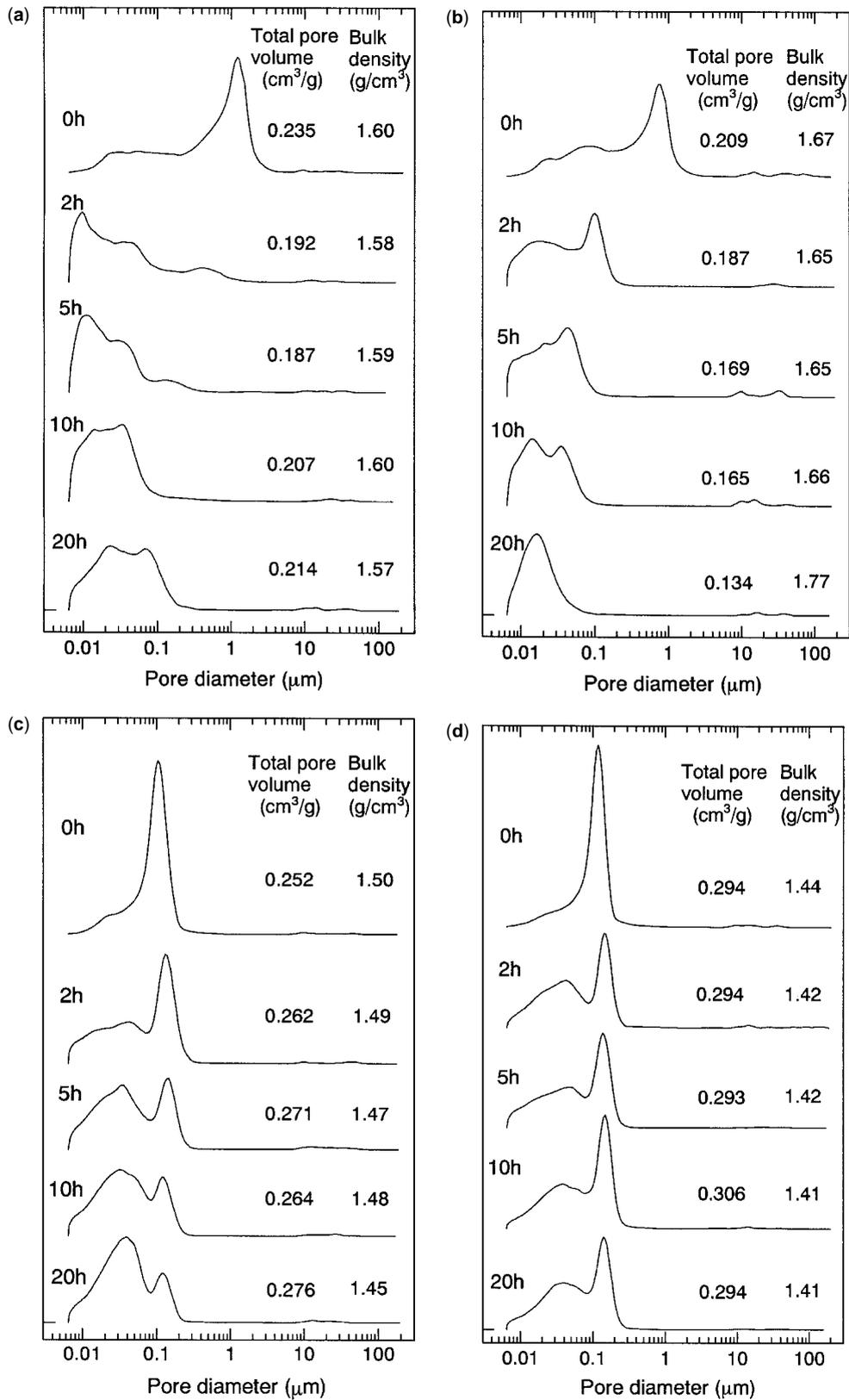


Figure 12. Pore size distributions of the specimen with processing time, a = Al/(Al + Si) of 0, b = 0.05, C = 0.45, and d = 0.50.

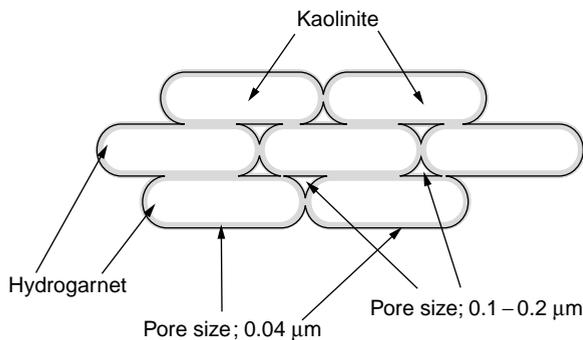


Figure 13. Estimated schematic figure for the formation of hydrogarnet under higher Al/(Al + Si).

of soil needs further investigations, it is believed that the strength development is attained through ultra-fine particles becoming uniformly dispersed within the densified body, that is, through a mechanism similar to the that of strength development in DSP (densified system particles—containing homogeneously arranged ultra-fine particles) (31–33).

As described previously, sufficient strength development is obtained through hydrothermal treatment of bodies that have been press-formed from a mixture of lime and the starting material, kaolinite. The strength is believed to be attained not by the filling up of the pores in the material but by the formation and dispersion of ultra-fine hydrogarnet particles within the press-formed, dense body. This solidification mechanism allows strength to be attained without destroying the agglomerated micropore structure. This is important from the point of view of the humidity regulation performance of the solidified bodies.

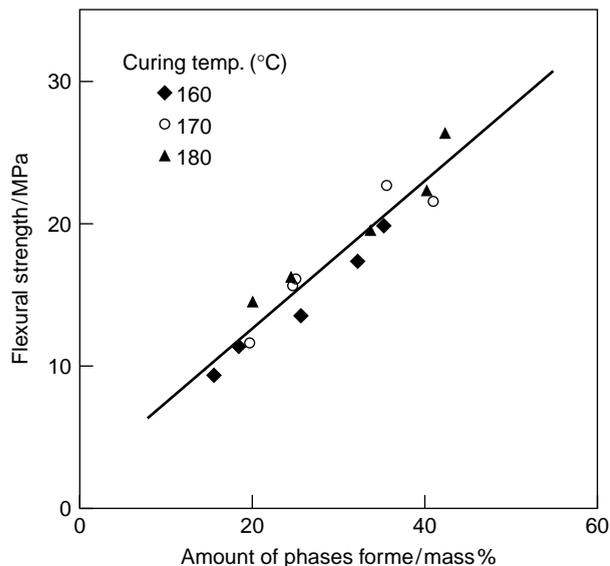


Figure 14. Examples for the generation of the flexural strength with amount of formed phases under the various curing time. The phases formed change with curing time and temperature. However, the strength seems to be controlled by the amount of the formed phases independently.

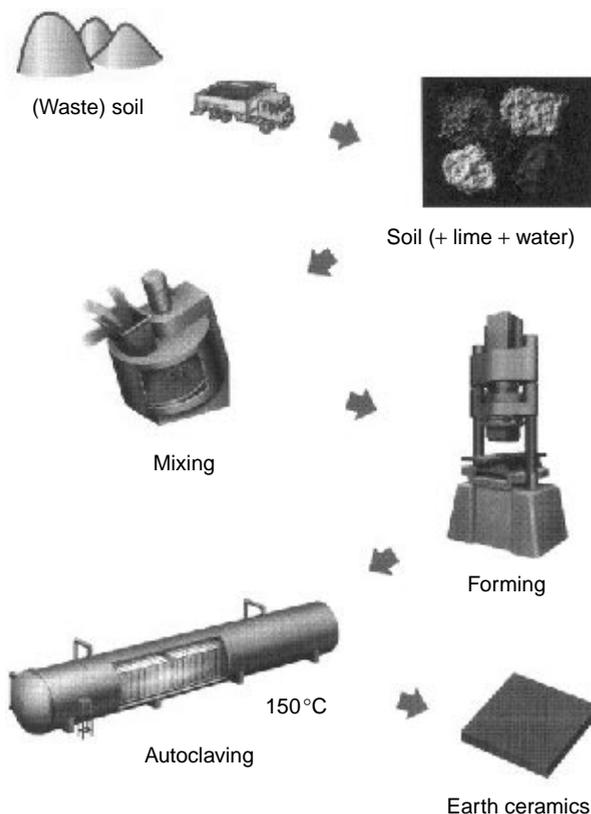


Figure 15. Actual processing of the earth ceramics manufacture

PERFORMANCE OF THE HYDROTHERMALLY SOLIDIFIED SOIL BODIES

Earth Ceramics

The industrial method of synthesizing hydrothermally solidified soil bodies is illustrated in Fig. 15. A little lime and water are added to the soil and mixed well. Since the treatment temperature is low, straw or other organic additives can be added if necessary in order to obtain higher strength or to enhance the finish. The mixture is then dry-pressed into tiles and then cured for a few hours at about 150°C in saturated steam pressure to obtain solidified bodies.

Among those industrial ceramics that utilize smaller amounts of energy (34) for their manufacture, ceramic tiles are considered to consume relatively little energy. The energy required for synthesizing earth ceramics is even lower, being about 2.7 GJ/m³ (35), which is only 1/6th that of the energy needed for ceramic tiles (Fig. 16).

Since there is little limitation with regard to the starting materials and the energy required for synthesis is small, it can be concluded from the point of view of nature's ecosystem that earth ceramics are materials with very small input and output.

Pore sizes in earth ceramics are concentrated in two regions: at around 0.05 μm corresponding to the initial pores at the time of press forming, and at around 0.01 μm (10 nm) reflecting the agglomerated structure of soil. This is about 1/10,000th of the pore diameters usually found in concrete

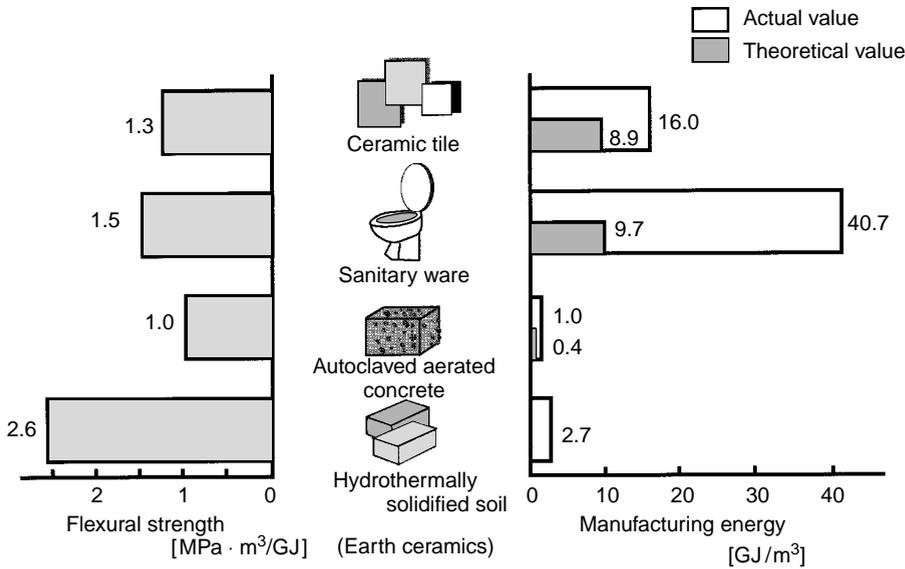


Figure 16. Energy consumption of the materials on processing.

and pottery (Fig. 17). The amount of water vapor absorbed at equilibrium when the relative humidity is varied from 40% to 80% at 25°C is shown in Fig. 18. Although the response in the case of earth ceramics is somewhat slower, it can be seen that they exhibit humidity absorbency properties as good as, or better than, that of wood because of the presence of micropores in the starting material (soil).

Living in a House of Soil

Earth ceramic tiles (size: 200 × 200 mm) were used as the flooring material for the living room of a highly airtight

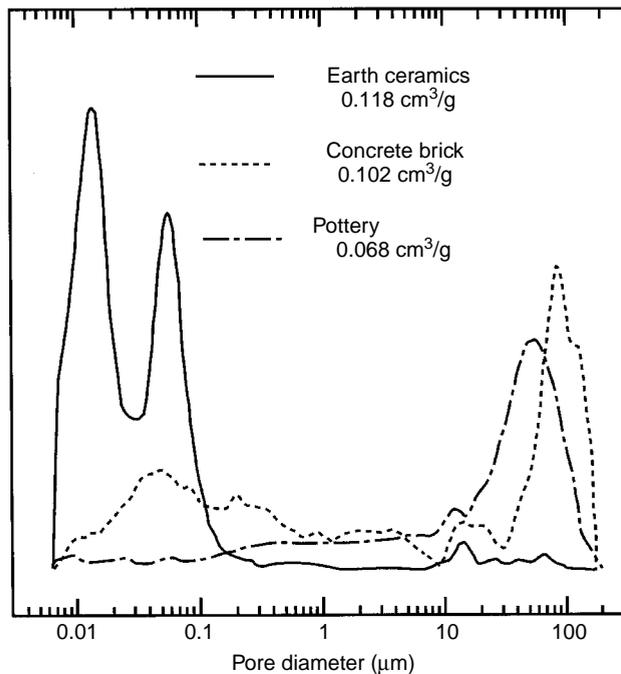


Figure 17. Pore size distribution of the earth ceramics.

and heat-insulated apartment (Fig. 19), and the changes in temperature and humidity were measured. An apartment in the same apartment complex with acrylic carpet flooring was used for comparison (reference apartment), the floor plan and family makeup being the same for both apartments. The measurements during the winter of 1997 (December to June) are shown in Fig. 20. The room with earth ceramic flooring exhibits very stable temperature variation. Since there were differences in the heating systems of the two apartments, the measurements were carried out after the heaters were switched off for the night (Table 2). The high heat insulation performance of earth ceramics was confirmed by the fact that the average rate of decrease in temperature (because of the difference between indoor and outdoor temperatures) in the earth ceramic floor apartment was 1.3°C compared to 5.7°C for the reference apartment. In the former apartment the humidity was also unaffected by the outdoor atmosphere and remained

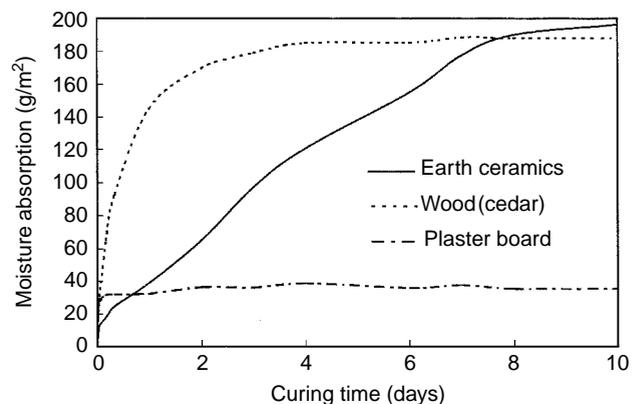


Figure 18. Variation of the moisture absorption capacity with time at 25°C when specimens (earth ceramics, wood, and plaster-board) were kept under relative humidity from 40% of the equilibrium condition to 80%.



Figure 19. Photo for the application of the earth ceramics flooring in the room.

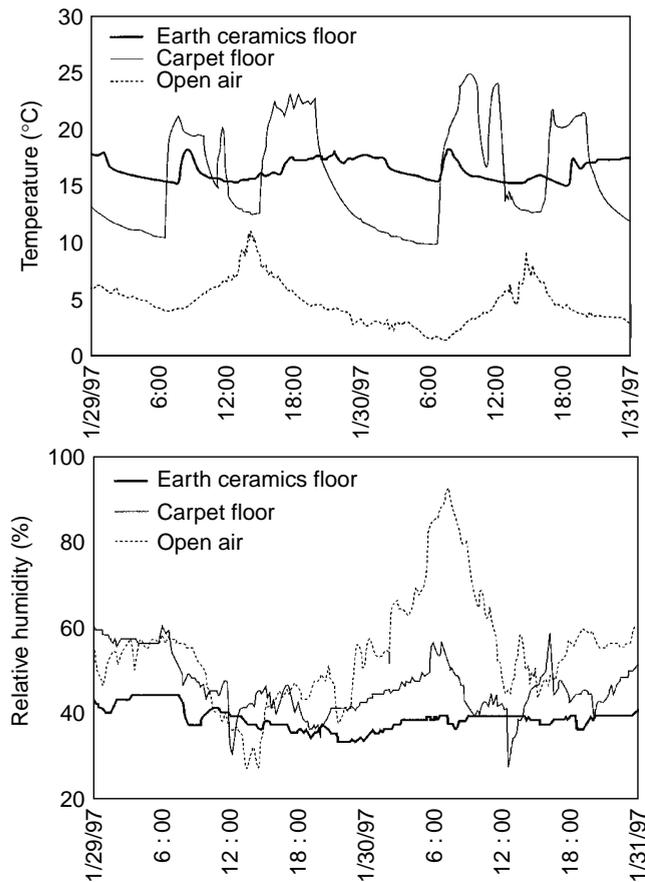


Figure 20. (a) Example of the variation of the temperature when the earth ceramics or carpet was used on the floor in winter. (b) Example of the variation of the relative humidity when earth ceramics or carpet was used on the floor in winter.

Table 2. Examples of the Temperature Variation after Turning off the Air Conditioner

	Earth Ceramics	Carpet
2/5	1.1°C	4.8°C
2/6	1.6	9.6
2/7	0.9	3.0
2/8	1.8	4.2
2/9	1.2	6.6
2/10	1.4	3.3
2/11	1.4	8.4
Mean	1.3°C	5.7°C

stable. In highly airtight and heat-insulated houses, the relative humidity increases as the temperature falls. The superior ability of earth ceramics for self-regulation of humidity is evident from the fact that the increase of relative humidity was 1.6%/°C in the reference apartment, but only 0.1%/°C in the earth ceramic floor apartment (Table 2). The variations of temperature and humidity measured at 6 am, 12 noon, and at 8 pm during one winter month are shown in Fig. 21. The temperature and humidity in the earth ceramic flooring were within the range 15 to 18°C and 40% to 50%, respectively. This shows that extremely stable and comfortable living environment can be obtained by the use of earth ceramics. The nighttime temperatures in the earth ceramic floor apartment were about four or five degrees lower compared to the reference apartment, but there was little recognizable difference in the effective temperature felt by the human body. This might be because the air temperatures near the ceiling, floor, and intermediate locations were about the same, the temperature difference being only about 0.5 to 1.0°C.

The measurements were continued for one year. It was found that compared to the reference apartment, the

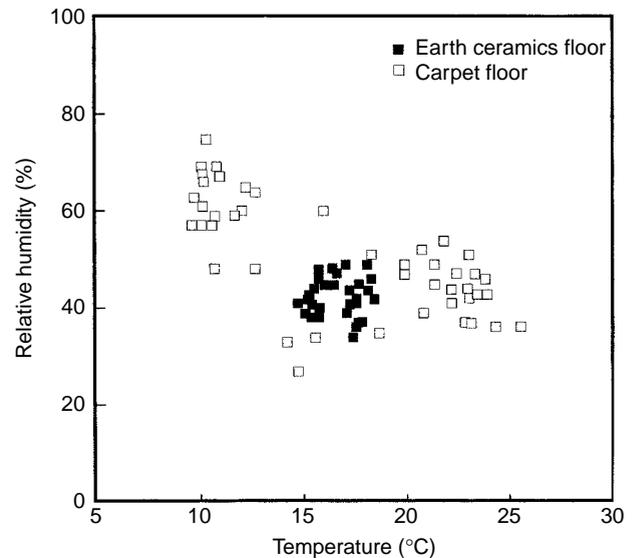


Figure 21. Variations of temperature and relative humidity during one winter month when the earth ceramics or carpet was used on the floor.

Table 3. Examples of the Relative Humidity Variation after Turning off the Air Conditioner

	Earth Ceramics	Carpet
2/5	0.0%/°C	1.7%/°C
2/6	-0.6	1.0
2/7	-0.8	1.0
2/8	1.4	1.4
2/9	0.0	1.8
2/10	-0.8	2.4
Mean	-0.1%/°C	1.6%/°C

variations of temperature and humidity were small throughout the year in the earth ceramic flooring apartment. In particular, the humidity was within the 40% to 70% RH range, which is the normal range of comfort for humans. Therefore, the use of humidifiers or dehumidifiers was not necessary, and the period of air-conditioner operation was short, resulting in low use of fossil energy. The amount of energy utilized for living (electricity, gas, water) in the apartment before and after the earth ceramic floor was installed is shown in Fig. 22 converted into an equivalent amount of CO₂. After remodeling, the amount of electricity thought to have been used for air-conditioning dropped sharply, and the seasonal fluctuation of energy consumption was controlled. On average, there was a 17% reduction of energy consumption that year. This energy consumption refers to the entire quantity required for living in the apartment (floor area = 72 m²). The energy

consumption reduction effect with respect to the living room (floor area = 32 m²) where the earth ceramics floor was laid was considerably greater.

In addition to the above beneficial effects, earth ceramics proved to be effective in controlling a chronic allergy suffered by the occupants of the apartment. Breeding tests with ticks, which are one of source of allergy, showed that at 25°C, 90% RH, 200 ticks multiply in number to 990, 1470, and 1895 in earth ceramics, vinyl cloth, and acrylic carpet, respectively. This clearly demonstrates the advantage of earth ceramics over other materials in restricting the breeding of allergy sources. It remains unclear at present, however, whether the low breeding rate in earth ceramics is due to the good humidity regulation characteristics of the material or whether it is the result of changes in the pH of the surface by humidity absorption.

NEW FUNCTIONAL MATERIALS

It is thought that hydrothermally solidified soil materials attain strength through the ultra-fine hydrogarnet particles becoming uniformly dispersed within the densified press-formed body, similar to DSP materials. This type of material synthesis holds many possibilities as the process for producing new functional materials. Such possibilities need to be investigated through further experiments. For example, another functional material could be tried in place of the ultra-fine dispersed particles.

From the point of view of humidity regulation, the author has investigated here only the process of hydrothermal synthesis. There are many other promising possibilities of using low-energy processes for producing new materials. One is the use of natural porous materials. Sepiolite ($Mg_5Si_8O_{20}(OH)_2 \cdot 8H_2O$) contains micropores of about 1 nm and mesopores that are a few nm in size. Allophane ($1-2SiO_2 \cdot Al_2O_3 \cdot nH_2O$) is an amorphous substance aluminosilicate formed during the weathering of volcanic glass, which is the major constituent of volcanic ash. It is widely distributed in nature in the form of hollow spheres of 3–5 nm in diameter (36, 37). As shown in the water vapor absorption/desorption isotherms of Fig. 23, both sepiolite and allophane exhibit high humidity absorption/desorption ability. Even at relative humidities less than 40%, they show this high ability. This is thought to result from the disordered surface structure and micropores smaller than 1 nm formed by the adsorbates. By adding a small quantity of binder to allophane-rich soil (Kanuma-soil/Japan), dry-forming the material to shape and firing at about 900°C, it is possible to obtain a solidified material that has high humidity regulation ability (trade name: Eco-carat). The humidity absorption/desorption characteristics of Eco-carat at 40% to 80% RH is shown in Fig. 24. Although there is limitation in the choice of material, the micropore volume of Eco-carat is around three times that of earth ceramics, and it shows extremely high humidity regulation performance. However, since these micropores disappear at about 1000°C because of phase changes, it is not possible to raise the firing temperature. With this restricted firing temperature, one cannot expect sufficient strength to develop by the characteristic sintering mechanism of

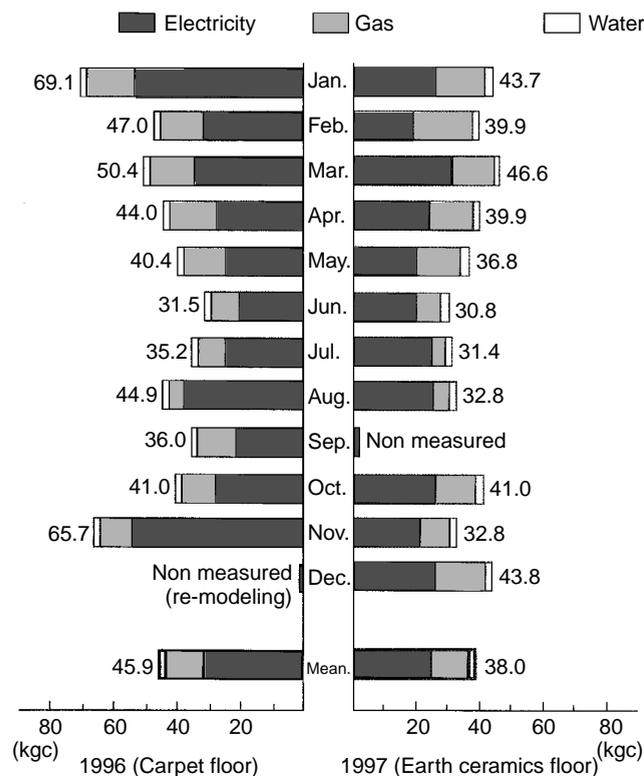


Figure 22. Amount of energy utilized for living (electricity, gas and water) before (carpet) and after the use of the Earth Ceramics floor.

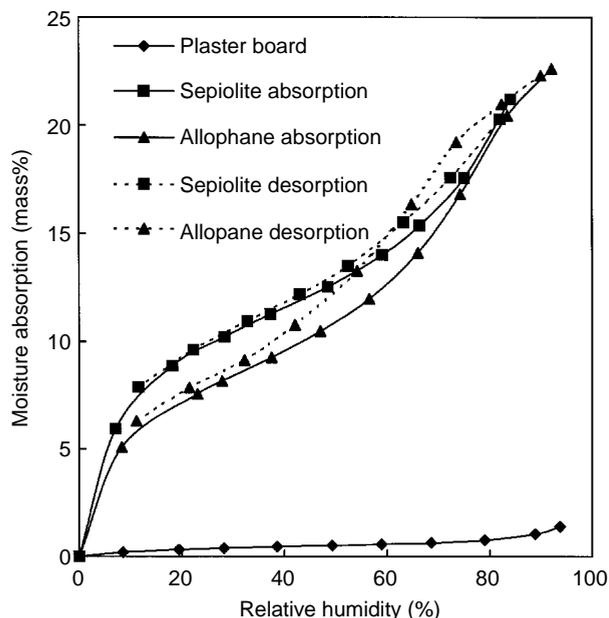


Figure 23. Water vapor absorption/desorption isotherms of sepiolite, allophane, and plasterboard.

ceramics. But because of the high humidity regulation performance, sufficient performance is obtained even if thin material is used. In actual practice, the material has been laid on interior walls to good effect.

CONCLUSIONS

The present age is one in which ignoring the global environment will have serious consequences for humankind. In maintaining, as far as possible, the inherent highly advanced properties and abilities of nature, it is important to develop technologies that convert, using the least amount of energy possible, these gifts of nature into forms that can be utilized in the human ecosystem. Earth ceramics can

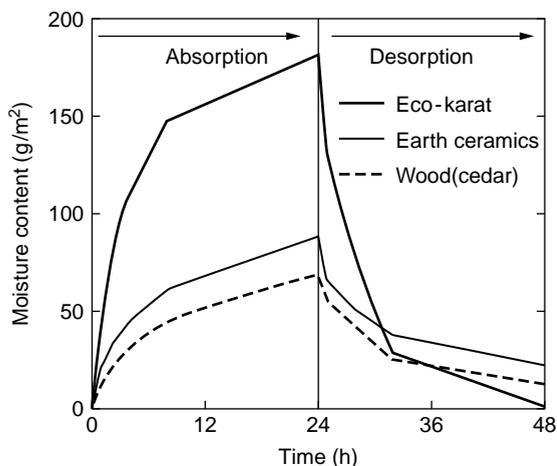


Figure 24. Humidity absorption/desorption characteristics of Eco-carat at 40% to 80% RH.

be obtained using almost any kind of soil including waste soil, and the temperature required for their solidification is low enough to allow waste industrial heat to be utilized for the solidification process. When used in the human ecosystem, earth ceramics exhibit energy-saving characteristics such as self-detection and self-regulation of temperature and humidity. After use, they can be returned to nature without burdening the natural ecosystem.

The development of new values of life and living by changing today's synthetic materials into smarter ones that minimize the load on Mother Earth and creation of such technology and culture should be considered as our greatest responsibility toward the next generation.

ACKNOWLEDGMENT

The author would like to thank Dr. H. Maenami and Mr. O. Watanabe, INAX Corp. Japan and Prof. Dr. T. Mitsuda, Univ. of East Asia for the helpful discussions.

BIBLIOGRAPHY

1. D.H. Meadows, D.L. Meadows, and J. Randers, *Beyond the limits*, Chelsea Green Publishing Company, Vermont, 1992, pp. 104–140.
2. L.R. Brown and H. Kane, *Full House*, W.W. Norton & Company, New York, 1994, pp. 35–106.
3. E.U. Weizsacker, *Erdpolitik*, Wissenschaftliche Buchgesellschaft, Darmstadt, Germany, 1989, pp. 68–121.
4. J. Tibbetts, *Green House*, Environmental Health Perspectives, **104**(10): 1036–1039 (1996).
5. H. Komine, Effects from house air quality on health, *Annual of Housing Research Foundation*, **23**: 5–17 (1997).
6. '99 Handbook of Energy and Economic Statistics in Japan, The Energy data and Modeling Center, 1999, pp. 32–33.
7. T. Uemura, J. Kohara and S. Tokoro, *Materials and Structure of the Wall for Humidity Control* (in Japanese), Shoukokusha, Japan, 1991, pp. 18–20.
8. C. Arai, T. Mizutani, Y. Murase, T. Hanakawa, and Y. Sano, Measurement of Pore Distribution by Water Vapour Adsorption, *Soc. Powder Tech. Japan*, **20**(3): 115–121 (1983).
9. V.G. Carter and T. Dale, *Topsoil and Civilization*, University of Oklahoma Press, Norman, OK, 1974, pp. 10–31.
10. S. Iwata, *Ecological Life* (in Japanese), Ienohikariyoukai, Japan, 1991, pp. 12–13.
11. G.E. Bessey, *The History and Present Day Development of the Autoclaved Calcium Silicate Building Products Industries*, Society of Chemical Industry, pp. 3–6, London, UK, 1965.
12. G.E. Bessey, *Sand-Lime Brick*, National Building Studies Special Report No. 3, 1–21 (1948).
13. P.D. Rademaker, H. Hibino, T. Mitsuda, *Electron Micrographs of Calcium Hydrates*, Annual Report of the Ceramics Research Lab. Nagoya Institute of Technology, **1**: 33 (1991).
14. H.F.W. Taylor, *The Chemistry of Cements*, Academic Press, New York, 1964, pp. 168–232.
15. S. Sohmiya, *Handbook for Hydrothermal Science* (in Japanese), Gihodoh, Japan, 1997, pp. 292–320.
16. F.H. Wittmann, *Advances in Autoclaved Aerated Concrete*, A.A. Balkema, Rotterdam, 1992, pp. 11–34.

17. T. Mitsuda, Synthesis of Tobermorite from Zeolite, *Mineral. J.*, **6**(3): 143–58 (1970).
18. T. Mitsuda and H.F.W. Taylor, Influence of Alumina on the Conversion of Calcium Silicate Hydrate Gels into 11Å Tobermorite at 90°C and 120°C, *Cem. Concr. Res.*, **5**(3): 203–209 (1975).
19. S.A.S. El-Hemaly, T. Mitsuda and H.F.W. Taylor, Synthesis of Normal and Anomalous Tobermorites, *Cem. Concr. Res.*, **7**(4): 429–38 (1977).
20. G.L. Kalousek, Crystal Chemistry of Hydrous Calcium Silicates: I. Substitution of Aluminum in the Lattice of Tobermorite, *J. Am. Ceram. Soc.*, **40**: 74–80 (1957).
21. E. Passaglia and R. Rinaldi, Katoite, a New Member of the $\text{Ca}_3\text{Al}_2(\text{SiO}_4)_3\text{-Ca}_3\text{Al}_2(\text{OH})_{12}$ Series and a New Nomenclature for the Hydrogrossular Group of Minerals, *Bull. Mineral.*, **107**: 605–18 (1984).
22. J.L. Larosa-Thompson and M.W. Grutzeck, C-S-H, Tobermorite and Coexisting Phases in the System $\text{CaO-Al}_2\text{O}_3\text{-SiO}_2\text{-H}_2\text{O}$, *World Cem.*, **27**(1): 69–74 (1996).
23. D.S. Klimesch and A. Ray, Hydrogarnet Formation during Autoclaving at 180°C in Unstirred Metakaolin–Lime–Quartz Slurries, *Cem. Concr. Res.*, **28**(8): 1109–17 (1998).
24. D.S. Klimesch and A. Ray, Effects of Quartz Particle Size on Hydrogarnet Formation during Autoclaving at 180°C in the $\text{CaO-Al}_2\text{O}_3\text{-SiO}_2\text{-H}_2\text{O}$ System, *Cem. Concr. Res.*, **28**(9): 1309–16 (1998).
25. D.S. Klimesch and A. Ray, Effects of Quartz Particle Size and Kaolin on Hydrogarnet Formation during Autoclaving, *Cem. Concr. Res.*, **28**(9): 1317–23 (1998).
26. I. Stebnicka-Kalicka, Application of Thermal Analysis to the Investigation of Phase Composition of Autoclaved Cement Pastes and Mortars, *Therm. Anal.* **1**: 369–74 (1980).
27. S.A. Abo-El-Enain, N.A. Gabar, and R.Sh. Mikhail, Morphology and Microstructure of Autoclaved Clinker and Slag-Lime Pastes in Presence and in Absence of Silica Sand, *Cem. Concr. Res.*, **7**(3): 231–38 (1977).
28. N. Isu, S. Teramura, H. Ishida, and T. Mitsuda, Influence of Quartz Particle Size on the Chemical and Mechanical Properties of Autoclaved Aerated Concrete (II) Fracture Toughness, Strength and Micropore, *Cem. Concr. Res.*, **25**(2): 249–54 (1995).
29. T. Mitsuda, K. Sasaki, and H. Ishida, Phase Evolution During Autoclaving Process of Aerated Concrete, *J. Am. Ceram. Soc.*, **75**(7): 1853–63 (1992).
30. O. Watanabe, K. Kitamura, H. Maenami, and H. Ishida, Hydrothermal Reaction of Silica Sand Complex with Lime, *J. Am. Ceram. Soc.*, (2000) in press.
31. L. Hjorth, Microsilica in Concrete, *Nordic Concr. Res.*, **1**: 1–18 (1982).
32. L. Hjorth, Development and Application of High-density Cement Based Materials, *Phil. Trans. R. Lond.*, **A310**, 167/73 (1983).
33. S. Brunauer, J. Skalny, I. Odler, and M. Yudenfreund, Hardend Portland Cement Pastes of Low Porosity, *Cem. Concr. Res.*, **3**: 279–93 (1973).
34. M.F. Ashby, *Materials Selection in Mechanical Design*, Pergamon Press, New York, 1992, p. 245.
35. H. Shin and T. Kurushima, Thermodynamic Consideration on Energy Consumption for Processing of Ceramics, *Bull. Ceram. Soc. Japan*, **32**(12): 981–84 (1997).
36. G.W. Brindley and G. Brown, Crystal Structures of Clay Minerals and their X-Ray Identification, *Mineralogical Soc. Monograph*, **5**: 104–109, 405–407 (1980).
37. B. Velde, *Development in Sedimentology*, Elsevier, New York, 1985, pp. 225–56.

SOUND CONTROL WITH SMART SKINS

C.R. FULLER

Virginia Polytechnic Institute
and State University
Blacksburg, VA

INTRODUCTION

In the last decade, there has been an increased interest in developing methods for the active control of sound radiation from vibrating structures (1). In one promising method, termed active structural acoustic control (ASAC), actuators are attached directly to the structure and are used to modify its structural vibration characteristics (spatial and temporal) in order to minimize the sound radiation (1). In ASAC, the actuators tend to be compact and thus cover only a very small portion of the structure; their effect is achieved because of the distributed elastic response of the structure. This technique has worked well for a number of applications, usually where the structure has a reasonable mobility and a low modal density of response. In some applications, however, the structure is quite massive or stiff (e.g., an electrical transformer casing), and it is extremely difficult to elicit the necessary control field response with practical control actuators. In this article we discuss a variant of the ASAC approach in which the control inputs come from a *smart* or *active skin* that covers all or most of the vibrating surface. A schematic of the smart skin approach is shown in Fig. 1.

The objective of the smart skin is to locally change the radiation impedance (the resistive component) of the structure in order to control the total radiated power in contrast to the conventional ASAC, which alters the dynamic response of the host structure. The sound radiation levels are directly coupled to the normal displacement w_{sk} of the smart skin. Thus modification of the transfer function between the structural displacement w_s and the smart skin displacement w_{sk} will lead to a change in sound radiation. This modification can occur via a decrease in the amplitude of w_{sk} , thus decreasing the sound levels, or via a change in the amplitude distribution of w_{sk} , causing the normal skin surface to be an inefficient sound radiation over an extended area. Since it does not drive the host structure,

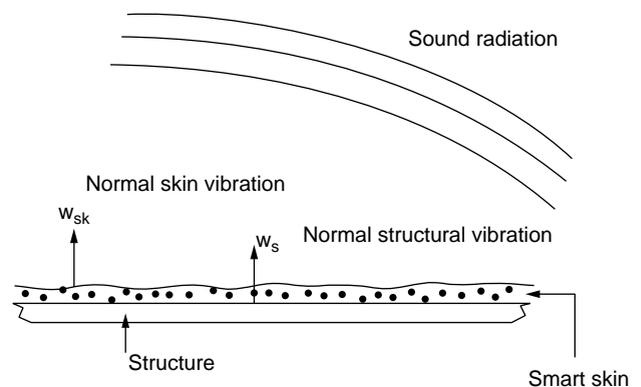


Figure 1. Concept of a smart skin for sound radiation control.

an active skin is suitable for heavy or stiff structures with low mobility. The negative trade-off, however, of the active skin approach, is that since it covers the structure completely, it may require many independent sections if the structural response is complex and near or above the sound radiation critical frequency (1). For illustrative purposes, two different skins are described: one based on a composite of acoustic foam and the piezoelectric polymer PVDF and the other on ceramic piezoelectric elements arranged in a motion amplification configuration.

SMART FOAM SKIN

Acoustic foam is commonly employed as a form of passive liner or skin in order to reduce sound radiation. For example, in aircraft interiors, a layer of acoustic foam is located between the inside surface of the fuselage and the interior trim panels. Generally, the foam completely covers the fuselage surface, and its acoustic purpose is to reduce noise transmitting through the fuselage to the interior. However, it is well known that passive treatments such as foam work well at high frequencies (>1000 Hz), and their performance is poor at low frequencies. Here we describe the use of a smart foam skin, which is a hybrid of the materials of passive foam and active piezoelectric elements (2). The objective is to develop a flexible active skin that combines good high-frequency passive performance with the low-frequency performance of an active system.

The smart skin has to be of such a form that it can be extended to cover complex, distributed surfaces similar to conventional acoustic foam. A typical configuration of the smart foam is shown in diagrammatic form in Fig. 2. The smart foam can be seen to consist of conventional acoustic foam with embedded layers of the flexible polymer piezoelectric material, PVDF (1). Two aspects of the embedded PVDF are significant. First, the PVDF is curved in order to couple the predominantly in-plane piezoelectric effect

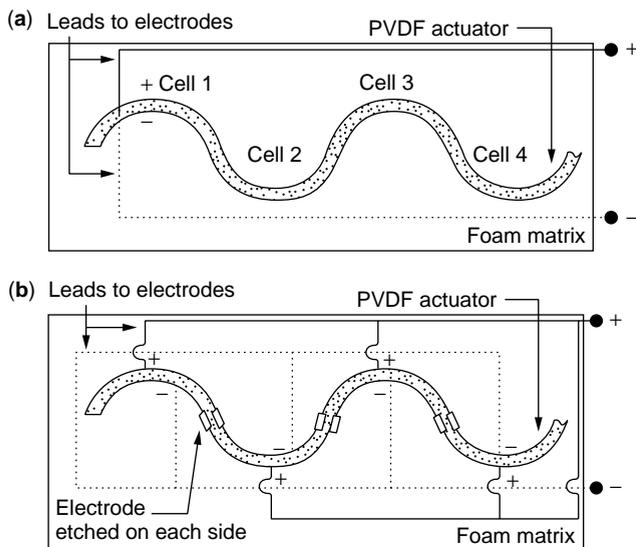


Figure 2. PVDF actuator configuration: (a) Original and (b) parallel.

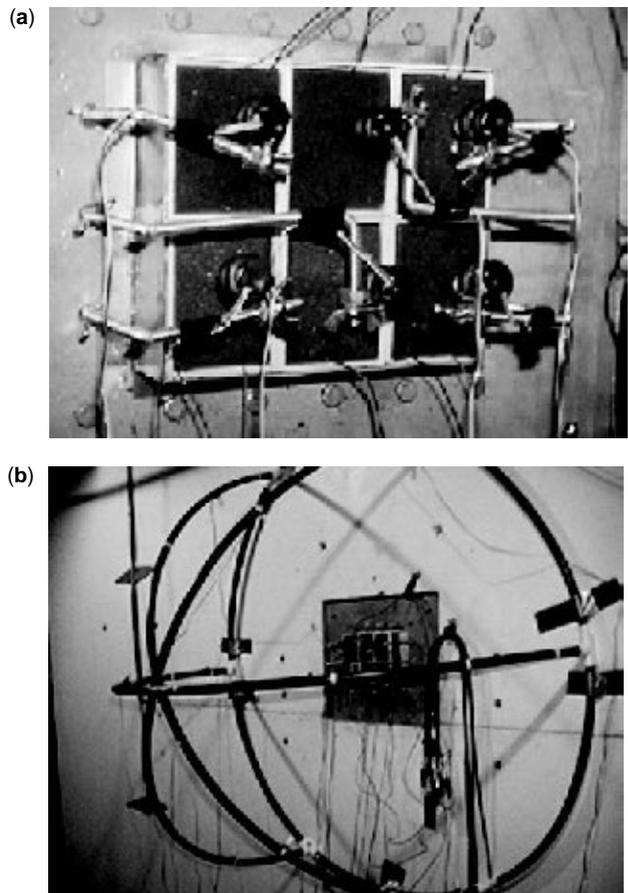


Figure 3. (a) Actuator and error sensor configuration. (b) Spherical dome for power measurements.

with the required normal motion (in effect creating a motion amplifier as discussed below), and second, as shown in Fig. 2(b), alternate half-wave sections of the PVDF can be wired out of phase. This second aspect is designed to increase the radiation efficiency of the active element (and thus its control authority) by causing all the PVDF sections to move in the same direction when a voltage is applied across the PVDF electrodes.

Figure 3(a) shows a picture of a smart foam active skin covering a plate of dimensions $170 \times 50 \times 1.5$ mm(3). The smart foam skin can be seen to be comprised of six independently controllable skin cells constructed as described above. In addition, each cell has a lightweight balsa wood wall used to increase the normal displacement by constraining the foam edge. Above each smart foam cell is a microphone located close to the foam surface and used as an error sensor in a MIMO feedforward Filtered-x LMS control approach (1) (see Fig. 11 for a typical feedforward LMS arrangement). The plate and smart foam skin is located in a rigid baffle located in an anechoic chamber. Figure 3(b) shows a hemispherical array of ten microphones located over the plate and in the radiated sound field. The microphones are used to estimate the total radiated sound power from the plate-smart foam system. The disturbance to the plate was provided by a piezoelectric patch actuator (1), bonded to the back. The

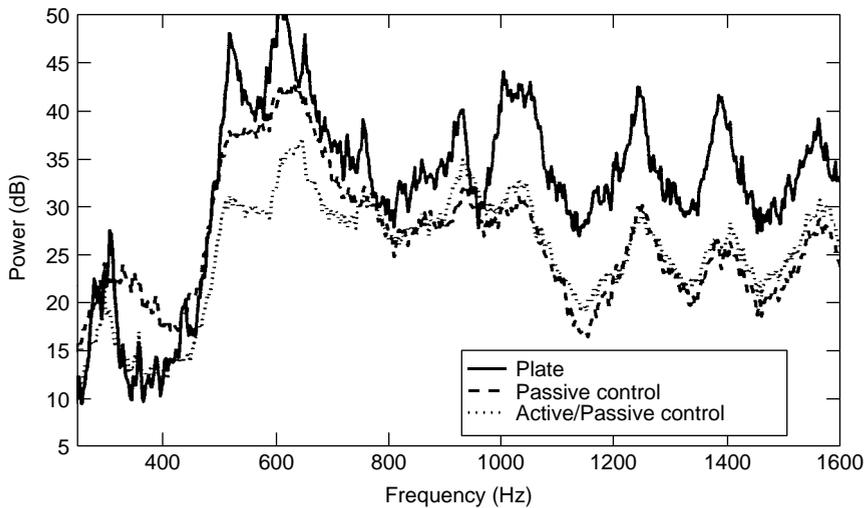


Figure 4. Radiated power for a broadband 1I1O case using multiple smart foam modules operating in phase.

reference signal for the LMS control approach was taken from the internal signal generator used to drive the disturbance (termed internal reference).

Figure 4 presents the radiated power with and without control, when all the smart skin cells are wired together in phase as a single channel of control. The error signal is provided by a single microphone located close to the smart foam surface and at the plate-foam center. Also shown is the passive effect of the smart skin when it is located on the plate but not activated. It is apparent that the passive effect of the skin is good at high frequencies above 1000 Hz but is limited to resonant frequencies of the base plate below this value. Turning on the active control provides reasonable attenuation at low frequencies, though there are some frequency ranges where the control is negligible, for instance near 900 Hz. In this case the smart skin transfer function is reduced in level. We now extend the controller so that the six smart skin modules can be controlled independently with a six by six LMS control arrangement. Figure 5 presents these new results for the low frequency range. For the results of Figure 5 three different reference signals control configuration are also studied: one using

an internal reference, one using an external reference signal taken from an accelerometer located on the plate (representing a more realistic arrangement), and an external reference signal with feedback (FB) from the active component of the smart skin removed (1). It is apparent that much improved performance is achieved over the SISO case of Fig. 4, particularly for the internal reference case, due to the multicell active skin being able to match the complex radiation impedance load near 900 Hz, for example (3). In this case the smart skin transfer function is also modified in a distributed manner.

Using an external reference signal also provides reasonable attenuation; however, it is reduced from the internal case implying that the system is acausal (1). Some of the lost performance is recovered when feedback removal is employed, indicating that the smart foam vibration has some input to the plate system.

Recently the smart foam skin has been used to demonstrate control of interior noise in aircraft (4). Figure 6 shows a smart foam skin covering four panels in the crown section of the fuselage of a Cessna business jet. The application is focused toward reducing cockpit noise due

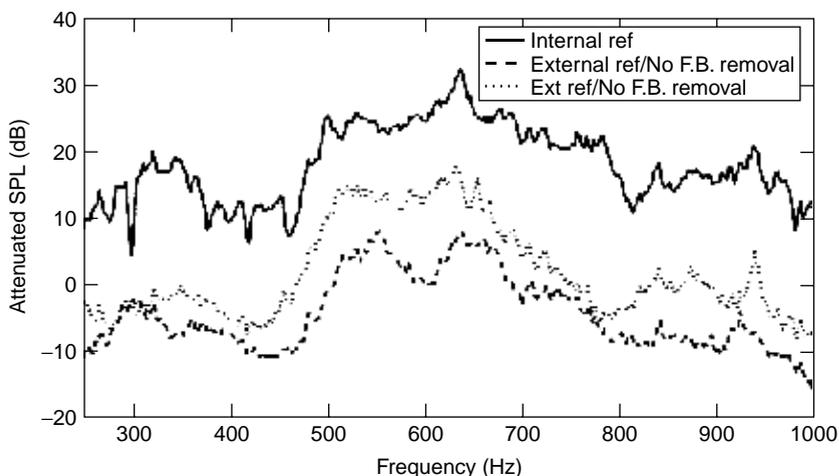


Figure 5. Attenuated SPL for broadband 6I6O case using multiple-independent smart foam modules.

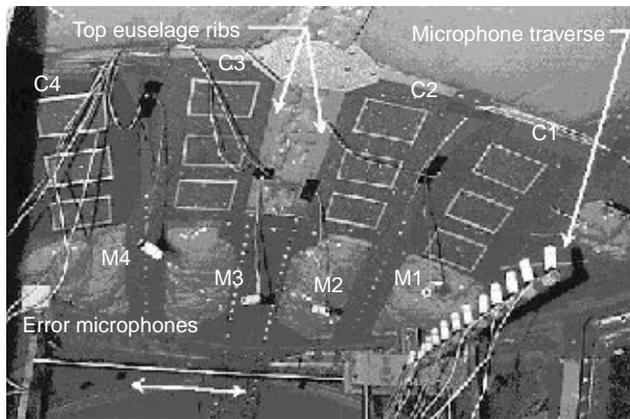


Figure 6. Cessna crown panel control arrangement.

to exterior flow separation over the crown of the aircraft. Error microphones were located as shown at the ear locations of the crew and a microphone traverse was used to measure the sound pressure levels in a plane at the crew head height. The flow noise disturbance was simulated by an exterior speaker located just over the crown of the aircraft and driven by band-limited white noise. A four by four feedforward LMS control approach was implemented using a realistic reference signal taken from an interior mounted accelerometer located on the fuselage at the aircraft crown (i.e., just under the excitation location). Figure 7 presents tabular results of the attenuation achieved at the error sensors (near the crews' ears) with an excitation band of 500 to 900 Hz. The reference speaker refers to the use of a reference signal from the signal driving the disturbance. The reference accelerometer refers to the use of a fuselage-mounted accelerometer as a reference sensor and in this case attenuation of the order of 2 to 4 dB are achieved. The global attenuation measured using the microphone traverse was found to be 2.5 dB with the active skin turned on. However, the active skin also provides a passive attenuation of 4 dB when it is installed over the bare fuselage panels and not turned on. Thus the total global attenuation of the smart foam active skin is around 6.5 dB, a significant difference. It also apparent from Fig. 7 that one of the main limitations on achievable attenuation is the causality of the controller when using an accelerometer as a reference signal. When the reference signal is taken from the speaker drive signal the control path delay is less, and the performance increases markedly. The results do, however, demonstrate the potential of the smart foam skin in reducing structurally radiated sound in a realistic application.

PIEZOELECTRIC DOUBLE AMPLIFIER SMART SKIN

Piezoelectric transducers tend to be high-force, low-displacement devices (1). In contrast, active noise control applications in air require high-displacement actuators, particularly at very low frequencies. Thus much of the work in developing piezoelectric based actuators for active noise control applications has been in designing devices that amplify their displacement. This amplification

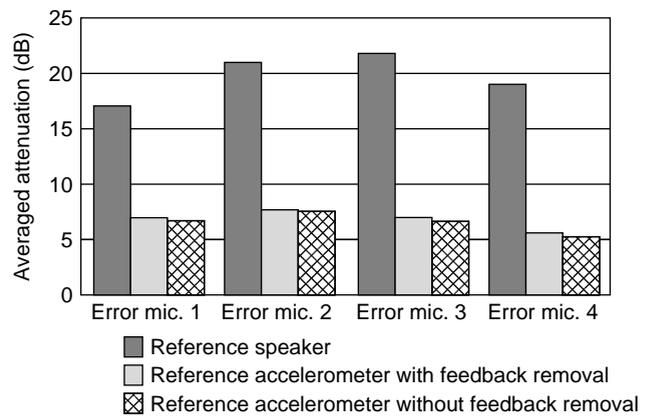


Figure 7. Averaged attenuation microphones for band-limited 500 to 900 Hz excitation.

is usually based on a geometric lever-type principle, and thus results in lower output force. More explicitly, the actuators are designed to have the correct source impedance relative to their load. In our application, the load is air with a relatively low impedance, thus the device needs to have a low source impedance for maximum power output.

Figure 8 shows a schematic diagram of a piezoelectric double amplifier actuator, which is the basis of the second active skin concept (5). The legs of the element consist of piezoelectric bimorphs or unimorphs. In this case, the piezoelectric transducers are manufactured from the ceramic material PZT (1). These devices are amplifiers in that due to their asymmetry, small in-plane motions are amplified to larger transverse tip motions at the top of the legs. The tops of the legs are connected to a triangular or curved stiff, lightweight diaphragm as shown. Thus as the legs move in, the diaphragm is squeezed upward. Since the diaphragm axis is transverse to the tip motion, very small tip motions cause very large diaphragm motions (i.e., amplify it) in a vertical direction. Thus the complete structure comprises a double amplifier actuator and gives amplification ratios of diaphragm to piezoelectric element in-plane deflection of the order of 20 : 1. The whole configuration can be built in heights typically ranging from

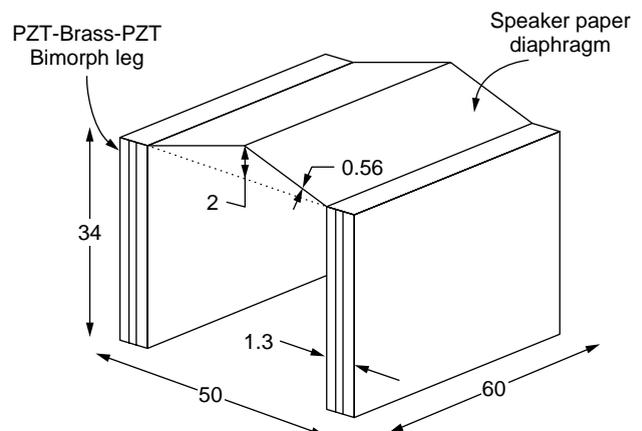


Figure 8. The active skin element. (domains in mm)

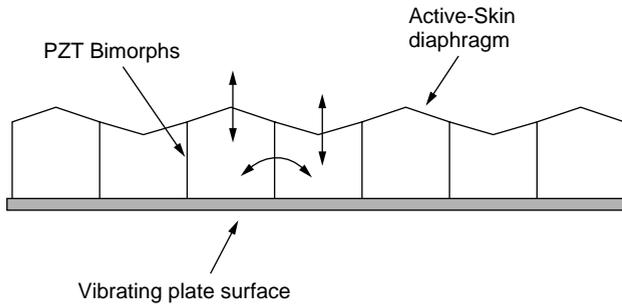


Figure 9. Smart skin constructed from piezoelectric double-amplifier elements.

3 to 6 cm, leading to a fairly compact device. In constructing an active skin of such devices, a number of them are positioned to completely cover the surface of a structure as shown in Fig. 9. The devices can be either located directly on the structure as shown or positioned just above it with a small air gap. In addition, the devices can be wired together as one channel of control or independently controlled, depending on the complexity of the base structural response.

Figure 10 shows an actual device designed and constructed by the Materials Research Laboratory at Pennsylvania State University. The device is 50×60 mms, 34 mms high, and was found to have a maximum cover displacement of $300 \mu\text{m}$ at 100 Hz. Figure 11 shows six of the devices arranged to completely cover the surface of a 170×150 mm aluminum plate of 1.5 mm thickness. In this test arrangement, the active skin cells are located on a perforated aluminum sheet which is located 5 mm from the surface of the radiating plate. Thus the active skin has a small air gap between its bottom surface and the radiating surface of the structure (5). Small accelerometers located on each active cell diaphragm are also apparent in Fig. 11. These accelerometers are used to provide time domain estimates of the radiated pressure in the far-field from the

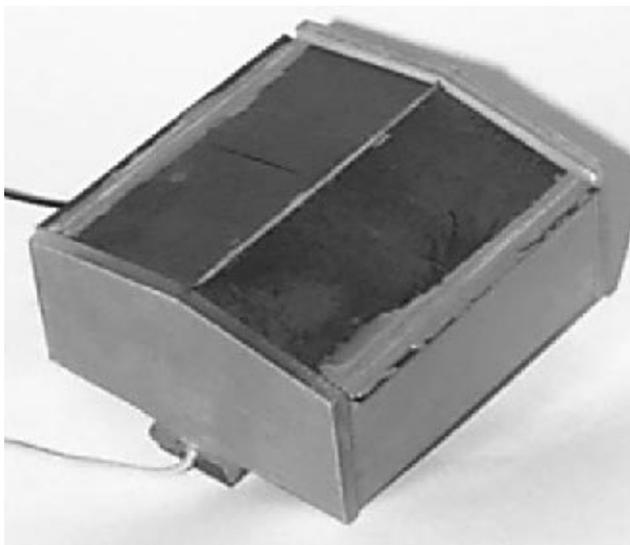


Figure 10. A single active-skin cell.

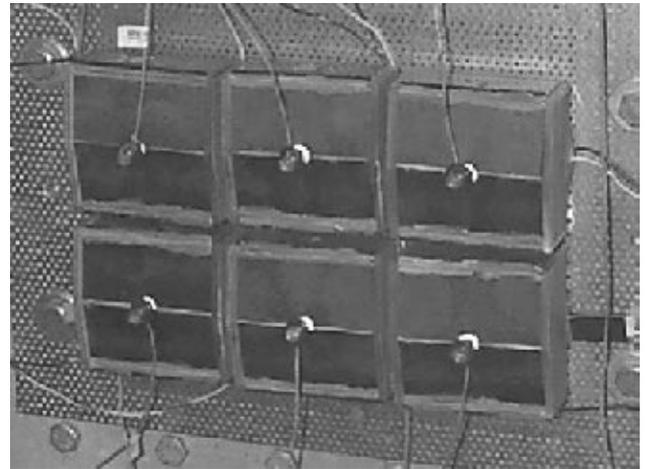


Figure 11. The active-skin in a top-mounted SAS configuration.

measured surface vibration data, termed structural acoustic sensing (SAS) and described in (6). Such approaches allow integration of the sensors into the smart skin itself.

The test plate and the active cells were mounted in a rigid baffle located in the anechoic chamber at VAL. A noise disturbance to the plate was provided by a small shaker attached to the back of the plate and driven with band-limited random noise. The radiated sound from the plate-skin structure was measured using an array of 16 microphones located on a hemispherical tube structure as described above and a microphone traverse that could measure the sound directivity in the horizontal midplane of the plate. The total radiated power from the plate could be calculated from the 16 pressure levels measured by the microphone hemispherical array (5).

Figure 12 depicts a schematic of the experimental rig and the control arrangement. The control approach used was the Filtered -x LMS algorithm (1) implemented on a TMS40 DSP. The shaker was driven with band limited noise of 175 to 600 Hz. The Filtered -x algorithm was

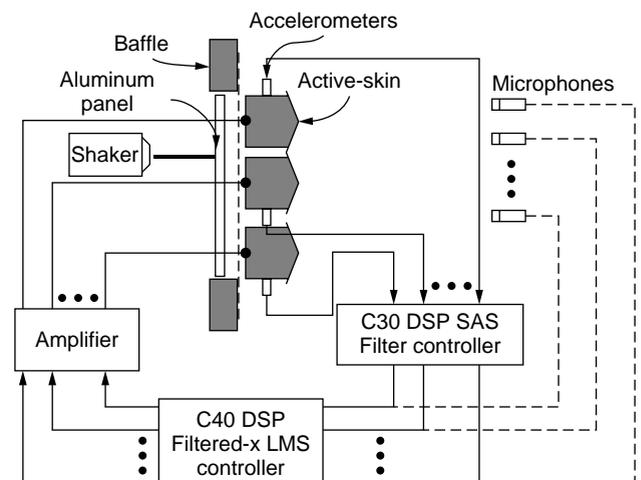


Figure 12. The active-skin experimental setup.

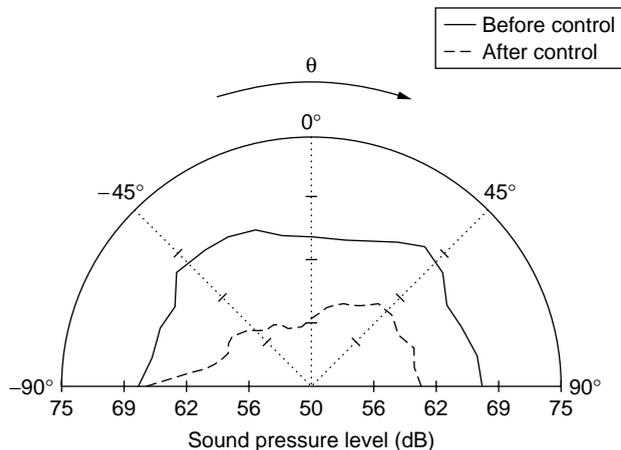


Figure 13. Total in-plane acoustic directivity (SPL), top-mounted accelerometer configuration with microphone error sensing.

executed with a 2000 Hz sample rate, and 175 and 96 tap FIR filters were used for the control and system identification paths, respectively. Since six independent cells were located on the structure to comprise the active skin, a six by six controller was implemented (5). Two tests were performed using different error sensors. In the first test, six microphones evenly distributed over the microphone array were used as conventional pressure error sensors located in the radiated far field. In the second test, the diaphragm accelerometer signals were used in the structural acoustic sensing approach, described in (6), to estimate the pressures at the same locations as the previous error microphones. These estimates were then used as error signals for the LMS algorithm.

Figure 13 presents experimental results of the directivity of the total radiated sound power measured using the far-field microphone traverse before and after the control using the active skin elements. It is apparent that the active skin provides global sound pressure level attenuation of the order of 10 dB, which is impressive since the excitation band encompasses multiple modes of vibration of the radiating plate (5). Figure 14 shows the corresponding radiated power versus frequency. Good control is seen over the complete bandwidth of 170 to 600 Hz except around 350 and 530 Hz, where anti-resonances occur in the plate-active skin system. The overall sound power reduction for the results of Fig. 14 is 10.9 dB. Further experiments were conducted using the accelerometers in the SAS approach, and the results are presented in Fig. 15. Good attenuation is evident across the frequency band, except near the system anti-resonance. The overall reduction is now 9.5 dB, which is still impressive. Thus the results demonstrate that it is possible to utilize an active skin that can provide significant attenuation of sound radiated from a structure vibrating in complex response shapes. The successful use of the accelerometers is significant in that it shows that an active skin with completely integrated actuators and sensors can be constructed to provide very significant broadband attenuation of sound radiated from structures under broadband excitation (5).

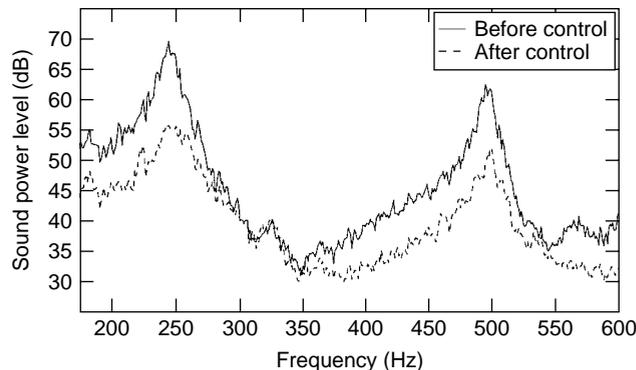


Figure 14. Radiated sound power spectra, top-mounted accelerometer configuration with microphone error sensing.

SMART SKINS FOR SOUND REFLECTION CONTROL

It should be noted that the above mentioned smart skin approaches could also be used to absorb sound impinging on structures by coating the structure with the smart skin. However, in this application, a modified sensing approach is needed in which the reflected or scattered wave components are independently (than the total pressure field) sensed and minimized by the controller. Fuller et al. (7) discuss such approaches using the smart foam noted above, and a combination of two microphones located near the smart foam surface are used to separate out the reflected wave information from the total pressure field (7). Figure 16 shows a schematic of the experimental testing in a plane wave acoustic standing wave tube. The noise is generated by a speaker at the right end of the tube and impinges on the smart foam. The two microphones are used to separate out the reflected and incident wave responses from the total pressure field. The reflected wave signal is used as error information to the LMS controller. The controller thus provides a control signal to the smart foam to minimize the reflected signal.

Figure 17 presents the measured intensity of the incident and reflected wave intensities versus frequency with the control off and on. With the control off, the incident and reflected intensities are almost equal at low frequencies

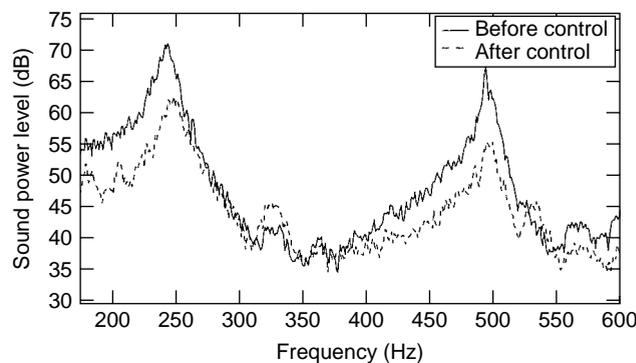


Figure 15. Radiated sound power spectra, top-mounted accelerometer configuration with SAS error sensing.

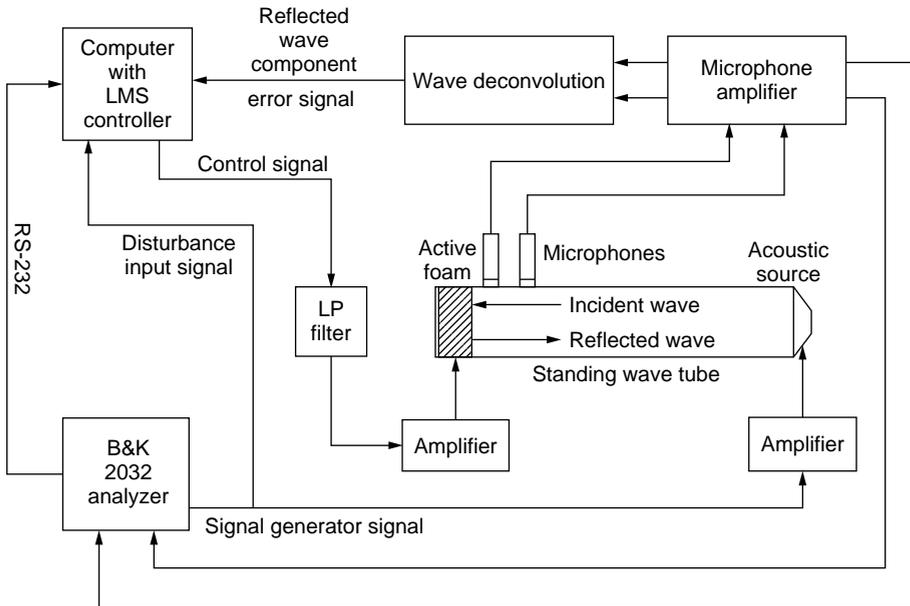


Figure 16. Smart skin reflection control experimental arrangement.

below 300 Hz, implying that the smart foam is acting like a rigid surface with very little sound absorption. Above 300 Hz, as is expected, the foam provides increasing passive sound absorption, and the reflected intensity is less than the incident. When the active control is turned on, the incident intensity remains the same, but the smart skin leads to a significant reduction in reflected sound energy below 300 Hz. This reduction in reflected sound due to the smart skin is apparent over the complete frequency range of Fig. 17. The two microphones can also be used to measure the acoustic impedance of the smart foam. When the control is turned on at low frequencies, the normal acoustic impedance of the foam falls from very large values to be almost identical to the characteristic impedance of the air. Thus the active element in conjunction with the controller

of the smart foam have modified the smart foam dynamics so that it looks like a perfectly sound absorbing surface.

ADVANCED CONTROL APPROACHES FOR SMART SKINS

The conventional control approaches used with a smart skin can be divided into two types; multi-channel feedforward, which is generally used when access to a coherent reference signal is available, and multiple input-multiple output state space feedback methods, which are often used when such a convenient reference signal is not available. These approaches are summarized in (1). As discussed above, the smart skin approach relies on covering a major part of the structure with independently controllable

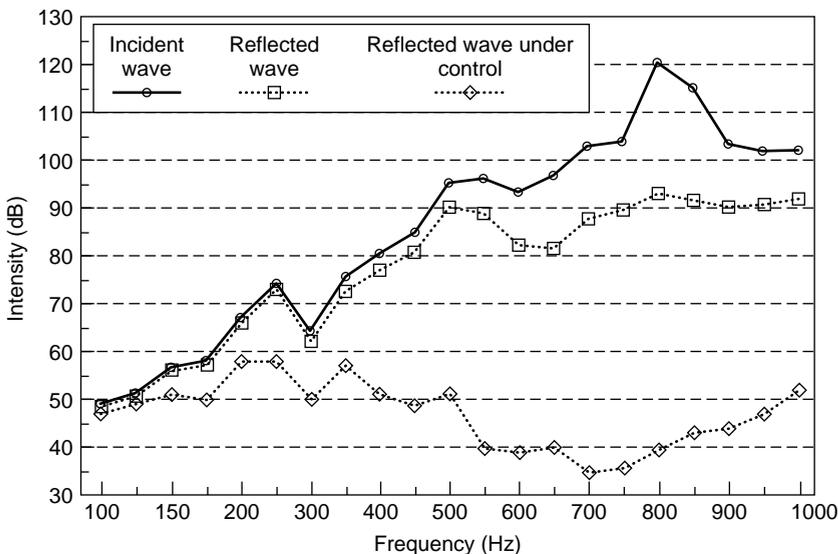


Figure 17. Reflection control using a smart skin.

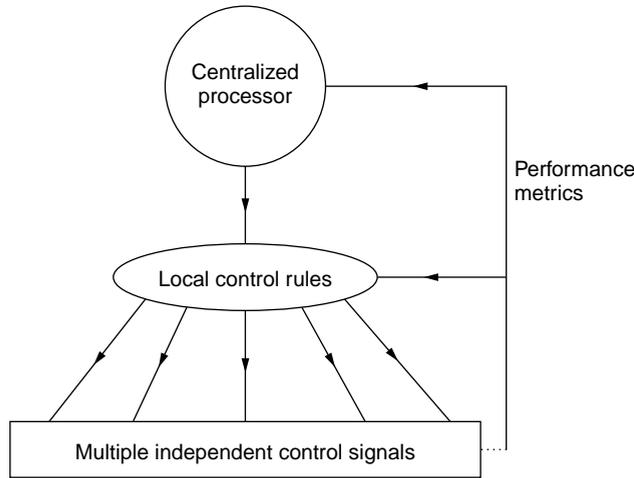


Figure 18. Biological control approach.

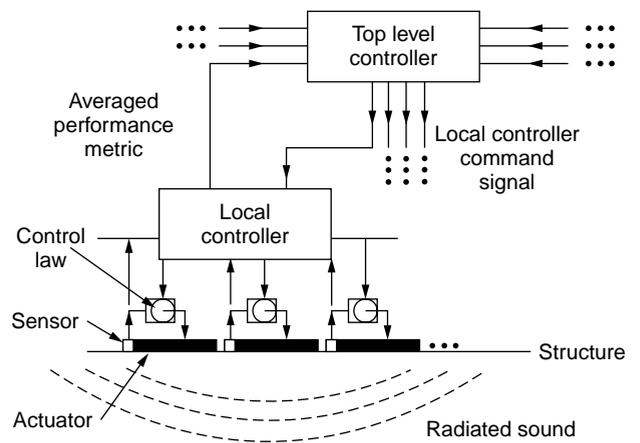


Figure 20. Schematic of a BIO controller arrangement.

elements. It can thus be seen that when the structure is large and/or the frequency of interest high (or wavelength relative to the structure short), many smart skin elements are required, implying a control approach with a very high number of control channels. In this case, the conventional approaches are likely to be unsuitable due mainly to computational limits on the control processor and stability/performance aspects. There are two different approaches suitable for high sensor/actuator count systems (8,9). Both approaches are hierarchical and are inspired by biological systems of muscle control. They are thus termed BIO controllers.

In the first approach, the smart skin elements are arranged into groups of “slave” actuators under the control of a “master” actuator. A schematic of the controller is shown in Fig. 18. A top-level centralized controller is used to send signals to the master actuators. Simple local control laws are used to modify and apply the same signal to nearby slave actuators. For example a very simple local law discussed in (8) would be take the same control signal, apply it to an in-phase, out-of-phase, or off-phase

slave actuator via simple analog switches and keep the setting that gives the lowest cost function value. Figure 19 shows a block diagram realization of such a control system for a feedforward approach. The process then continues to the next slave actuator, and so on, in a predetermined pattern. For the system of Fig. 19 the top-level controller could be digital, while the local control changes occur via simple analog-switching circuits. The approach in effect takes many independent actuators and connects them together via the local controller to create a suboptimal distributed actuator driven by one (or few) channel of control from the top-level centralized controller. The net result of such approaches is a large reduction in control channels to the top-level digital controller, and thus the computational overhead requirements are vastly reduced. The BIO approach in effect takes advantage of some limited knowledge of the dynamics of the distributed system to be controlled in order to reduce the extensive number crunching required in fully coupled optimal approaches.

In the second approach, local analog feedback loops are closed around individual smart skin elements and associated sensors as shown in Fig. 20. The analog local feedback loops have programmable feedback gains that are adapted by a higher-level digital controller in order to minimize a global cost function (obtained from an array of sensors) such as radiated sound power from the structure covered by the smart skin (9). Such approaches have been used to control sound radiation from very large structures. As with all feedback approaches, stability is an important issue. Thus work has also been performed to increase the stability margins via using directional feedback sensors to partially de-couple each local feedback loops. In addition, specialized distributed actuators are used that rolloff in level in the higher-frequency regions where the local open loop transfer function becomes non-minimum phase.

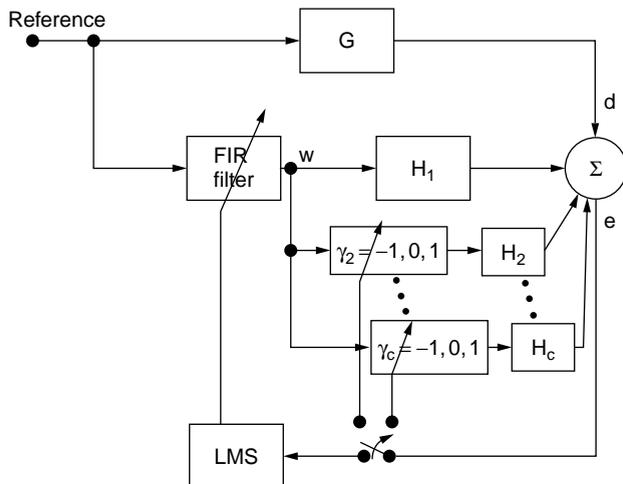


Figure 19. BIO controller with phase local control law.

CONCLUSION

The results presented have demonstrated the high potential for the implementations of a smart skin approach for reducing sound radiated from vibrating structures when the radiating structure is massive, stiff (i.e., low mobility), or the source vibration pattern is complex. The smart skin

has also demonstrated the possibility of combining active and passive control approaches in order to increase the control bandwidth and the efficiency of the active portion. A configuration has been demonstrated that further shows that the error sensors can be integrated directly into the skin and still result in a far-field sound reduction.

BIBLIOGRAPHY

1. C.R. Fuller, S.J. Elliott, and P.A. Nelson, *Active Control of Vibration*. Academic Press, San Diego, CA, 1996.
2. C.A. Gentry, C. Guigou, and C.R. Fuller. *JASA* **101**(4): 1771–1778 (1997).
3. C.A. Gentry, C. Guigou, and C.R. Fuller. Submitted to *JASA*, 1999.
4. C. Guigou and C.R. Fuller. *Proc. SPIE Smart Structures and Materials Conf.*, San Diego, CA, SPIE Vol. 3044, pp. 68–78, 1997.
5. B.D. Johnson, M.S. Thesis. VPI& SU. Blacksburg, VA, 1997.
6. J.P. Maillard and C.R. Fuller. *JASA*, **98**(5): 2613–2621 (1995).
7. C.R. Fuller, M.J. Bronzel, C.A. Gentry, and D.E. Whittington. *Proc. NOISE-CON 94*, pp. 429–436, 1994.
8. C.R. Fuller and J.P. Carneal. *JASA*, **93**(6): 3511–3513 (1993).
9. M. Kidner and C.R. Fuller, *Proc. 8th Conf. on Nonlinear Vibrations, Stability and Dynamics of Structures*. Blacksburg, VA, July 2000.

SPIN-CROSSOVER MATERIALS

JOEL S. MILLER
University of Utah, Chemistry
Salt Lake City, UT

Smart materials respond to their environment as illustrated by photochromic eyeglasses, that darken upon exposure to ultraviolet light to attenuate additional ultraviolet light. Hence, materials that have fast reversible responses to environmental stimuli are sought as components of smart systems. Similar to photochromic materials, thermochromic materials reversibly respond to heat and exhibit substantial color changes upon small changes in temperature. Spin-crossover materials(1) are a class of thermochromic materials that possess fast, reversible color changes amenable to display and memory devices (2). These color changes can also be induced by light (photochromic) or pressure (piezochromic) as well as heat. Due to the nature of the mechanism of their thermo-, photo-, or piezochromic responses (i.e., redistribution of the electron density at a metal ion site within the molecule), they are extremely fast and reversible. As a consequence of the (1) fast color change, (2) strong contrast between colors, and (3) the intermolecular interactions within the solid, the differing colors can be maintained for a long period of time, and (4) due to the lack of moving parts (i.e., no bond breaking or forming), these materials are completely recyclable and amenable to fast, low power-consuming, high-data-density display (2,3) and storage devices and “smart” materials and systems of the future.

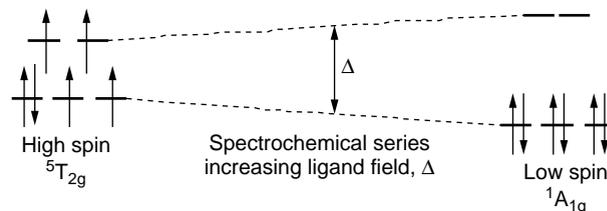


Figure 1. Switchover from a high-spin ($S = 2$) state to a low-spin ($S = 0$), as a Fe(II) state compound is cooled, requires ligands in the middle of the spectrochemical series. This transition can also be induced by light or pressure.

Thermochromism results from transition-metal complexes, such as Fe(II), which can be thermally stimulated to change from a colored low-spin electronic state to a frequently colorless high-spin state (1a) (Fig. 1). The high-spin ${}^5T_{2g}$ ground state for Fe(II) has a $t_{2g}-e_g$ splitting Δ , of $<11,000\text{ cm}^{-1}$, and the low-spin ${}^1A_{1g}$ excited state for Fe(II) has a Δ of $>21,000\text{ cm}^{-1}$. $\Delta \sim 16,000\text{ cm}^{-1}$ for Fe(II) surrounded with six unsaturated nitrogen-bound ligands, FeN_6 , can be induced to switch between the high-spin and low-spin states. Upon switching between the high- and low-spin states on cooling, FeN_6 has a significant decrease in Fe–N distances by $19 \pm 5\text{ pm}$, and an increase in the magnetic susceptibility χ due to a change of four in the number of unpaired electrons. From a thermodynamic perspective, the enthalpy ΔH is $10 \pm 6\text{ kJ/mol}$, and the entropy ΔS is $52 \pm 13\text{ J/K}\cdot\text{mol}$; hence, the transition is entropy-driven (1a). Additionally and importantly, the color changes from deeply colored red/purple to colorless upon switching to the high-spin state (see later). Concomitantly, the unit cell typically changes significantly.

The color change of the spin state switch is similar to that of liquid crystal displays (LCD) prevalent in digital watches; however, as a consequence of the mechanism, the thermochromic metal complexes change colors much faster without degradation upon cycling with respect to LCDs (3). Due to the change from low to high spin, this class of materials is referred to as spin-crossover materials. In addition to the technologically important color changes, spin-crossover materials also exhibit a small, but measurable change in magnetic susceptibility. This sharp transition in the change in the magnetic properties is illustrated by the temperature dependence of the magnetic susceptibility–temperature product for $\text{Fe}(\text{o-phenanthroline})_2(\text{NCS})_2$, which undergoes a first-order phase transition from a low- to a high-spin state at -97°C (4), (Fig. 2).

Materials that can be easily and reversibly stimulated to change colors for an innumerable number of cycles have been exploited for display devices. Liquid crystal displays (LCD) found in digital watches, are a common example (3). Materials that have greater switching speeds, sharper contrast, and enhanced stability enabling more duty cycles may lead to improved display and memory devices in the future. Spin-crossover materials can exhibit sharp color changes from small changes in temperature (i.e., they are dramatically thermochromic). As a consequence of the thermochromic mechanism (redistribution of the electron density within the molecule without either bond breaking

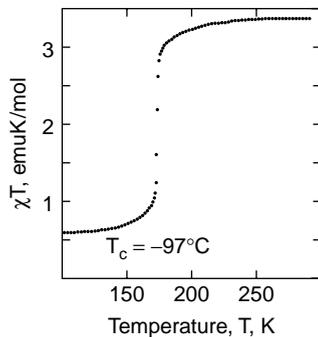


Figure 2. Temperature dependence of the magnetic susceptibility–temperature product for $\text{Fe}(\text{o-phenanthroline})_2(\text{NCS})_2$, which undergoes a first-order phase transition from a low- to a high-spin state at 176 K (-97°C) (4a).

or forming), they are extremely fast and recyclable and hence are candidates for high-data-density display and storage devices of the future.

For display/memory devices, it is necessary that the transition temperature (T) is near room temperature, $\sim 22^\circ\text{C}$. This is, however, insufficient because the ambient temperature fluctuates and hence the transition needs to be effected over a broad temperature range, $17 \pm 27^\circ\text{C}$. To achieve this, the system must exhibit history-dependent behavior (hysteresis) such that the transition temperature for color change upon increasing temperature ($T\uparrow$) exceeds the transition temperature for color change upon decreasing temperature ($T\downarrow$) ideally by at least 50°C , that is, $T\uparrow - T\downarrow > 50^\circ\text{C}$. Molecules cannot exhibit hysteretic effects, but in a solid or film, interactions between molecules can lead to hysteretic effects. Hysteresis has been reported for $\text{FeL}_2(\text{NCS})_2$ ($L = (\text{N}_2(\text{CH})_2\text{N}-)_2$), where $T\uparrow = -128.7^\circ\text{C}$ and $T\downarrow = -149.5^\circ\text{C}$ (Fig. 3). (5) Thus, although the transition and $T\uparrow - T\downarrow$ temperatures are too low to be practical, the necessary phenomena have been demonstrated, and new systems that exhibit higher temperatures are needed.

Using a mixture of triazole, $\text{HN}(\text{CH})_2\text{N}_2$ (trz), and aminotriazole, $\text{H}_2\text{NN}(\text{CH})_2\text{N}_2$ (H_2Ntrz) ligands coordinated with Fe(II), a polymer of $[\text{Fe}(\text{trz})_{3-3x}(\text{H}_2\text{Ntrz})_{3x}]$

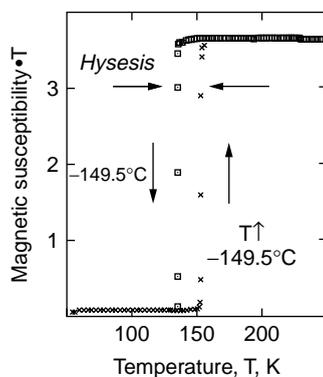


Figure 3. Temperature dependence of the magnetic behavior of $\text{FeL}_2(\text{NCS})_2$ showing the low-moment (purple) behavior below -149.5°C ($T\uparrow$) and high-moment (colorless) behavior above -128.7°C ($T\downarrow$) (5).

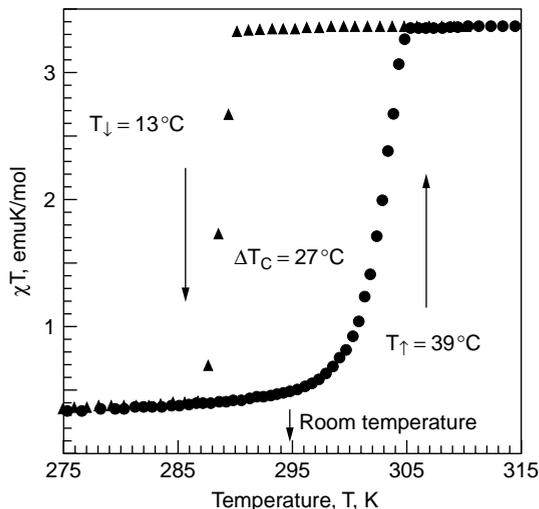


Figure 4. Temperature dependence of the magnetic behavior of $\text{Fe}(\text{trz})_{2.85}(\text{H}_2\text{Ntrz})_{0.15}|(\text{ClO}_4)_2 \cdot n\text{H}_2\text{O}$ showing the low-moment (purple) (Figs. 5 and 6) behavior below 39°C ($T\uparrow$) and high-moment (colorless) behavior above 13°C ($T\downarrow$) (6).

$(\text{ClO}_4)_2 \cdot n\text{H}_2\text{O}$ composition has been isolated, which for $x = 0.05$ exhibits $T\uparrow = 39^\circ\text{C}$ and $T\downarrow = 13^\circ\text{C}$ (2,6,7) (Fig. 4). These values bracket room temperature and demonstrate the feasibility of room temperature applications. In addition to the change in magnetic behavior, the color concomitantly as with hysteresis occurs (Fig. 5), from purple to colorless at 21°C (Fig. 6). Solid solutions of triazole and aminotriazole can be blended to lead to a systematic change in the transition temperatures: $T\uparrow = 296 - 160x$ and $T\downarrow = 313 - 180x$ in units of Kelvin.

Smart materials for future applications need to respond to environmental stimuli, and spin-crossover materials (1) are a moderately large class of materials that respond to heat, light, and/or pressure. This summary focuses on the use of heat to change the electronic structure of a material, which in turn leads to substantial and reversible color, magnetic, and structural changes. Most of the materials discussed in this context are inorganic coordination complexes demonstrating that (1) reversible first-order transitions occur, (2) such materials exhibit the technologically important property of hysteresis, and (3) both the transitions and hysteresis can occur at room temperature.

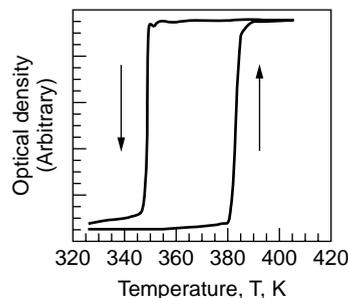


Figure 5. Temperature dependence of the optical density of $\text{Fe}(\text{trz})_{2.85}(\text{H}_2\text{Ntrz})_{0.15}|(\text{ClO}_4)_2 \cdot n\text{H}_2\text{O}$ at 520 nm showing hysteresis (purple) (1a).

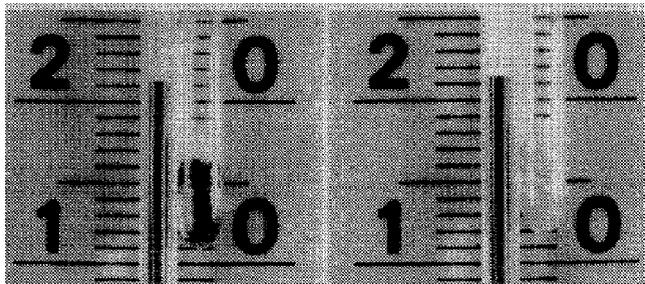


Figure 6. Dramatic color change for the dark purple low-spin state of $[\text{Fe}(\text{trz})_{2.85}(\text{H}_2\text{Ntrz})_{0.15}](\text{ClO}_4)_2 \cdot n\text{H}_2\text{O}$ below 21°C to the colorless high-spin state at 21°C (6).

ACKNOWLEDGMENTS

The author acknowledges continued partial support by the Department of Energy Division of Materials Science (Grant Nos. DE-FG02-86ER45271.A000, DE-FG03-93ER45504, and DEFG0296ER12198) and helpful discussions with Prof. O. Kahn.

BIBLIOGRAPHY

1. (a) P. Gülich, A. Hauser, and H. Spierling, *Angew. Chem.* **33**: 2024 (1994). (b) E. König, G. Ritter, and S.K. Kulshreshtha, *Chem. Rev.* **85**: 219 (1985), P. Gülich, *Struct. Bond.* **44**: 83 (1981).
2. (a) O. Kahn, E. Codjovi, Y. Garia, P.J. van Koningsbruggen, R. Lapouyade, and L. Sommier, *ACS. Symp. Ser.* **644**: 298 (1996). (b) O. Kahn and C.J. Martinez, *Science* **279**: 44 (1998). O. Kahn, J. Kröber, and C. Jay, *Adv. Mater.* **4**: 718 (1992).
3. C. Esher and R. Wingen, *Adv. Mater.* **4**: 189 (1992). R. Bissell, N. Boden, *Chem. Brit.* **31**: 38 (1995).
4. (a) B. Gallois, J.-A. Real, C. Hauw, and J. Zarembowitch, *Inorg. Chem.* **29**: 1152 (1990). (b) M. Sorai and S. Seki, *J. Phys. Chem. Sol.* **35**: 555 (1974).
5. W. Vreugdenhil, J.H. van Dieman, R.A.G. de Graaff, J.G. Haasnoot, J. Reedijk, A.M. van der Kraan, O. Kahn, and J. Zarembowitch, *Polyhedron* **9**: 2971 (1990).
6. J. Kröber, E. Codjovi, O. Kahn, O. Groliere, and C. Jay, *J. Am. Chem. Soc.* **115**: 9810 (1993).
7. O. Kahn, and E. Codjovi, *Philos. Trans. R. Soc. London. A* **354**: 359 (1996).

THERMORESPONSIVE INORGANIC MATERIALS

JOHN S.O. EVANS
University of Durham
Durham, UK

INTRODUCTION

The encyclopedia of Chemical Technology (1) defines a smart material as one that “responds to its environment in a timely manner” and “receives, transmits, or processes a stimulus and responds by producing a useful effect. . . .” One of the most common everyday stimuli is, of course, temperature. The effect of temperature on the vast majority of materials is well known—as materials are heated, they initially expand in volume before eventually melting, subliming, or decomposing. Thermal expansion is often viewed as a deleterious property. Engineers have to build expansion gaps into structures such as railway lines and bridges; the design of components in a car engine must be somewhat compromised so that they function both during start-up on a cold morning and when the engine is hot; historically, clock pendulums had to be carefully engineered to prevent slow running on warm days. Thermal expansion can, however, also be put to good use. It has long been known that the walls of bowed buildings can be pulled back into shape by a cooling iron bar; steel tires can be shrink-fitted onto the wheels of railway carriages. Although these examples represent a technologically useful response to the stimulus of increased (or decreased) temperature, they do not quite fall into the category of “smart.” However, the very simple concept of coupling together two materials, one that has a large and one a smaller coefficient of thermal expansion to produce a bimetallic strip can certainly be considered to create a smart composite body. Here, the strains induced by the higher expansion of one material cause the strip to bend as temperature is increased, leading to a simple temperature-sensing/responsive device.

The vast majority of materials known and used in technological applications have a positive coefficient of thermal expansion; the reasons for this behavior are discussed later. Certain materials, however, display the opposite property and contract in volume when heated. These materials thus have a negative coefficient of thermal expansion, α_l (vide infra), which has led to their somewhat confusing description as “negative thermal expansion” (NTE) materials. The properties of these materials and the origin of these effects form the main topic of this article.

Materials that have this unusual thermoresponse have a number of important technological applications. Applications related to specific materials are discussed later; more general comments are appropriate here. The first major area of application is in producing composite bodies. By mixing normal materials with a negative thermal expansion phase, one can achieve a composite that has a precisely controlled positive, negative, or even zero coefficient of

expansion. Second, certain of the materials discussed later can be chemically doped to control their expansion properties. Thus, one can envisage a single material that could be adjusted to have zero overall expansion. Such materials are of use where repeated thermal shock might lead to mechanical failure (an everyday example is oven-table cookware) or in optical devices such as high precision mirrors where any thermal expansion might distort optical properties. Materials that have strong intrinsic thermal contraction are most likely to be used as compensators for the positive expansion of other phases. For example, there is obvious interest in the electronics industry in producing circuit boards and heat sinks whose expansion properties match those of silicon; the dental industry would like ceramic fillings whose properties match those of teeth; the expansion properties of materials used in optical components such as refractive index gratings and optic fibers must be precisely controlled.

The observation that normal materials expand when heated can be explained at the most simple level by referring to Fig. 1 which shows a potential energy well for a typical diatomic molecule. As molecules are warmed, one excites higher and higher energy vibrational levels. Because interatomic potentials are typically asymmetrical (bonds are more extensible than compressible), this leads to an increase in bond length with temperature. A more formal explanation for the thermal expansion of solids is given by the Gruneisen relationship that relates the thermal expansion of a material [$\alpha_v = (1/v)(\partial V/\partial T)_P$ to its volume (V), specific heat at constant volume (C_V) and isothermal compressibility ($K = -(1/v)(\partial^2 V/\partial P^2)_T$]

$$\alpha_v = \frac{\gamma C_v K}{V}. \quad (1)$$

The Gruneisen parameter γ is defined by $\gamma = -d(\ln v)/d(\ln V)$, where v is a vibrational frequency and reflects the anharmonicity of a typical potential in a solid.

The frequencies of most modes in a normal crystal decrease as the volume increases, and the Gruneisen parameter of normal materials is typically positive in the range of 1 to 3 and only weakly temperature-dependent. Thus, the thermal expansion of normal materials may be expected to have a temperature dependence similar to their specific heat capacity. Therefore, one would expect thermal expansion to be zero at absolute zero, to increase rapidly as temperature is increased, and to approach a constant value at temperatures above the Debye temperature of the material. This simple behavior is at least approximated in many materials. In “negative thermal expansion” materials considered in the remainder of this article, there will always be an underlying expansive component caused by vibrational modes that tend to increase bond distances. In certain circumstances, however, these modes may be dominated by other more exotic effects. It is well known that strong chemical bonds (for example, those between highly charged elements and oxygen: Si–O, W–O, etc.) expand significantly less than weaker bonds (e.g., Na–O, K–O, etc.) (2). Thus

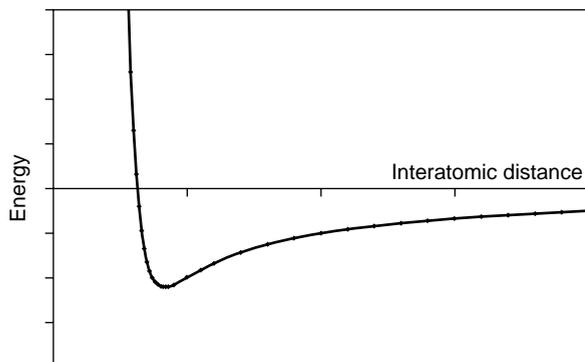


Figure 1. A typical interatomic potential well.

materials in which the unusual effects dominate normal expansion will often contain strongly bonded atoms.

There are two simple expressions commonly found in the literature to describe the thermal expansion of materials. The linear coefficient of thermal expansion may be defined by the differential form $\alpha_l = 1/l(dl/dT)$ or by the average coefficient of thermal expansion $\alpha_l = (l - l_0)/(l_0(T - T_0))$. Because the expansion coefficient generally changes as a function of temperature, it is important to quote the temperature range for any value of α_l . For anisotropic materials, α_l has been defined as $1/3\alpha_V$. Values of α_l for typical materials are quoted in Table 1 and are plotted in terms of % length extension for selected examples in Fig. 2.

ORIGINS OF NEGATIVE THERMAL EXPANSION

Phonons

The previous section described how thermal expansion can be related to the specific heat capacity of a material and, in turn to the population of vibrational modes. In general, this

Table 1. Thermal Expansion Coefficients of Common Materials

Material	$\alpha (\times 10^{-6}) \text{ K}^{-1}^a$	Temperature (K)	Ref.
Si	+2.45	373–223	3
Cu	+16.64	293	4
Ice	-5	45	5
Ice	0	63	5
Ice	+55	250	5
NaCl	+39.6	293	4
Al ₂ O ₃	+9.15	1000	6
α -Quartz	42	223–373	7
β -Quartz	0	575–1000	7
Fused quartz	0.5	300	4
Invar	0.07	278–303	8
NaTi ₂ (PO ₄) ₃	+4.2	300–1000	9
NbTi(PO ₄) ₃	-2.3	300–1000	9
ZrW ₂ O ₈	-9.1	0–350	10
ZrV ₂ O ₇	-7.1	400–500	11
Sc ₂ (WO ₄) ₃	-2.2	0–450	12
Y ₂ (WO ₄) ₃	-7	15–1373	13
ALPO-17	-17	18–300	76

^a α Values are, in general, temperature-dependent. Values are quoted at a specific temperature or across a given temperature range.

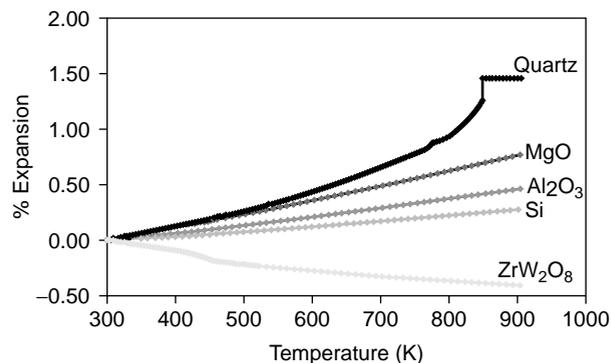


Figure 2. Thermal expansion curves for selected materials.

leads to positive thermal expansion. However, it was first pointed out in the 1950s that certain vibrational modes can lead to the opposite effect and negative thermal expansion (14–16). The simplest example of such an effect can be readily understood by referring to Fig. 3. If one considers an metal–oxygen–metal linkage in a structure, a longitudinal vibration (e.g., along an M–O–M bond) will lead to an extension of the M–M distance. During a transverse vibration, however, if the M–O distance remains essentially unchanged, the M–M distance decreases. This can be understood with respect to the Gruneisen constant for such a mode. A simple analogy for a transverse mode in a M–O–M linkage is a guitar string. If one plucks a note (excites a transverse vibrational mode) on a guitar string, then stretches the string using the tuning forks, the note moves to a higher frequency. Therefore, the Gruneisen parameter, $\gamma = -d(\ln v)/d(\ln V)$ for such a mode is negative. The contribution to thermal expansion for such a mode will also be negative.

Equation (1) is, of course, a simplification of the properties of a real material. In a real structure, different vibrational modes have different energies and are populated at different temperatures. In reality, one must consider the actual population of modes and use the average Gruneisen parameter at a given temperature where $\gamma_{av} = \sum c_i \gamma_i / \sum c_i$, and c_i weights the contribution of a mode to the overall specific heat C_V . In general, transverse modes have lower energy than longitudinal modes and are preferentially populated at low temperature. Thus,

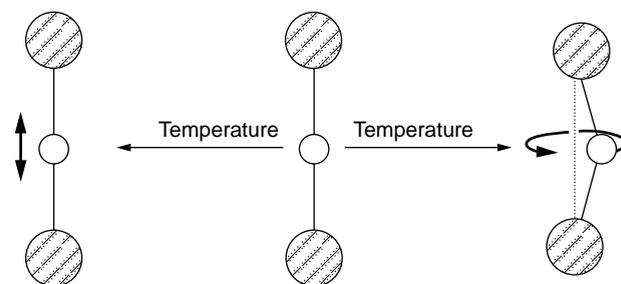


Figure 3. Schematic representation of the effect of different types of vibrational modes on thermal expansion. Shaded circles represent metal atoms and the open circle an oxygen atom. Longitudinal vibrations tend to expand metal–metal distances; transverse vibrations can lead to contraction.

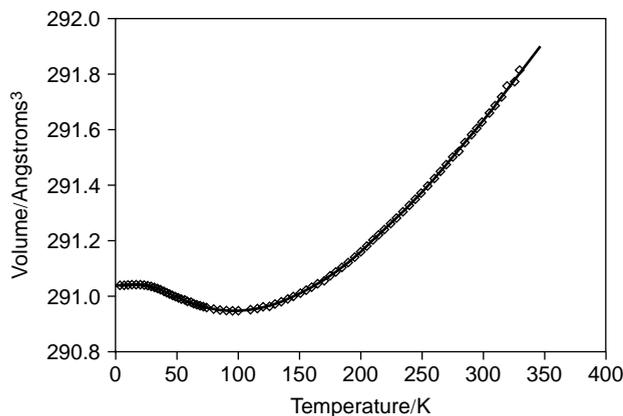


Figure 4. Thermal expansion curves of CuFeS_2 , using data of Knight (17). The solid line represents a fit to a simple two Debye type model. At low temperatures, the population of modes that have negative Gruneisen parameters leads to negative thermal expansion. At higher temperatures, modes that have a positive Gruneisen parameter dominate.

they may dominate the overall Gruneisen parameter at low temperatures and lead to negative thermal expansion. The population of low-energy transverse modes is the origin of negative thermal expansion at very low temperatures in a number of materials. Examples of this phenomenon are shown in Figs. 4 and 5. Figure 4 shows the cell volume of CuFeS_2 derived from powder neutron diffraction data as a function of temperature (17). Here, the thermal expansion is essentially zero at low temperatures [in accordance with expectations of Eq. (1)], is negative at temperatures below 100 K, where modes that have negative Gruneisen parameters presumably dominate, & approaches a constant positive value at high temperatures. Similar phenomena occur in many materials, including compounds that have rock salt, diamond, and zinc blende structures. Barron and co-workers have provided extensive data in this field (4). Interesting correlations can be made between the magnitude of the negative thermal expansion and factors such as the openness of the lattice and the rigidity (covalency) of the chemical bonds.

In the context of phonon-induced negative thermal expansion, the properties of ice are worth mentioning. The

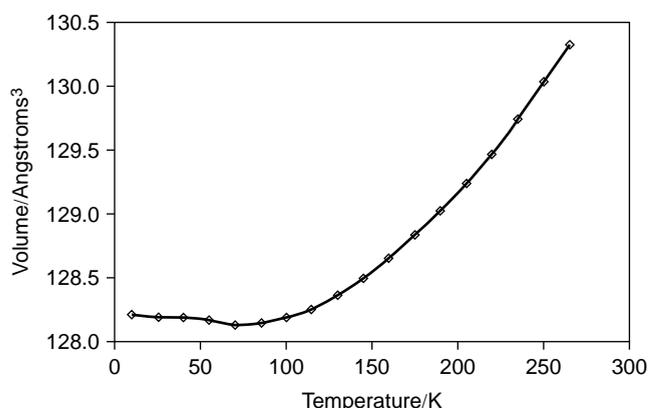


Figure 5. Thermal expansion of ice using data of Rottger (5).

increase in volume of water when freezing is one of the best known examples of an unusual thermoresponsive effect. Both the normal hexagonal and cubic forms of ice have a density (0.92 g/cm^3) that is significantly lower than that of water. This leads to the costly winter phenomenon of frozen pipes bursting then leaking during a thaw, and it also means that ice floats on water. For this reason, ponds and lakes always freeze from the top down, giving rise to a protective layer of ice on top of the water. This process is essential for the survival of aquatic life. This expansion of water can also be incorporated into simple, cheap devices to monitor the temperature history of frozen food and other commodities. Water can be sealed in a thin glass vial, and surrounded by paper impregnated by a water-soluble dye. When the water freezes, it expands and breaks the glass vial. Subsequent warming of the vial causes the water to melt, flow into the dye, and color the surrounding area. Thus, one can monitor whether the temperature of a commodity has ever exceeded 0°C during its storage/shipment. Less well known is the fact that ice shows negative thermal expansion in the solid state at temperatures below about 60 K. Figure 5 shows the temperature dependence of the cell volume of ice as determined by synchrotron X-ray diffraction (5). This cell volume contraction is again caused by the population of low-energy ($<50 \text{ cm}^{-1}$) transverse vibrational modes at low temperatures (5,18).

Rigid Unit Modes

The previous section described how vibrational modes may cause negative thermal expansion at low temperatures in certain materials. The population/importance of individual modes are intimately related to a material's structure. Vibrational modes can, however, give rise to NTE across far more extensive temperature regions in certain categories of materials.

Figure 6 shows the structure of quartz (7). Its thermal expansion properties are included in Fig. 2. At low

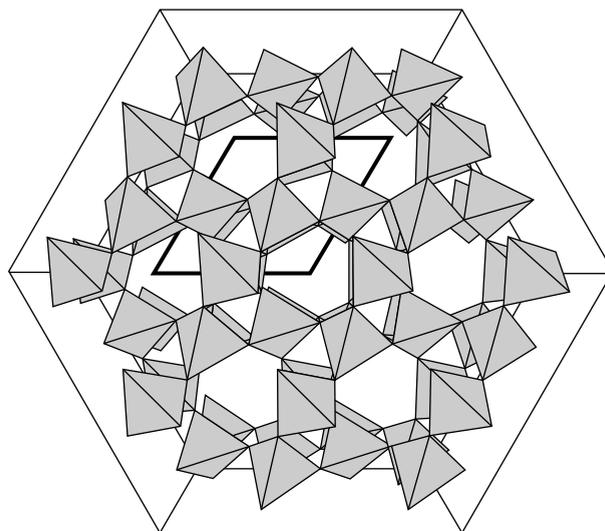


Figure 6. Structure of quartz viewed down the c axis showing corner-sharing SiO_4 tetrahedra. The ab plane of the unit cell is shown in bold lines.

temperatures, quartz has a positive coefficient of thermal expansion; at high temperatures ($>573^{\circ}\text{C}$), the thermal expansion becomes low or even negative. The ideas that led to an explanation of this phenomenon can be traced to the structural description of phase transitions in quartz given by Megaw (19). The structure of the high temperature (β) form of quartz can be described as a network of corner-sharing SiO_4 tetrahedra that form paired helical chains parallel to the crystallographic c axis (a 6_4 or 6_2 screw axis). The intertwined chains give rise to the hexagonal channels (important in the structure of β -eucryptite described later) through the structure that are visible in Fig. 6. The structure of the low temperature (α) form of quartz can be derived from that of β quartz by a coupled rotation of these tetrahedra through an angle θ about the $\langle 100 \rangle$ axes. In the more symmetrical β -quartz, θ is 0° ; in α -quartz, θ is 16.3° at room temperature (20). This coupled static rotation of essentially rigid polyhedral groups leads to a rapid reduction in volume on cooling and a large positive expansion coefficient on warming. Similar dynamic rotations of such groups at high temperature can explain the negative thermal expansion of β -quartz.

Dove and co-workers built on the concepts of rigid tetrahedral building blocks to produce models for negative thermal expansion in terms of such distortions (21). Figure 7 shows a schematic representation of a simple two-dimensional structure that can be considered a logical extension to the 1-D schematic of Fig. 3. If this figure represents a hypothetical 2-D structure containing metal atoms strongly bound to a square coordination environment of oxygen atoms, it can be readily appreciated that the structure has inherent flexibility. Certain distortions of the structure are possible in which (strong) M–O distances and (rigid) O–M–O bond angles remain unchanged. Such distortions are likely to be energetically facile and thus are good candidates as low energy vibrational modes of the material. Dove and co-workers termed such modes “rigid unit modes” (RUMs) and developed methodologies for identifying them in real 3-D structures.

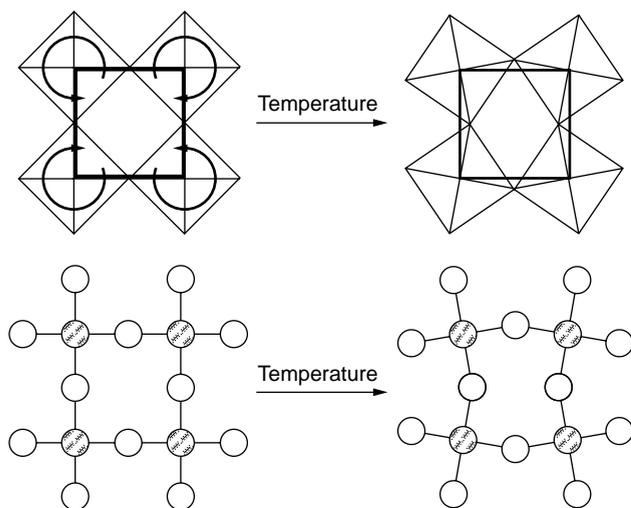


Figure 7. Schematic representation of rigid unit mode vibrations in a framework oxide structure as polyhedral and atomic views. Metal atoms are shaded; oxygen atoms are open circles.

It can also be seen that this type of mode is directly analogous to the negative Gruneisen parameter transverse vibration of Fig. 3. In fact, the area of the unit cell of Fig. 7 is proportional to the tilt angle θ . The thermal average of this angle $\langle \theta \rangle_T$ increases with temperature so that the overall cell area $A(T)$ is given by

$$A(T) = A_0 \cos^2 \theta \approx (1 - \langle \theta^2 \rangle_T). \quad (2)$$

If one assumes simple harmonic motion and the principle of equipartition of energy, $A(T)$ can be related to the temperature T , the moment of inertia I of the rigid body, and the vibrational frequency ν by

$$A(T) = A_0 \left(1 - \eta_A \frac{k_B T}{I \nu^2} \right), \quad (3)$$

where k_B is the Boltzmann constant and η_A is a constant specific to the mode in question (22).

A more detailed description of the thermal expansion properties of β -quartz has recently been given in which the ideas of RUMs are developed within the standard Gruneisen theory of thermal expansion (22). This work emphasized the importance of optical modes of the type shown in Fig. 7 but also showed that acoustic modes can make an important contribution to negative thermal expansion.

Concepts of RUMs can also be used to understand displacive phase transitions of the type displayed by quartz and many other framework materials. It is interesting to note, and it is a general feature of many framework materials that display negative thermal expansion, that displacive phase transitions and NTE behaviour are intimately linked. Many framework materials possess a large positive coefficient of thermal expansion as they are warmed. This may be viewed as evidence that the rigid units are “unwinding” in the structure. Once materials reach their high symmetry (maximum volume) structures, they display negative thermal expansion.

Phase Transitions

As suggested by Fig. 1, chemical bonds usually expand as temperature increases. Consequently, the majority of materials have positive expansion coefficients. There are, however, unusual circumstances, often in the region of structural phase transitions, in which at least average bond distances can actually decrease as temperature increases. The ideas of bond valence which originate in the work of Pauling and were developed extensively by Brown and O’Keeffe (23,24) showed that the contribution of a given bond to the valence sum of any atom depends approximately exponentially on the bond length $v = \exp(r_0 - r/0.37)$, where r_0 is a constant for a given E–X combination of elements and v is the valence contribution due to bond length r . This observation can be thought of at the simplest level as short bonds being proportionately stronger than long bonds. From the form of this relationship, one can readily appreciate that if a metal coordination environment changes from being distorted to being more regular at a phase transition, the average bond length decreases.

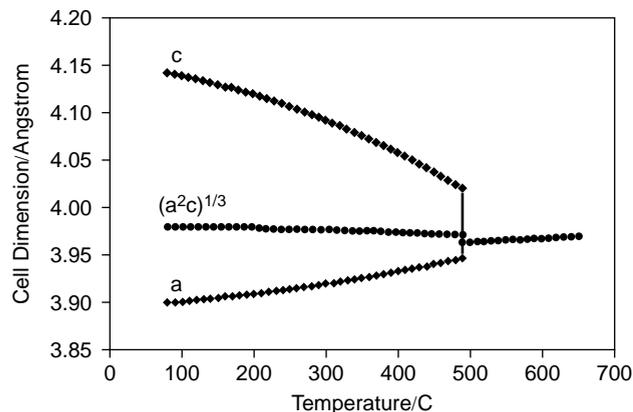


Figure 8. Cell parameters of PbTiO_3 as it approaches its tetragonal–cubic phase transition.

For example, the bond length r in an undistorted MO_6 octahedron is always shorter than the average bond distance in a distorted octahedron that has, for example, three short and three long bonds.

One example of this effect will be discussed here. Figure 8 shows the temperature dependence of the cell parameters of PbTiO_3 as it approaches its ferroelectric–paraelectric phase transition at 490°C (25). PbTiO_3 contains highly distorted Ti octahedra at room temperature, Ti–O bond lengths are 1.766, 4×1.979 , and 2.390 Å (26). When warmed, these octahedra become less distorted, and the average Ti–O bond length decreases from 2.012 to 1.983 Å (this represents a bond length expansion coefficient of $\alpha_l = -3 \times 10^{-5} \text{K}^{-1}$). The overall coefficient of thermal expansion of PbTiO_3 from 30– 400°C is $\alpha_l = -3.3 \times 10^{-6} \text{K}^{-1}$ (27). Presumably, the contraction of the Ti–O bond lengths in this material dominates the normal expansion of other bonds.

Negative volume changes at phase transitions have also been described in materials such as $\text{Co}_2\text{P}_2\text{O}_7$. As described later, such phases have been included as fillers in glass bonding frits to help match thermal expansion coefficients of materials to be joined. Control of displacive phase transitions is also important in the thermal expansion properties of $\text{ZrV}_{2-x}\text{P}_x\text{O}_7$ and $\text{A}_2(\text{MO}_4)_3$ families and will be discussed later.

Magnetic Transitions

Magnetostrictive phenomena in the region of magnetic phase changes can also give rise to materials of low thermal expansion. The “normal” phonon-driven positive thermal expansion of materials that have a significant magnetoelastic coupling can be compensated for by a large contraction driven by changes in the magnetic structure. This is the case for alloys such as “Invar”, $\text{Fe}_{0.65}\text{Ni}_{0.35}$ (see later), and transition metals such as Cr and α -Mn. Correctly processed, Invar has been quoted as having a thermal expansion coefficient of $0.02 \times 10^{-6} \text{K}^{-1}$, though across a restricted temperature range (28,29). Materials such as $\text{Lu}_2\text{Fe}_{17}$ and Y_2Fe_{17} also exhibit negative thermal expansion below approximately 400 K (30).

Linking Themes

Perhaps the best framework in which all of these phenomena may be described is to consider the change in the internal pressure of a solid caused by heating. This internal pressure, will in turn, gives rise to a volume change. The volume coefficient of thermal expansion can be related to the isothermal compressibility K and the internal pressure by

$$\alpha_V = \frac{1}{V} \left(\frac{\partial V}{\partial T} \right)_P = K \left(\frac{\partial P}{\partial T} \right)_V = K \left(\frac{\partial S}{\partial V} \right)_T. \quad (4)$$

Thus the sign of the coefficient of thermal expansion can be related to the sign of $(\partial S/\partial V)_T$. The entropy of normal materials (or the amount of structural disorder) decreases as pressure increases and volume decreases; materials that have NTE, however, increase in disorder as volume is reduced.

A further link between many of the materials of the next section is their proximity to a structural phase transition. Many of the Invar phases described in the next section are close to a structural phase transition to a body-centered structure. Cubic zirconium tungstate discussed later is thermodynamically unstable at temperatures below 1000°C with respect to more condensed binary phases and undergoes a number of phase transitions as a function of applied pressure and temperature. These instabilities are relevant to its unusual properties. Interestingly, metallic plutonium, which has negative thermal expansion in its δ and η phases has no less than six allotropes (α , β , γ , δ , η , ϵ) at atmospheric pressure; the NTE δ phase has an atomic volume 26% larger than the α .

Thus, it is possible to identify several criteria that might be expected to give rise to unusual thermoresponsive effects in solids:

1. proximity of the material to a structural phase transition
2. proximity of the material to a magnetic or electronic transition
3. the presence of a framework structure containing strong metal anion bonds and two coordinate anions
4. the absence of interstitial cations in a framework structure
5. framework structures in which volume-reducing displacive phase transitions are frustrated

Finally, it is interesting to note that it is possible to make composites that exhibit negative thermal expansion without relying on any of these mechanisms. Sigmund and Torquato showed that it is theoretically possible to produce three-phase composites that exhibit isotropic negative thermal expansion by engineering a topologically specific combination of two materials whose thermal expansion is positive (one high, one low) and empty space (31). In theory, such composites can be designed so that when heated, the bimaterial interfaces of the composite bend into the void space and the overall composite contracts—the material literally folds in on itself.

MATERIALS THAT DISPLAY NEGATIVE THERMAL EXPANSION

Invar

Probably the first technologically exploited example of an “unusual” thermoresponsive effect occurred in 1897 when Swiss physicist Charles Edwarde Guillaume working at the International Bureau of Weights, Sevres, discovered that an Fe/Ni alloy of approximate composition $\text{Fe}_{0.65}\text{Ni}_{0.35}$ showed almost zero thermal expansion across a broad temperature range. This material was christened “Invar,” and because of its discovery, Guillaume was awarded the Nobel prize for Physics in 1920. Invar is used in manufacturing many high precision devices. Perhaps the most widespread current application of Invar is in television and computer screens where it is used as a mask to prevent the electron beam from unintentionally hitting phosphor spots of the wrong color that would lead to blurred images. The low expansion of Invar is essential in this application due to the possibility of heating by the electron beam. Historically, Invar was used in clock pendulums where its low expansion prevented losses in accuracy as temperature changes. Other areas of application include waveguide tubes, surveying tapes, molds for shaping composites in aircraft products, and thermostat controls. Invar is also used in both tankers and tubing for liquid natural gas facilities. In typical applications, Invar has an expansion less than one-tenth that of steel, yet retains strength and is typically cheaper than a material such as Ti. Guillaume is also credited with the invention of “Elinvar” which has extremely small changes in elasticity and hence, mechanical strength as a function of temperature. Thus, this material was used in producing watch springs.

It was realized very quickly that the unusual expansion properties of Invar were intimately connected to its magnetic properties. Invar itself is a disordered face-centered cubic alloy that has a magnetic phase transition at around 500 K. Invar has an extremely low coefficient of thermal expansion below this temperature and a more normal one above it.

The connection between magnetism and thermal expansion led Weiss to propose the “ 2γ state” model to explain the Invar effect (32). There are two different magnetic states in this model, separated by a low energy barrier: a ferromagnetic ground state in which magnetic spins are parallel and the volume is larger, and an antiferromagnetic (antiparallel spins) state at a slightly higher energy, but with a smaller volume. As temperature is increased, the higher energy state becomes increasingly populated, and its smaller volume compensates for the usual expansion effects caused by thermal vibrations. Very recently, van Schilfgaarde et al. described ab initio calculations of the volume (and hence temperature) dependence of the magnetic and thermodynamic properties of Invar (33). Their work describes how noncollinear spin arrangements are crucially important in the Invar system. The magnetic structure undergoes a continuous transition from a high volume ferromagnetic state to a disordered noncollinear spin arrangement at lower volumes. This noncollinearity leads to an unusual dependence of the binding energy as a

function of volume and hence, to unusual thermal expansion properties.

Many other materials (both crystalline and disordered) exhibit the Invar effect, including Fe_3Pt , Cr, $\alpha\text{-Mn}$, $\text{Lu}_2\text{Fe}_{17}$, and Y_2Fe_{17} . There is also a related phenomenon known as the anti-Invar effect in which a magnetic phase transition leads to an anomalously high thermal expansion. Such a property could be used as an activator in a micromachine.

ZrW₂O₈

The thermal expansion properties of ZrW_2O_8 are perhaps the most unusual of any crystalline material. ZrW_2O_8 can be prepared as a simple cubic phase that displays isotropic negative thermal expansion from 2 to more than 1000 K (34). The magnitude of this effect is significant; the coefficient of thermal expansion from 2 to 350 K is $-9.07 \times 10^{-6} \text{ K}^{-1}$ (10). No other material displays such large isotropic contraction across such a wide temperature range. Even though this material has been known since the late 1950s, only in the last few years have its structure and unusual properties been fully understood.

The structure of $\alpha\text{-ZrW}_2\text{O}_8$, determined by powder neutron diffraction, is shown in Figure 9. It is a cubic material (space group $P2_13$) and can be described as a network of corner-sharing ZrO_6 octahedra and WO_4 tetrahedra. Each octahedron shares all six corners with a WO_4 tetrahedron; each tetrahedron shares only three of its four corners with an octahedron leaving one oxygen strictly

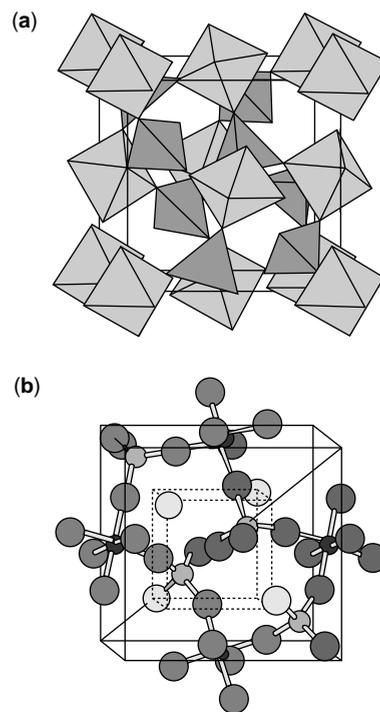


Figure 9. The structure of cubic ZrW_2O_8 . (a) A polyhedral representation of corner-sharing ZrO_6 octahedra and WO_4 tetrahedra. (b) A ball-and-stick view of a section of the framework structure. Zr small dark circles, W small light circles, oxygen large grey circles. The one-coordinate oxygen is shaded a lighter gray.

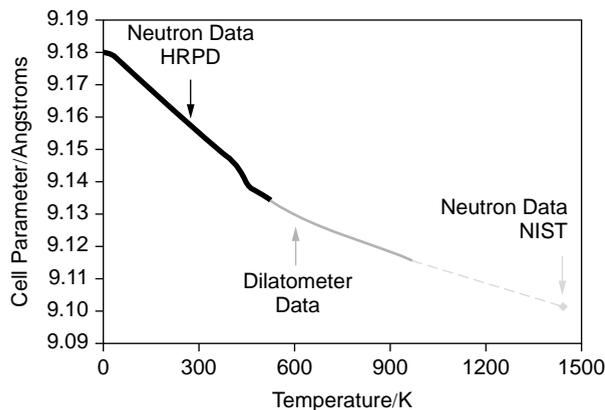


Figure 10. Thermal expansion behavior of ZrW_2O_8 . The plot shows the cubic unit cell parameter of ZrW_2O_8 from 2 to 1443 K. Data are derived from powder neutron diffraction and dilatometric studies (see text for details). The dilatometric data, where the physical length of a ceramic bar is measured, have been scaled to the cell parameter derived from neutron diffraction data at 298 K. Regions of overlap have been omitted for clarity. ZrW_2O_8 is kinetically stable up to 1050 K, unstable in the dotted region of the curve, and thermodynamically stable at 1443 K.

one-coordinate. The temperature dependence of the cubic unit cell of ZrW_2O_8 is shown in Fig. 10. Data points from 2 to 520 K were determined by powder neutron diffraction, where the unit cell parameter is measured directly. Data from 520 to 960 K were determined by dilatometric measurements on a ceramic block of the material, whereas the single data point at 1443 K was determined by neutron diffraction. The only region of this graph where ZrW_2O_8 is thermodynamically stable is the single data point at 1443 K. Below approximately 1050 K, however, ZrW_2O_8 is metastable for extended periods of time. The origin of this negative thermal expansion can be directly traced to the structure of the ZrW_2O_8 framework. Dove and co-workers showed that the topology of the structure is such that ZrW_2O_8 has intrinsic flexibility and can support a large number of rigid unit modes (35). Thus, ZrW_2O_8 is a three-dimensional analog of the simple 2-D model of Fig. 7. Their calculations predict that ZrW_2O_8 will support a large number of low-energy vibrations that tend to contract the lattice.

Experimental evidence for this model has been obtained by a number of techniques. David et al. showed how information about the Gruneisen parameter weighted vibrational density of states can be extracted directly from the cell parameter information of Fig. 10 by using maximum entropy techniques (Fig. 11) (36). This work shows that the negative thermal expansion can be attributed to a family of modes whose energy ranges from 3 to 8 meV. Ernst et al. used inelastic neutron scattering to probe the phonon density of states directly (Fig. 12) (37). These measurements show that there are a significant number of low-energy phonons in ZrW_2O_8 , and again led the authors to conclude that modes in the energy range of 1.5–8.5 meV are most important for NTE. Specific heat capacity measurements by Ramirez and Kowach again highlighted the importance of low-energy vibrations in ZrW_2O_8 (38). Recent calculations

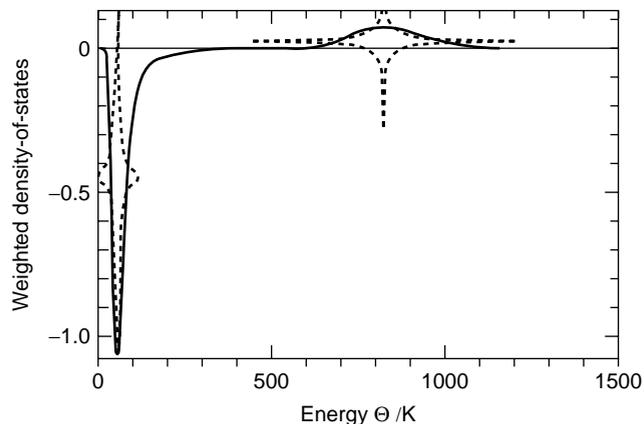


Figure 11. A maximum entropy reconstruction of the Gruneisen parameter weighted phonon density of states extracted from the cell parameter data of Fig. 10. The negative peak around 55 K (4.7 meV) corresponds to a family of low-energy modes that tend to contract the lattice; those centered at 850 K (73 meV) tend to expand the lattice.

provide further support for the importance of these vibrations (39).

It has been shown that the structure of ZrW_2O_8 gives this material a number of other highly unusual properties as a function of both applied temperature and pressure. As the material is warmed to 450 K, it undergoes a phase transition to $\beta\text{-ZrW}_2\text{O}_8$ in which the lighter shaded oxygen atoms of Fig. 9 become dynamically disordered (34,40). Oxygen migration at such low temperatures is highly unusual. Under an applied pressure of around 0.3 GPa, ZrW_2O_8 undergoes a further phase transition to $\gamma\text{-ZrW}_2\text{O}_8$, which also involves the migration of oxygen atoms to new sites in the structure (41–43). When heated to 100°C, the γ form reverts to the low pressure α form. This series of phase transitions it has been suggested, is a potential method of energy storage. At higher pressures,

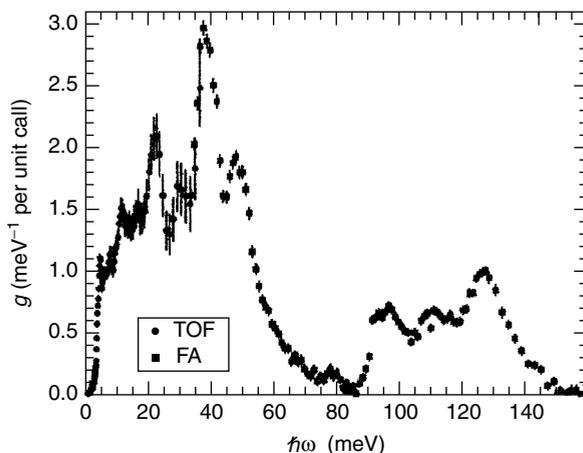


Figure 12. Generalised phonon density of states $g(\omega)$ of ZrW_2O_8 at 300 K. Data recorded by inelastic neutron scattering using time-of-flight (<40 meV) and filter-analyzer spectroscopy (>40 meV) [reproduced from article by Ernst et al., with permission(37)].

the material becomes amorphous (44). Recently Wilkinson et al. successfully prepared a new room temperature polymorph—trigonal ZrW_2O_8 . Taken together, these observations suggest that the unusual thermoresponses of ZrW_2O_8 are related to the metastability of the material, its unusual topology, and its proximity to a number of different phase transitions.

Other materials such as HfW_2O_8 and $\text{ZrW}_{2-x}\text{Mo}_x\text{O}_8$ have been prepared recently and characterized (45–48).

AM_2O_7 and Related Phases

It has been known for a number of years that members of the cubic AM_2O_7 family (e.g., ZrV_2O_7 and ZrP_2O_7) exhibit unusual thermal expansion properties. Their structure (Fig. 13) is closely related to that of ZrW_2O_8 and again contains corner-sharing ZrO_6 octahedra and AO_4 tetrahedra. In contrast to ZrW_2O_8 , however, the tetrahedra form A_2O_7 pyrophosphate/vanadate units and thus share all four of their corners; therefore, all oxygens are two-coordinate. This basic structural type is known for a range of A metals, including $A = \text{Si, Ge, Ti, Re, Mo, Nb, Sn, Zr, Mo, W, Pb, Ce, U, and Th}$, and for $M = \text{P, V}$. The structural properties of many of these materials as a function of temperature are complex. At high temperatures, they have the simple cubic structure of Fig. 13. When cooled, however, many members of this family undergo displacive phase transitions. In ZrP_2O_7 , the unit cell parameter as a function of temperature shows a single phase transition to what is apparently a $3 \times 3 \times 3$ cubic supercell at room temperature (49). ZrV_2O_7 , on the other hand, undergoes two

clear phase transitions, initially to an incommensurately modulated $\sim 3 \times 3 \times 3$ superstructure, then, finally to an ordered $3 \times 3 \times 3$ superstructure (50). Both SiP_2O_7 and TiP_2O_7 show similar behavior (51,52).

Thus, these structures show the classic behavior of framework materials, as described previously: a strong positive coefficient of thermal expansion as they approach a displacive phase transition, followed by low or negative thermal expansion thereafter. High temperature coefficients of thermal expansion are $\alpha_l = -7.1 \times 10^{-6}$ (400–500 K) and $+5.4 \times 10^{-6} \text{ K}^{-1}$ (600–700 K) for ZrV_2O_7 and ZrP_2O_7 respectively.

Interestingly, Sleight and co-workers showed how the formation of solid solutions of the type $\text{ZrV}_{2-x}\text{P}_x\text{O}_7$ can be used to suppress these phase transitions and extend the range of negative thermal expansion to include room temperature (53). Figure 14 contains dilatometric data for selected members of the family. Materials such as ZrVPO_7 have low or negative thermal expansion at room temperature and below. This represents an important way in which thermal expansion properties may be effectively controlled and material properties fine-tuned.

NZP Phases

The NZP or Nasicon family of materials is one of the most widely investigated for controllable low or negative thermal expansion properties. This family derives its name from the parent composition $\text{NaZr}_2(\text{PO}_4)_3$ and has general formula $\text{M}^1\text{M}^2\text{A}_2(\text{BO}_4)_3$, where $\text{M}^1, \text{M}^2, \text{A}$, and B can be a variety of metal cations. The name Nasicon is usually reserved for the $\text{Na}_{1+x}\text{Zr}_2(\text{SiO}_4)_x(\text{PO}_4)_{3-x}$ family that has been widely investigated as a solid-state electrolyte due to its high ionic conductivity [see, for example, Brevail et al. for a review (54)] Nasicon derives its name from Na superionic conductor.

The structure of NZP is shown in Fig. 15. It can be described as a framework of fully corner-sharing ZrO_6 octahedra and PO_4 tetrahedra. Two octahedra are linked by three PO_4 tetrahedra to form $\text{Zr}_2(\text{PO}_4)_3^-$ units. These groups align in chains that run parallel to the c axis and are cross-linked via PO_4 groups to adjacent units to form a 3-D framework. The framework created contains a number of interstitial sites where charge balancing M^1 and M^2 cations are located. Within the $\text{Zr}_2(\text{PO}_4)_3^-$ units, there is a trigonal prismatic site that is usually empty. There are approximately octahedral sites (site symmetry -3) between adjacent $\text{Zr}_2(\text{PO}_4)_3^-$ units in the same chain that are occupied by Na in the parent structure. In addition, there are three more approximately octahedral sites between the chains. This network of full and empty sites provides the low-energy migration pathways that give these phases their ionic conduction properties.¹

¹The description here is of the rhombohedral $R\bar{3}c$ form of the material. Several NZP compositions [e.g., $\text{LiZr}_2(\text{PO}_4)_3$ at room temperature and $\text{Na}_{1+x}\text{Zr}_2(\text{SiO}_4)_x(\text{PO}_4)_{3-x}$ for $1.8 < x < 2.2$] undergo a phase transition to a monoclinic structure when cooled from elevated temperatures. This phase transition is displacive and involves static rotations of the polyhedra. The essential features of the structure remain unchanged, though thermal expansion properties are influenced.

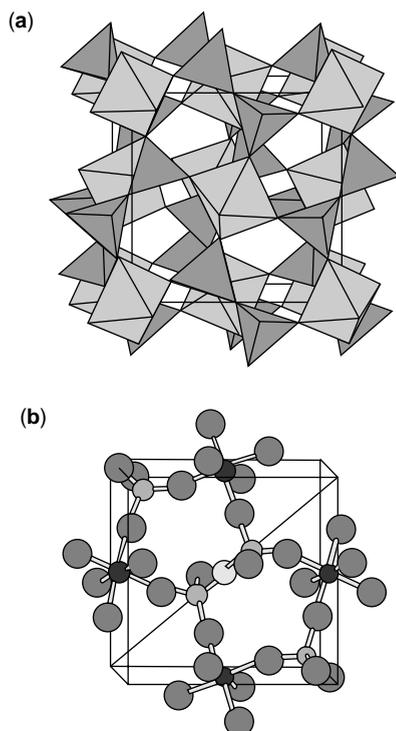


Figure 13. Structure of ZrV_2O_7 . (a) ZrO_6 octahedra share corners with VO_4 tetrahedra. (b) VO_4 tetrahedra share corners to form V_2O_7 pyrovanadate groups.

Figure 14. Dilatometric data recorded for various members of the $ZrV_{2-x}P_xO_7$ family (53). (data provided by the authors).

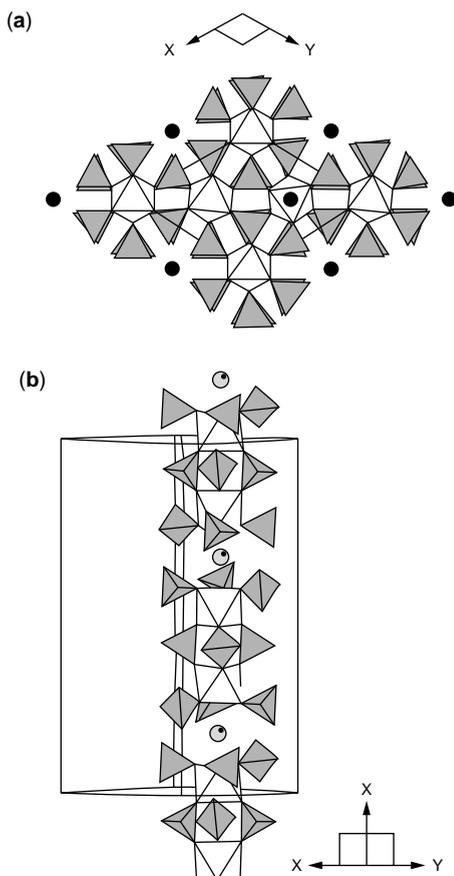
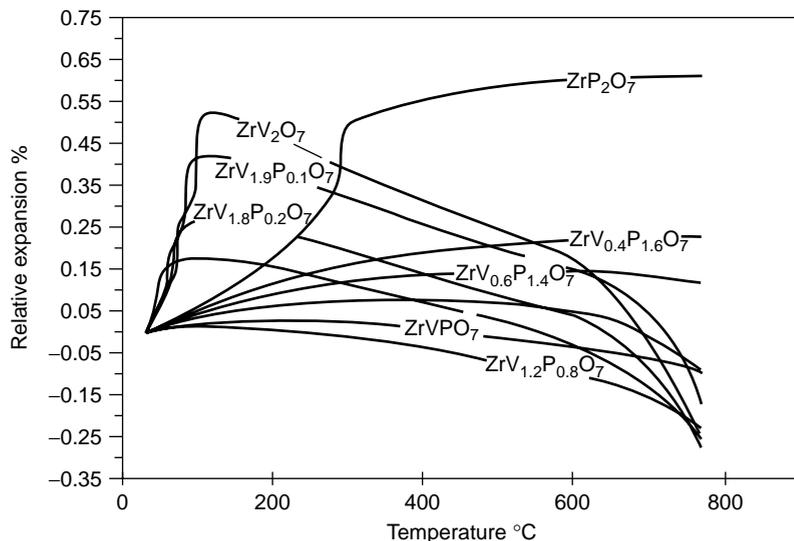


Figure 15. The structure of NZP materials. ZrO_6 octahedra share corners with PO_4 tetrahedra to form a 3-D framework. Na cations (circles) sit in interstitial sites in this framework. (a) A view down the c axis of the material. (b) A view perpendicular to the c axis emphasizing the chains of polyhedra in the structure. The “lantern” group of two octahedra and three tetrahedra linked together is also present in the structure of $Sc_2(WO_4)_3$ (Fig. 16).

The NZP family meets many of the criteria laid out previously. The structure contains a framework of corner-sharing polyhedra that have two-coordinate bridging oxygen atoms and a significant number of “vacant” interstitial sites that can accommodate vibrational modes of the polyhedra.

Interest in this family of materials was originally stimulated by a paper by Boilot et al. (55) who described dilatometric data for $Na_{1+x}Zr_2(SiO_4)_x(PO_4)_{3-x}$ materials, where x ranges from 1 to 3, and gave the prescient statement that “the compound $x = 1$ displays an important shrinkage which could allow this material to be useful when expansion is undesirable.” Stemming from this original comment, there has been a considerable body of work on these materials, most notably by Roy, Agrawal, McKinstry, and their co-workers at Pennsylvania State University.

Diffraction measurements show that the low overall thermal expansion in the simple $MZr_2(PO_4)_3$ phases ($M = Li, Na, K, Rb, Cs$) is due to significant anisotropy in the thermal expansion of the a and c axes; the c axis typically shows an expansion, and the a axis shows a contraction. Lenain et al. (56) related this fact to the observation that similar changes in cell parameters are observed as one substitutes successively larger alkali metals at the M site. An elegant model was proposed to explain this anisotropy based on geometrically coupled rotations of the polyhedral units of the structure. Expansion of the M^1 site causes coupled displacements of the framework polyhedra. The inference from this work was that increasing temperature has an effect comparable to increasing the size of the alkali metal. Sleight and Khosrovani (57) took a more quantitative approach and used a combination of the temperature dependence of $M-O$ distances from the literature and the purely geometric distance least squares (DLS) modeling technique to rationalize the properties of these phases. Variable temperature diffraction studies of $LiGe_2(PO_4)_3$ by Alami and more recently by Lightfoot and co-workers (58–61) support these models and confirm the basic ideas. The overall change in cell parameters as a function of

Table 2. Thermal Expansion Coefficients of Selected NZP Related Materials. Thermal Expansion Coefficients for These Materials Are Highly Dependent on Sample Preparation Methods^a

	$\alpha(\text{diffraction}) \times 10^{-6} \text{ K}^{-1}$	$\alpha(\text{dilatometry}) \times 10^{-6} \text{ K}^{-1}$	Temperature Range
NaZr ₂ P ₃ O ₁₂	3.7	-3	25–500
KZr ₂ P ₃ O ₁₂	0.3	-7.5	25–500
Ca _{0.5} Zr ₂ P ₃ O ₁₂	-0.07	-2.11	25–500
Sr _{0.5} Zr ₂ P ₃ O ₁₂	2.0	3.2	25–500
Ca _{0.25} Sr _{0.25} Zr ₂ P ₃ O ₁₂	-0.1	1.4	rt–500
NaTi ₂ P ₃ O ₁₂	4.28	-4.55	27–1000
NbZrP ₃ O ₁₂	-1.5	-	rt–800

^aData in this table have been selected from those in an extensive review (62) and are examples where the results of both diffraction and dilatometry were presented in the same study.

temperature is caused by changes in the size and shape of the M¹ site which causes the octahedra and tetrahedra to undergo geometrically coupled rotations along with minor internal distortions (octahedra typically distort to a greater extent than tetrahedra).

One of the most technologically significant features of the NZP structure is the wide range of both iso and aliovalent cation substitutions possible at the M¹, M², A, and B sites. The M cations can be substituted by a range of species, including H⁺, Li⁺, K⁺, Rb⁺, Cs⁺, Ag⁺, NH₄⁺, Mg²⁺, Ca²⁺, Sr²⁺, and Ba²⁺; the Zr site by Na⁺, Mg²⁺, Mn²⁺, Ni²⁺, Cu²⁺, Zn²⁺, Al³⁺, Co³⁺, Fe³⁺, In³⁺, Sc³⁺, Y³⁺, Yb³⁺, Nd³⁺, Ti⁴⁺, Ge⁴⁺, Sn⁴⁺, Hf⁴⁺, Th⁴⁺, U⁴⁺, and Nb⁵⁺; and the P site can be substituted by Si⁴⁺, Ge⁴⁺, As⁵⁺, and S⁶⁺. This combination of doping possibilities leads to an essentially infinite family of materials whose low thermal expansion properties are controllable (see Table 2). In terms of cation substitution, NZP is one of the most flexible inorganic materials known. Another potential use of these materials is in nuclear waste storage. Roy et al. showed that NZP will accommodate all of the ions in a normal nuclear fuel reprocessing scheme to form a single-phase ceramic; this has clear environmental implications.

The vast majority of NZP phases actually expand at the unit cell level, even though they show strong bulk contraction (as measured, for example, by dilatometry). NaTi₂(PO₄)₃, for example, has a bulk expansion reportedly as low as $-4.55 \times 10^{-6} \text{ K}^{-1}$ (62), yet the intrinsic cell volume change from the same study is $+3.2 \times 10^{-6} \text{ K}^{-1}$ (22–1000°C). This discrepancy between intrinsic and extrinsic behavior can again be related to the marked anisotropy of expansion of these phases; the overall properties of a ceramic body depend strongly on intergrain interactions. The anisotropy can lead to microcracking in bulk specimens due to stresses induced by different expansions of neighbouring grains in the ceramic body and can frequently lead to hysteresis in the thermal expansion of composites. The healing of microcracks gives an additional component to the contraction on warming; their formation on cooling an expansion. Breval et al. (63) showed that introducing Si in NZP phases can lead to the formation of glassy grain boundaries that soften at relatively low temperature and lead to further discrepancy between extrinsic and intrinsic properties.

An important consequence of the possibility of aliovalent doping in NZP is the ability to prepare materials where the M¹ site is either fully occupied [e.g., NaZr₂(PO₄)₃],

50% occupied [e.g., Ca_{0.5}Zr₂(PO₄)₃], 33% occupied [e.g., La_{0.33}Zr₂(PO₄)₃], or fully vacant [e.g., NbTi(PO₄)₃]. Because the behavior of the M¹ site influences polyhedral tilts in these materials to such an extent, one has a relatively straightforward method of controlling thermal expansion properties. For example, though phases such as NaZr₂(PO₄)₃ show a positive expansivity of the c and negative expansivity of the a axes, this situation is reversed in certain materials such as Ca_{0.5}Zr₂(PO₄)₃, with the c axis contracts, and a expands. The presence of a partially vacant and therefore compressible M¹ site provides a flexible buffer to accommodate expansion of the full sites.

Ca_{0.5-x}Sr_xZr₂(PO₄)₃ is one system that has been investigated in some detail. This interest was again prompted by the observation that in Ca_{0.5}Zr₂(PO₄)₃, the a axis contracts and c expands whereas in Sr_{0.5}Zr₂(PO₄)₃ a expands and c contracts.² Workers at Penn State were successful in synthesizing Ca_{0.25}Sr_{0.25}Zr₂(PO₄)₃ in which α_a and α_c remained nearly constant when heated to 500°C. Acoustic emission studies showed that microcracking in such a material when cooled can be greatly reduced (64). The overall coefficient of thermal expansion of such a composite was reported to be $-0.1 \times 10^{-6} \text{ K}^{-1}$ between 298 and 873 K.

Another interesting phase is NbTi(PO₄)₃ which has fully vacant M sites. It is unusual among the NZP family because it shows intrinsic negative thermal expansion; the a-axis contraction dominates the c-axis expansion. Woodcock et al. (65) provided detailed thermal expansion data for this compound and found that α_l (defined as $1/3 \alpha_V$) ranges from -3 at room temperature to around $+0.5 \times 10^{-6} \text{ K}^{-1}$ at 1000°C. The larger alkali metal (K and Rb) systems have also been reported to show intrinsic contraction. LiZr₂(PO₄)₃ exhibits sharp contraction [$\alpha_l = -10.1$ to $-22.2 \times 10^{-6} \text{ K}^{-1}$] between 20 and 155°C (66).

Agrawal and Roy also demonstrated that low expansion composite bodies can be made between NZP members and other ceramics such as MgO, ZnO, Nb₂O₅, ZrSiO₄. Another interesting application of NZP phases in producing smart high-tech materials was described by Agrawal and co-workers (67). Carbon-carbon composites have a number of properties that make them ideal for certain

²The observation that Ca_{0.5}Zr₂(PO₄)₃ shows the behavior typical of an ordered R3c structure rather than the behavior of a disordered R3 structure has been rationalized in terms of the vacancy site in Ca_{0.5}Zr₂(PO₄)₃ which is smaller and thus less compressible than in the Sr compound.

applications in space vehicles. These include a high strength to weight ratio, good strength retention at high temperature, and good resistance to thermal shock. They do, however, have one major drawback—they are readily oxidized at high temperatures. A number of materials have thus been tested as potential protective coatings. Most, however, suffer from poor thermal mismatch to C–C composites, which leads to their ultimate failure. Agrawal et al. showed that members of the NZP family whose coefficients of thermal expansion match those of the composites can be successfully used as coatings. $\text{Ba}_{0.875}\text{Zr}_2(\text{SiO}_4)_{0.175}(\text{PO}_4)_{2.825}$, for example, can be successfully hot pressed around composite bars, and the resultant body can be heated to 1200°C without any significant oxidation of carbon.

The $\text{Sc}_2(\text{WO}_4)_3$ Family

Negative thermal expansion has also been reported in the $\text{A}_2(\text{MO}_4)_3$ family of materials with $\text{Sc}_2(\text{WO}_4)_3$ related structures. These materials can be described as containing AO_6 octahedra that share all six of their corners with MO_4 tetrahedra (Fig. 16). These materials crystallize in the orthorhombic space group $Pnca$ at high temperatures, as originally described by Abrahams and Bernstein, though several members of the family undergo a volume-reducing displacive phase transition to a monoclinic structure when cooling (68,69). These materials meet many of the criteria laid out previously (strong metal-to-oxygen bonds, rigid polyhedra, no interstitial cations), and it is not surprising that they show low or negative expansion coefficients. It is also relevant to emphasize their structural similarities to several of the other materials described before. Corner-sharing octahedra and tetrahedra are common to the ZrW_2O_8 , ZrV_2O_7 , and NZP families of materials. In particular, the $\text{A}_2(\text{MO}_4)_3$ family is closely related to the interstitial cation-free members of the NZP family such as $\text{NbTi}(\text{PO}_4)_3$; the basic building blocks of the two structures are identical, but their 3-D connectivity differs.

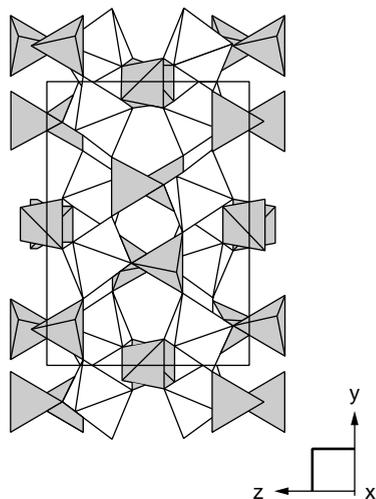


Figure 16. The structure of $\text{Sc}_2(\text{WO}_4)_3$. ScO_6 octahedra (white) share corners with WO_4 tetrahedra (shaded).

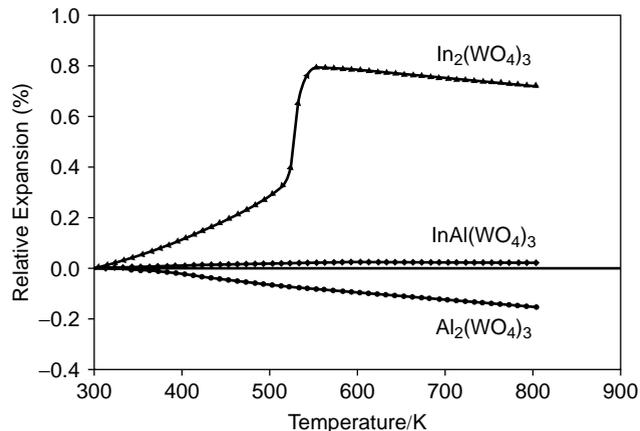


Figure 17. Thermal expansion properties of selected members of the $\text{In}_{2-x}\text{Al}_x(\text{WO}_4)_3$ family of materials.

This family of materials also has a high degree of chemical flexibility. The A^{3+} site can be doped by using a variety of elements ranging in size from Al^{3+} ($r = 0.535 \text{ \AA}$) to Gd^{3+} ($r = 0.938 \text{ \AA}$). Additionally, aliovalent doping using, for example, Zr^{4+} on the A site and P^{5+} on the M site to produce phases such as $\text{Zr}_2(\text{PO}_4)_2(\text{WO}_4)$ is possible. These substitutions again lead to controllable thermal expansion properties (70–73). $\text{Y}_2(\text{WO}_4)_3$, which contains the largest A^{3+} cation studied crystallographically, has the most negative coefficient of thermal expansion (72). The monoclinic to orthorhombic phase transition can also be influenced by the A-site electronegativity. $\text{In}_2(\text{WO}_4)_3$ undergoes a monoclinic to orthorhombic transition at around 600 K (Fig. 17); $\text{Al}_2(\text{WO}_4)_3$ is orthorhombic at all temperatures above 300 K. By preparing solid solutions such as $\text{AlIn}(\text{WO}_4)_3$ it is possible to produce ceramic bodies displaying near zero net expansion (Fig. 17). Expansion coefficients for selected materials are included in Table 3.

It is worth noting that, like the NZP materials, most members of the $\text{Sc}_2(\text{WO}_4)_3$ family display anisotropic expansion. Intrinsic expansion (measured by diffraction methods) can be very different in bulk properties. $\text{Sc}_2(\text{WO}_4)_3$, for example, has a cell volume expansion coefficient $\alpha_V = -6.5 \times 10^{-6} \text{ K}^{-1}$ corresponding to $\alpha_l = -2.2 \times 10^{-6} \text{ K}^{-1}$ but can show linear expansion coefficients as low as $-11 \times 10^{-6} \text{ K}^{-1}$ as a bulk ceramic body.

Zeolites

Zeolites are another group of materials that meets many of the criteria laid out previously for unusual expansivity. Zeolites can be described as frameworks built up from corner-sharing SiO_4 and AlO_4 tetrahedra and have charge-balancing species (atomic or molecular) located in voids of the frameworks. Thus, their general formula can be expressed as $\text{M}_{x/n}[(\text{AlO}_2)_x(\text{SiO}_2)_{2-x}]_y\text{H}_2\text{O}$, where M^{n+} cations occupy cavity sites in the framework. There are around 200 aluminosilicate frameworks known of which around 40 are natural minerals. In addition to the zeolites, there are a significant number of materials, often called zeotypes, that contain elements such as P, Be, Ga, and Zn at the tetrahedral sites. Of particular interest in the

Table 3. Thermal Expansion Coefficients of Selected $A_2(MO_4)_3$ Materials^a

	α (diffraction) $\times 10^{-6} K^{-1}$	α (dilatometry) $\times 10^{-6} K^{-1}$	Ref.
Sc ₂ (WO ₄) ₃	-2.2 (10–450)	-11 (303–1073)	12
Sc ₂ (MoO ₄) ₃	-2.1 (180–300)	-5 (303–1073)	87
In ₂ (WO ₄) ₃		+10 (303–803)	70
AlIn(WO ₄) ₃		+0.5 (303–803)	70
Al ₂ (WO ₄) ₃		-3 (303–803)	70
Y ₂ (WO ₄) ₃	-7 (15–1373)		13
Lu ₂ (WO ₄) ₃	-6.8 (400–900)		72
Zr(PO ₄) ₂ (WO ₄)	-3.0 (100–400)	-5.5 (303–873)	70

^aValues are derived from dilatometric data on bulk samples and diffraction data. Linear coefficients of thermal expansion from diffraction measurements are defined as $1/3\alpha_V$. Temperature range in K is given in parentheses.

search for negative thermal expansions are the pure silica frameworks [i.e., $x = 0$ in the preceding general formula] in which there are no interstitial cations. Therefore, they can be described as (metastable) polymorphs of silica. Cation-free frameworks are also found in the Al/P framework materials such as AlPO₄. The thermal expansion properties of the zeolites provide an interesting example in which computer modeling predicted an unusual property before it was experimentally measured. In 1993, Parker and co-workers used atomistic simulations to predict that certain zeolites should contract as temperature increases (74). These properties were confirmed by subsequent diffraction experiments. In 1995, they modeled 18 different zeolite structures and predicted that all but two ought to contract as temperature increases. This behavior was subsequently confirmed by diffraction measurements. Negative thermal expansion has since been found in a number of zeolites, including pure SiO₂ polymorphs of faujasite and chabazite ITQ-1, ITQ-3, ITQ-9, SSZ-23, ZSM-S, AlPO-5, and ALPO-17 (75–80). Park et al. (79) emphasized the difference in thermal expansion properties of synthesized and calcined zeolites. Typical coefficients of thermal expansion are included in Table 4.

β -Eucryptite/Spodumene

β -Eucryptite, LiAlSiO₄, can be described as a “stuffed-quartz” structure (20). Half of the Si atoms of quartz are replaced by Al atoms, and charge-balancing Li⁺ cations

reside in the sixfold channels visible in Fig. 6 (at the cell origin). Al/Si atoms are ordered parallel to the *c* axis leading to a doubling of the unit cell *c* axis, and Li ordering in the various tetrahedral sites of the channels leads to doubling of the *a* and *b* axes. β -Eucryptite has been well studied due to its low or negative thermal expansion; α_V is $-6 \times 10^{-6} K^{-1}$ from 20 to 523 K and $+0.29 \times 10^{-6} K^{-1}$ from 523 to 873 K (82). This low expansion is achieved via significant anisotropy; the *c* axis contracts, and the *a* axis remains approximately constant (<298 K) or expands. As a result, β -eucryptite and other structurally similar compounds in the Li₂O–Al₂O₃–SiO₂ (LAS) system are used extensively in low-expansion glass ceramics. Products based on these systems include oven-to-table cookware and high precision engineering components.

Various workers have proposed mechanisms to explain the low expansion of β -eucryptite. Recently Xu et al. compared the expansion properties of ordered and disordered Al/Si samples (82). They concluded that a variety of mechanisms determine the overall expansion properties, including tilting of the Si/Al tetrahedra, tetrahedral distortion, and disordering of the Li atoms. This Li ion disordering also gives rise to significant ionic conductivity parallel to the *c* axis of the material. Several authors have discussed how the population of different Li sites at different temperatures may influence thermal expansion properties. At low temperatures, however, the principal mechanisms underlying anisotropic expansion are tetrahedral tilting and the expansion of an unusually short O–O bond which arises

Table 4. Expansion Coefficients of Zeolites^a

Zeolite	α (diffraction) $\times 10^{-6} K^{-1}$	Temperature Range	Ref.
Chabazite	-16	293–873	81
ITQ-4	-4.3	95–510	81
Si-faujasite	-4.2	25–573	77
ALPO-17	-17	18–300	76
ITQ-1	-12.1	323–773	75
ITQ-3	-11.4	323–823	75
SSZ-23	-10.3	323–773	75
MFI	-15.1	393–975	79
AFI (AlPO ₄ - 5)	-14.5	424–774	79
DOH	-3.1	573–996	79
MTN	-5.0	463–1002	79
DDR	-8.7	492–1185	79

^aAll values are defined as $1/3\alpha_V$ derived from cell parameter data in the literature.

from edge sharing of LiO_4 and SiO_4 tetrahedra; tetrahedra are described as lengthening parallel to c and shortening perpendicular to this axis.

A second important member of the LAS family is β -spodumene ($\text{LiAlSi}_2\text{O}_6$). Whereas β -eucryptite can be described as a stuffed quartz structure, β -spodumene can be described as derived from keatite, a high pressure form of SiO_2 that has not been found in nature. Keatite is a tetragonal material built from corner-sharing SiO_4 tetrahedra and contains five-, seven-, and eight membered rings. β -Spodumene is derived from this structure and contains Al/Si disordered across two tetrahedral sites and Li disordered across four sets of split tetrahedral sites. β -Spodumene again shows highly anisotropic behavior; the c -axis expands and the a and b axes contract as temperature increases. This effect has again been explained by coupled rotations of tetrahedral groups.

One interesting application which has been demonstrated for β -eucryptite, and for which materials such as ZrW_2O_8 are also being investigated, is for athermalizing fiber Bragg gratings (FBGs). FBGs are created in optical fibers by introducing a periodic modulation into the refractive index of a material and can be manufactured so that they have precisely controlled optical transmission and reflection properties. As such they can be used to control the wavelength of light, combine lights of different wavelengths, and split certain wavelengths from a fiber. These are important properties needed, for example, in wavelength-division multiplexing (WDM) for optical communication networks.

One problem of FBGs is the thermal variation of the Bragg wavelength λ_B , which is influenced by both the temperature dependence of the refractive index (frequently the dominant term) and the thermal expansion of the materials used. A typical temperature dependence in GeO_2 -doped SiO_2 at 1550 nm is 0.012 nm K^{-1} . This temperature dependence can be eliminated by actively controlling the temperature of the unit, though this is an expensive and complex option. Passive control can be achieved by mounting the grating on a negative expansion support. Early designs relied on using two materials that have different (positive) expansion to achieve this effect. By using an NTE mount, both the reliability and ease of manufacture may be improved. Corning, for example, reported a system based on a β -eucryptite glass ceramic substrate in which the temperature dependence of λ_B is reduced from 0.012 to 0.001 nmK^{-1} . In this application, the microstructure (determined by processing conditions) of the NTE support is extremely important. By correct processing, it is possible to reduce the expansion coefficient of β -eucryptite ceramics from their crystallographic value of $\sim -0.4 \times 10^{-6} \text{ K}^{-1}$ to as low as $-7 \times 10^{-6} \text{ K}^{-1}$ (293–393 K). Achieving this high a negative coefficient is important because the temperature dependence of λ_B can be approximated by

$$\frac{d\lambda_B}{dT} = 2\Lambda \left[\frac{dn}{dT} + n\alpha \right], \quad (5)$$

where Λ is the grating spacing. For an average refractive index of $n = 1.461$ and a value of $dn/dT = 11 \times 10^{-6} \text{ K}^{-1}$, a thermal expansion coefficient around $-7.5 \times$

10^{-6} K^{-1} is required to compensate for refractive index changes (83).

A second interesting technical challenge in the development of such devices is the production of frits for bonding fibers to low or negative thermal expansion substrates. For the device described before, Corning used $\text{SnO-ZnO-P}_2\text{O}_5$ inorganic frits containing $\text{Co/Mg}_2\text{P}_2\text{O}_7$ fillers (84). These fillers themselves undergo a martensitic transition that is accompanied by a significant decrease in volume. This volume change can be used to adjust the expansion properties of the frit to match the components it must bond.

Cordierite

Cordierite ($\text{Mg}_2\text{Al}_4\text{Si}_5\text{O}_{18}$) has been widely investigated for its low thermal expansion properties (85). The structure can again be described in terms of a framework of corner-sharing Al/SiO₄ tetrahedra that form four- and six-membered rings. There is an additional octahedral site that is occupied by Mg. The Al and Si atoms in cordierite are ordered in six-membered tetrahedral rings that have an overall orthorhombic symmetry. In many synthetic samples, however, there is no such order, and the symmetry is hexagonal; such materials have been called indialites. The structure of cordierite is closely related to the mineral beryl ($\text{Al}_2\text{Be}_3\text{Si}_6\text{O}_{18}$) in which the tetrahedral sites are occupied by a mixture of Be and Si. Like many silicates, a variety of cations can be doped into this structural type. Synthetic solid solutions have been investigated using elements such as Ga, Ge, and Mn; cation substitution again leads to controllable changes in thermal expansion properties. Cation substitutions in the beryl framework are also evidenced in commercially important phases such as emerald (Cr/Fe substitution) and related minerals such as aquamarine (blue green), morganite (pink), and heliodor (yellow).

These materials (similarly to β -eucryptite and β -spodumene) achieve a low overall volume expansion (typically $< 2 \times 10^{-6} \text{ K}^{-1}$ between 298 and 873 K) via significant structural anisotropy. Most substituted cordierites expand in the a - b plane and contract along the c axis. The magnitude of c -axis contraction typically decreases as T increases. In cordierite itself, c reaches a minimum value around 500°C. This behavior can again be ascribed to coupled rotations of the constituent polyhedra. Cordierite materials are used extensively as low-expansion catalyst supports, especially in automotive catalytic converters.

Selected Commercially Available Materials

Zerodur is a glass ceramic material developed by Schott Glass for various applications where controlled low thermal expansion is required (86). An initial glassy material is produced that contains predominantly $\text{SiO}_2/\text{Al}_2\text{O}_3/\text{P}_2\text{O}_5/\text{Li}_2\text{O}$, though it also has smaller amounts of cations such as Zn, Mg, Ti, Zr, Na, and As. This glass is then subjected to careful heat treatment to produce a composite glass-ceramic containing approximately 70–78% crystalline material. The crystalline component has the β -quartz structure and displays a negative coefficient of thermal expansion which compensates for the positive expansion of the glassy component. Correctly processed, Zerodur can have a coefficient of thermal expansion as low

as $0 \pm 0.02 \times 10^{-6} \text{ K}^{-1}$. Zerodur is routinely fabricated in blocks weighing several tons, and still larger blocks can be manufactured for special applications. Circular mirror blanks for high powered telescopes up to 8.6 m in diameter that weigh 45 tonnes have been manufactured. Zerodur also has extremely low permeability to He which has led to its use as building frames for laser gyroscopes which are used for angle measurements in aircraft, helicopters, and in spacecraft such as the Ariane rocket. Other areas of application include temperature stable distance spacers in lasers, supports in microlithography, and in the imaging optics for microchip manufacture.

In the 1960s Corning developed "ULE", an ultra low expansion titanium silicate glass. The average coefficient of thermal expansion between 5 and 35°C is certified as $0 \pm 0.03 \times 10^{-6} \text{ K}^{-1}$. The thermal expansion properties of ULE are unchanged by thermal cycling, regardless of the heating rates employed. ULE lightweight mirror blanks can also be produced that are reduced in weight up to 95% by using a frit bonding process to produce a "honeycomb" structure whose cell walls are as thin as 1.3 mm. A fusion-based process leading to an 80% weight reduction in which small pieces of ULE are welded together was used to manufacture the 2.4-m mirror blank for the Hubble Space Telescope.

BIBLIOGRAPHY

- J.A. Harvey, in *Encyclopedia of Chemical Technology*, 4 ed., Wiley.
- R.M. Hazen and C.T. Prewitt, *Am. Mineral.* **62**: 309–315 (1977).
- Y. Okada and Y. Tokumaru, *J. Appl. Phys.* **56**: 314–320 (1984).
- T.H.K. Barron, J.G. Collins and G.K. White, *Adv. Phys.* **29**: 609–730 (1980).
- K. Rottger, A. Endriss, J. Ihringer, S. Doyle and W. F., Kuhs, *Acta Crystallogr B* **50**: 644–648 (1994).
- A., Schauer, *Can. J. Phys.* **43**: 524–531 (1965).
- D., Taylor, *Br. Ceram. Trans. J.* **83**: 129–134 (1984).
- R. Roy, D.K. Agrawal, and H.A. McKinstry, *Annu. Rev. Mater. Sci.* **19**: 59–81 (1989).
- D. Taylor, *Br. Ceram. Trans. J.* **90**: 64–69 (1991).
- J.S.O. Evans, W.I.F. David, and A.W. Sleight, *Acta Crystallogr. B* **55**: 333–340 (1999).
- J.S.O. Evans, unpublished results (1999).
- J.S.O. Evans, T.A. Mary, and A.W. Sleight, *J. Solid State Chem.* **137**: 148–160 (1998).
- P.M. Forster and A.W. Sleight, *J. Inorg. Mater.* **1**: 123–127 (1999).
- M. Blackman, *Proc. Phys. Soc. B* **70**: 827 (1957).
- T.H.K. Barron, *Ann. Phys.* **1**: 77 (1957).
- G.K. White, *Contemporary Phys.* **34**: 193–204 (1993).
- K.S. Knight, unpublished results (1999).
- S.M. Bennington, J.C. Li, M.J. Harris, and D.K. Ross, *Physica B* **263**: 396–399 (1999).
- H.D. Megaw, in *Crystal Structures a Working Approach*. W.B. Saunders: Philadelphia, 1973.
- P.J. Heaney, In "Silica Physical Behavior, Geochemistry & Materials Applications" *Rev. Miner.*, P.J. Heaney, C.T. Prewitt, and G.V. Gibbs, eds., 1994; Vol. 24.
- M.T. Dove, M. Gambhir, K.D. Hammonds, V. Heine, and A.K.A. Pryde, *Phase Transitions* **58**: 121–143 (1996).
- P.R.L. Welche, V. Heine, and M.T. Dove, *Phys. Chem. Miner.* **26**: 63–77 (1998).
- I.D. Brown and D. Altermatt, *Acta Crystallogr. B* **41**: 244–247 (1985).
- N.E. Brese and M. O'Keeffe *Acta Crystallogr. B* **47**: 192–197 (1991).
- G.A. Rossetti, J.P. Cline, and A. Navrotsky, *J. Mater. Res.* **13**: 3197–3206 (1998).
- A.M. Glazer and S.A. Mabud, *Acta Crystallogr. B* **34**: 1065 (1978).
- D. Taylor, *Br. Ceram. Trans. J.* **84**: 181–188 (1985).
- W.F. Schlosser, G.M. Graham, and P.P.M. Meincke, *J. Phys. Chem. Solids* **32**: 927–938 (1971).
- S. Chikazumi, *J. Magn. Magn. Mater.* **10**: 113–119 (1979).
- D. Gignoux, D. Givord, F. Givord and R. Lemaire, *J. Magn. Magn. Mater.* **10**: 288–293 (1979).
- O. Sigmund and S. Torquato, *J. Mech. Phys. Solids* **45**: 1037–1067 (1997).
- R.J. Weiss *Proc. R. Soc. London. A* **82**: 281–288 (1963).
- M. van Schilfgaarde, I. A. Abrikosov and B. Johansson, *Nature* **400**: 46–49 (1999).
- T.A. Mary, J.S.O. Evans, T. Vogt, and A.W. Sleight, *Science* **272**: 90–92 (1996).
- A.K.A. Pryde, K. D. Hammonds, M.T. Dove, V. Heine, J.D. Gale and M.C. Warren, *J. Phys. C* **8**: 10973–10982 (1996).
- W.I.F. David, J.S.O. Evans, and A.W. Sleight, *Europhys. Lett.* **46**: 661–669 (1999).
- G. Ernst, C. Broholm, G.R. Kowach, and A.P. Ramirez, *Nature* **396**: 147–149 (1998).
- A.P. Ramirez and G.R. Kowach, *Phys. Rev. Lett.* **80**: 4903–4906 (1998).
- R. Mittal and S.L. Chaplot, *Phys. Rev. B* **60**: 7234–7237 (1999).
- J.S.O. Evans, T.A. Mary, T. Vogt, M.A. Subramanian, and A.W. Sleight, *Chem. Mater.* **8**: 2809–2823 (1996).
- J.S.O. Evans, Z. Hu, J.D. Jorgensen, D.N. Argyriou, S. Short, and A.W. Sleight, *Science* **275**: 61–65 (1997).
- Z. Hu, J.D. Jorgensen, S. Teslic, S. Short, D.N. Argyriou, J.S.O. Evans, and A.W. Sleight, *Physica B* **241**: 370–372 (1997).
- J.D. Jorgensen, Z. Hu, S. Teslic, D.N. Argyriou, S. Short, J.S.O. Evans, and A.W. Sleight, *Phys. Rev. B* **59**: 215–225 (1999).
- C.A. Perottoni and J.A.H. Da Jornada, *Science* **280**: 886–889 (1998).
- J.D. Jorgensen, Z. Hu, S. Short, A.W. Sleight, and J.S.O. Evans, *J. Appl. Phys.* **89**: 3184–3188 (2001).
- C. Lind, A.P. Wilkinson, Z.B. Hu, S. Short, and J.D. Jorgensen, *Chem. Mater.* **10**: 2335 (1998).
- J.S.O. Evans, P.A. Hanson, R.M. Ibberson, U. Kameswari, N. Duan, and A.W. Sleight, *J. Am. Chem. Soc.* **122**: 8694–8699 (2000).
- U. Kameswari, A.W. Sleight, and J.S.O. Evans, *Int. J. Inorg. Mater.* **3**: 333–337 (2000).
- R.L. Withers, Y. Tabira, J.S.O. Evans, I.J. King, and A.W. Sleight, *J. Solid State Chem.* **157**: 186–192 (2001).
- R.L. Withers, J.S.O. Evans, J. Hanson, and A. W. Sleight, *J. Solid State Chem.* **137**: 161–167 (1998).
- E. Tillmanns, W. Gebert, and W.H. Baur, *J. Solid State Chem.* **7**: 69–84 (1974).
- J. Sanz, J.E. Iglesias, J. Soria, E.R. Losilla, M.A.G. Aranda, and S. Bruque, *Chem. Mater.* **9**: 996–1003 (1997).

53. V. Korthuis, N. Khosrovani, A.W. Sleight, N. Roberts, R. Dupree, and W.W. Warren, *Chem. Mater.* **7**: 412–417 (1995).
54. E. Breval, H.A. McKinstry, and D.K. Agrawal, *Br. Ceram. Trans. J.* **93**: 239–251 (1994).
55. J.P. Boilot, J.P. Salonie, G. Desplanches, and D. Le Potiel *Mater. Res. Bull.* **14**: 1469 (1979).
56. G.E. Lenain, H.A. McKinstry, J. Alamo, and D.K. Agrawal, *J. Mater. Sci.* **22**: 17–22 (1987).
57. N. Khosrovani and A.W. Sleight, *Int. J. Inorg. Mater.* **1**: 3–11 (1999).
58. M. Alami, R. Brochu, J.L. Soubeyroux, P. Gravereau, G.L.F., and P. Hagenmuller, *J. Solid State Chem.* **90**: 185–193 (1991).
59. D.A. Woodcock, P. Lightfoot, and C. Ritter, *Chem. Commun.* 107–108 (1998).
60. P. Lightfoot, D.A. Woodcock, J.D. Jorgensen, and S. Short, *Int. J. Inorg. Mater.* **1**: 53–60 (1999).
61. D.A. Woodcock, P. Lightfoot, and R.I. Smith, *J. Mater. Chem.* **9**: 2631–2636 (1999).
62. E. Breval and D.K. Agrawal, *Br. Ceram. Trans. J.* **94**: 27–32 (1995).
63. E. Breval, H.A. McKinstry, and D.K. Agrawal, *J. Am. Ceram. Soc.* **81**: 926–932 (1998).
64. V. Srikanth, E.C. Subbarao, D.K. Agrawal, C.Y. Huang, R. Roy and G.V. Rao, *J. Am. Ceram. Soc.* **74**: 365–368 (1991).
65. D.A. Woodcock, P. Lightfoot, and R.I. Smith, *Mater. Res. Soc. Proc.* **547**: 191–196 (1999).
66. J. Alamo and J.C. Rodrigo, *Solid State Ionics* **32–33**: 70–76 (1989).
67. D.K. Agrawal, G. Harshe, E. Breval, and R. Roy, *J. Mater. Res.* **11**: 3158–3163 (1996).
68. S.C. Abrahams, *J. Chem. Phys.* **45**: 2745–2752 (1966).
69. A.W. Sleight and L.H. Brixner, *J. Solid State Chem.* **7**: 172–174 (1973).
70. J.S.O. Evans, T.A. Mary, and A.W. Sleight, *J. Solid State Chem.* **133**: 580–583 (1997).
71. J.S.O. Evans and T.A. Mary, *Int. J. Inorg. Mat.*, **2**: 143–151 (2001).
72. P.M. Forster, A. Yokochi, and A.W. Sleight, *J. Solid State Chem.* **140**: 157–158 (1998).
73. T.A. Mary and A.W. Sleight, *J. Mater. Res.* 912–915 (1999).
74. J.W. Couves, R.H. Jones, and S.C. Parker, *J. Phys. Condens. Mater.* **27**: L329–L332 (1993).
75. D.A. Woodcock, P. Lightfoot, P.A. Wright, L.A. Villaescusa, M.J. DiazCabanas, and M.A. Cambor, *J. Mater. Chem.* **9**: 349–351 (1999).
76. M.P. Attfield and A.W. Sleight, *Chem. Mater.* **10**, 2013–2019 (1998).
77. M.P. Attfield and A.W. Sleight, *Chem. Commun.* 601–602 (1998).
78. P. Tschaufeser and S.C. Parker, *J. Phys. Chem.* **99**: 10600 (1995).
79. S.H. Park, R.W.G. Kunstleve, and H. Graetsch, *Stud. Surf. Sci. Catal.* **105**: 1989–1994 (1997).
80. P. Lightfoot, D.A. Woodcock, M.J. Maple, L.A. Villaescusa, and P.A. Wright, *J. Mater. Chem.* **11**: 212–216 (2001).
81. D.A. Woodcock, P. Lightfoot, L.A. Villaescusa, M.-J. Diaz-Cabanas, M.A. Cambor, and D. Engberg, *Chem. Mater.* **11**: 2508–2514 (1999).
82. H. Xu, P.J. Heaney, D.M. Yates, V.D. R.B., and M.A. Bourke, *J. Mater. Res.* **14**: 3138–3151 (1999).
83. G.H. Beall, K. Chyung, and J.E. Pierson, *Proc. 18th Int. Congr. Glass*, (1998) 1738–1743.

84. R. Morena and G.L. Francis, *Proc. 18th Int. Congr. Glass*, (1998) 1402–1407.
85. D.K. Agrawal, V.S. Stubican, and Y. Mehrotra, *J. Am. Ceram. Soc.* **69**: 847–851 (1986).
86. Sohott Glass Technologies Inc., in *Zerodur—Precision from Glass Ceramics*, 1999.
87. J.S.O. Evans and T.A. Mary, *Int. J. Inorg. Mater.* **2**: 143–151 (2000).

TRIBOLUMINESCENCE, APPLICATIONS IN SENSORS

KOJI FUJITA

Kyoto University
Sakyo-ku, Kyoto, Japan

KATSUHISA TANAKA

Kyoto Institute of Technology
Sakyo-ku, Kyoto, Japan

TSUGUO ISHIHARA

Hyogo, Prefectural Institute of Industrial Research
Suma-ku, Kobe, Japan

KAZUYUKI HIRAO

Kyoto University
Sakyo-ku, Kyoto, Japan

INTRODUCTION

Triboluminescence is the emission of photons caused by applying mechanical energy to a solid. The word Triboluminescence, coined by Wiedemann (1) in 1895, has its root in the Greek “tribein” to rub. The word has broad general usage and covers a variety of mechanical methods of excitation, spectroscopic origins of luminescence, and mechanisms of excitation.

Triboluminescence (TL) has very long history. To the best of our knowledge, TL was first reported by Francis Bacon in the early seventeenth century and is mentioned in *The Advancement of Learning* (2,3). He observed that lumps of sugar emit light when they are scraped. At present, it is known that nearly one-half of all inorganic compounds and between one-quarter to one-third of all organic compounds exhibit TL. Spectroscopic examination and characterization of the emitted light are useful for determining the excited-state origins of TL. The best general spectroscopic characteristics that can be expected from a crystal is that the TL is similar to photoluminescence (PL) at the same temperature. TL and PL spectra are expected to be identical if the sole effect of a fracture is a population of excited electronic states. The importance of perturbations to the crystal and/or the emitting centers caused by fracture is evidenced by the differences between PL and TL spectra. For instance, some crystals show TL at room temperature but do not exhibit PL at that temperature. More commonly, a triboluminescent crystal exhibits PL, but the TL spectrum contains features that are absent from the PL spectrum. These features may be emission bands that do not occur in the PL spectrum or changes in the relative intensities of bands in the TL spectrum compared to those in the PL spectrum.

A number of different mechanisms are required to account for the phenomenon of TL, although its details are still poorly understood. In this review, the mechanisms of different types of TL along with their spectroscopic experimental examples are given first. Then, our experimental results on TL in inorganic crystals that contain rare-earth ions are presented. Finally, the potential applications of TL are briefly described.

CLASSIFICATION OF TL

The mechanisms by which TL is excited have not yet been well clarified. Thus, in addition to elucidating excited-state origins and spectral features that are possible due to advances in TL spectroscopy, classification of the mechanisms of TL is required to understand the phenomenon. The four mechanisms of triboexcitation that are currently under active consideration can be broadly categorized as electrically induced, thermally induced, chemically induced and mechanically induced. Each consists of a broad range of physical processes. As will be seen, the problem is not an absence of mechanisms that can explain the excitation, but rather is too many variable possibilities. The purpose of this section to present briefly the mechanisms currently under consideration.

Electrically Induced TL

The electrical mechanisms that have been proposed all require that mechanical energy generates free electrons, which leads to the emission of photons by electron collision with molecules, recombination luminescence of cations and anions, or electroluminescence (EL).

Piezoelectricity in a crystal from pressure requires that the crystal be noncentrosymmetric (4). When a piezoelectric crystal is cleaved or fractured, one of its newly created surfaces becomes positively charged and the other surface of the crack is negatively charged (Fig. 1). Generally, all

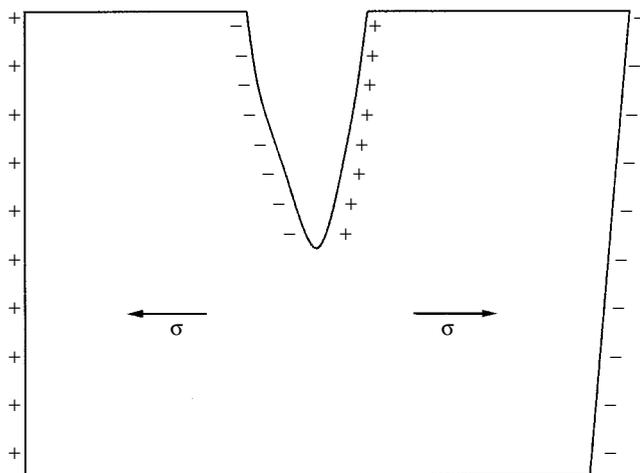


Figure 1. Schematic illustration of charge separation in the crack of a piezoelectric crystal. The crystal is subjected to a tensile stress σ directed along a polar axis of the crystal.

Table 1. Examples of Triboluminescent Materials Induced by Piezoelectricity

Materials	References
Ditriphenylphosphine oxide manganese bromide	(7,8,10)
Triethylammonium tetrakis (dibenzoylmethide) europate	(9,10)
Sucrose	(11)
Tartaric acid	(6)
Uranyl nitrate hexahydrate	(5,12)
Acenaphthene	(10,13)
Phthalic anhydride	(10,13)
Phenanthrene	(10)
Hexaphenylcarbodiphosphorane	(14,15)
<i>N</i> -Isopropyl carbazole	(16,17)
<i>N</i> -Acetyl anthranilic acid	(10,14)
Doped zinc sulfide (ZnS)	(18)

piezoelectric crystals exhibit the phenomenon of TL, and the crystals that do not show TL are nonpiezoelectric (5). Furthermore, several polymorphic crystals that belong to the piezoelectric point group exhibit TL, whereas crystals that do not belong to the piezoelectric point group do not show TL (5). Walton et al. (6) calculated the voltage generated by the fracture of a piezoelectric crystal under the assumption that the crystal is subjected to a tensile stress directed along a polar axis of the crystal and estimated whether or not the voltage is sufficient to initiate gas discharge. The piezoelectric constant in this direction is generally of the order of 10^{-12} C/N, and the stress near the tip of a crack is of the order of 10^8 N/m². Thus, if it is assumed that the relaxation time of the stress is less than the time it takes for the fracture to spread across the crystal, the charge density of the newly created surfaces is of the order of 10^{-4} C/m². The electric field between the oppositely charged density of newly created surfaces is ρ/ϵ_0 , that is, nearly 10^7 V/m, where ρ is the charge density and ϵ_0 is the permittivity constant. Such a field has sufficient magnitude to cause dielectric breakdown of the ambient gas. Several examples of materials where triboluminescence is, induced by piezoelectricity during fracture are tabulated in Table-1 (5–18). The TL of the first two materials is so intense that it can be seen in daylight.

The TL spectrum of sucrose provided clear evidence of emission from nitrogen gas molecules. The nitrogen emission observed corresponds to ${}^3\Pi_u \rightarrow {}^3\Pi_g$ fluorescence (19). A well-defined vibronic band structure is observed, as shown in Fig. 2 (11). A minimum energy of 8.9×10^4 cm⁻¹ is needed to excite nitrogen to its ${}^3\Pi_u$ excited state. The ultraviolet part of the nitrogen gas discharge can excite the PL of surrounding materials; the TL of uranyl nitrate hexahydrate, for instance, has such an origin (5,12). On the other hand, it was found that the TL spectrum of Cu, Ag or Mn-doped ZnS that is piezoelectric corresponds to the PL spectrum that has a small shift in the energy of the peak maxima (18). All of these shifts in TL spectra were in the same direction as those observed upon applying hydrostatic pressure. The spectrum of Cu-doped ZnS exhibits two bands and matches the EL spectrum more closely than the PL spectrum.

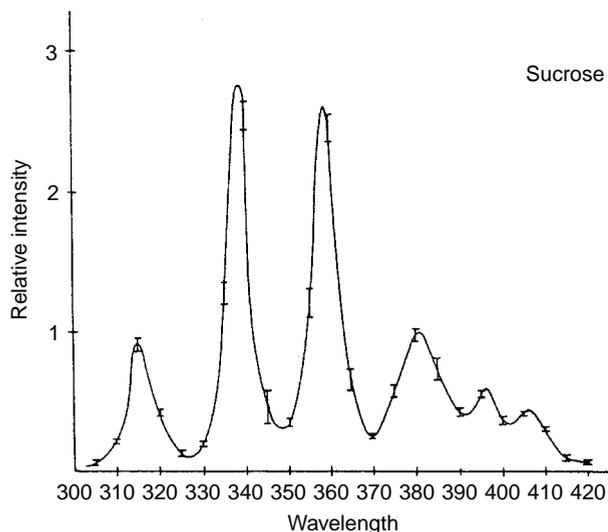


Figure 2. Triboluminescent spectrum of sucrose, corresponding to emission from a nitrogen molecule [from (11)].

In several cases, the crystals and glasses whose structures possess a center of symmetry and are nonpiezoelectric exhibit TL. This means that processes other than piezoelectricity are also responsible for charging newly created surfaces. In alkali halides such as LiF and NaCl without piezoelectricity, direct evidence of positive and negative surface charges on fresh fracture surfaces was provided by Wollbrandt et al. (20) using an electrostatic probe that had a 100- μm spatial resolution. Such electrical activity was so strongly supported by the observation that electromagnetic radiation in the radio range was detected coincident with the fracture of LiF, NaCl, and MgO (21,22). One possible source for creation of charged surfaces is the motion of charged dislocations intersected by the fracture surface (23,24). Another source is internal electrification at cleavage or shear planes in the crystal (23,24). If the charge density is sufficiently high, microdischarge takes place in the vicinity of the crack tip, where catastrophic ionization results in dielectric breakdown of the ambient atmosphere. The electrical discharge yields emission of photons, electron, and other particles that accompanies the fracture of solids, including metals, semiconductors, and insulators (25–27). The emission of photons, electrons, and other particles can also excite other luminescent centers, such as impurity ions or defect centers. Even in the absence of microdischarging, local concentrations of surface charge accelerate previously emitted electrons toward the surface. Because the energy of emitted electrons is in the keV range (28), electron bombardment can be sufficiently energetic to cause further emission of photons, electrons, and other particles.

In the 1965 Matsushiro earthquake swarm, where many cracks and fissure zones were formed at the surface (29), luminescence, as well as an anomalous change in the earth's potential, was observed (30). Such surface cracking causes emission of charged particles such as the exoelectron, which is responsible for exciting luminescence, as suggested by Brady and Rowell (31). They recorded the emission spectra of photons induced by the fracture of rocks

such as granite and basalt in argon, helium, air, water, and vacuum and found that the emission consists of line spectra corresponding to the excitation of the ambient atmosphere. Although many rock materials contain minerals, known as piezoelectrics, the basalt was free from any strong piezoelectric minerals. In addition, the spectrum of basalt in argon was identical to that produced by granite in argon. Deformation of piezoelectric minerals does occur during fracture of rocks, but these facts imply that the piezoelectric mechanism is not associated with TL in these minerals. It was proposed that an exoelectron bombardment mechanism is responsible for exciting the ambient atmosphere.

Moreover, photons, electrons, and other particles are emitted by the fracture of silica glass (32–34). Silica glass has excellent optical properties that can be used for practical applications. One of the superior properties of silica glass as an optical material is high transparency, which makes it possible to use silica glass as windows, lenses, prisms, waveguides, and fibers for optical telecommunications. Kawaguchi (34) investigated the time-resolved TL of silica glass in vacuum and nitrogen atmosphere. Two emission bands at 1.9 and 2.7 eV were observed in the spectra. The 1.9-eV band peaks around $5\mu\text{s}$ and decays around $100\mu\text{s}$ after the instant of fracture. The 2.7-eV band rises in about $50\mu\text{s}$, peaks around $500\mu\text{s}$, and decays in several tens of milliseconds after fracture. The energy position and the time response of the two bands were similar to those in PL. The 2.7-eV band was ascribed to the luminescence of oxygen vacancies on the fracture surface of the silica glass, and the 1.9-eV band was related to relaxation of nonbridging oxygen hole centers. A plausible mechanism by which TL can be excited is that the defect centers created at fracture are excited by collisions of emitted electrons and ions. Another mechanism is that the defect centers are excited directly during defect formation by mechanical energy supplied.

Chemically Induced TL

Chemical reaction takes place between atoms and ions liberated during the fracture of crystals, and subsequently, the release of energy gives rise to luminescence. Moreover, molecules of the surrounding gases are absorbed or adsorbed on newly created surfaces when the solid is fractured, and the release of energy in this process causes luminescence. Kasemo and Walden (35) reported spontaneous emission of photons and electrons during chemisorption of chlorine on sodium. Kasemo (36) also observed luminescence during chemisorption of oxygen on aluminum and magnesium surfaces.

Thermally Induced TL

When fracture occurs, plastic work is transformed into heat. The heat produced near the crack tips either stimulates the defect centers and causes luminescent excitation or leads to blackbody radiation or incandescent emission if it is very high. Although incandescent emission does not involve luminescence, it is included here to account for different processes of light emission during deformation or fracture.

Wick (37) pointed out that the TL of X- or γ -ray irradiated synthetic fluorites (CaF_2) that contain different rare-earth ions occurs by deformation-induced thermoluminescence. The TL spectra were composed of emission bands characteristic of trivalent rare-earth ions such as Dy and Tb. Deformation-induced thermoluminescence was also observed for X- or γ -ray irradiated alkali halides (6). On the other hand, by solving the thermal conduction equation, it was shown that temperatures of the order of 10^4 K can be sustained within the plastically strained regions in the vicinity of crack tips (6). Fracture of alkali halide crystals such as LiF, NaF, and NaCl in vacuum was accompanied by flashes of high temperature luminosity, that is, incandescent TL (6). The blackbody curves were fitted to the spectra by assuming a temperature of 10^4 K for LiF and NaF and of 6×10^3 K for NaCl. Because these temperatures are higher than the melting points of the samples, it was postulated that the time taken by the crack to grow by a length equal to the diameter of the local hot spot is less than the time for the heat to be conducted away through the body of the crystal.

Chapman and Walton (38) measured TL spectra of a variety of glasses and of crystalline quartz cut slowly by a rotating diamond-impregnated saw blade and found that the TL spectra resemble the emission of a blackbody radiator at a temperature that corresponds to that of the heated material. Emission temperatures were estimated at around 1850 K for armor plate glass, 2100 K for Pyrex glass, 2400 K for soda lime glass, 2300 K for high-density lead glass, and 2800 K for cut quartz. Because the TL spectra of quartz and glasses did not contain incandescent emission during fracture, it seems that the blackbody emission produced during cutting of these materials by an diamond-impregnated saw blade is attributable to the heat produced by the friction between the blades and the samples.

Mechanically Induced TL

There is a possibility that thermally activated electronic transitions occur in large strain or high stress regions. This effect is the inverse of thermal radiationless quenching of excited states. When displacements along the configuration coordinate fluctuate greatly, the potential curves of the ground and excited states cross each other (Fig. 3), and the subsequent transition of excited electrons to the ground state gives rise to luminescence. The mechanism of radiationless transitions that lead to a population of excited electronic states in large strain or high stress regions near a crack tip was theoretically advanced by Lin et al. (39). Recently, Li and his co-workers (40,41) investigated TL during cleavage of semiconductors, and found that when the number of atomic bonds that are broken on an average cleaved pair of surfaces is about 10^{15} , the number of emitted photons is nearly 10^8 – 10^9 . The observation that the number of emitted photons is much less than the number of broken bonds reveals that the cleavage process is mainly nonradiative. It appears that the TL of Si, Ge, InP, and other semiconductors is caused by radiationless transitions via thermal excitation, where the interatomic distance among some of the atoms of broken bonds is very

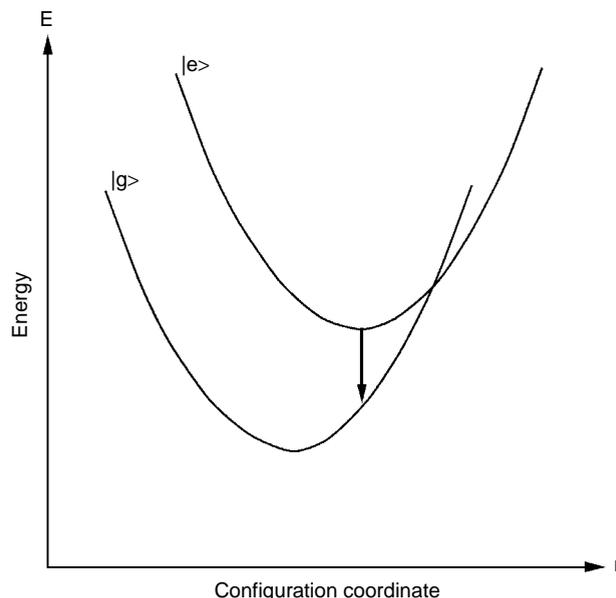


Figure 3. Schematic illustration of TL originating from an excited electronic state via radiationless transition. The $|g\rangle$ and $|e\rangle$ represent the potential energy curves of ground and excited states, respectively.

large and therefore, the thermal movement of electrons from the ground state to the excited state is possible.

On the other hand, Chapman and Walton (42) observed TL for single crystals of fluorite (CaF_2) doped with Tb^{3+} , Dy^{3+} , Sm^{3+} , or Eu^{3+} . The crystals were chosen because the possibility of piezoelectrically induced TL can be ruled out. When the crystals were cut by a diamond-impregnated circular saw, the TL spectra show significantly more structure than the room-temperature PL spectra. Although the excited-state origins of TL are rare-earth ions, it is likely that emitting rare-earth ions are located in the vicinity of the tip of a growing crack and/or on the faces of the virgin crack, where the applied stress is concentrated in a severely distorted crystal lattice. Hence, the rare-earth ions are subjected to a significantly distorted environment for time of the order of an atomic vibrational period. This, in turn, leads to changes in the Frank–Condon factors and thus to enhanced vibrational spectral components. The perturbing effect of high stress or large strain on electronic structures unambiguously appears in TL spectra. However, questions remain as to whether or not the TL of fluorites doped with rare-earth ions originates from the population of excited electronic states by radiationless transitions because the microdischarge of gas molecules takes place in nonpiezoelectric crystals, as mentioned before. Photons that originate from the microdischarge may excite the rare-earth ions, and subsequently yield TL. Measurements of TL in fluorites without rare earths are needed to reveal the detailed mechanisms of TL.

TL OF OXIDE CRYSTALS DOPED WITH RARE EARTHS

Rare-earth-doped inorganic crystals and glasses have optical properties that are interesting and important from

fundamental and practical viewpoints. The optical properties of rare-earth-doped inorganic materials are applicable to the development of optoelectronics devices such as lasers, optical amplifiers, optical memories, and optical modulators. As mentioned in the previous section, some inorganic crystals doped with rare-earth ions exhibit TL. For instance, TL was observed in single crystalline fluorite (CaF_2) doped with trivalent rare-earth ions, and it was shown that the rare-earth ions are responsible for TL (42). As for oxide materials, our research group found that polycrystalline hexacelsians ($\text{BaAl}_2\text{Si}_2\text{O}_8$ and $\text{SrAl}_2\text{Si}_2\text{O}_8$) doped with rare-earth ions exhibit TL caused by the $4f-4f$ and/or $4f-5d$ electronic transitions of rare-earth ions (43–45). Intense TL was also observed in polycrystalline $\text{Sr}_3\text{Al}_2\text{O}_6:\text{Eu},\text{Dy}$ and $\text{SrAl}_2\text{O}_4:\text{Eu}^{2+}$, as revealed by Akiyama et al. (46) and Xu et al. (47,48), respectively. In addition, Sage et al. (49) demonstrated TL in a rare-earth complex and urged its application to sensing of structural damage and fracture.

In the next section, we describe our recent experimental results for TL in polycrystalline oxides doped with rare-earth ions.

TL of Hexacelsians Doped with Rare Earths

The fact that polycrystalline barium hexacelsians ($\text{BaAl}_2\text{Si}_2\text{O}_8$) doped with rare-earth ions exhibit TL was accidentally discovered by Ishihara et al. (43). The polycrystalline samples were prepared from reagent-grade BaCO_3 , Al_2O_3 , SiO_2 , and rare-earth oxides. After the raw materials were mixed thoroughly, the mixture was heated above the melting point of $\text{BaAl}_2\text{Si}_2\text{O}_8$. A densified body of polycrystalline $\text{BaAl}_2\text{Si}_2\text{O}_8$ doped with rare-earth ions was obtained by cooling from its liquid state. The crystal structure of $\text{BaAl}_2\text{Si}_2\text{O}_8$ is schematically illustrated in Fig. 4. The structure consists of layers of silicate and aluminate structural units between which Ba^{2+} layers are inserted.

TL was measured at room temperature while pressing the polycrystalline samples with a pressure device. TL spectra were obtained by using a CCD detector equipped with a multichannel analyzer. Photoluminescence (PL) spectra were measured using a fluorescence spectrophotometer for comparison. As an example, TL (solid circles) and PL (solid line) spectra of $\text{BaAl}_2\text{Si}_2\text{O}_8:\text{Dy}^{3+}$ are shown in Fig. 5. The emission lines at 483, 576, and 662 nm in both TL and PL spectra are ascribable to the ${}^4\text{F}_{9/2}-{}^6\text{H}_{15/2}$, ${}^4\text{F}_{9/2}-{}^6\text{H}_{13/2}$, and ${}^4\text{F}_{9/2}-{}^6\text{H}_{11/2}$ transitions of Dy^{3+} , respectively. The peak positions and the relative intensities of the emission lines in the TL spectrum are almost identical to those in the PL spectrum. It is unambiguous that the TL is caused by the electronic transitions of the doped Dy^{3+} ions. Figure 6 shows TL (solid circles) and PL (solid line) spectra of $\text{BaAl}_2\text{Si}_2\text{O}_8:\text{Tb}^{3+}$. The excitation wavelength for PL is 350 nm. All of the emission lines observed in the TL spectrum are assigned to the ${}^5\text{D}_4-{}^7\text{F}_J$ and ${}^5\text{D}_3-{}^7\text{F}_J$ transitions of Tb^{3+} , as indicated in the figure. The emission lines due to the ${}^5\text{D}_3-{}^7\text{F}_4$ and ${}^5\text{D}_3-{}^7\text{F}_5$ transitions in the TL spectrum do not appear in the PL spectrum, presumably because the excitation wavelength (350 nm) is not suitable for emission from the ${}^5\text{D}_3$ state. In fact, these emission lines become

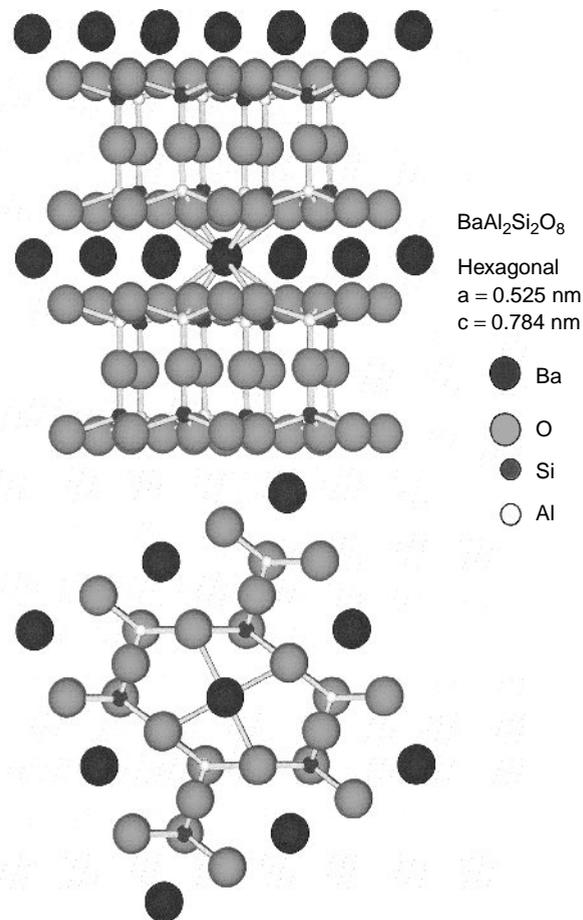


Figure 4. Schematic illustration of crystal structure of $\text{BaAl}_2\text{Si}_2\text{O}_8$.

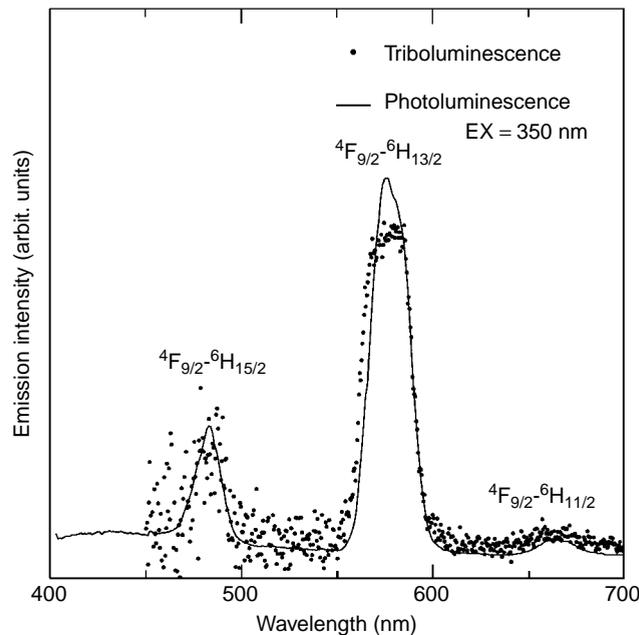


Figure 5. TL (closed circles) and PL (solid line) spectra of a $\text{BaAl}_2\text{Si}_2\text{O}_8:\text{Dy}^{3+}$ polycrystal.

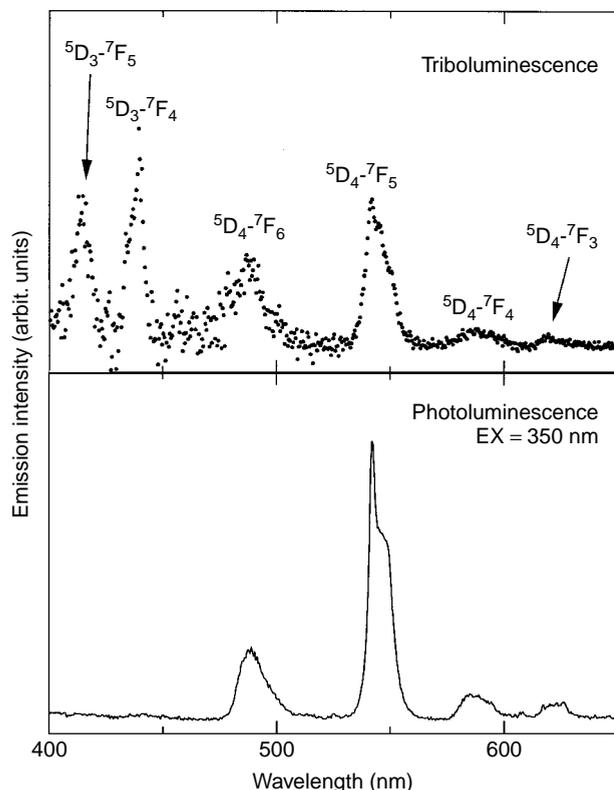


Figure 6. TL (closed circles) and PL (solid line) spectra of a $\text{BaAl}_2\text{Si}_2\text{O}_8:\text{Tb}^{3+}$ polycrystal.

visible in the PL spectrum when the excitation wavelength is changed to 250 nm. Similar TL and PL spectra were obtained for polycrystalline $\text{SrAl}_2\text{Si}_2\text{O}_8:\text{Tb}^{3+}$. Thus far, TL characterized by the electronic transitions of doped rare-earth ions was observed in $\text{BaAl}_2\text{Si}_2\text{O}_8$ doped with Eu^{2+} , Sm^{2+} , Sm^{3+} , Yb^{2+} , or Ce^{3+} and $\text{SrAl}_2\text{Si}_2\text{O}_8$ doped with Eu^{2+} or Dy^{3+} , in addition to the examples previously mentioned. In our early work on the TL of hexacelsians, we argued that the emission peak of TL is shifted to a longer wavelength compared with PL for 4f–5d transitions (44). However, the difference in the peak position between TL and PL spectra is explained mainly in terms of the incomplete sensitivity correction of the detector used for measuring TL, in addition to the fact that the S/N ratio in the TL spectra is very low.

It is also known that TL is observed in $\text{BaAl}_2\text{Si}_2\text{O}_8$ without intentional dopants. The TL spectrum of $\text{BaAl}_2\text{Si}_2\text{O}_8$ without intentional dopants manifests intense emission lines at around 315, 335, 355, and 380 nm, as shown in Fig. 7. Whether or not a crystal exhibits TL depends on its crystal structure. Table 2 shows the relationship among crystal structure, piezoelectricity, and the presence of TL, as summarized by Chandra (5). The hexacelsian has a space group of symmetry D_{6h} ($P6/mmm$) which precludes piezoelectricity according to this table. Nonetheless, TL is observed for a hexacelsian without intentional dopants, as shown in Fig. 7. Brady and Rowell (31) proposed that the light emission observed in earthquakes is caused by exo-electron excitation of the ambient atmosphere. Nakayama (50) found that many charged particles are emitted from

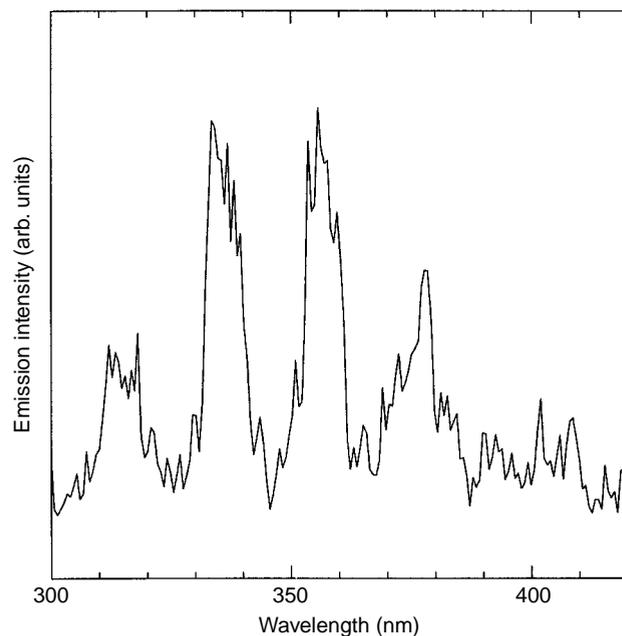


Figure 7. TL spectrum of a $\text{BaAl}_2\text{Si}_2\text{O}_8$ polycrystal without any intentional dopants.

mica when it is scratched and argued that triboemission originates from a discharge from ambient atmosphere in the vicinity of the frictional surface. A comparison between Figs. 2 and 7 reveals that the TL spectrum observed for $\text{BaAl}_2\text{Si}_2\text{O}_8$ without intentional dopants is ascribable to

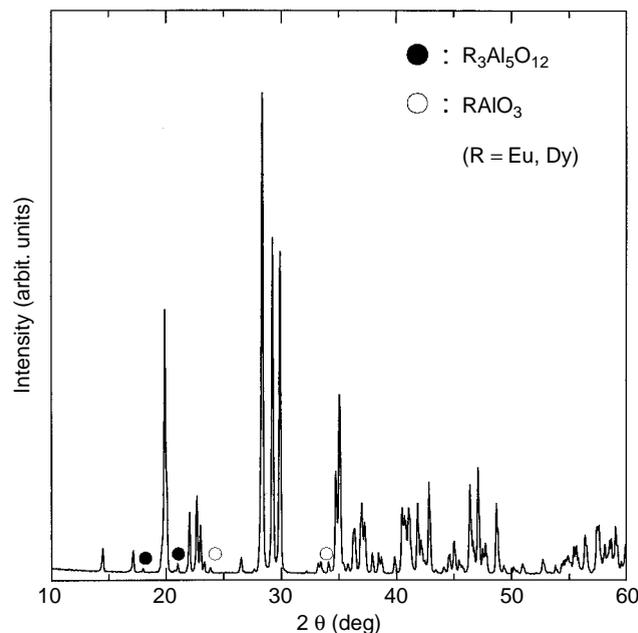


Figure 8. X-ray diffraction pattern of a $\text{SrAl}_2\text{O}_4:\text{Eu}^{2+}, \text{Dy}^{3+}$ polycrystalline sample prepared by the conventional solid-state reaction. All of the diffraction lines but those indicated by closed and open circles, which are ascribable to $\text{R}_3\text{Al}_5\text{O}_{12}$ and RAlO_3 (R corresponds to Eu^{3+} and/or Dy^{3+}), are assigned to SrAl_2O_4 .

Table 2. Relationship among Crystal Structure, Piezoelectricity, and TL proposed by Chandra (5)

Crystal System	Point Group		Piezoelectricity (present: + absent: -)	Triboluminescence (present: + absent: -)
	International Symbol	Schoenflies Symbol		
Triclinic	1	C_1	+	+
	1	C_1	-	-
Monoclinic	2	C_2	+	+
	$m = 2$	C_s	+	+
	$2/m$	C_{2h}	-	-
Orthorhombic	222	D_2	+	+
	$mm2$	C_{2v}	+	+
	mmm	D_{2h}	-	-
Tetragonal	4	C_4	+	+
	4	S_4	+	+
	$4/m$	C_{4h}	-	-
	422	D_4	+	+
	$4mm$	C_{4v}	+	+
	$42m$	D_{2d}	+	+
	$4/mmm$	D_{4h}	-	-
Trigonal	3	C_3	+	+
	3	S_6	-	-
	32	D_3	+	+
	$3m$	C_{3v}	+	+
	$3m$	D_{3d}	-	-
Hexagonal	6	C_6	+	+
	6	C_{3h}	+	+
	$6/m$	C_{6h}	-	-
	622	D_6	+	+
	$6mm$	C_{6v}	+	+
	$6m2$	D_{3h}	+	+
	$6/mmm$	D_{6h}	-	-
Cubic	23	T	+	+
	$m3$	T_h	-	-
	432	O	+	+
	$43m$	T_d	+	+
	$m3m$	O_h	-	-

the light emission brought about by the discharge from nitrogen molecules. The discharge presumably results from cleavage at the Ba^{2+} layer of the $BaAl_2Si_2O_8$. Doped rare-earth ions are excited by photons caused by the discharge from nitrogen molecules because a TL was measured in air. The excited rare-earth ions lead to luminescence due to $4f-5d$ and/or $4f-4f$ transitions.

TL of Alkaline-Earth Aluminates Doped with Rare Earths

As described before, intense TL was observed from polycrystalline $Sr_3Al_2O_6:Eu,Dy$ and $SrAl_2O_4:Eu^{2+}$ (46–48). Rare-earth-doped strontium aluminate crystals are very interesting because some of them exhibit long lasting phosphorescence (51). In particular, $SrAl_2O_4$ doped with Eu^{2+} and Dy^{3+} shows intense phosphorescence that lasts a very long time. The long lasting phosphorescence is a phenomenon where a solid irradiated by UV or white light beforehand continues to emit light even after the excitation ceases. In this section, we mention our experimental results for TL of polycrystalline $SrAl_2O_4:Eu^{2+},Dy^{3+}$, Dy^{3+} -doped $(Sr,Ba)Al_2O_4$, and $(Sr,Ca)Al_2O_4$ polycrystals.

An interesting application of the TL of $SrAl_2O_4:Eu^{2+}$ for sensing stress was demonstrated by Xu et al. (47,48) and is cited in the following section.

Figure 8 shows the X-ray diffraction pattern of polycrystalline $SrAl_2O_4:Eu^{2+},Dy^{3+}$ prepared by the conventional solid-state reaction. Although very weak diffraction lines ascribed to $R_3Al_5O_{12}$ and $RAIO_3$ (R corresponds to Eu^{3+} and/or Dy^{3+}) are observed, as indicated by solid and open circles in the figure, almost all of the diffraction lines are attributable to $SrAl_2O_4$. In addition, although Dy^{3+} and Eu^{3+} form the crystalline phases previously mentioned, some of the Dy^{3+} and Eu^{2+} ions are incorporated into the $SrAl_2O_4$ phase and replace the Sr^{2+} ions because the sample exhibits long lasting phosphorescence.

Figure 9 shows photographs of TL taken at various periods using a video camera after uniaxial compressive stress was applied to the $SrAl_2O_4:Eu^{2+},Dy^{3+}$ polycrystal. In the figure, the periods calculated from the shutter velocity of the video camera are represented below each photograph as times of 0s to 11/30s, respectively. No luminescence was found before the fracture of the sample (at 0s). A green emission is clearly observed in the photographs after the

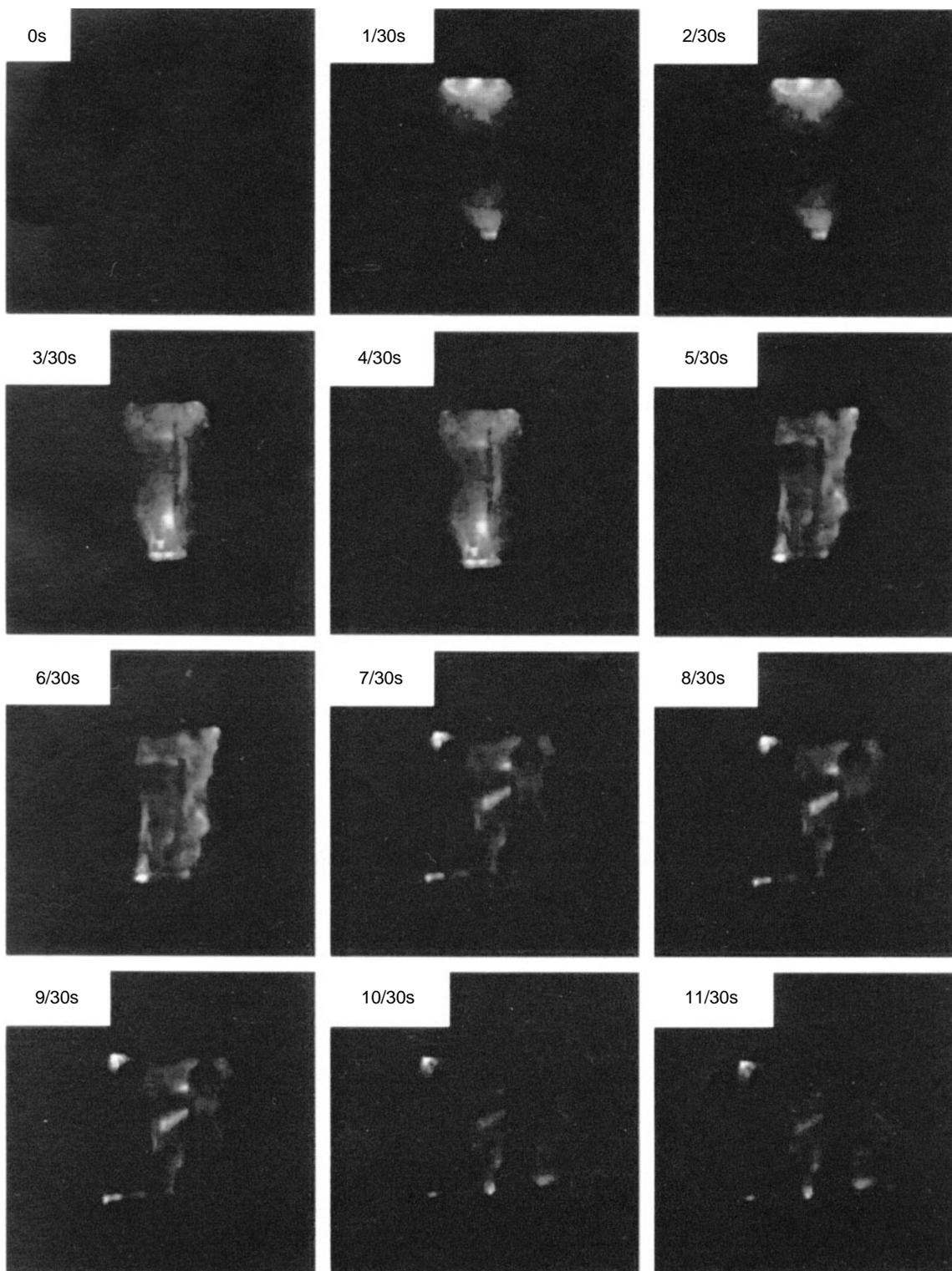


Figure 9. Photographs of the TL of a $\text{SrAl}_2\text{O}_4\text{:Eu}^{2+}, \text{Dy}^{3+}$ polycrystal. The photographs were taken at various periods after a uniaxial compressive stress was applied to the sample. In (48), a real-time image of stress-induced luminescence was also demonstrated for $\text{SrAl}_2\text{O}_4\text{:Eu}^{2+}$. Here, it should be stressed that our TL data are based on the fracture of solids, whereas nondestructive deformation of the crystal gives rise to TL in (48).

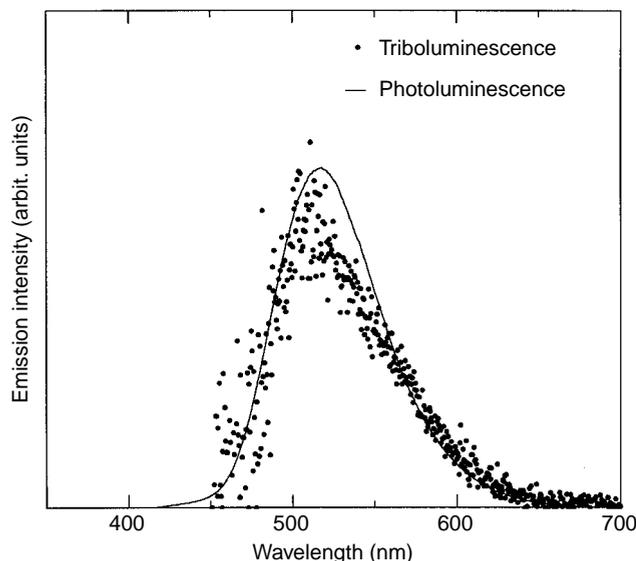


Figure 10. TL (closed circles) and PL (solid line) spectra of a $\text{SrAl}_2\text{O}_4:\text{Eu}^{2+},\text{Dy}^{3+}$ polycrystal.

fracture of the sample, although the stress required for the fracture could not be estimated by using our equipment. Because the TL was measured after the long lasting phosphorescence ceased, the luminescence in Fig. 9 is undoubtedly caused by fracture of the sample. In Fig. 9, it is observed that the part that exhibits green emission within the sample varies from place to place with time. This time dependence results from the possibility that the stress is not applied uniformly to the polycrystalline sample.

TL (solid circles) and PL (solid line) spectra of $\text{SrAl}_2\text{O}_4:\text{Eu}^{2+},\text{Dy}^{3+}$ are shown in Fig. 10. The excitation wavelength for PL is 330 nm. A broad band is observed at around 500–520 nm in both spectra. The wavelength region of this broad band coincides with the green emission. A comparison of the broad band between TL and PL spectra reveals that the peak position and the profile are similar to, indicating that the origin of the luminescence is the same as that of TL. This broad band is ascribable to the $4f^{65d}-4f^7$ transition of the doped Eu^{2+} ion. Emission due to the $4f-4f$ transitions of Dy^{3+} is not observed in the TL nor the PL spectrum.

In PL, light at a wavelength of 330 nm excites Eu^{2+} ions, and radiative decay in Eu^{2+} is observed. It is thought that Dy^{3+} plays the role of acceptor for an electron or a positive hole in long lasting phosphorescence (51,52). As for TL, emission of electrons, ions, and photons takes place when triboluminescent solids are fractured. Hence, the following model is proposed for the mechanism of TL in a $\text{SrAl}_2\text{O}_4:\text{Eu}^{2+},\text{Dy}^{3+}$ polycrystal. Initially, photons emitted by the fracture of SrAl_2O_4 excite Eu^{2+} ions. Some excited electrons relax to the ground state of Eu^{2+} and radiate. Other electrons and positive holes also formed by the excitation are trapped at defect sites relevant to Dy^{3+} , and luminescence due to recombination of the hole and electron at the Eu^{2+} site takes place via long lasting phosphorescence. We speculate that the former process is dominant because the emission does not last for a long period,

as indicated in Fig. 9. The intensity of green emission decays rapidly (within one second or so), as found in the photographs in Fig. 9. Another possible cause of TL is fracture- and/or deformation-induced thermoluminescence (53). The frictional heat generated by the fracture and/or deformation of a crystal stimulates an electron and trapped at a defect site in advance, and the recombination of the released electron and a hole trapped beforehand at the Eu^{2+} site (i.e., Eu^{3+}) yields the emission due to Eu^{2+} . It is commonly known that the europium ion can be in its trivalent state (Eu^{3+}) in many solids.

Note that another mechanism was proposed by Xu et al. (47) for TL based on the nondestructive deformation of $\text{SrAl}_2\text{O}_4:\text{Eu}^{2+}$. They argued that a positive hole trapped at a certain localized level in the energy gap is released by the movement of a dislocation which is caused by the deformation of the crystal, and that the recombination of the hole with an electron trapped at the Eu^{2+} site (i.e., Eu^+) gives rise to the emission of photons as TL. Nonetheless, the monovalent state of europium ion (Eu^+) is unusual from the chemical viewpoint, and the existence of Eu^+ is unclear, to date. Thus, the mechanism of TL in this material is controversial.

To clarify the influence of crystal fracture on the local structure and on the luminescent properties of doped rare-earth ions in alkaline earth aluminates, we have measured TL in Dy^{3+} -doped $(\text{Sr},\text{Ba})\text{Al}_2\text{O}_4$ and $(\text{Sr},\text{Ca})\text{Al}_2\text{O}_4$ polycrystals. Dy^{3+} was selected because the coordination state relevant to Dy^{3+} , including the coordination symmetry around Dy^{3+} and the electronic state of the chemical bond between Dy^{3+} and a ligand, can be readily deduced from the relative intensity of the emission lines assigned to the $4f-4f$ transitions of Dy^{3+} . Besides, one can alter the ligand field around Dy^{3+} systematically by using a solid solution such as $(\text{Sr},\text{Ba})\text{Al}_2\text{O}_4$ and $(\text{Sr},\text{Ca})\text{Al}_2\text{O}_4$ as the host material. It is known that broad composition ranges of solid solution are present and the compositional dependence of the lattice parameter manifests a monotonic variation in the $\text{SrAl}_2\text{O}_4-\text{BaAl}_2\text{O}_4$ and $\text{SrAl}_2\text{O}_4-\text{CaAl}_2\text{O}_4$ systems (54). According to X-ray diffraction analysis, the substitution of Sr by Ba in the $(\text{Sr},\text{Ba})\text{Al}_2\text{O}_4$ system does not lead to any change in the crystal structure when the amount of Ba that replaces Sr is less than 40 mol%. The structure of $\text{Sr}_{1-x}\text{Ba}_x\text{Al}_2\text{O}_4$ polycrystals prepared by the conventional solid-state reaction is mainly monoclinic ($\alpha\text{-SrAl}_2\text{O}_4$ structure) at least for $x < 0.4$, although the hexagonal phase (BaAl_2O_4 structure) exists mainly when the concentration of Ba^{2+} is larger than 40 mol%. On the other hand, for the $\text{SrAl}_2\text{O}_4-\text{CaAl}_2\text{O}_4$ system, a drastic change in the X-ray diffraction pattern, that is, crystal structure, is observed when the composition changes from $\text{Sr}_{0.9}\text{Ca}_{0.1}\text{Al}_2\text{O}_4$ to $\text{Sr}_{0.8}\text{Ca}_{0.2}\text{Al}_2\text{O}_4$, whereas the crystal structure of SrAl_2O_4 is very similar to that of $\text{Sr}_{0.9}\text{Ca}_{0.1}\text{Al}_2\text{O}_4$. The $\text{Sr}_{1-x}\text{Ca}_x\text{Al}_2\text{O}_4$ polycrystals take pseudohexagonal and monoclinic structures for $x > 0.2$ and $x < 0.2$, respectively.

TL spectra of $(\text{Sr},\text{Ba})\text{Al}_2\text{O}_4$ and $\text{Sr}_{0.9}\text{Ca}_{0.1}\text{Al}_2\text{O}_4$ doped with Dy^{3+} are shown in Fig. 11. Three emission lines are observed in all of the spectra. These lines are assigned to the $4f-4f$ transitions of Dy^{3+} ; the emission lines at around 480, 575, and 660 nm are attributable to the ${}^4\text{F}_{9/2}-{}^6\text{H}_{15/2}$, ${}^4\text{F}_{9/2}-{}^6\text{H}_{13/2}$, and ${}^4\text{F}_{9/2}-{}^6\text{H}_{11/2}$ transitions of Dy^{3+} ,

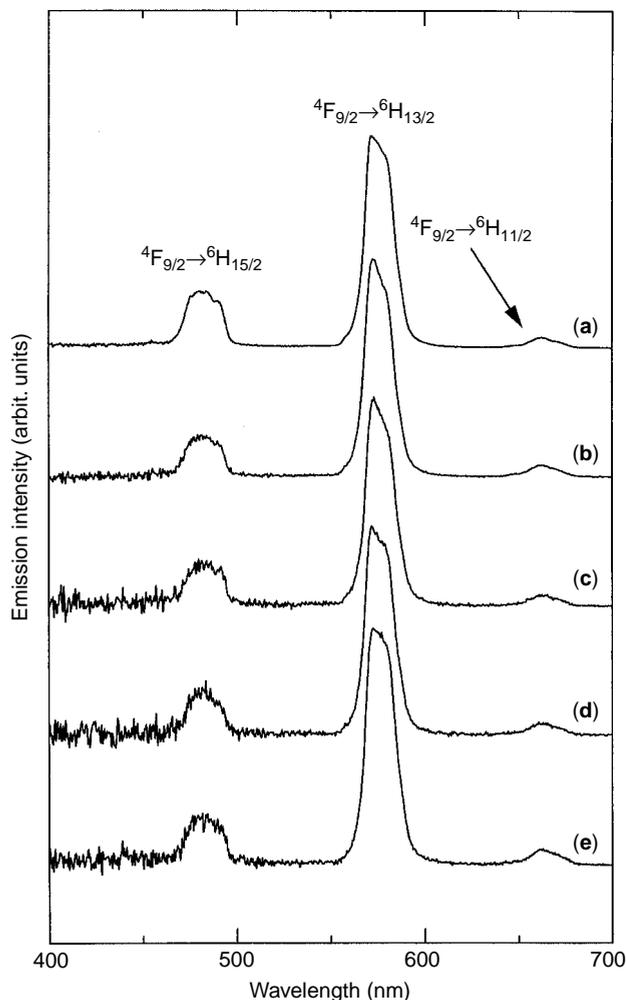


Figure 11. TL spectra of Dy³⁺-doped (a) Sr_{0.6}Ba_{0.4}Al₂O₄, (b) Sr_{0.8}Ba_{0.2}Al₂O₄, (c) Sr_{0.9}Ba_{0.1}Al₂O₄, (d) SrAl₂O₄, and (e) Sr_{0.9}Ca_{0.1}Al₂O₄. The assignment of emission lines is indicated in the figure.

respectively, as indicated in the figure. These samples exhibit PL as shown in Fig. 12. These spectra were obtained under excitation at 350 nm. All of the emission lines in the figure are attributed to the 4f–4f transitions of Dy³⁺, as indicated in the figure. Noted that TL was barely observed in Sr_{0.8}Ca_{0.2}Al₂O₄:Dy³⁺ and Sr_{0.6}Ca_{0.4}Al₂O₄:Dy³⁺, although these compounds exhibit PL under excitation at 350 nm. Presumably, this occurs because the crystal structure of Sr_{0.8}Ca_{0.2}Al₂O₄ and Sr_{0.6}Ca_{0.4}Al₂O₄, a pseudohexagonal structure, is different from those of the other compounds; the other crystalline phases take a monoclinic structure, as mentioned before.

Comparison of Figs. 11 and 12 suggests that the intensity ratio of the 480-nm emission (⁴F_{9/2}–⁶H_{15/2} transition) to the 575-nm emission (⁴F_{9/2}–⁶H_{13/2} transition) is almost independent of the composition for the TL spectra, whereas the relative intensity of the 480-nm emission to the 575-nm emission varies with composition for the PL spectra. Besides, the intensity ratio of 480-nm emission to 575-nm emission is smaller in the TL than in the PL spectra.

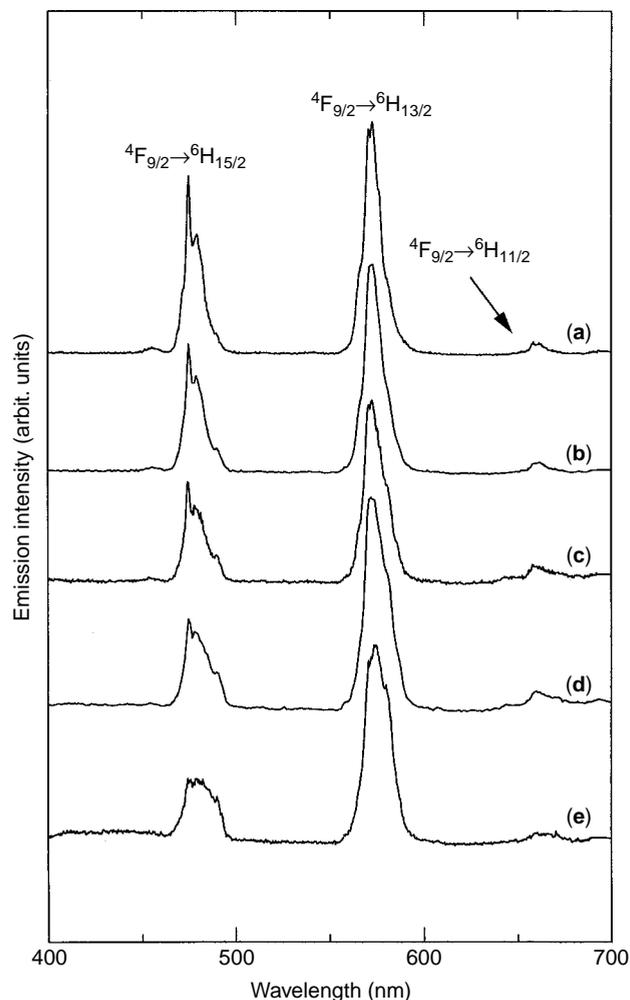


Figure 12. PL spectra obtained under excitation at 350 nm for Dy³⁺-doped (a) Sr_{0.6}Ba_{0.4}Al₂O₄, (b) Sr_{0.8}Ba_{0.2}Al₂O₄, (c) Sr_{0.9}Ba_{0.1}Al₂O₄, (d) SrAl₂O₄, and (e) Sr_{0.9}Ca_{0.1}Al₂O₄. The assignment of emission lines is indicated in the figure.

To visualize this relationship, the integrated intensity was evaluated for the 480-nm and 575-nm emissions, and the intensity ratio was plotted against the mean ionic radius of the alkaline earth in the crystal in Fig. 13. The open and closed circles correspond to TL and PL, respectively. The ratio $I_{(480\text{nm})}/I_{(575\text{nm})}$ increases as the mean ionic radius increases for PL, whereas the ratio is almost independent of the mean ionic radius for TL. The intensity ratio is also smaller for TL than for PL. Two possibilities are suggested to explain this phenomenon. One is the effect of self-absorption (55); the photons emitted from an optically active center such as rare-earth ions are reabsorbed by the other active center in TL. Because TL usually occurs within the bulk, photons are reabsorbed until they come out of the bulk. A comparison among TL, PL, and excitation spectra for the transition that corresponds to 480 nm is shown for SrAl₂O₄:Dy³⁺ in Fig. 14. The TL, PL, and excitation spectra are represented by closed circles, solid line, and open circles, respectively. An overlap of the PL and the excitation spectra is observed at around 475 nm. Although it is

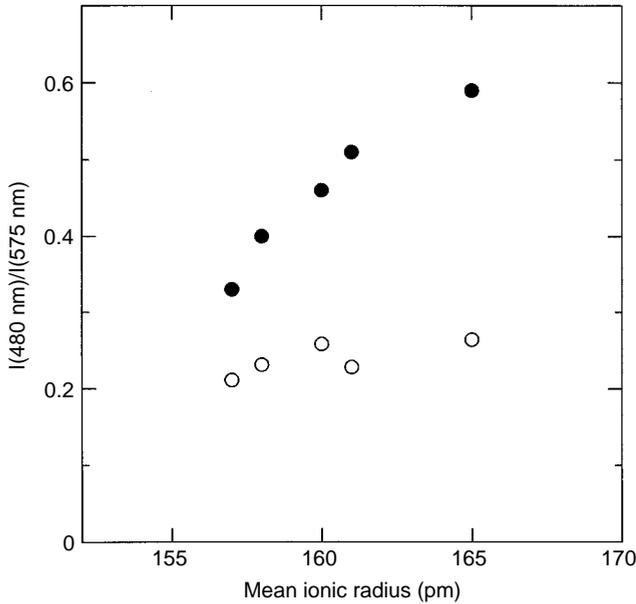


Figure 13. Variation of integrated emission intensity ratio, $I_{(480 \text{ nm})}/I_{(575 \text{ nm})}$ vs. the mean ionic radius of the alkaline earth in the crystals. Open and closed circles denote the TL and PL, respectively.

difficult to compare the integrated intensity of these emission and absorption lines quantitatively, the possibility of self-absorption cannot be ruled out.

Another possibility for the difference in relative emission intensity between TL and PL spectra involves the characteristics of the ligand field around Dy^{3+} . It is well known that the emission intensity for each electronic transition of a rare-earth ion can be approximately connected to ligand fields via the Judd–Ofelt theory (56,57). According to this theory, the radiative transition probability for

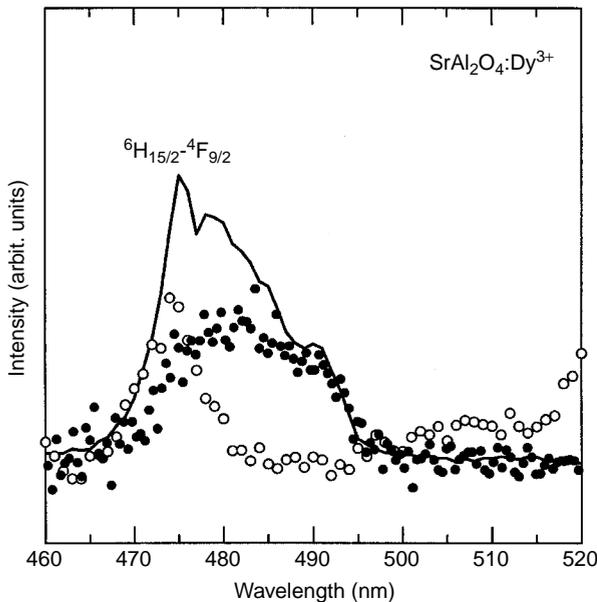


Figure 14. TL (closed circles), PL (solid line), and excitation (open circles) spectra at around 480 nm for $\text{SrAl}_2\text{O}_4:\text{Dy}^{3+}$.

the electric dipole transition A is expressed as

$$A = \left(\frac{64\pi^2 e^2}{3h(2J+1)} \right) \left(\frac{1}{\lambda^3} \right) \left(\frac{n(n^2+2)^2}{9} \right) S(aJ:bJ'), \quad (1)$$

where h is Planck's constant, J is the total angular momentum, n is the refractive index, e is the elementary electric charge, λ is the peak wavelength of the emission line, and $S(aJ:bJ')$ is the line strength for the transition between the $|a\rangle$ and $|b\rangle$ states. The line strength $S(aJ:bJ')$ is expressed by

$$S(aJ:bJ') = \sum_{t=2,4,6} \Omega_t |\langle aJ \| U^{(t)} \| bJ' \rangle|^2, \quad (2)$$

where $\langle aJ \| U^{(t)} \| bJ' \rangle$ is the reduced matrix element that can be calculated for each electronic transition of rare-earth ions. The reduced matrix element is independent of the difference in the environment in which the rare-earth ion is located. On the other hand, because the parameter Ω_t includes the distance between the rare-earth ion and the ligand, the charge of the ligands, and so forth, Ω_t varies depending on the kinds of ligand fields and the rare-earth ions. The line strength in Eq. (2) is proportional to the integrated intensity of the lines in the optical absorption and emission spectra, so that the Ω parameter which represents the characteristics of ligand fields is reflected by the integrated intensity of absorption and emission lines. It is thought that Ω_2 is relevant to the coordination symmetry of ligands and Ω_6 is an indicator of the covalency of the chemical bond between a rare-earth ion and a ligand (58,59). As for the Dy^{3+} , the integrated intensity ratio of 480-nm to 575-nm emissions correlates with Ω_6/Ω_2 . This suggests that the integrated intensity ratio becomes small when the coordination symmetry for Dy^{3+} is low (60). Hence, the variation of the relative integrated intensity for PL shown in Fig. 13 indicates that the ligand field for Dy^{3+} changes as the size of alkaline earth ion changes, as expected. Presumably, this variation is based on the average coordination structure around Dy^{3+} because Dy^{3+} ions can occupy at least two different positions at Sr^{2+} sites in the SrAl_2O_4 crystal (61). On the other hand, the fact that the relative integrated emission intensity is almost independent of the mean ionic radius of the alkaline earth in TL suggests that the average coordination symmetry for Dy^{3+} which gives rise to the TL does not change even though the kind of alkali-earth ion is varied. Furthermore, the smaller value of $I_{(480 \text{ nm})}/I_{(575 \text{ nm})}$ in TL spectra compared with PL spectra indicates that the average coordination symmetry for Dy^{3+} , which brings about the TL, is low. Consequently, it is suggested that the TL observed is the emission of photons from Dy^{3+} placed in a distorted site such as a fractured surface and/or the vicinity of a crack tip.

APPLICATIONS OF TL

Recent findings for rare-earth-doped materials that show intense TL have attracted attention because of their potential application for sensing structural damage, fracture, and deformation. We describe one example of sensing deformation demonstrated by Xu et al. (47,48) using the $\text{SrAl}_2\text{O}_4:\text{Eu}^{2+}$ polycrystal. They mixed the $\text{SrAl}_2\text{O}_4:\text{Eu}^{2+}$

powder with an optical epoxy resin to make a composite disk 25 mm in diameter and 15 mm thick. A compressive stress was applied in the direction of the diameter. As a result, an intense green emission was observed. Besides, the emission intensity was distributed within the disk, and the distribution of emission intensity fairly coincides with the distribution of deformation calculated by using a theory of elastics. Similarly, Xu et al. (62) showed that an artificial skin that consists of piezoluminescent film (ZnS) could act as a self-diagnosing material to detect mechanical stress remotely by visible light emission. Sage et al. (49) also proved the effectiveness of TL for sensing structural damage. They prepared composite materials doped with Eu, Tb, and Mn complexes that exhibit intense TL. They successfully showed that these materials are useful for real-time monitoring of the magnitude and the location of structural damage. These experiments demonstrate that the distribution of stress and damage within a solid can be visualized by using TL materials.

BIBLIOGRAPHY

- G. Wiedemann and F. Schmidt. *Ann. Phys. Chem. (Leipzig)*, **54**(4):604–625 (1895).
- F. Bacon. *Of the Advancement of Learning*. 1605, J.M. Dent and Sons, London, 1915.
- E.N. Harvey. *A History of Luminescence*. American Philosophical Society, Philadelphia, PA, 1957, Chap. 10.
- C. Kittel. *Introduction to Solid State Physics*. 6e., Wiley, NY, 1986, Chap. 13.
- B.P. Chandra. *Phys. Status Solidi A* **64**(1):395–405 (1981).
- A.J. Walton. *Adv. Phys.* **26**(6): 887–948 (1977).
- F.A. Cotton, D.M.L. Goodgame, and M. Goodgame. *J. Am. Chem. Soc.* **84**(20): 167–172 (1962).
- G.E. Hardy and J.I. Zink. *Inorg. Chem.* **15**(12): 3061–3065 (1976).
- C.R. Hurt, N. McAvoy, S. Bjorklund, and N. Filipescu. *Nature (London)*, **212**(5058): 179–180 (1966).
- G.E. Hardy, J.C. Baldwin, J.I. Zink, W.C. Kaska, P. Liu, and L. Dubois. *J. Am. Chem. Soc.* **99**(11): 3552–3558 (1977).
- J.I. Zink, G.E. Hardy, and J.E. Sutton. *J. Phys. Chem.* **80**(3): 248–249 (1976).
- J.I. Zink. *Inorg. Chem.* **14**(3): 555–558 (1975).
- J.I. Zink. *J. Am. Chem. Soc.* **96**(21): 6775–6777 (1974).
- G.E. Hardy, W.C. Kaska, B.P. Chandra, and J.I. Zink. *J. Am. Chem. Soc.* **103**(5): 1074–1079 (1981).
- J.I. Zink and W.C. Kaska. *J. Am. Chem. Soc.* **95**(22): 7510–7513 (1973).
- N. Kitamura, O. Saravari, H.-B. Kim, and S. Tazuke. *Chem. Phys. Lett.* **125**(4): 360–363 (1986).
- R. Nowak, A. Krajewska, and M. Samoc. *Chem. Phys. Lett.* **94**(3): 270–271 (1983).
- P.A. Thiessen and K. Meyer. *Naturwissenschaften* **57**(9): 423–427 (1970).
- G. Herzberg. *Molecular Spectra and Molecular Structure*. Vol. 1, Spectra of Diatomic Molecules. 2e., Van Nostrand, Princeton, NJ, 1950.
- J. Wollbrandt, U. Bruckner, and E. Linke. *Phys. Status Solidi A* **77**(2): 545–552 (1983).
- L.M. Belyaev and Y.N. Martyshev. *Phys. Status Solidi* **34**(1): 57–62 (1969).
- J.T. Dickinson, L.C. Jensen, and A. Jahan-Latibari. *J. Vac. Sci. Technol. A* **2**(2): 1112–1116 (1984).
- B.P. Chandra, M.S.K. Khokhar, M. Elyas, and S.M.D. Rao. *J. Phys. D Appl. Phys.* **23**(1): 90–94 (1990).
- B.P. Chandra, S. Tiwari, M. Rammrakhiani, and M.H. Ansari. *Cryst. Res. Technol.* **26**(6): 767–781 (1991).
- S.C. Langford, J.T. Dickinson, and L.C. Jensen. *J. Appl. Phys.* **62**(4): 1437–1449 (1987).
- Y. Enomoto and H. Hashimoto. *Nature (London)*, **346**(6285): 641–643 (1990).
- K. Nakayama, N. Suzuki, and H. Hashimoto. *J. Phys. D Appl. Phys.* **25**(2): 303–308 (1992).
- J. Wollbrandt, E. Linke, and K. Meyer. *Phys. Status Solidi A* **27**(2): K53–K55 (1975).
- K. Nakamura and Y. Tsuneishi. *Bull. Earthquake Res. Inst. (Tokyo Univ.)* **44**: 1371–1384 (1966).
- T. Rikitake, Y. Yamazaki, Y. Hagiwara, K. Kawada, M. Sawada, Y. Sasai, T. Watanabe, K. Momose, T. Yoshino, K. Otani, K. Ozawa, and Y. Sanzai. *Bull. Earthquake Res. Inst. (Tokyo Univ.)* **44**: 363–408 (1966).
- B.T. Brady and G.A. Rowell. *Nature (London)*, **321**(6069): 488–492 (1986).
- J.I. Zink, W. Beese, J.W. Schindler, and A.J. Smiel. *Appl. Phys. Lett.* **40**(2): 110–114 (1982).
- J.T. Dickinson, S.C. Langford, L.C. Jensen, G.L. McVay, J.F. Kelso, and C.G. Pantano. *J. Vac. Sci. Technol. A* **6**(3): 1084–1089 (1988).
- K. Kawaguchi. *Phys. Rev. B* **52**(13): 9224–9228 (1995).
- B. Kasemo and L. Walden. *Solid State Commun.* **15**(3): 571–574 (1974).
- B. Kasemo. *Phys. Rev. Lett.* **32**(20): 1114–1117 (1974).
- F.G. Wick. *J. Opt. Soc. Am.* **29**(10): 407–412 (1939).
- G.N. Chapman and A.J. Walton. *J. Appl. Phys.* **54**(10): 5961–5965 (1983).
- S.H. Lin, D. Wutz, Z.Z. Ho, and H. Eyring. *Proc. Natl. Acad. Sci. USA* **77**(3): 1245–1247 (1980).
- D. Haneman and N. McAlpine. *Phys. Rev. Lett.* **66**(6): 758–761 (1991).
- D.G. Li, N.S. McAlpine, and D. Haneman. *Surf. Sci.* **303**(1–2): 171–178 (1994).
- G.N. Chapman and A.J. Walton. *J. Phys. C Solid State Phys.* **16**(28): 5543–5551 (1983).
- T. Ishihara, K. Tanaka, K. Hirao, and N. Soga. *Jpn. J. Appl. Phys.* **36**(6B): L781–L783 (1997).
- T. Ishihara, K. Tanaka, K. Fujita, K. Hirao, and N. Soga. *Solid State Commun.* **107**(12): 763–767 (1998).
- K. Tanaka, T. Ishihara, K. Fujita, and K. Hirao. *Mater. Res. Soc. Symp. Proc.* **604**: 323–328 (2000).
- M. Akiyama, C.-N. Xu, K. Nonaka, and T. Watanabe. *Appl. Phys. Lett.* **73**(21): 3046–3048 (1998).
- C.-N. Xu, T. Watanabe, M. Akiyama, and X.-G. Zheng. *Appl. Phys. Lett.* **74**(17): 2414–2416 (1999).
- C.-N. Xu, X.-G. Zheng, M. Akiyama, K. Nonaka, and T. Watanabe. *Appl. Phys. Lett.* **76**(2): 179–181 (2000).
- I. Sage, R. Badcock, L. Humberstone, N. Geddes, M. Kemp, and G. Bourhill. *J. Smart Mater. Struct.* **8**(4): 504–510 (1999).
- K. Nakayama. *Seidenki-Gakkai-Shi* **15**(6): 421–425 (1991) (in Japanese).
- T. Matsuzawa, Y. Aoki, N. Takeuchi, and Y. Murayama. *J. Electrochem. Soc.* **143**(8): 2670–2673 (1996).
- H. Takasaki, S. Tanabe, and T. Hanada. *J. Ceram. Soc. Jpn.* **104**(4): 322–326 (1996) (in Japanese).

53. T. Ishihara, unpublished data.
54. S. Ito, S. Banno, K. Suzuki, and M. Inagaki. *J. Ceram. Soc. Jpn.* **87**(7): 344–349 (1979).
55. J.P. Duignan, I.D.H. Oswald, I.C. Sage, L.M. Sweeting, K. Tanaka, T. Ishihara, K. Hirao, and G. Bourhill, (in press).
56. B.R. Judd. *Phys. Rev.* **127**(3): 750–761 (1962).
57. G.S. Ofelt. *J. Chem. Phys.* **37**(3): 511–520 (1962).
58. S. Tanabe, T. Ohyagi, N. Soga, and T. Hanada. *Phys. Rev. B* **46**(6): 3305–3310 (1992).
59. S. Tanabe, T. Ohyagi, S. Todoroki, T. Hanada, and N. Soga. *J. Appl. Phys.* **73**(12): 8451–8454 (1993).
60. S. Tanabe, T. Hanada, M. Watanabe, T. Hayashi, and N. Soga. *J. Am. Ceram. Soc.* **78**(11): 2917–2922 (1995).
61. V.A.-R. Schulze and H. Müller-Buschbaum. *Z. Anorg. Allg. Chem.* **475**: 205–210 (1981).
62. C.N. Xu, T. Watanabe, M. Akiyama, and X.G. Zheng. *Appl. Phys. Lett.* **74**(9): 1236–1238 (1999).

TRUSS STRUCTURES WITH PIEZOELECTRIC ACTUATORS AND SENSORS

RAFAEL BRAVO

Universidad del Zulia
Maracaibo, Venezuela

ANTHONY FARIA VAZ

Applied Computing Enterprises Inc.
Mississauga, Ontario, Canada
&

University of Waterloo
Waterloo, Ontario, Canada

MOHAMED DOKAINISH

McMaster University
Hamilton, Ontario, Canada

INTRODUCTION

Structural systems that have a built-in capability to sense external stimulus and respond to it depending on a predetermined criteria are commonly referred to as smart structures. An example of a smart structure is a flexible satellite antenna that can sense its own vibration and apply corrective action to dampen it out. A smart structure consists of four major components: the structure itself, sensors, actuators, and the control system. Smart structures can achieve performance objective that cannot be achieved by a traditional passive structure. Through the use of an active control system, a lightweight structural member can replicate the vibration suppression characteristics of a heavy structural member. To be effective, the design of the structure and the achievable control performance must be considered in an integrated manner. This approach enables the trade-offs among structural weight, geometry, and control performance to be optimized to achieve a specified operational characteristic.

Large flexible space structures (LFSS) have lightly damped vibrational modes that range from very low to high frequencies. The degradation to mission performance caused by structural flexibility must be eliminated. Conflicting requirements of structural stiffness and lightweight must be balanced by the structural engineer.

There are two possible ways of dealing with flexibility in a structure: use stiffer structural members, which increase weight, or augment the inherent damping in the structure. Due to weight restrictions, only the second approach is feasible. Viscoelastic materials can be used to passively augment the damping of the high-frequency modes. Unfortunately, these materials do not provide significant damping for low-frequency modes. As it is the low-frequency modes that most severely impact structural performance, active damping is required.

Active damping techniques make use of sensors and actuators to measure structural deflection and perform corrective action. This minimizes the maximum deflection amplitude and the settling time of the undesired oscillations. These techniques require a knowledge of the dynamic behavior of a structure. Complex structures are modeled mathematically using analytical techniques based on finite element analysis, or experimental model identification based on modal testing. In our research, we use analytical techniques to develop a dynamic model. The model is then validated by empirical modal tests.

One issue that arises, when modeling a dynamic system for control purposes, is the accuracy of the model behavior, compared to that of the actual structure. The dynamics of a LFSS are theoretically characterized by infinite-order eigenfunction expansions. In practice, some form of model reduction is used to develop a finite-order state-space model. The state space model is used for both controller design and simulation. The modeling error of dynamics takes two forms: neglected high-order modes and inaccurately modeled low-order modes. The errors associated with the low-order modes are due to uncertainty in the knowledge of mechanical parameters, and simplifications taken during the modeling process. The analytical modeling approaches assume an a priori damping factor of zero. Empirical tests must be used to determine the correct damping factors. Empirical testing of the LFSS in a terrestrial laboratory is problematic, due to the effects of gravity and the atmosphere, as the structures are intended for use in the zero gravity and vacuum of space. Accordingly, the controller design methodology must be able to cope with the inherent uncertainty in the model of the structural dynamics.

Active Vibration Control of Flexible Structures

The linear–quadratic–Gaussian control method, or LQG control, was developed in the 1960s and is capable of obtaining controllers for multiple input–multiple output systems such as the ones in consideration in this article. LQG controllers give good response when model dynamics are known exactly. However, LQG controllers do not necessarily cope well with model uncertainty (1). Furthermore, the infinite bandwidth of LQG controllers can excite unmodeled higher-order dynamics. This effect is termed the “spillover problem” (2). Allen et al. (3) applied the LQG method directly to the problem of vibration suppression in a truss structure. They employed a frequency-weighted LQG method in which the system dynamics are augmented with bandpass filters for the input and output signals. This compensated for the traditional deficiencies of LQG controllers: the bandpass filters restricted the control bandwidth to frequencies over which the structural

dynamics were well modeled. They demonstrated the limited bandwidth and robustness characteristics of their controller design on an experimental truss structure.

The H_∞ control design method (4–6) has recently been applied to the problem of control of large flexible space structures. This method enables the design of a controller that robustly stabilizes a system in the presence of bounded structured and unstructured uncertainties. Furthermore, the control design can also incorporate performance requirements that must be met in the presence of uncertainties. Current research is focused on the manner in which uncertainties and performance can be specified to obtain a useful controller design.

Buddie et al. (7) used a “weighted gap optimization” approach for the synthesis of the H_∞ controller, employing input and output weighting functions to “shape” the open loop response of the system. Experiments performed on both a flower-shaped structure and a truss structure successfully achieved an increased damping for particular modes while maintaining a specified level of robustness. Their experiments made use of momentum exchange actuators that were operated by linear dc motors. The accelerometer sensors were collocated at the actuator locations, and their signals are integrated to provide velocity.

A good survey on the subject of vibration problems associated with flexible spacecraft is provided by van Woerkom (8). In particular, he discusses the OGO series of satellites and the Hubble space telescope. He gives an overview of the various control techniques that were applied to stabilize these satellites.

Piezoelectric Materials in Smart Structures

The use of piezoelectric materials as sensors and actuators in flexible systems has been a topic of research since the mid-1980s. Hubbard and his collaborators (9,10) proposed the application of piezoelectric polymer (PVDF) as actuator for the active damping of beams with different boundary conditions. The sensors employed were tip-mounted linear and angular accelerometers for the case of cantilever beams and a base-mounted angular accelerometer for the case of the simply supported beam. They used several linear and nonlinear velocity–feedback control techniques to artificially increase the modal damping ratios of the flexible beams. These control laws were applied to rectangular and spatially varying distributed PVDF films, with voltage levels of up to ± 250 V. A significant level of damping increase was attained, particularly when using nonlinear on–off velocity–feedback control.

Crawley and de Luis (11) developed a quasi-static analysis procedure for modeling the interaction between rectangular piezoelectric materials and the structure to which they are attached. They considered the cases of

surface-bonded and embedded actuators, in bending and extension. Their predictions were compared to experiments performed on cantilevered beams of various materials and actuator placement. Lee and Moon (12) analyzed the effect of etching and splicing to shape the electrode of a piezoelectric film. Lee and Moon constructed a modal sensor that was able to observe a single mode of a flexible plate. Miller and his co-workers (13,14) showed that spatial filters could also be realized in terms of etched piezoelectric films. A state feedback control law can be implemented by etching piezoelectric films; this has the potential of simplifying the hardware required for control. Vaz (15–18) used the quasi-static analysis procedure of Crawley and de Luis to develop interaction equations for modeling the behavior of shaped piezoelectric sensors and actuators. The interaction equations were validated through experimental testing; (19–21). The interaction equations were generalized into a finite element framework in (7,22,23).

TRUSS STRUCTURE CONFIGURATION

A truss structure was designed to be a testbed for evaluating potential control techniques for LFSS using piezoelectric actuators and sensors. The design criteria for the selection of the final configuration were the following:

1. Have similarity to structures used in flexible structure literature.
2. Include at least 10 modes under 100 Hz.
3. Not include local bending of truss members in first 5 modes of vibration.
4. Meet cost, weight, and actuation power requirements.

The structure was constructed from polycarbonate tubing with 12.7 mm outer diameter, connected at the nodes by joining blocks. The polycarbonate material was selected because of its advantages over metals such as aluminum and steel in this application. The main selection criteria were cost, machinability, weight, strength, electrical properties, ease of assembly, and bonding of piezoelectric materials to the structure. Table 1 shows the mechanical properties and the dimensions of the polycarbonate tubes used in the structure, including the parameters needed for the finite element simulation. The configuration of the structure is shown in Fig. 1. The truss members represented by the dark lines undergo the most strain during vibration of the truss structure. These truss members are ideally suited to sensor and actuator placement.

The truss structure was vertically cantilevered at the bottom, and it was composed of six vertical bays, each a 300-mm cube. The thick lines represent the active truss

Table 1. Dimensions and Mechanical Properties of Polycarbonate Tubing

Material	Outside Diameter (mm)	Thickness (mm)	Young's Modulus (N/m ²)	Density (kg/m ³)
Polycarbonate	12.7	1.5875	2.35×10^9	1.2×10^3

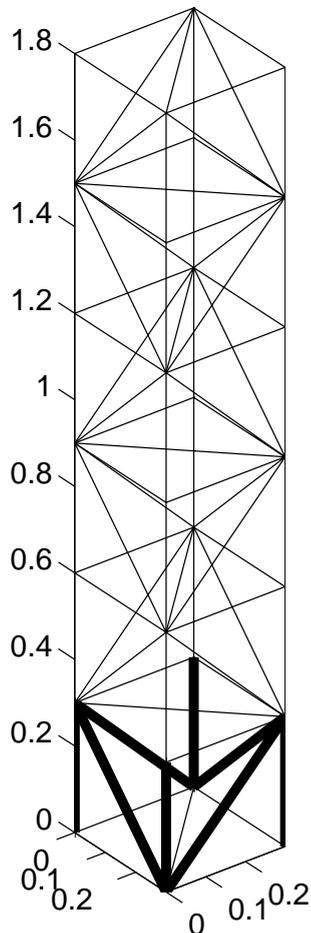


Figure 1. Configuration of the truss structure.

members, which are fitted with sensors and actuators. Not all the active truss members are to be used in the control design simultaneously. The design was iterated until the desired design criteria were met. The structure was approximately 1800 mm high. Figure 2 gives a detailed view of two bays of the structure, showing more clearly the topology of the truss. At each junction there is a joining block, also made of polycarbonate material.

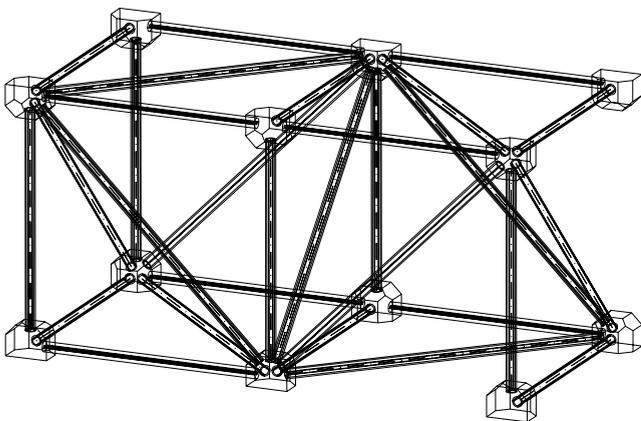


Figure 2. Detail of two bays.

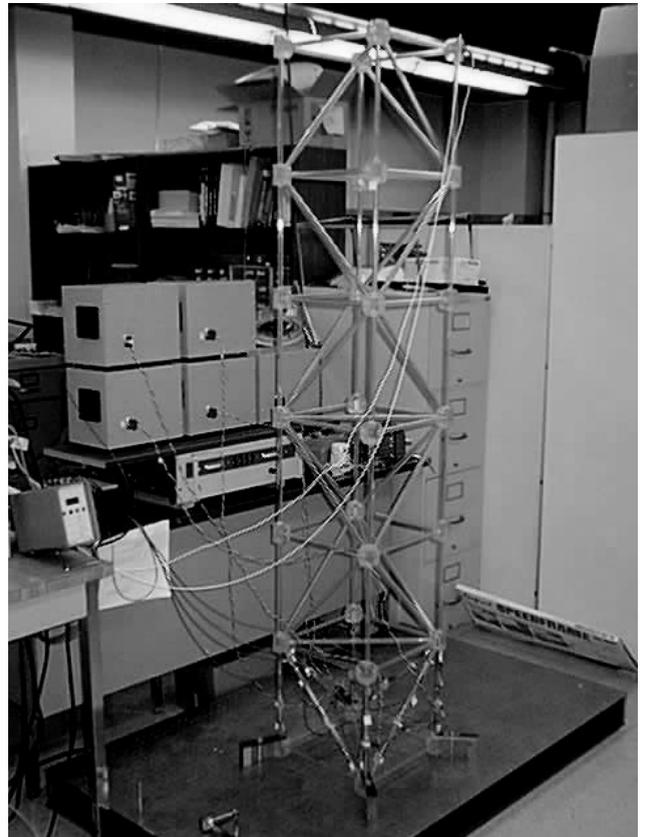


Figure 3. Experimental truss structure.

The experimental truss structure is shown in Fig. 3. The truss members are bonded to the junction blocks using a liquid plastic solvent containing methylene chloride. The associated electronics are shown in the background. The five gray boxes contain the custom designed high-voltage amplifiers. The amplifiers are powered by an HP 6521A dc power supply. Accelerometers are placed at the top of the structure. The lower portion of the truss structure is shown in closeup in Fig. 4. The active truss members are the lower truss members that have bonded piezoelectric films.

The active truss members are fitted with two identical polyvinylidene fluoride (PVDF) rectangular piezoelectric sensors. One sensor measures the open circuit voltage, which is proportional to the strain. The second measures the short circuit current, which is proportional to the strain rate. In the midspan of the active truss member, there is a piezoceramic actuator, which receives the control signal and transforms it into actuation forces.

FINITE ELEMENT AND MODAL ANALYSIS OF THE TRUSS STRUCTURE

In the finite element method, a continuous structure is discretized into a finite number of elements (24). This converts the dynamics of the continuous structure, with infinite number of degrees of freedom, to a finite-dimensional representation. The discretized structure has dynamics that are characterized by a system of ordinary differential

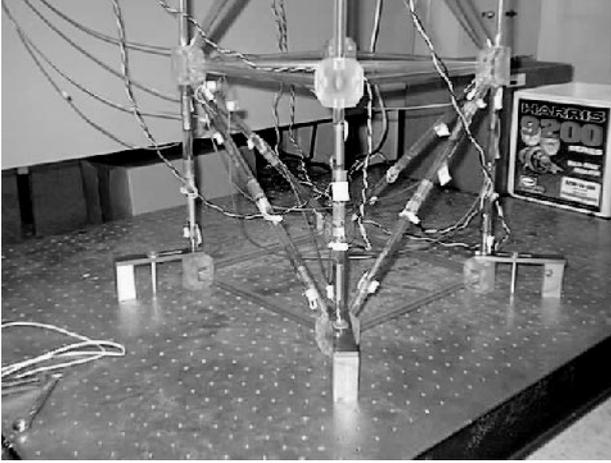


Figure 4. Active truss members.

equations, whereas the continuous structure dynamics are characterized by a system of partial differential equations.

Assembly of Finite Element Model

The truss members are modeled by an interconnected set of rod elements. A rod element has two nodes with one degree of freedom per node that corresponds to the displacement in the axial direction. Figure 5 shows a diagram of the rod element. The displacement varies linearly along the rod element. The nodes are labeled with a local numbering scheme $\mathcal{L}_{\text{NODE}} = \{1, 2\}$. The axial displacement of at local node i is denoted by ξ_i for $i = 1, 2$.

The mesh used for the finite element model with rod elements is shown in Fig. 6. The nodes are placed at the junctions, and are indicated by star symbols. The cross section of the rod elements is indicated by the ovals at midspan. Using rod elements is equivalent to assuming that the structure is pin-jointed, and thus the strain profile along each element is constant. For this reason, increasing mesh density, by partitioning the truss members with additional rod elements, would not alter the strain profile. Concentrated mass elements are added at each junction to simulate the effect of the extra weight of the junction blocks, and are denoted in by small squares coincident with the nodes. The finite element model has 83 rod elements and 28 nodes. Each node has 3 degrees of freedom; hence the structure has a total of 84 degrees of freedom. The structure is fixed at the base; this restricts 3 degrees of freedom at each of the four base nodes, for a total of 12 constrained degrees of freedom.

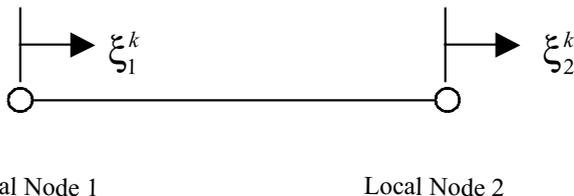


Figure 5. Rod element k .

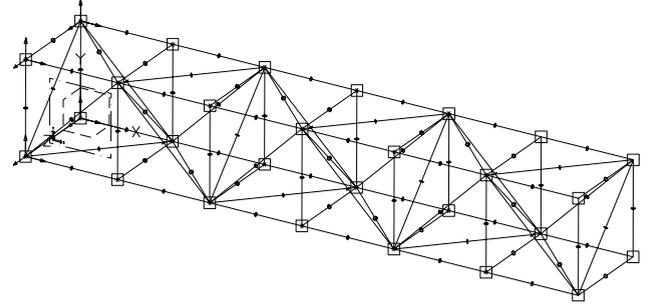


Figure 6. Finite element mesh for the rod element model.

The nodes are labeled with a global numbering scheme $\mathcal{G}_{\text{NODE}} = \{1, 2, \dots, n\}$, where $n = 28$. Each truss member is associated with a unique rod element. The rod elements are labeled with a global numbering scheme $\mathcal{G}_{\text{ELEM}} = \{1, 2, \dots, m\}$, where $m = 83$. The function G_{NODE} maps a global element number and a local node number to global node number; that is, $G_{\text{NODE}}: \mathcal{G}_{\text{ELEM}} \times \mathcal{L}_{\text{NODE}} \rightarrow \mathcal{G}_{\text{NODE}}$.

Each node has three coordinates that define its *undeformed* position with respect to a global coordinate frame. We let p_1^α , p_2^α , and p_3^α respectively denote the x , y , and z position coordinates of node α , $\alpha \in \mathcal{G}_{\text{NODE}}$. The 123 coordinate frame is an alternate representation of the x - y - z coordinate frame: 1 corresponds to x , 2 to y , and 3 to z . We consider a rod element k with nodes α and β , where $k \in \mathcal{G}_{\text{ELEM}}$, and $\alpha, \beta \in \mathcal{G}_{\text{NODE}}$. The projection of the length of rod element k on axis i of the global the coordinate coordinate frame is given by

$$\Delta p_i^k = |p_i^\alpha - p_i^\beta|, \quad i = 1, 2, 3. \quad (1)$$

The length of rod element k , denoted by L^k , is given by

$$L^k = \sqrt{(\Delta p_1^k)^2 + (\Delta p_2^k)^2 + (\Delta p_3^k)^2}. \quad (2)$$

The direction cosines for rod element k are computed as

$$\cos \theta_i^k = \frac{\Delta p_i^k}{L^k}, \quad i = 1, 2, 3. \quad (3)$$

When the truss structure is subjected to forces, the rod elements deform and the nodes move from their rest position. To illustrate this, we let q_1^α , q_2^α , and q_3^α , respectively, denote the x , y , and z coordinates of the displacement of the *deformed* position of node α , $\alpha \in \mathcal{G}_{\text{NODE}}$, from its *undeformed* position. We consider a rod element k with nodes α and β , where $k \in \mathcal{G}_{\text{ELEM}}$, and $\alpha, \beta \in \mathcal{G}_{\text{NODE}}$. The components of the deformation of rod element k , in the global coordinate frame, can be found by subtracting the node displacements as follows:

$$\Delta q_i^k = (q_i^\beta - q_i^\alpha) \operatorname{sgn} (p_i^\beta - p_i^\alpha), \quad i = 1, 2, 3. \quad (4)$$

The component of deformation Δq_i^k corresponds to an elongation if $\Delta q_i^k > 0$, and a compression if $\Delta q_i^k < 0$. The total

deformation $\Delta q_{\text{TOTAL}}^k$ of the rod element k is given by

$$\Delta q_{\text{TOTAL}}^k = \Delta q_1^k \cos \theta_1^k + \Delta q_2^k \cos \theta_2^k + \Delta q_3^k \cos \theta_3^k, \quad (5)$$

where $\cos \theta_i^k$ is the direction cosine of element k and axis i . The strain of element k , denoted by ε^k , is given by

$$\varepsilon^k = \frac{\Delta q_{\text{TOTAL}}^k}{L^k}. \quad (6)$$

Now, let q^α denote the displacement vector at node α , $\alpha \in \mathcal{G}_{\text{NODE}}$. The composite displacement vector q comprises the displacement vectors at each node of the structure. These vectors are defined as

$$q = \begin{bmatrix} q^1 \\ \vdots \\ q^n \end{bmatrix}, \quad q^\alpha = \begin{bmatrix} q_1^\alpha \\ q_2^\alpha \\ q_3^\alpha \end{bmatrix}. \quad (7)$$

Let Δq^k denote the deformation vector of element k , $k \in \mathcal{G}_{\text{ELEM}}$. The composite deformation vector Δq comprises the deformation vectors at each element of the structure. These vectors are defined as

$$\Delta q = \begin{bmatrix} \Delta q^1 \\ \vdots \\ \Delta q^m \end{bmatrix}, \quad \Delta q^k = \begin{bmatrix} \Delta q_1^k \\ \Delta q_2^k \\ \Delta q_3^k \end{bmatrix}. \quad (8)$$

The composite total deformation vector Δq_{TOTAL} comprises the total deformations of each element of the structure. In particular, Δq_{TOTAL} is defined as

$$\Delta q_{\text{TOTAL}} = \begin{bmatrix} \Delta q_{\text{TOTAL}}^1 \\ \vdots \\ \Delta q_{\text{TOTAL}}^m \end{bmatrix}. \quad (9)$$

The composite strain vector ε comprises the strains of each element of the structure. In particular, ε is defined as

$$\varepsilon = \begin{bmatrix} \varepsilon^1 \\ \vdots \\ \varepsilon^m \end{bmatrix}. \quad (10)$$

The $3m \times 3n$ matrix S is constructed so that

$$\Delta q = S q. \quad (11)$$

In particular, the ab element of S , denoted by $(S)_{ab}$, is given by

$$(S)_{ab} = \begin{cases} \text{sgn}(p_i^{\text{GNODE}(k,1)} - p_i^{\text{GNODE}(k,2)}) & \text{if some } i \in \{1, 2, 3\} \text{ and } k \in \mathcal{G}_{\text{ELEM}} \text{ satisfies} \\ & a = 3(k-1) + i \text{ and } b = 3[\text{GNODE}(k,1) - 1] + i, \\ \text{sgn}(p_i^{\text{GNODE}(k,2)} - p_i^{\text{GNODE}(k,1)}) & \text{if some } i \in \{1, 2, 3\} \text{ and } k \in \mathcal{G}_{\text{ELEM}} \text{ satisfies} \\ & a = 3(k-1) + i \text{ and } b = 3[\text{GNODE}(k,2) - 1] + i, \\ 0 & \text{otherwise,} \end{cases}$$

where $1 \leq a \leq 3m$ and $1 \leq b \leq 3n$.

The $m \times 3m$ block diagonal matrix R is constructed so that

$$\Delta q_{\text{TOTAL}} = R \Delta q. \quad (12)$$

In particular, R is given by

$$R = \text{diag} \left[(\cos \theta_1^1, \cos \theta_2^1, \cos \theta_3^1), \dots, (\cos \theta_1^m, \cos \theta_2^m, \cos \theta_3^m) \right]. \quad (13)$$

The $m \times m$ matrix N is constructed so that

$$\varepsilon = N \Delta q_{\text{TOTAL}}. \quad (14)$$

In particular, matrix N is given by

$$N = \text{diag} \left[\frac{1}{L^1}, \frac{1}{L^2}, \dots, \frac{1}{L^m} \right]. \quad (15)$$

Consider further an element k with local displacements ξ_1^k and ξ_2^k as shown on Fig. 5. The local displacements can be determined from the displacement vector q . In particular,

$$\xi_1^k = \cos \theta_1^k q_1^{\text{GNODE}(k,1)} + \cos \theta_2^k q_2^{\text{GNODE}(k,1)} + \cos \theta_3^k q_3^{\text{GNODE}(k,1)}, \quad (16)$$

$$\xi_2^k = \cos \theta_1^k q_1^{\text{GNODE}(k,2)} + \cos \theta_2^k q_2^{\text{GNODE}(k,2)} + \cos \theta_3^k q_3^{\text{GNODE}(k,2)}.$$

In accordance with the above expressions, a $2 \times 3n$ matrix G_e^k is constructed so the following relationship holds.

$$\begin{bmatrix} \xi_1^k \\ \xi_2^k \end{bmatrix} = G_e^k q. \quad (17)$$

In particular, the ab element of G_e^k , denoted by $(G_e^k)_{ab}$, is given by

$$(G_e^k)_{ab} = \begin{cases} \cos \theta_i^k & \text{if some } i \in \{1, 2, 3\} \text{ and } k \in \mathcal{G}_{\text{ELEM}} \text{ satisfies} \\ & a = 1 \text{ and } b = 3[\text{GNODE}(k,1) - 1] + i, \\ & \text{or} \\ & a = 2 \text{ and } b = 3[\text{GNODE}(k,2) - 1] + i, \\ 0 & \text{otherwise,} \end{cases}$$

The following mass and stiffness matrices are obtained from finite element modeling (24): the rod element k has a consistent mass matrix M_e^k , and a consistent stiffness

matrix K_e^k given by

$$M_e^k = \frac{\rho^k A_{CS}^k L^k}{6} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \quad \text{and} \quad K_e^k = \frac{E^k A_{CS}^k}{L^k} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}. \quad (18)$$

Element k has the parameters density ρ^k , cross-sectional area A_{CS}^k , Young's modulus of elasticity E^k , and length L^k . The mass matrix of the complete structure, M , is given by

$$M = \sum_{k=1}^m (G_e^k)^T M_e^k G_e^k. \quad (19)$$

The matrix M is positive definite. To incorporate the effect of the junction block masses, concentrated mass elements are added at the nodes. The stiffness matrix of the complete structure, K , is given by

$$K = \sum_{k=1}^m (G_e^k)^T K_e^k G_e^k. \quad (20)$$

The matrix K is positive semidefinite.

The result of a finite element discretization of a structure is a system of ordinary differential equations, which are called the *free system model*:

$$M\ddot{q} + Kq = f, \quad (21)$$

where q is the displacement vector and f is a vector of actuation forces applied to the structure. The vector f is defined in a later section where the operation of piezoelectric film actuators is explained. Equation (21) does not incorporate the boundary conditions associated with the structure. Accordingly, Eq. (21) has $n = 84$ degrees of freedom.

In the structure shown in Fig. 1, the bottom four nodes are clamped to the support base. Hence the displacements at nodes 1, 2, 3, and 4 are zero; that is, $q_i^\alpha = 0$ for $i = 1, 2, 3$ and $\alpha = 1, 2, 3, 4$. Hence there are 12 restricted degrees of freedom. There are $\bar{n} = 72$ constrained degrees of freedom. The constrained displacement vector \bar{q} only includes the components q_i^α , $i = 1, 2, 3$, and $\alpha \in \mathcal{G}_{\text{NODE}}$, that are not forced to be zero. The $\bar{n} \times n$ matrix P maps q to \bar{q} by $\bar{q} = Pq$. For our truss structure, the corresponding P matrix is given by

$$P = [0_{\bar{n} \times 12}, I_{\bar{n} \times \bar{n}}]. \quad (22)$$

Note that the following relationship also holds: $q = P^T \bar{q}$. The constrained mass matrix \bar{M} and the constrained stiffness matrix \bar{K} are given by

$$\bar{M} = PMP^T \quad \text{and} \quad \bar{K} = PKP^T. \quad (23)$$

The matrices \bar{M} and \bar{K} can be from obtained from M and K , respectively, by deleting the rows and columns associated with the clamped nodes. The *constrained system model* is

$$\bar{M}\ddot{\bar{q}} + \bar{K}\bar{q} = \bar{f}, \quad (24)$$

where $\bar{f} = Pf$.

Modal Analysis

The constrained system model, Eq. (24), is used to model the dynamics of the structure. The vibrational mode shapes and their frequencies are obtained from an eigenvalue analysis of Eq. (24). Since \bar{M} is positive definite, the following eigenvalue problem has \bar{n} solutions:

$$\omega_j^2 \bar{M} \hat{v}_j + \bar{K} \hat{v}_j = 0 \quad \text{for } j = \{1, 2, \dots, \bar{n}\}. \quad (25)$$

Order the eigenvalues ω_j^2 and eigenvectors \hat{v}_j so that $\omega_j^2 \leq \omega_{j+1}^2$. Construct the modal matrix Φ so that $\Phi = [\hat{v}_1, \dots, \hat{v}_{\bar{n}}]$. Construct the modal frequency matrix Ω so that $\Omega = \text{diag}[\omega_1^2, \dots, \omega_{\bar{n}}^2]$. The vectors v_j , $j = \{1, 2, \dots, \bar{n}\}$, are normalized according to the following relationships:

$$\begin{aligned} \Phi^T \bar{M} \Phi &= I, \\ \Phi^T \bar{K} \Phi &= \Omega. \end{aligned} \quad (26)$$

The modal matrix Φ defines a coordinate transformation from the modal coordinate vector η to the physical coordinate vector \bar{q} , that is,

$$\bar{q} = \Phi \eta. \quad (27)$$

Substituting this in the system model and premultiplying by Φ^{-1} , the system model yields the *modal model* of the system:

$$\Phi^T \bar{M} \Phi \ddot{\eta} + \Phi^T \bar{K} \Phi \eta = \Phi^T \bar{f},$$

which is equivalent to

$$\ddot{\eta} + \Omega \eta = \Phi^T \bar{f}. \quad (28)$$

To account for structural damping, a diagonal modal damping matrix H can be added.

$$\ddot{\eta} + H\dot{\eta} + \Omega \eta = \Phi^T \bar{f}. \quad (29)$$

The matrix H has the following form

$$H = \text{diag}[2\zeta_1\omega_1, 2\zeta_2\omega_2, \dots, 2\zeta_{\bar{n}}\omega_{\bar{n}}], \quad (30)$$

where ζ_j is the damping factor for mode j , $0 \leq \zeta_j \leq 1$, $j = \{1, 2, \dots, \bar{n}\}$. The form of the model in equation (30) has the advantage that it is uncoupled, since the matrices Ω and H are both diagonal. The damping factors ζ_j , $j = \{1, 2, \dots, \bar{n}\}$, can be determined from either material properties or experimentation.

The finite element modeling program I-DEAS (25) is used to analyze the truss structure. The program allows a structure to be defined in a graphical manner. The program constructs a representation of Eq. (24) once the material properties and boundary conditions are defined. The modes of the structure are obtained by solving Eq. (25) to obtain the mode shapes \hat{v}_j and frequencies ω_j for $j = \{1, 2, \dots, \bar{n}\}$. The graphical representations of the first four mode shapes are given in Fig. 7.

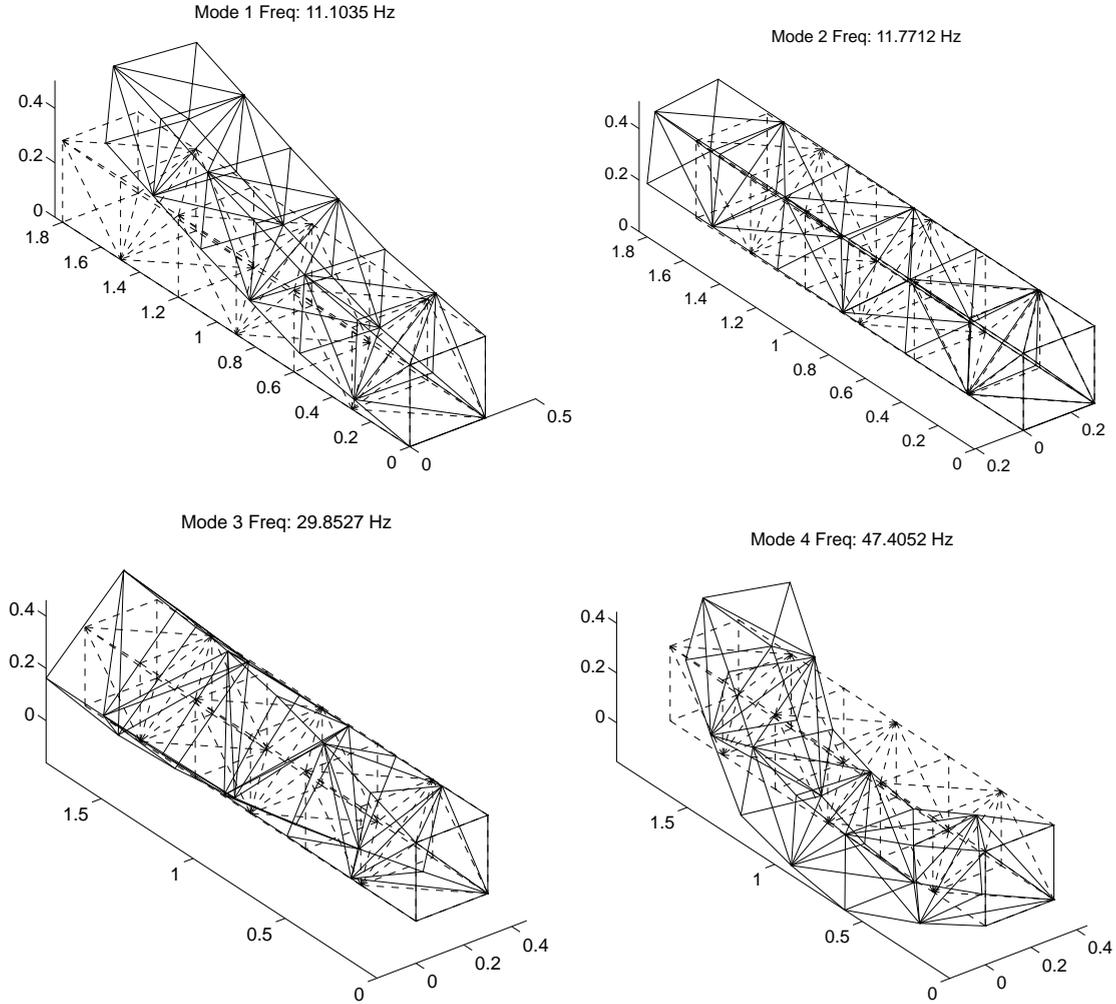


Figure 7. Mode shapes and frequencies of the structure.

ACTUATOR AND SENSOR CONSTRUCTION

Table 2 shows some of the relevant mechanical and piezoelectric properties of the piezoelectric materials used in this research. For sensor construction, polyvinylidene fluoride (PVDF) piezoelectric film was selected. This material was chosen because it produces relatively large voltages for a given strain (high g_{31} coefficient). For actuator construction, the piezoceramic material lead zirconate titanate (PZT) was selected. Piezoceramics produce a relatively high strain for a given electric field (high d_{31} coefficient).

The sensors and actuators piezoelectric films are attached to the tubular truss member as illustrated on Fig. 8. The actuators are cylindrical shells of ceramic PZT material, split longitudinally to facilitate the installation in the active truss members. The sensors are rectangular pieces of PVDF. The sensors are installed in pairs: one PVDF film measures strain, whereas the other measure strain rate. The dimensions of the sensors and actuators are given in Table 3. The length is measured along the axis of the truss member; the width refers to the dimension along the

circumference of the member, and the thickness is measured across the radius, as shown on Fig. 8. A sensor film has length L_s^k , width W_s^k , and thickness t_s^k , whereas an actuator has length L_a^k , width W_a^k , and thickness t_a^k .

For later convenience, we introduce the following notation: Let $S_{\text{SEN}} = \{1, 2, \dots, r\}$ be the index set for the piezoelectric film sensor pairs on a truss member. The function SEN maps a sensor number to a global element number; that is, SEN: $S_{\text{SEN}} \rightarrow \mathcal{G}_{\text{ELEM}}$. Let $S_{\text{ACT}} = \{1, 2, \dots, \bar{r}\}$ be the index set for the piezoelectric film actuators on a truss member. The function ACT maps an actuator number to a global element number; that is, ACT: $S_{\text{ACT}} \rightarrow \mathcal{G}_{\text{ELEM}}$.

FORMULATION OF THE STATE SPACE DYNAMIC MODEL

In this section the dynamic interaction equations for the piezoelectric sensors and actuators with the finite element model are derived. The interaction equations are used to formulate the state space model of the dynamic equations. The state space model is the canonical form for designing control systems and simulation.

Table 2. Properties of Piezoelectric Materials

Material	Application	d_{31} (C/N)	g_{31} (V m/N)	Young's Modulus E (N/m ²)	Permittivity ϵ (F/m)
PVDF	Sensor	23×10^{-12}	216×10^{-3}	2×10^9	106×10^{-12}
BM527 (PZT)	Actuator	215×10^{-12}	7×10^{-3}	6.9×10^{10}	3.54×10^{-8}

Observation Matrix for the Piezoelectric Sensor

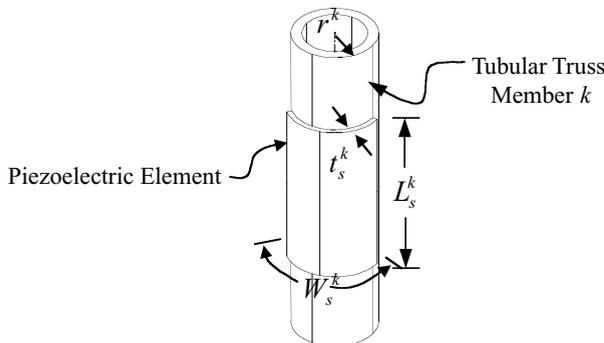
The open circuit voltage $v_{oc}(t)$ of a piezoelectric film sensor is proportional to the strain of the film. The short circuit current $i_{sc}(t)$ of a piezoelectric film sensor is proportional to the strain rate of the film. Let $v_{oc}^k(t)$ and $i_{sc}^k(t)$ denote the open circuit voltage and short circuit current associated with the piezoelectric film attached to the truss member k , $k \in \mathcal{G}_{ELEM}$. Lee and Moon (12) developed equations relating the deformation of a piezoelectric film and the resulting open circuit voltages and short circuit currents. Adapting these equations yields the following equations for the open circuit voltage $v_{oc}^k(t)$ and short circuit current $i_{sc}^k(t)$:

$$v_{oc}^k(t) = -\frac{d_{31s}^k E_s^k W_s^k}{C_s^k} \int_0^{L_s^k} \epsilon_s^k(x, t) dx, \quad (31)$$

$$i_{sc}^k(t) = -d_{31s}^k E_s^k W_s^k \int_0^{L_s^k} \frac{\partial \epsilon_s^k(x, t)}{\partial t} dx, \quad (32)$$

where d_{31s}^k is the transverse piezoelectric charge to stress ratio, E_s^k is the Young's modulus of the film, W_s^k is the width of the piezoelectric film, C_s^k is the capacitance of the film, and $\epsilon_s^k(x, t)$ is the strain profile of the piezoelectric film. Note that the subscript s denotes parameters of the sensor film, while superscript k denotes the rod element associated with the truss member to which the sensor is attached.

Modeling the truss member with rod elements yields a strain profile that is constant along the tubular truss member. Hence the strain for element k can be represented as $\epsilon^k(t)$, and determined from Eq. (6). Due to the negligible thickness of the PVDF sensor films, and the hardness of


Figure 8. Piezoelectric film attached to tubular truss member.

the epoxy, we assume $\epsilon_s^k(x, t) = \epsilon^k(t)$. Accordingly, the expressions for the open circuit voltage and closed circuit current simplify as

$$v_{oc}^k(t) = \gamma_{V_{oc}}^k \epsilon_s^k(t), \quad \gamma_{V_{oc}}^k = -\frac{d_{31s}^k E_s^k W_s^k L_s^k}{C_s^k}, \quad (33)$$

$$i_{sc}^k(t) = \gamma_{I_{sc}}^k \dot{\epsilon}_s^k(t), \quad \gamma_{I_{sc}}^k = -d_{31s}^k E_s^k W_s^k L_s^k. \quad (34)$$

>From the preceding equations, it is easy to see that measurements of open circuit voltage and short circuit current can be employed to obtain the strain and strain rate in a truss member. By using Eqs. (6), (10), and (11), the above-mentioned voltage and currents can be related to the displacement of vector q in Eq. (7).

Let the composite vector of open circuit voltages $V_{oc}(t)$, and the composite vector of short circuit currents $I_{sc}(t)$ be given by

$$V_{oc}(t) = \begin{bmatrix} V_{oc}^{SEN(1)} \\ \vdots \\ V_{oc}^{SEN(r)} \end{bmatrix} \quad \text{and} \quad I_{sc}(t) = \begin{bmatrix} I_{sc}^{SEN(1)} \\ \vdots \\ I_{sc}^{SEN(r)} \end{bmatrix}. \quad (35)$$

Matrices $M_{V_{oc}}$ and $M_{I_{sc}}$ of size $r \times m$ can be constructed so that

$$V_{oc} = M_{V_{oc}} \epsilon \quad \text{and} \quad I_{sc} = M_{I_{sc}} \dot{\epsilon}. \quad (36)$$

In particular, the ab elements of matrices $M_{V_{oc}}$ and $M_{I_{sc}}$ are given by

$$(M_{V_{oc}})_{ab} = \begin{cases} \gamma_{V_{oc}}^{SEN(a)} & \text{if } b = SEN(a), \\ 0 & \text{otherwise,} \end{cases} \quad (37)$$

$$(M_{I_{sc}})_{ab} = \begin{cases} \gamma_{I_{sc}}^{SEN(a)} & \text{if } b = SEN(a), \\ 0 & \text{otherwise,} \end{cases} \quad (38)$$

where $1 \leq a \leq r$ and $1 \leq b \leq m$.

Control Matrix for the Piezoelectric Actuator

In the following derivation, perfect bonding between a piezoelectric film and the truss member is assumed. This is a valid assumption due to the hardness of the epoxy used to bond the piezoelectric films. The manner in which the

Table 3. Dimensions of Sensors and Actuators

Application	Material	Length (mm)	Width (mm)	Thickness (mm)
Sensor	PVDF	42	17.0	28×10^{-3}
Actuator	BM527 (PZT)	25.4	39.9	1.0

strain in a substructure varies when it is bonded to a piezoelectric film is given by a coupled set of partial differential equations analyzed in (11,16). For the case of perfect bonding, with a substructure undergoing axial deformation, the following simplification is possible. The piezoelectric strain $\varepsilon_a(\bar{\xi})$ at the *normalized coordinate* $\bar{\xi}$, where $\bar{\xi} = -1$ denotes one end of the piezoelectric film, and $\bar{\xi} = +1$ denotes the other end, is given by the formula

$$\varepsilon_a(\bar{\xi}) = \frac{\Psi}{\Psi + 2} \left[\frac{\varepsilon_{\text{SUB}}^+ + \varepsilon_{\text{SUB}}^-}{2} - \left(\frac{\varepsilon_{\text{SUB}}^+ - \varepsilon_{\text{SUB}}^-}{2} \right) \bar{\xi} + \frac{2}{\Psi} \Lambda \right] \quad (39)$$

The strain of the substructure at $\bar{\xi} = -1$ and $\bar{\xi} = +1$ are denoted by $\varepsilon_{\text{SUB}}^-$ and $\varepsilon_{\text{SUB}}^+$, respectively. The stiffness ratio between the piezoelectric film and the substructure is denoted by Ψ . The effective strain of due to the piezoelectric effect is denoted by Λ . Although the strain of the piezoelectric film is the same as the strain of the substructure underneath the film, the strain of the substructure is different in locations not covered by the film.

We will now adapt Eq. (39) to the case where the substructure is a truss member modeled by a rod element. In such a case, the substructure strains at either end of the piezoelectric film are identical; this corresponds to $\varepsilon_{\text{SUB}}^- = \varepsilon_{\text{SUB}}^+$. Hence the strain over the piezoelectric film is uniform. The strain of a piezoelectric film bonded to the truss member associated with rod element k is denoted by ε_a^k , and given by

$$\varepsilon_a^k = \left(\frac{\Psi^k}{\Psi^k + 2} \right) \varepsilon^k + \left(\frac{2}{\Psi^k + 2} \right) \Lambda_a^k, \quad (40)$$

where ε^k is the strain of the truss member k on the portion *not* underneath the piezoelectric film, and Λ_a^k is the effective strain due to the piezoelectric effect. The effective strain Λ_a^k is given by

$$\Lambda_a^k = \left(\frac{d_{31a}^k}{t_a^k} \right) v_a^k, \quad (41)$$

where v_a^k is the voltage applied to the piezoelectric film associated with element k . The stiffness ratio for element k is given by

$$\Psi^k = \frac{E^k t^k}{E_a^k t_a^k}, \quad (42)$$

where E^k and E_a^k are, respectively, Young's modulus of elasticity for the truss member k and the piezoelectric film actuator attached to it; t^k and t_a^k are, respectively, the thickness of truss member k and the piezoelectric film actuator attached to it.

When there is perfect bonding between a piezoelectric film and its substructure, the force exerted by the piezoelectric on the structure is concentrated at the ends of the piezoelectric film (11,16). Let g_e^k denote the force exerted by the ends of piezoelectric film on the truss member associated with element k . Summing the stress at the end of

the piezoelectric yields

$$g_e^k = E_a^k A_{aCS}^k (\varepsilon_a^k - \Lambda_a^k), \quad (43)$$

where A_{aCS}^k is the cross-sectional area of the piezoelectric actuator film bonded to truss member k . The transmitted forced can be rearranged as

$$g_e^k = g_{e\text{STIFF}}^k + g_{e\text{ACTIVE}}^k, \quad (44)$$

where

$$g_{e\text{STIFF}}^k = \mu_{\text{STIFF}}^k \varepsilon^k, \quad (45)$$

$$g_{e\text{ACTIVE}}^k = \mu_{\text{ACTIVE}}^k v_a^k, \quad (46)$$

and

$$\mu_{\text{STIFF}}^k = E_a^k A_{aCS}^k \left(\frac{\Psi^k}{\Psi^k + 2} \right), \quad (47)$$

$$\mu_{\text{ACTIVE}}^k = -E_a^k A_{aCS}^k \left(\frac{\Psi^k}{\Psi^k + 2} \right) \left(\frac{d_{31a}^k}{t_a^k} \right). \quad (48)$$

The force g_e^k has two components. The component $g_{e\text{STIFF}}^k$ is the passive force due to the intrinsic elasticity of the piezoelectric film, which augments the stiffness of the truss member. The component $g_{e\text{ACTIVE}}^k$ is the active force contributed by the piezoelectric effect that varies with the applied voltage v_a^k .

Equations (45) and (46) only define the forces $g_{e\text{STIFF}}^k$ and $g_{e\text{ACTIVE}}^k$ for truss members k that have a piezoelectric film actuator; that is, $k = \text{ACT}(j)$ for some $j \in \mathcal{S}_{\text{ACT}}$. We extend the definition to all truss members in the following manner. For all $k \in \mathcal{G}_{\text{ELEM}}$,

$$g_{e\text{STIFF}}^k = \begin{cases} \mu_{\text{STIFF}}^k \varepsilon^k, & \text{if } k = \text{ACT}(j) \text{ for some } j \in \mathcal{S}_{\text{ACT}}, \\ 0 & \text{otherwise;} \end{cases}$$

and

$$g_{e\text{ACTIVE}}^k = \begin{cases} \mu_{\text{ACTIVE}}^k v_a^k, & \text{if } k = \text{ACT}(j) \text{ for some } j \in \mathcal{S}_{\text{ACT}}, \\ 0 & \text{otherwise.} \end{cases}$$

Let the composite vector of passive element forces $g_{e\text{STIFF}}^k$, and the active element forces $g_{e\text{ACTIVE}}^k$ be given by

$$g_{e\text{STIFF}} = \begin{bmatrix} g_{e\text{STIFF}}^1 \\ \vdots \\ g_{e\text{STIFF}}^m \end{bmatrix} \quad \text{and} \quad g_{e\text{ACTIVE}} = \begin{bmatrix} g_{e\text{ACTIVE}}^1 \\ \vdots \\ g_{e\text{ACTIVE}}^m \end{bmatrix}. \quad (49)$$

Let the composite vector of applied voltages V_a be given by

$$V_a = \begin{bmatrix} v_a^{\text{ACT}(1)} \\ \vdots \\ v_a^{\text{ACT}(\bar{r})} \end{bmatrix}. \quad (50)$$

The $m \times m$ matrix Y_{STIFF} is constructed so that

$$g_{e\text{STIFF}} = Y_{\text{STIFF}} \varepsilon, \quad (51)$$

where strain vector ε is given by Eq. (10). In particular, the ab element of matrix Y_{STIFF} is given by

$$(Y_{\text{STIFF}})_{ab} = \begin{cases} \mu_{\text{STIFF}}^{\text{ACT}(a)} & \text{if } a = b \text{ and } a = \text{ACT}(j) \text{ for some} \\ & j \in \mathcal{S}_{\text{ACT}}, \\ 0 & \text{otherwise,} \end{cases}$$

for $1 \leq a \leq m$ and $1 \leq b \leq m$.

The $m \times \bar{r}$ matrix Y_{ACTIVE} is constructed so that

$$g_{e\text{ACTIVE}} = Y_{\text{ACTIVE}} V_a. \quad (52)$$

In particular, the ab element of matrix Y_{ACTIVE} is given by

$$(Y_{\text{ACTIVE}})_{ab} = \begin{cases} \mu_{\text{ACTIVE}}^{\text{ACT}(b)} & \text{if } a = \text{ACT}(b) \text{ and } b \in \mathcal{S}_{\text{ACT}}, \\ 0 & \text{otherwise,} \end{cases}$$

for $1 \leq a \leq m$ and $1 \leq b \leq \bar{r}$.

In the context of finite element model, Eq. (21), the applied force f has two components:

$$f = f_{\text{STIFF}} + f_{\text{ACTIVE}}, \quad (53)$$

where

$$f_{\text{STIFF}} = S^T R^T Y_{\text{STIFF}} N R S q \quad (54)$$

and

$$f_{\text{ACTIVE}} = S^T R^T Y_{\text{ACTIVE}} V_a. \quad (55)$$

Matrices S , R , and N are defined in Eqs. (11), (13), and (15), respectively.

State Space Model Construction

In this section, we construct a state space model for the modal model of Eq. (29). The canonical form of a state space is shown below.

$$\begin{aligned} \dot{x} &= Ax + Bu, \\ y &= Cx + Du. \end{aligned} \quad (56)$$

The state x of the modal model is the composite vector of mode amplitudes η and mode amplitude rates $\dot{\eta}$. The input vector u is the vector of voltages applied to piezoelectric actuators V_a . The output vector y is the composite vector of open circuit voltages and short circuit currents for the piezoelectric film sensors. In particular, x , u , and y are given by

$$x = \begin{bmatrix} \eta \\ \dot{\eta} \end{bmatrix}, \quad u = V_a, \quad \text{and} \quad y = \begin{bmatrix} V_{oc} \\ I_{sc} \end{bmatrix}.$$

The dimension of the modal amplitude vector η is the number of modes in the model. Although the constrained system model in Eq. (29) has $\bar{n}=72$ modes, we can construct a model with a reduced set of modes. This is typically done to simplify the computational complexity of control design. For a \bar{n} mode model, $1 \leq \bar{n} \leq \bar{n}$, state space matrices are A , B , C , and D are given by

$$A = \begin{bmatrix} 0 & I_{\bar{n} \times \bar{n}} \\ -\Omega_{\bar{n}} & -H_{\bar{n}} \end{bmatrix} + \Phi_{\bar{n}}^T S^T R^T Y_{\text{STIFF}} N R S P^T \Phi_{\bar{n}}, \quad (57)$$

$$B = \Phi_{\bar{n}}^T S^T R^T Y_{\text{ACTIVE}}, \quad (58)$$

$$C = \begin{bmatrix} M_{voc} N R S P^T \Phi_{\bar{n}} & 0 \\ 0 & M_{isc} N R S P^T \Phi_{\bar{n}} \end{bmatrix}, \quad (59)$$

$$D = 0_{2r \times \bar{r}}, \quad (60)$$

where

$$\Phi_{\bar{n}} = [\hat{v}_1, \hat{v}_2, \dots, \hat{v}_{\bar{n}}], \quad (61)$$

$$\Omega_{\bar{n}} = \text{diag} [\omega_1^2, \omega_2^2, \dots, \omega_{\bar{n}}^2], \quad (62)$$

$$H_{\bar{n}} = \text{diag} [2\zeta_1 \omega_1, 2\zeta_2 \omega_2, \dots, 2\zeta_{\bar{n}} \omega_{\bar{n}}]. \quad (63)$$

The modal matrix $\Phi_{\bar{n}}$ contains the first \bar{n} eigenvectors of Eq. (25).

Two state space models are used for modeling the dynamics of the truss structure. A low-order *design model* is used for controller design, and a high-order *evaluation model* is used to simulate the controller performance. The design model retains the first five modes ($\bar{n} = 5$) and assumes no damping ($\zeta_j = 0$, $j = 1, 2, \dots, 5$). This is consistent with the fact that the true damping of structure is not known until it is tested. The evaluation model retains the first 30 modes ($\bar{n} = 30$) and uses the estimated modal damping factors ($\zeta_j = 0.004$, $j = 1, 2, \dots, 30$) obtained from empirical tests on the structure.

CONTROL DESIGN AND SIMULATIONS

LFT Framework for H_∞ Feedback Control Design

A robust H_∞ feedback controller has been designed for the truss structure. To obtain the controller using this technique, the dynamic model of the system has to be written in lower linear fractional transformation (LFT) form. This formulation framework is convenient because it allows for the inclusion of the effects of the sensor and control disturbances, as well as loop-shaping weights on the inputs and outputs. Figure 9 shows the block representation of the lower LFT form of the control problem. The signal z is the *exogenous output*, y is the *controlled output*, w is the *exogenous input*, and u is the *control input*. The relationship

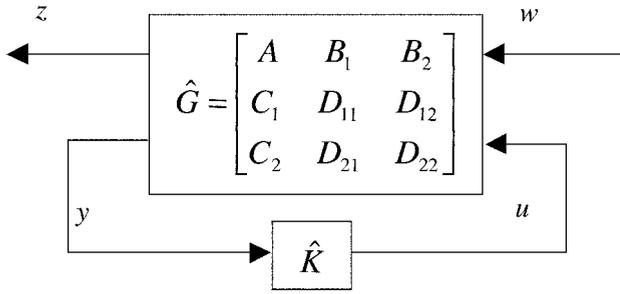


Figure 9. Lower LFT representation of control problem.

between these signals can be represented in the following manner:

$$\begin{aligned} \dot{x} &= Ax + B_1 w + B_2 u, \\ z &= C_1 x + D_{11} w + D_{12} u, \\ y &= C_2 x + D_{21} w + D_{22} u. \end{aligned} \quad (64)$$

In the *LFT* framework, it is desired to minimize the H_∞ norm from w to z . The nature of these signals and their relationship to controller design is explained later.

Consider the state space model of the truss dynamics in Eq. (56). The model has the matrix transfer function $\hat{P}(s)$ that is given by

$$\hat{P}(s) = C(sI - A)^{-1}B + D \quad (65)$$

Figure 10 shows the conventional block diagram of the closed loop control system, including the sensor noise n and disturbances d to the control. In order to put the problem in the *LFT* framework, we aggregate the unwanted inputs n and d into an exogenous input signal w . We also aggregate the signals that we wish to penalize, the state x and the control input u , into an exogenous output signal z . The resulting exogenous input and output vectors are

$$w = \begin{bmatrix} d \\ n \end{bmatrix}, \quad z = \begin{bmatrix} x \\ u \end{bmatrix}. \quad (66)$$

Different frequency ranges in the input signals and outputs can be penalized through the use of loop-shaping filters. This is illustrated in Fig. 11 for a system in lower *LFT* form with an input filter \hat{W}_1 and an output filter \hat{W}_2 .

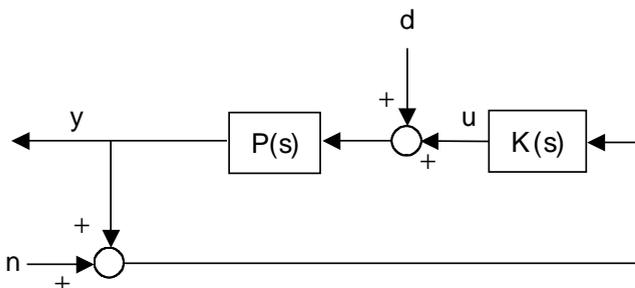


Figure 10. Block diagram of the closed loop system.

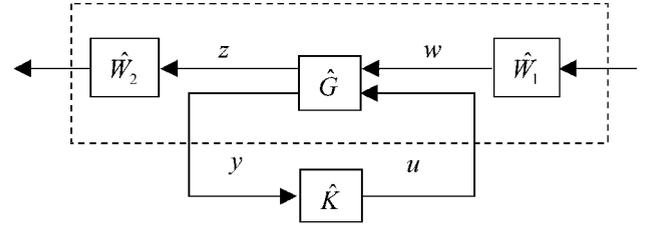


Figure 11. Shaping of the open loop.

The system model can be uncertain over different frequency ranges. For example, uncertainty in the high-frequency region can result from modal truncation. Uncertainty in the low frequency can result from low-frequency disturbances being applied to the system due to ground movements. The uncertainty can be handled by specifying appropriate filters in \hat{W}_1 , such as a bandpass filter, to introduce robustness to the perturbations in the model dynamics. The output filter \hat{W}_2 can be used to introduce performance requirements by placing weights on different elements of the control signal and state space. These specification techniques are explained by Zhou et al. (6)

The dynamic controller that minimizes the H_∞ norm of the closed loop is given by (26)

$$\hat{K} = C_k(sI - A_k)^{-1}B_k + D_k \quad (67)$$

The state space matrices that form the controller are found from the solution of the following matrix inequality

$$H_{X_L} + Q^* J^* P_{X_L} + P_{X_L} J Q < 0, \quad (68)$$

where the unknowns are X_L and J . The matrices in equation (65) are defined as follows.

$$J = \begin{bmatrix} A_k & B_k \\ C_k & D_k \end{bmatrix}, \quad H_{X_L} = \begin{bmatrix} \bar{A}^* X_L + X_L \bar{A} & X_L \bar{B} & \bar{C}^* \\ \bar{B}^* X_L & -I & D_{11}^* \\ \bar{C} & D_{11} & -I \end{bmatrix},$$

$$\bar{A} = \begin{bmatrix} A & 0 \\ 0 & 0_{n_k \times n_k} \end{bmatrix}, \quad \bar{B} = \begin{bmatrix} B_1 \\ 0 \end{bmatrix}, \quad \underline{B} = \begin{bmatrix} 0 & B_2 \\ I & 0 \end{bmatrix},$$

$$\underline{C} = \begin{bmatrix} 0 & I \\ C_2 & 0 \end{bmatrix}, \quad \bar{C} = [C_1 \quad 0],$$

$$\underline{D}_{21} = \begin{bmatrix} 0 \\ D_{21} \end{bmatrix}, \quad \underline{D}_{12} = [0 \quad D_{12}],$$

$$P_{X_L} = \begin{bmatrix} \bar{B}^* X_L & 0 \\ \underline{D}_{12}^* \end{bmatrix}, \quad Q = \begin{bmatrix} \underline{C} & \underline{D}_{21} & 0 \end{bmatrix}.$$

Gahinet (26) shows that if a controller with the structure in Eq. (67) exists, then X_L is given by

$$X_L = \begin{bmatrix} X & X_2 \\ X_2^* & I \end{bmatrix}, \quad X - Y^{-1} = X_2 X_2^*, \quad (69)$$

where X and Y are the solutions to the following linear matrix inequalities.

$$\begin{bmatrix} N_y & 0 \\ 0 & I \end{bmatrix}^* \begin{bmatrix} A^*X + XA & XB_1 & C_1 \\ B_1^*X & -I & D_{11}^* \\ C_1 & D_{11} & -I \end{bmatrix} \begin{bmatrix} N_y & 0 \\ 0 & I \end{bmatrix} < 0, \quad (70)$$

$$\begin{bmatrix} N_x & 0 \\ 0 & I \end{bmatrix}^* \begin{bmatrix} AY + YA^* & B_1 & YC_1^* \\ B_1^* & -I & D_{11}^* \\ C_1Y & D_{11} & -I \end{bmatrix} \begin{bmatrix} N_x & 0 \\ 0 & I \end{bmatrix} < 0, \quad (71)$$

The matrices N_x and N_y are full rank and satisfy the following.

$$\text{Im}(N_y) = \text{Ker}[B_2^* \quad D_{12}^*],$$

$$\text{Im}(N_x) = \text{Ker}[C_2 \quad D_{21}].$$

Once X_L has been found, the matrix inequality given in Eq. (68) becomes linear. Accordingly, Eq. (65) can be solved using LMI solution techniques to obtain the matrix J . The controller matrices (A_k, B_k, C_k, D_k) are obtained as partitions of J . The numerical solutions to the above LMI problems can be obtained using the LMI Lab toolbox (27) in MATLAB (28).

Controller Design for Vibration Suppression of Truss

In the simulations, an impact test was performed. Figure 12 shows the location of the impact. Sensor-actuator pairs were placed on the four truss members indicated by thick lines. This selection is made due to the high axial strains that these truss members undergo compared to other members in the first five vibrational modes. All the considered modes are controllable and observable with this sensor and actuator configuration.

The controller for the truss is designed using a five mode reduced order model of the dynamics. The truss members with both sensors and actuators are shown in Fig. 12. The state space matrices $A, B, C,$ and D are, respectively, given by Eqs. (57) to (60) with $\tilde{n} = 5$. The corresponding LFT representation is obtained as follows:

$$\begin{aligned} B_1 &= [0.1B \quad 0_{10 \times 4}], \quad B_2 = B, \\ C_1 &= \begin{bmatrix} Q \\ 0_{4 \times 10} \end{bmatrix}, \quad C_2 = C, \\ D_{11} &= 0_{14 \times 8}, \quad D_{12} = \begin{bmatrix} 0_{10 \times 4} \\ R \end{bmatrix}, \\ D_{21} &= 0.01 \times [0_{4 \times 4} \quad I_{4 \times 4}], \quad D_{22} = D. \end{aligned}$$

The Q and R are weights on the state and control are given by

$$Q = \text{diag}[20 \quad 20 \quad 30 \quad 40 \quad 50 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0] \quad \text{and} \\ R = 1 \times 10^4 I_{4 \times 4}.$$

This choice corresponds to a 10% actuation disturbance and 1% sensor noise. The control input is heavily weighted to avoid saturation of the control input. The modal amplitudes are penalized, whereas modal amplitude rates are

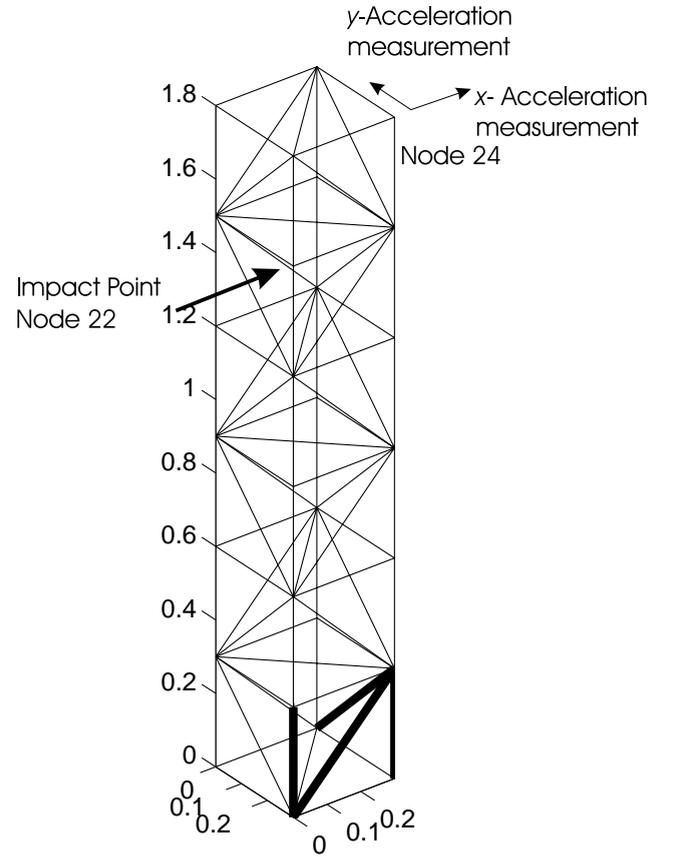


Figure 12. Impact test: Location of disturbance and active truss members.

not. This LFT formulation provides reasonable trade-off between performance and robustness to sensor noise and amplifier nonlinearities. Figure 13 shows a maximum singular value plot. The decreased peaks of the closed loop indicate the increased damping of the first five modes of the closed loop system relative to the open loop system. It also shows that outside this range, the frequency response

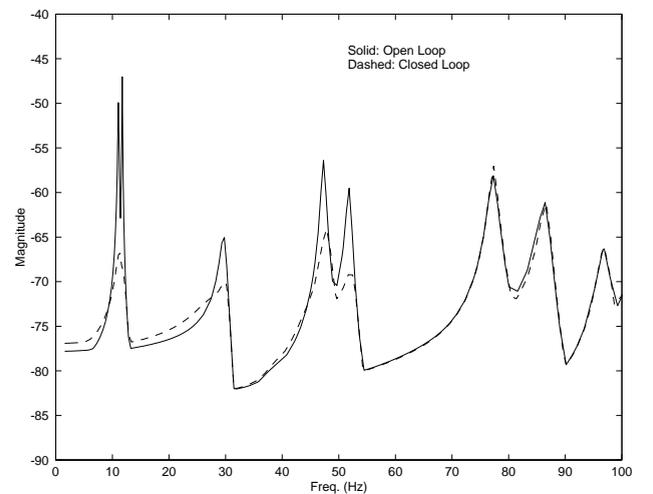


Figure 13. Maximum singular value plot for H_∞ control.

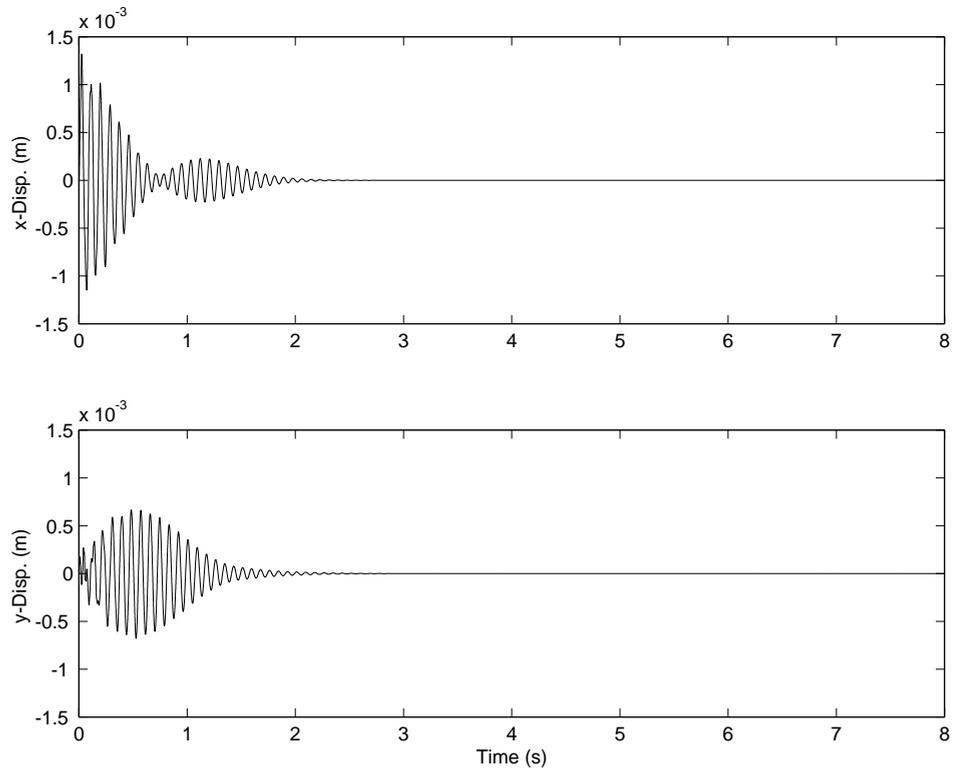


Figure 14. Displacement of node 24 during H_∞ control (simulation).

of the system is almost unchanged. This shows that the controller does not destabilize high frequency unmodeled modes.

The high-order evaluation model is used for the dynamics of the truss structure. The model includes the first 30 modes of vibration. The simulated actuation voltage was limited to a range of ± 265 volts to match the voltage swing available from the high voltage amplifiers. Simulink (29) was used to simulate the closed loop response of the truss with a dynamic controller. The displacement and control signal response are shown on Figs. 14 and 15.

EXPERIMENTAL RESULTS

The architecture of the truss structure testbed is shown on Fig. 16. The feedback controller is performed by a high-performance DSP card in a Pentium-based computer. The control logic is specified using the Simulink (29) graphical interface, which runs in the Matlab (28) environment. The control logic is compiled into a real-time executable by and Wincom (30) software, which also runs in the Matlab environment. The executable code is downloaded by the computer to a Quanser MultiQ-3 I/O card. The MultiQ card has a DSP and supports 8 input ADC and 8 output DAC channels with a 12 bit resolution at a sampling frequency of 15 kHz. Piezoelectric film actuators and sensors are bonded to the truss structure as shown in Fig. 4. Each piezoceramic actuator is driven by a custom-designed high-voltage amplifier, with a maximum voltage output of 560 V. A HP 6521A dc high-voltage power supply is used

to power the high-voltage amplifiers. For calibration purposes, strain gauges are fitted to some of the truss members. Two accelerometers are installed at the top corner of the structure to measure accelerations in the X and Y directions as shown in Fig. 12. The accelerations are integrated to obtain displacement measurements.

Two types of experiments are performed: impact tests and shaker tests. The impact test uses an impact hammer to strike the truss structure, and thereby impart an initial deflection. The shaker test uses one of the active truss member actuators as a disturbance source. A disturbance signal is applied to the piezoelectric actuator to cause resonant vibrations in the truss structure. The open loop and closed loop responses of the structure are compared for both tests.

Impact Test

The impact test is performed by using an impact hammer to strike the location shown on Fig. 12. The impact hammer is equipped with an internal accelerometer. By integrating the accelerometer signal, the initial deflection resulting from the impact can be determined. Since the dominant time constant of the truss dynamics is much lower than the impact duration, the impact behaves like it imparts an initial deflection to the truss structure. The open loop and closed loop responses of the structure can then be compared.

The open loop tests are carried out to have a reference against which to compare the performance of the closed loop controlled system. The results were also used to calibrate the modal damping ratios in the test model used

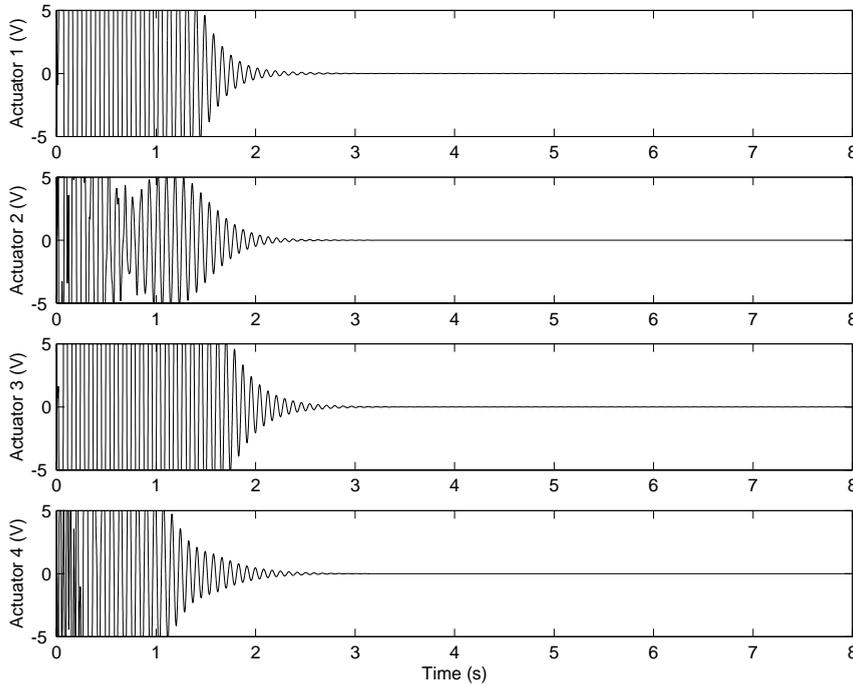


Figure 15. Control signals during H_∞ control (simulation).

in the simulations. Figure 17 shows the displacement response of the system to the impact, obtained from double integration of the accelerometer signals. The small inherent structural damping of the system is visible. The structure oscillates without control for 8 seconds. Figure 18 shows the frequency response of the open loop system. The dominant resonant peak corresponds to the second mode of vibration, at 13.3 Hz.

The H_∞ controller design explained earlier was implemented on the testbed. The displacement and control signal are shown in Figs. 19 and 20, respectively. The settling time of the vibrations is dramatically shorter for the closed loop case. The oscillations damp out in less than 2.5 seconds. The improvement in performance is easier to notice

in the frequency response plot, shown in Fig. 21. The maximum resonant peak, at 13.3 Hz, is reduced by 25 dB when compared to the open loop case.

Shaker Test

In the shaker experiment, a continuous sinusoidal disturbance signal is applied to a piezoceramic actuator on the structure. The active truss member, whose actuator is used as a disturbance source, is shown on Fig. 22. The sinusoidal disturbance signal is adjusted to the dominant resonant frequency of 13.3 Hz with the maximum amplitude of 560 V. The acceleration measurements were taken at node 24 in both the X and Y directions. Double integration

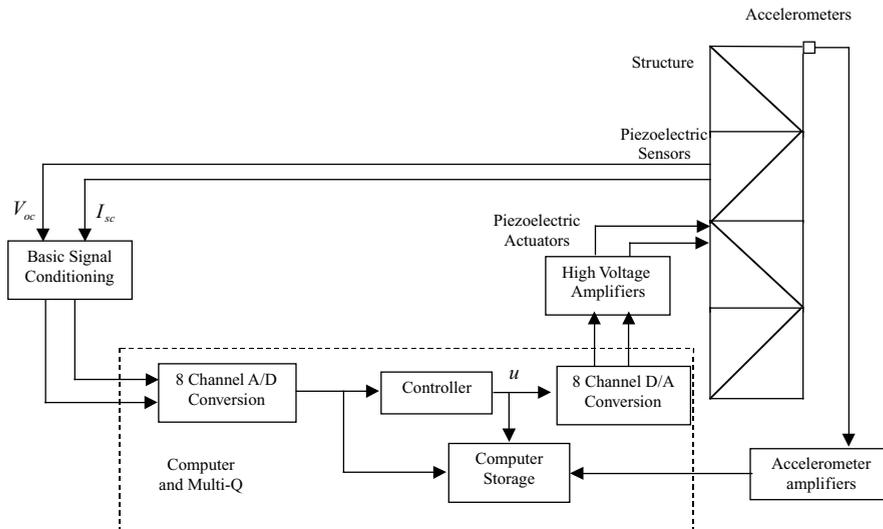


Figure 16. Schematic of testbed system architecture.

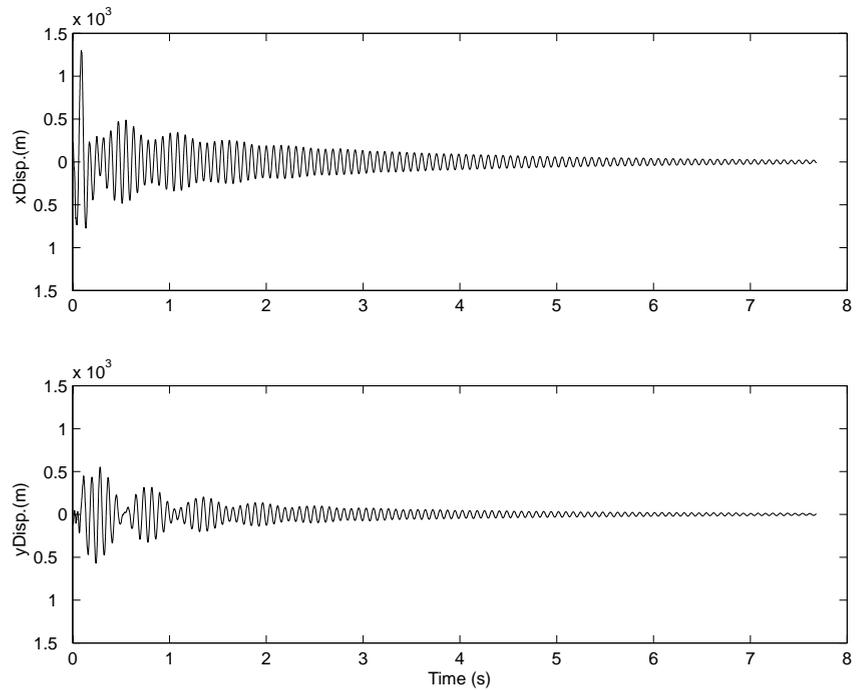


Figure 17. Displacement of node 24 of open loop (experimental).

of the acceleration signals yields displacement measurements at node 24. The resonant vibrations are allowed to build up to a steady state before the H_∞ controller is applied.

The displacement at node 24 in the X and Y directions is shown in Fig. 23. The control signals applied to actuators 1 to 4 are shown in Fig. 24. The first 4.6 seconds of both responses shows the open loop resonant behavior of the structure when no control is applied. At 4.6 seconds, the H_∞ controller is applied. The oscillation amplitude is decreased to its steady state value about 1.4 seconds later. The steady state oscillation amplitude is about a factor 10 smaller than the open loop amplitude.

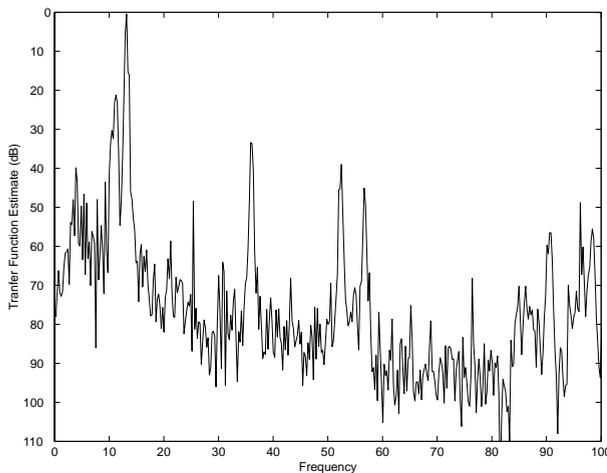


Figure 18. Frequency response of open loop (experimental).

The experiments match well with the numerical simulation, confirming the findings obtained earlier the optimal H_∞ controller is effective in providing a considerable increase in the amount of vibration suppression, due to its ability to target the control energy on a few specific modes of vibrations, as stipulated by the designer. The attenuation provided by the control law for the first five modes of the nominal structure, compared to the open loop case is shown in Table 4.

CONCLUSION

The use of piezoelectric film actuators and sensors for active vibration suppression in truss structures has been investigated. Techniques for modeling truss dynamics and the design robust controllers for suppressing vibrations have been developed. The techniques have been successfully used on an experimental truss structure. The truss structure was designed to be representative of the complex dynamics of the large flexible space structures under consideration by the Canadian Space Agency.

The dynamics of truss structures with piezoelectric actuators and sensors are modeled using a two-step procedure. First, a finite element model of the truss dynamics, without piezoelectric actuators and sensors, is developed. Second, the interaction equations for modeling the coupling dynamics of piezoelectric film actuators and sensors to the finite element model of a truss are developed. This two-step analysis procedure allows the sensor and actuator placement to be optimized for a given finite element model. The dynamic modeling methodology has been validated by tests on an experimental truss structure.

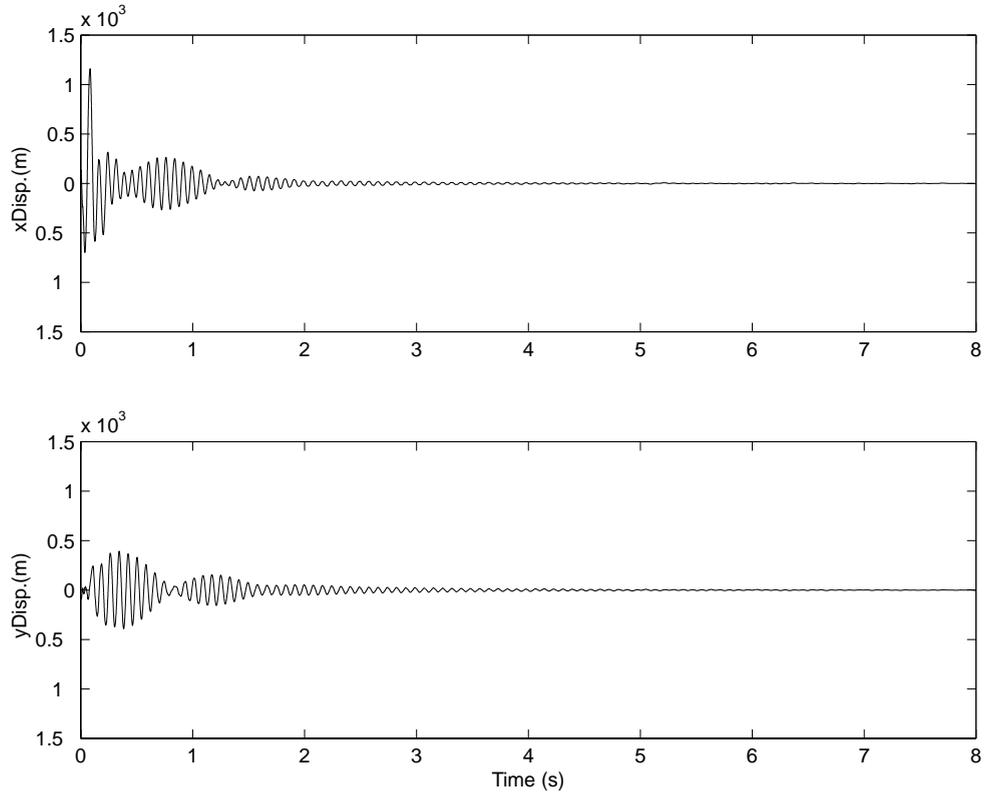


Figure 19. Displacement of node 24 during H_∞ control (experimental).

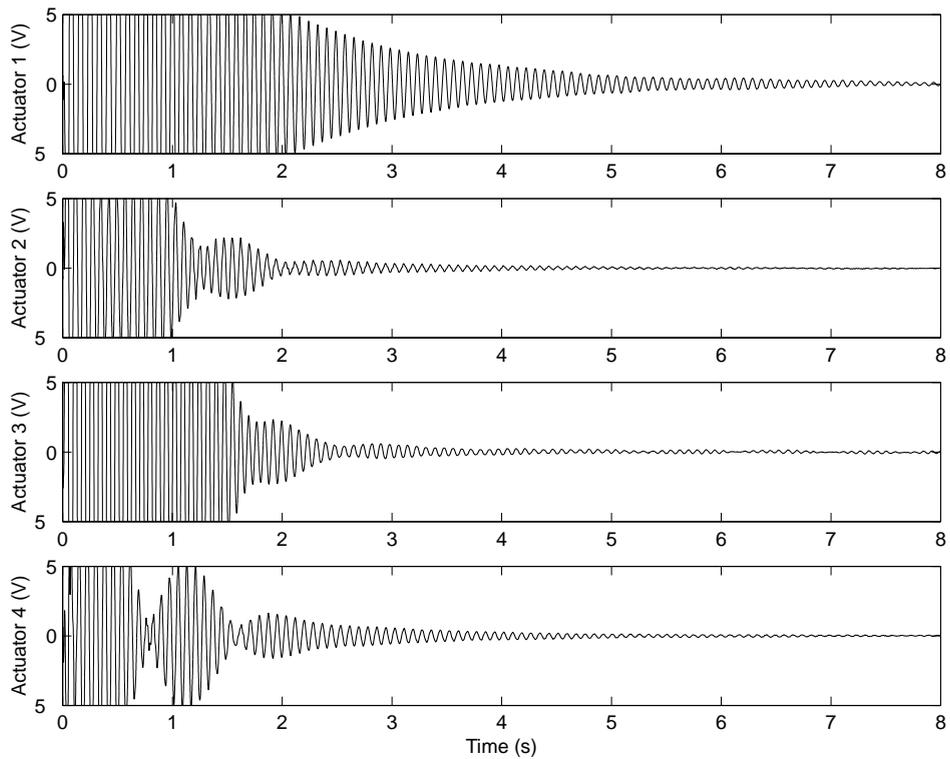


Figure 20. Control signals during H_∞ control (experimental).

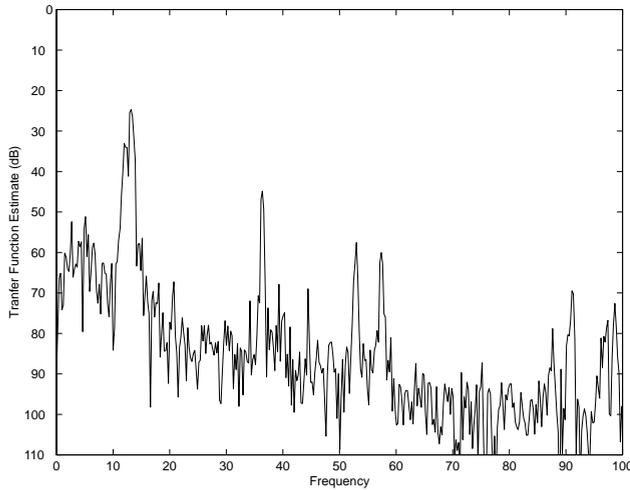


Figure 21. Frequency response during H_∞ control (experimental).

Flexible structures are difficult to control because their dynamics are characterized by a large number of vibrational modes. To reduce computational complexity, controller design is typically performed using a reduced order model. The H_∞ controller design procedure yields a controller that concentrates the control energy on the modes included in the design model. The design procedure accounts for sensor noise and disturbances resulting from nonlinearities in the amplifier and piezoelectric actuators. Control input saturation can be avoided by using a high penalty on the control energy during the controller design. The H_∞ controller performance is analyzed using a high-order evaluation model. The simulations showed that the H_∞ controller provides significant increase in damping to the modes included in the design model, but does not affect the higher-order excluded modes. This behavior is ideal, as it ensures that the structure will not become unstable through the excitation of higher-order modes.

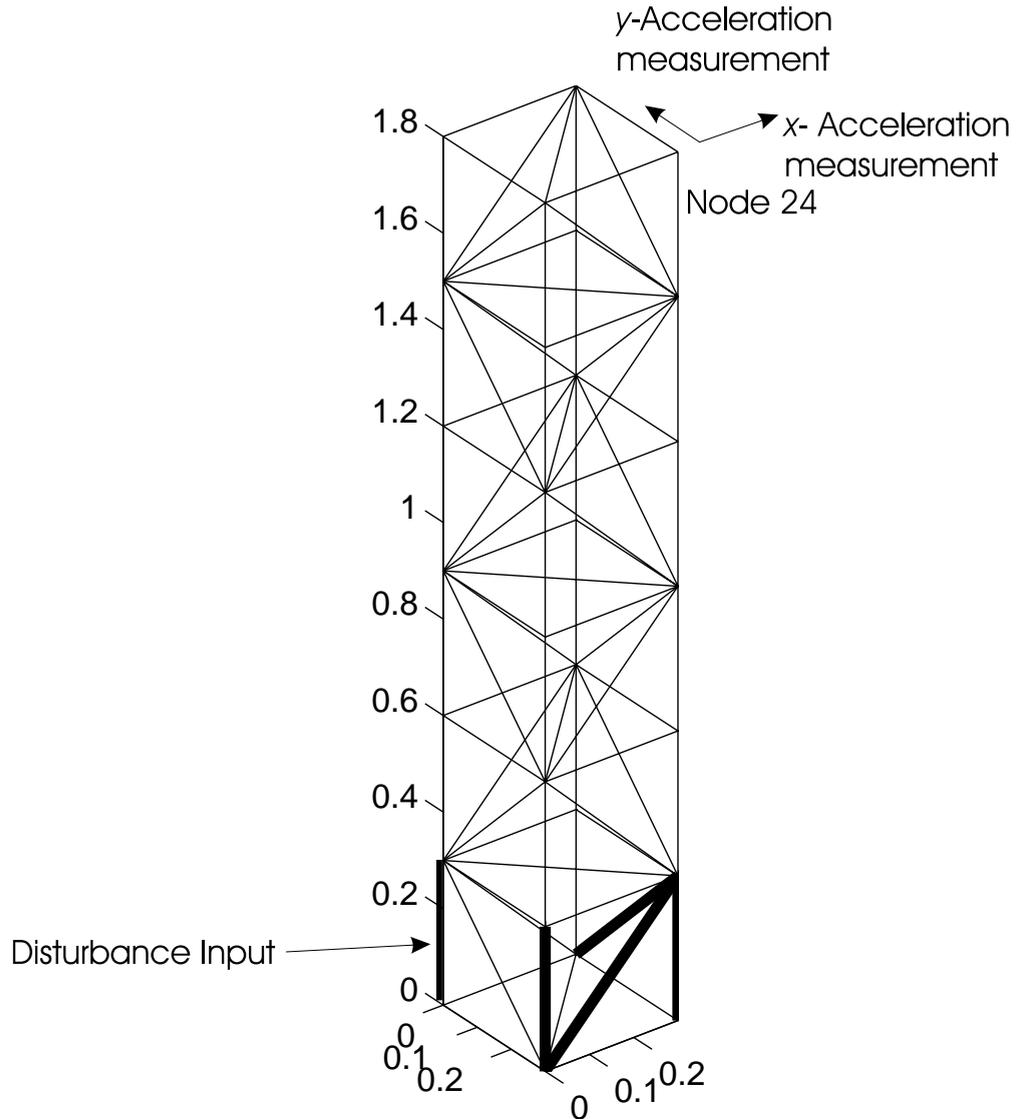


Figure 22. Application point for the continuous disturbance.

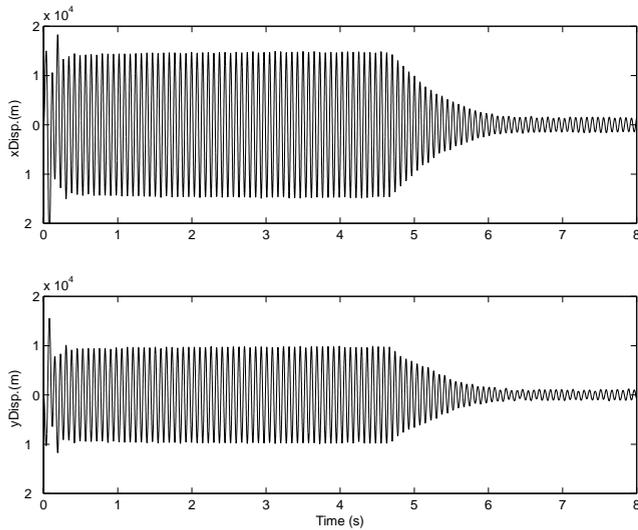


Figure 23. Displacement of node 24 during continuous disturbance test with H_∞ control (experimental).

Experimental results with the truss structure have confirmed the validity of the simulations. Two tests have been performed, an impact test and shaker test. Comparisons between the open loop and the closed loop responses show that the H_∞ controller significantly decreases the vibrational mode amplitudes. The controller targets its efforts on the modes retained in the design model.

Piezoelectric materials are ideally suited for constructing actuators and sensors for vibration suppression in flexible structures. Polyvinylidene fluoride (PVDF) is ideally suited for sensor construction. It is lightweight, flexible, and provides a high voltage for a given strain. Piezoceramic materials are suited to actuator construction. Piezoceramics are stiff, rugged, and provide relatively

Table 4. Mode Attenuation

Mode	Attenuation (dB)
1	13
2	25
3	12
4	29
5	15

large strains when subjected to an electric field. The predominately linear behavior of the piezoelectric materials, and the simple manner in which they can be integrated into a structure, make them a good choice for actuators and sensors in a vibration suppression control system.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Steven Yeung, Mr. Howard Reynaud, Dr. George Vukovitch, and Dr. Yan-Ru Hu at the Canadian Space Agency for their technical and financial assistance.

This work has been supported by the Canadian Space Agency under contracts 9F009-0-4140, 9F011-0-0924, 9F028-5-5106, and 9F028-6-6162, and the Natural Sciences and Engineering Research Council of Canada.

BIBLIOGRAPHY

1. J.C. Doyle. *IEEE Trans. Autom. Contr.* **AC-23**(4): 756–757 (1978).
2. M.J. Balas. *IEEE Trans. Autom. Contr.* **27**(3): 522–535 (1982).
3. J.J. Allen and J.P. Lauffer. *J. Dyn. Syst. Meas. Contr.* **119**: (September 1997).

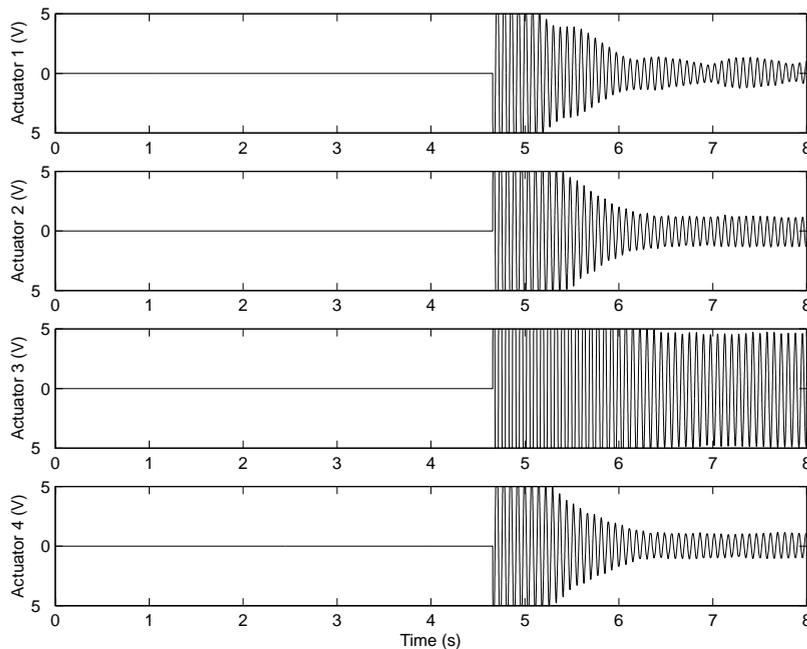


Figure 24. Control signals during continuous disturbance test with H_∞ control (experimental).

4. J.C. Doyle, K. Glover, P.P. Khargonekar, and B.A. Francis. *IEEE Trans. Autom. Contr.* **AC-34**(8): 831–847 (1989).
5. B.A. Francis. *A Course in H_∞ Control Theory*. Lecture Notes in Control and Information Series, Vol. 88. 1987.
6. K. Zhou, J.C. Doyle, and K. Glover. *Robust and Optimal Control*. Prentice Hall, Englewood Cliffs, NJ, 1995.
7. S.A. Buddie and T.T. Goergiu, Ü. Özgüner and M.C. Smith. *Int. J. Contr.* **58**(1): 1–19 (1993).
8. P. Van Woerkom. *Control-Theory Adv. Tech.* **9**(3): 639–669 (1993).
9. T. Bailey and J.E. Hubbard. *J. Guid. Contr. Dyn.* **8**: 605–611 (1985).
10. J.M. Plump, J.E. Hubbard, and T. Bailey. *J. Dyn. Syst. Meas. Contr.* **109**: 133–139 (1987).
11. E.F. Crawley and J. de Luis. *AIAA J.* **25**(10): 1373–1385 (1987).
12. C.K. Lee and F.C. Moon. *J. Appl. Mech.* **57**(6): 434–441 (1990).
13. D.W. Miller, S.A. Collins, and S.P. Peltzman. *31st AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conf.*, Long Beach, CA, 1990.
14. D.W. Miller and M.C. van Schoor. *Proc. 1st Joint US/Japan Conf. on Adaptive Structures*, Maui, Hawaii, 1990, pp. 304–331.
15. A.F. Vaz. *Modelling Piezoelectric Behaviour for Actuator and Sensor Applications*. DSS Contract 9F009-0-4140. Canadian Space Agency, February 1991.
16. A.F. Vaz. *Theoretical Development for an Active Vibration Damping Experiment*. DSS Contract 9F009-0-4140. Canadian Space Agency, March 1991.
17. A.F. Vaz. *Empirical Verification of Interaction Equations for Flexible Structures with Bonded Piezoelectric Films*. DSS Contract 9F011-1-0924. Canadian Space Agency, September 1991.
18. A.F. Vaz. *IEEE Trans. Instrumentation and Measurement*, **24**(2): (1998).
19. R. Bravo, T. Farr, and A.F. Vaz. *Analysis and Testing of Flexible Structure Dynamics*. 9F028-6-6162. Canadian Space Agency, December 1997.
20. T. Farr. M.Eng. Thesis. Mechanical Engineering Department. McMaster University, Canada, 1997.
21. A.F. Vaz and T.J. Farr. *IEEE Conf. on Instrumentation and Measurement*, Ottawa, Ontario, April 21–23, 1997.
22. R. Bravo. Ph.D. Thesis. Mechanical Engineering Department. McMaster University, Canada, 2000.
23. R. Bravo, S. Leatherland, and A.F. Vaz. *Final Report on Design and Construction of a Truss Structure Testbed*. 9F028-8-2094/001-XSD. Canadian Space Agency, April 1999.
24. W. Weaver and P.R. Johnston. *Structural Dynamics by Finite Elements*. Prentice Hall, Englewood Cliffs, NJ, 1987.
25. SDRC Corp. *I-DEAS Master Series Student Guide*. On World Press, 1996.
26. P. Gahinet. *Automatica* **32**(7): 1007–1014 (1996).
27. P. Gahinet, A. Nemirovski, A. Laub, and M. Chilali. *LMI Control Toolbox for Use with MATLAB*. Mathworks, Natick, MA, 1995.
28. Mathworks Inc. *Matlab: Language of Technical Computing, Matlab Function Reference*. Release 11, Natick, MA, 1999.
29. Mathworks Inc. *Simulink: Dynamics System Simulation for Matlab, Using Simulink*. Simulink 3, Release 11, Natick, MA, 1999.
30. Quanser Consulting Inc. *MultiQ-3 Programming Manual*. Hamilton, Ontario, Canada, 1999.
31. D.C. Hanselman. *Mastering MATLAB 5: A Comprehensive Tutorial and Reference*. Prentice Hall, Englewood Cliffs, NJ, 1998.
32. R. Bravo, A.F. Vaz, S. Leatherland, and M. Dokanish. *1999 CANSMART Workshop*, Canadian Space Agency, St. Hubert, Quebec, September 13–14, 1999.

VIBRATION CONTROL

SEUNG-BOK CHOI
Inha University
Inchon, Korea

INTRODUCTION

The insatiable demand for high performance on various dynamic systems quantified by high-speed operation, high control accuracy, and lower energy consumption has triggered vigorous research on vibrational control of distributed flexible structures and discrete systems. Numerous control strategies for conventional electromagnetic actuators have been proposed and implemented to suppress unwanted vibration. However, the successful empirical realization of electromagnetic actuators may be sometimes very difficult under certain conditions due to hardware limitations such as saturation and response speed. This difficulty can be resolved by employing smart material actuators in vibrational control. As is well-known, smart material technology features actuating capability, control capability, and computational capability (1). Therefore, these inherent capabilities of smart materials can execute specific functions autonomously in response to changing environmental stimuli. Among many smart material candidates, electrorheological(ER) fluids, piezoelectric materials, and shape-memory alloys (SMA) are effectively exploited for vibrational control in various engineering applications. A viable vibrational control algorithm can be optimally synthesized by integrating control strategies, and actuating technology, and sensing technology, as shown in Fig. 1. The design philosophy presented in Fig. 1 contains a very large number of decisions and design parameters for the characteristics of controllers, actuators, and sensors. Furthermore, the designer seeking a global optimal solution for the synthesis of a closed-loop smart structure system must also address other crucial decisions concerning the time delay of a high-voltage/current amplifier, the speed of the signal converter, and the microchip hardware of the control software. In this article, two different flexible smart structures fabricated from ER fluids and piezoelectric materials are introduced, and vibrational control techniques for each smart structure are presented. In addition, vibrational control methodology for a passenger vehicle under various road conditions is given by adopting an ER damper, followed by vibrational control of a flexible robotic manipulator that features piezoceramic actuators.

VIBRATIONAL CONTROL OF SMART STRUCTURES

ER Fluid-Based Smart Structures

Significant progress has been made in developing smart structures that incorporate electrorheological(ER) fluids. Typically, this class of smart structures features an autonomous actuating capability that makes them ideal

for vibrational control applications in variable service conditions and in unstructured environments. This may be accomplished by controlling the stiffness and energy dissipation characteristics of the structures. This, of course, is possible due to the tunability of rheological properties of ER fluids by the intensity of the electric field. The development of ER fluid-based smart structures was initiated by Choi et al. (2). They completed an experimental study of a variety of shear configurations based on sandwich beam structures. Gandhi et al. (3) suggested using an ER fluid as an actuator to suppress deflections of the flexible robotic arm structures by avoiding resonance. In this work, a phenomenological governing equation was derived by assuming that the structures are viscoelastic materials. A passive control scheme for obtaining a desired transient response was developed on the basis of experimentally obtained phenomenological governing equation, in which field-dependent modal properties were used as pseudocontrol forces (4). Vibrational control logic to minimize the tip deflections of an ER fluid-based cantilever beam structures was illustrated by field-dependent responses in the frequency domain (5). Coulter and Duclos (6) suggested a methodology for replacing a conventional viscoelastic material by an ER fluid. Following the formulation of an analytical model for ER fluid-embedded structures via the conventional sandwich beam theory, they presented a feasibility that the controllability of the complex shear modulus of an ER fluid itself can be used to obtain the desired responses of the structures. Rahn and Joshi (7) developed dynamics for an ER fluid-based on the complex shear modulus of the ER fluid and also theoretically suggested a feedback controller for transient vibration control. Oyadiji (8) developed a theoretical equation to predict the field-dependent frequency response by treating the ER fluid layer as a constrained damping layer and verified its validity by experiment. Choi et al. (9) presented a dynamic model for an ER fluid-based smart beam, in which the complex shear modulus of the ER fluid itself, measured by a rotary oscillation test, was taken into consideration. To validate the methodology, the predicted elastodynamic properties, such as damped natural frequencies and loss factors, were compared with those measured. Gong and Lim (10) experimentally investigated the vibrational properties of sandwich beam structures in which an ER fluid layer was partially or fully filled as a constraint damping layer. Yalcintas and Coulter (11) proposed a vibrational model based on thin-plate theory, and the transverse vibration response of a nonhomogeneous ER smart beam was investigated. In addition, the vibrational control capacity of an ER beam was illustrated by emphasizing mode shape control associated with an on and off state of the electric field. On the other hand, Choi and Park (12) controlled vibration of ER smart beam structures by using a closed-loop control. The vibrational control technique was empirically realized by activating a field-dependent fuzzy controller.

In this article, the field-dependent fuzzy control scheme is introduced after briefly explaining the typical block

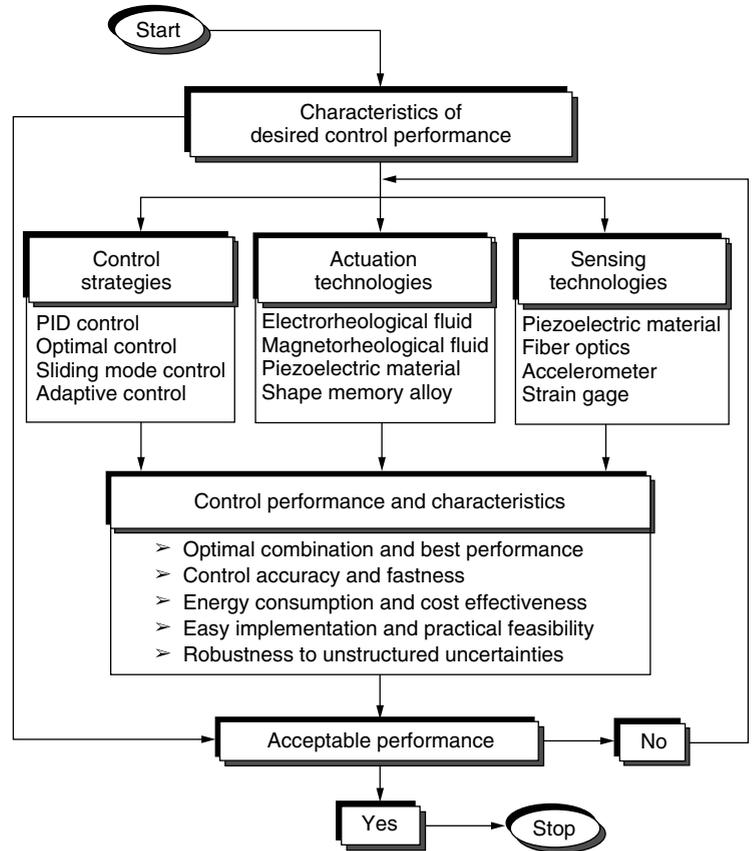


Figure 1. An algorithm for synthesizing a closed-loop smart structure system.

diagram for vibrational control of ER fluid-based smart structures, shown in Fig. 2. The control system consists of a set of sensors, signal converters, microprocessor, high-voltage amplifier, and control algorithm. Most of the sensors currently available such as accelerometers can be adapted to measure the dynamic response of ERfluid-based smart structures. The microprocessor which includes A/D (analog–digital) and D/A(digital–analog) signal converters plays a very important role in closed-loop control time. The

microprocessor should have at least 12 bits to realize control software and also should take into account a high sampling frequency up to 10 kHz. The high-voltage amplifier should have enough power to generate the required ER effect in smart structures. Furthermore, the response time of the high-voltage amplifier to the source input control voltage should be fast enough not to delay the control action of the feedback control system. Typically, a smart structure consists of two host(face) structures, insulators, and an ER

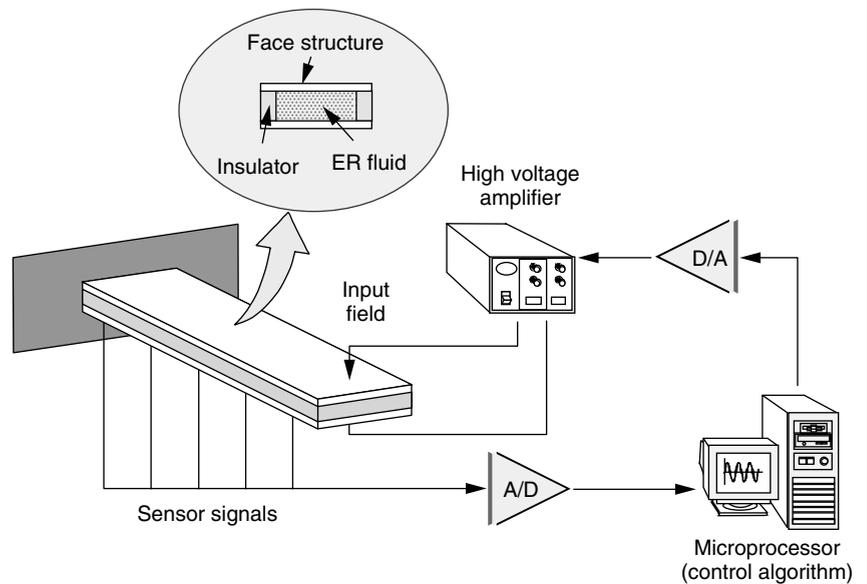


Figure 2. Schematic diagram for controlling the vibration of an ER fluid-based smart structure.

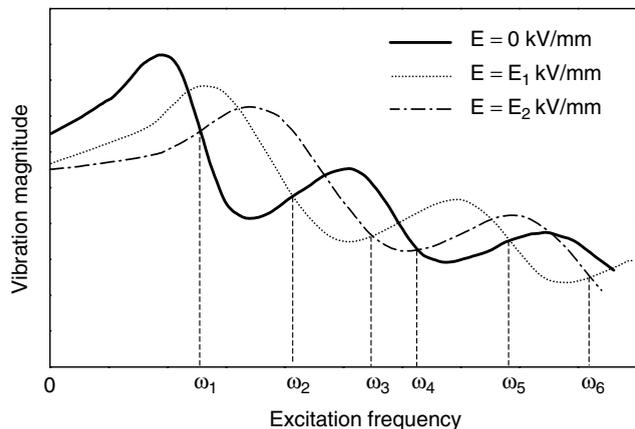


Figure 3. Field-dependent frequency responses of an ER beam structure.

fluid layer, as shown in Fig. 2. For the composite laminate, the lay-up angle of the laminates can be selected as a design parameter to investigate the effect of the ER fluid for different stiffnesses. The insulator is a seal to maintain the integrity of the structure and is also used to adjust the volume fraction of the ER fluid relative to the total volume of the structure. For certain elastodynamic purposes, a smart structure can be constructed that consists of multilayers of ER fluids whose rheological properties are different.

The elastodynamic properties of an ER fluid-based smart structure vary with the level of the electric field, as shown in Fig. 3. This implies that the natural frequency of each vibrational mode can be adjusted by tailoring the electric field and that, consequently, vibration in real time can be effectively suppressed in the presence of resonant disturbances (excitations). In other words, the desired response for minimizing the vibrational magnitude can be obtained by selecting the lowest envelope in the frequency range considered. The desired electric field corresponding to the desired response can be expressed as a fuzzy control algorithm (12): if $\omega_i \leq \omega < \omega_{i+1}$, then $E_d = E_j$. The variable ω denotes the disturbance frequency, and E_d is the desired electric field. Note that the variation potential of elastodynamic properties with respect to the applied field may be different upon operating conditions such as the magnitude of excitation when are altered. Therefore, the frequency bandwidth and the corresponding desired field for the control algorithm should be modified. From fuzzy logic, the control field for the case shown in Fig. 3 is determined as follows: if $0 \leq \omega < \omega_1$, then $E_d = E_2$; if $\omega_1 \leq \omega < \omega_2$, then $E_d = 0$; if $\omega_2 \leq \omega < \omega_3$, then $E_d = E_1$; if $\omega_3 \leq \omega < \omega_4$, then $E_d = E_2$; if $\omega_4 \leq \omega < \omega_5$, then $E_d = 0$; and if $\omega_5 \leq \omega < \omega_6$, then $E_d = E_1$. Figure 4 presents the tip deflection of a cantilevered ER beam in the frequency domain, which has been experimentally obtained by implementing the fuzzy control logic (12). It is evident from this figure that there are effective vibrational suppressions in the neighborhood of the resonant frequencies. However, a small nonzero vibrational magnitude exists across a broad frequency range. This indicates that ER fluids do not provide an actuating force but change the stiffness and the damping properties to avoid resonance. To improve vibrational control performance of the fuzzy control logic,

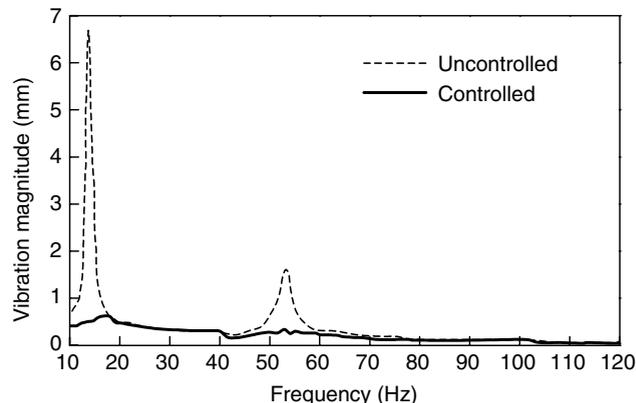


Figure 4. Forced vibrational responses of a cantilevered ER beam.

appropriate membership functions for the excitatory magnitudes and frequencies can be used to determine the electric fields desired.

On the other hand, it is known that an ER fluid contained in a distributed parameter structural system under continuous and periodic small deformations remains in the preyield state, which shows viscoelastic properties represented by a complex shear modulus (7,9). The complex shear modulus G_r^* of an ER fluid is expressed by $G_r^* = G_r^s + iG_r^l$, where $i = \sqrt{-1}$. Here, G_r^s is defined as the storage shear modulus (in-phase), a measure of the energy stored, relating to the stiffness of the structure that contains the ER fluid. G_r^l is the loss shear modulus (out-of-phase), a measure of the energy dissipated. The shear loss factor is the ratio of the energy lost to the energy stored in a cycle of deformation and denotes the damping characteristic of the ER fluid-embedded smart structure. The complex shear modulus of the ER fluid is normally measured by employing the oscillation mode of an electrorheometer (9). The measured complex shear modulus is integrated with a sixth-order partial differential equation, which is obtained by adopting conventional sandwich beam theory (13). Then, the field-dependent elastodynamic properties of the structure such as the natural frequency are determined through a finite-element model which is governed by $[M]\{\ddot{u}(t)\} + [C(E)]\{\dot{u}(t)\} + [K(E)]\{u(t)\} = \{f(t)\}$. The global mass, damping, and stiffness matrices are denoted by $[M]$, $[C(E)]$, and $[K(E)]$, respectively. Clearly, both the stiffness matrix $[K(E)]$ and damping matrix $[C(E)]$ are functions of the electric field (E) applied to the ER fluid domain. Thus, these matrices can be tuned as functions of the electric field. The variable $\{u(t)\}$ is a displacement vector, (\cdot) is the time derivative, and $\{f(t)\}$ represents the external (or disturbance) force vector. By introducing modal coordinates and also adopting mode shape characteristics of the smart structure, the finite-element model can be rewritten in a typical form of state space representation as follows (4): $\dot{x}(t) = Ax(t) + A(E)x(t) + Bf(t)$. The state vector $x(t)$ represents modal coordinates and the matrix B indicates the influence matrix of the disturbance. A represents the system matrix in the absence of an electric field, and $A(E)$ denotes the additional system matrix due to the electric field. This implies that the desired response of the

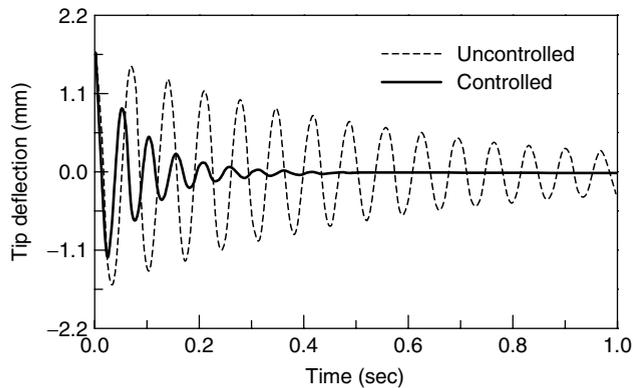


Figure 5. Transient vibrational responses of a cantilevered ER beam.

structure can be achieved by tuning the field-dependent $A(E)$. In transient vibrational control without an external disturbance, the desired eigenvalues of the system, which directly indicate the desired natural frequencies and damping ratios of the system, can be obtained by adjusting the intensity of the electric field in the matrix $A(E)$. One of the effective control algorithms for achieving this goal is a so-called pseudostate feedback controller proposed by Choi et al. (4). In this method, the state equation is modified to fit a PD (proportional-derivative) controller in which the proportional gain is related to the field-dependent natural frequency, and the derivative gain is related to the field-dependent damping ratio. In addition, we can easily shift the desired eigenvalues of the system to avoid resonant phenomena by employing this control algorithm. Figure 5 presents the transient vibrational control response of a cantilevered ER beam (4). The first mode eigenvalues of the structure are calculated from $-1.7313 \pm i91.167$ in the absence of an electric field. However, the desired eigenvalues of $-11.44 \pm i122.114$ are achieved by employing appropriate control parameter, which indicate the intensity of the electric field.

An ER beam structure for vibrational control can easily be extended to an ER plate structure. In the vibrational control of flexible plate structures, the significance of mode shape control is no less important than vibrational magnitude control. When we consider large flexible structures such as aircraft wings and helicopter blades, the mode shape is directly related to lift distribution and stability due to internal and external disturbances and other aeroelastic problems in a stringent environment. Therefore, much research on the mode shape control of plate structures have been undertaken by using smart material actuators (14). Choi et al. (15) proposed an ER plate and investigated its field-dependent mode shapes. Figure 6 presents the measured mode shape of an ER plate which has clamped-clamped boundary conditions (15). It is clearly observed that the magnitude of each mode shape is efficiently suppressed by applying a control electric field. Note that we can also control the mode shape in part of the plate structure (15) by partitioning the ER plate and applying an electric field to the specific portion. By doing this, we may alter the twist/camber of an airfoil in the aircraft wing, which in turn controls the lift distribution, to produce desirable performance by real-time control.

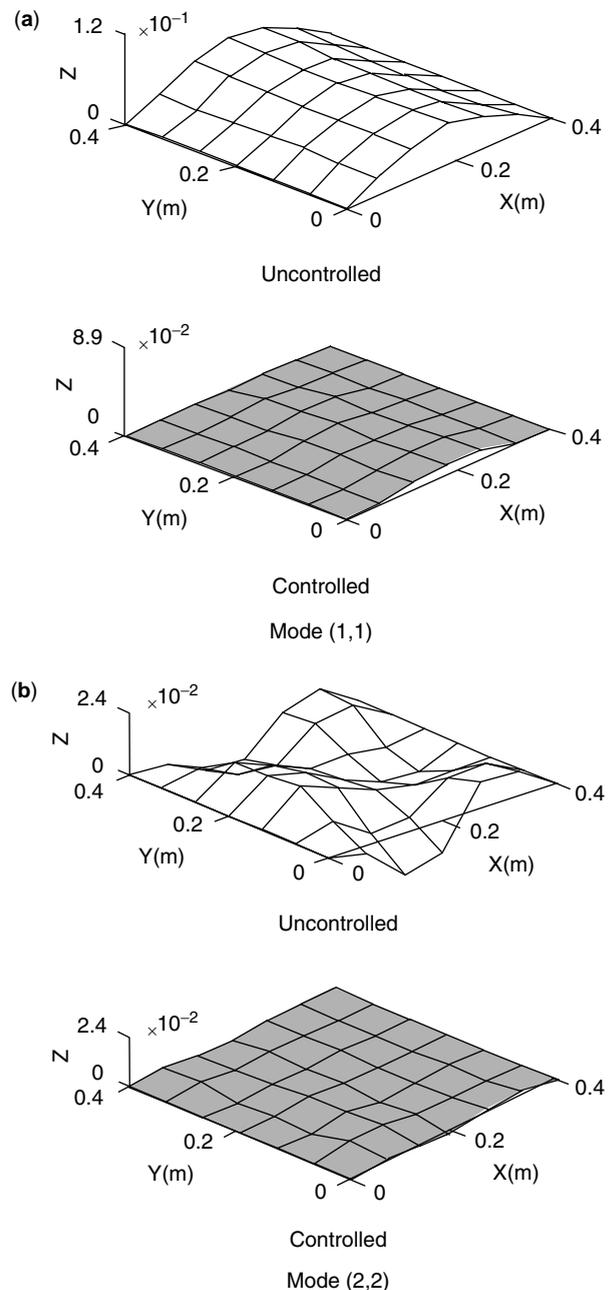


Figure 6. Measured mode shapes of an ER plate.

Smart Structures That Feature Piezoelectric Actuators and Sensors

So far, many natural and synthetic materials that exhibit piezoelectric properties have been proposed and developed. Natural materials include quartz, ammonium phosphate, paraffin, and bone; synthetic materials include lead zirconate titanate (PZT), barium titanate, lead niobate, lithium sulfate, and polyvinylidene fluoride (PVDF). Among these materials, PZT and PVDF are the most popular and commercially available. Both classes of materials are available in a broad range of properties suited to vibrational control applications as actuators or sensors. One of the salient properties of a piezoelectric material is that it responds very fast to voltage, and hence has a wide control

bandwidth. In addition, we can fabricate simple, compact, low-power devices that feature a set of piezoelectric actuators or/and sensors. Applications that use piezoelectric materials include vibrational control of flexible structures such as beams, plates, and shells; noise control of cabins; positional control of structural systems such as flexible manipulators; vibrational control of discrete systems such as engine mounts; ultrasonic motors; and various type of sensors, including accelerometers, strain gauges, and sound pressure gauges. The successful development of a technology that incorporates piezoelectric materials involves several issues. When we fabricate smart structures that use piezoelectric actuators and sensors, we must consider, the fabrication method (surface bonding or embedding), the curing temperature when embedding, insulation between piezoelectric layers, and the harness of electric wires. The important issues to considered in modeling piezoelectric-based smart structures include structural dynamics, actuator dynamics, sensor dynamics, the bonding effect, the hysteresis phenomenon, the optimal location of actuators and sensors, and the number of actuators and sensors.

The control technique for vibrational control of piezoelectric-based smart structures is very similar to that of a conventional vibrational control system, except that it uses a voltage amplifier, as shown in Fig. 7. The response time of the voltage amplifier, which normally has an amplification factor of 200, should be fast enough so that it does not deteriorate the dynamic bandwidth of the piezoactuators. The microprocessor that has A/D(analog to digital) and D/A(digital to analog) signal converters needs to have at least a 12-bit memory, and also needs to account for high sampling frequency up to 10 kHz. Most of the currently available control algorithms for piezoactuators are realized in an active manner. Therefore, a wide range of control techniques has been proposed for using piezoelectric material to control the vibration of flexible structures actively. Bailey and Hubbard (16) applied a piezofilm as an active vibrational damper for distributed

structural systems. Simulations and experimental investigations of transient vibrational control of a cantilever beam were conducted. They derived two types of controllers based on Lyapunov stability: a constant-amplitude controller (CAC) and a constant-gain controller (CGC). Favorable vibrational suppression was achieved by implementing these two controllers. It has been also shown that the CAC is more effective than the CGC for the same maximum voltage. However, when the CAC is employed, undesirable residual vibration is generated in the settled phase due to the excessive supply of control voltages from the inevitable time delay of the hardware system. Baz and Poh (17) proposed a modified independent modal space control method to suppress actively the unwanted vibration of a flexible beam structure that features piezoelectric actuators. The effects of the bonding layer material and the actuator location on the vibrational control performance were evaluated by numerical simulation. Tzou and Gadre (18) derived a physical model for vibrational control, in which a piezofilm slab was sandwiched between two other plates. The effectiveness of active vibrational control has been demonstrated by implementing CGC. Tzou (19) also applied a piezofilm for vibrational control of arbitrarily shaped shells. Control performance of the distributed systems was successfully evaluated through computer simulations by using the CAC and the CGC. Baz et al. (20) integrated the independent modal space control method and the positive position feedback method. Vibrational control performance was enhanced by argumenting the so-called time sharing strategy, and its effectiveness was validated by showing multimode controllability by a single piezoelectric actuator. On the other hand, Choi et al. (21) proposed a multistep constant-amplitude controller (MCAC) to reduce undesirable chattering in the settled phase. They experimentally demonstrated the effectiveness of the MCAC by comparing the vibrational control response of the CAC. Choi and Kim (22) also proposed a new type of discrete-time, fuzzy, sliding mode controller to reduce unwanted

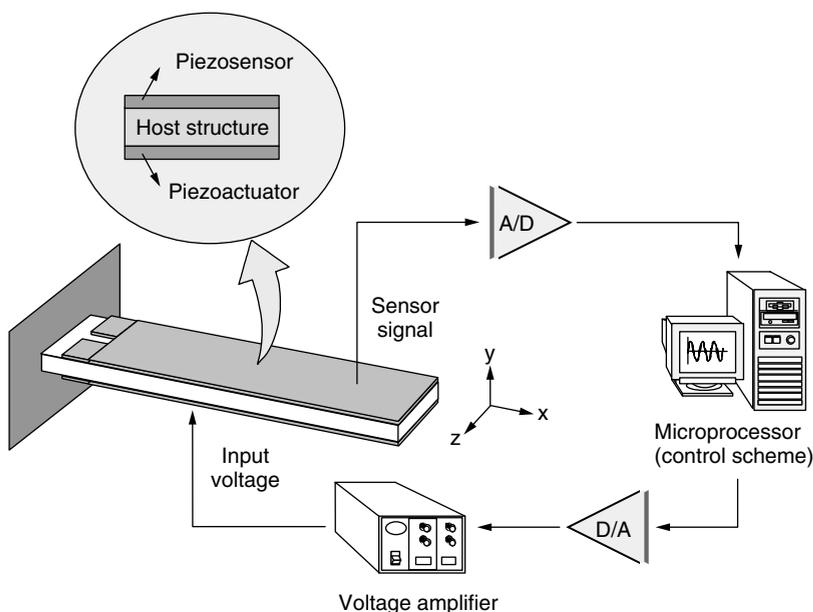


Figure 7. Schematic diagram for vibrational control of a smart structure that features a piezoactuator and a sensor.

vibrational magnitude favorably in the settled phase. Yang and Lee (23) developed three neural networks for smart structures that feature a PZT actuator and sensor, one for system identification, the second for on-line state estimation, and the third for vibrational suppression. The effectiveness of the proposed neural networks was demonstrated by experimentally undertaking transient vibrational control of a cantilevered beam structure. Meyer et al. (24) proposed two control methodologies for vibrational control of large flexible structures: positive position feedback (PPF) and linear quadratic Gaussian (LQG). It has been shown that PPF is effective in providing high damping for a particular mode, and LQG is very effective in meeting specific requirements such as minimization of tip motion. On the other hand, it is generally known that flexible structures are easily subjected to parameter variations in practice. Therefore, a robust vibrational control technique that can guarantee favorable structural performance under system uncertainties needs to be developed. Tang et al. (25) proposed an active-passive hybrid piezoelectric method that used a sliding mode controller to suppress unwanted vibration of flexible structures. A robust sliding mode controller that compensates for parameter variations such as material frequency and the hysteretic nonlinearity of the piezoactuator was designed and successfully implemented; it substantially reduced the vibrational magnitude. Choi et al. (26) formulated a robust quantitative feedback theory (QFT) controller to suppress the vibration of a flexible structure subjected to parameter variations and hysteretic nonlinearity. It has been demonstrated through experiment that the QFT controller is very effective for robust vibrational control of piezoelectric-based smart structures.

In this article, the CAC, CGC, and MCAC schemes, which are relatively easy to implement and very effective for vibrational control of piezoelectric-based smart structures, are introduced by considering the simple cantilevered beam structure shown in Fig. 7. From the figure, it is seen that the control objective is to reduce the vibration in the y direction by activating the piezoactuator. To stabilize the structural system, a positive-definite Lyapunov function F , which is basically a measure of the energy (potential and kinetic) in the system, is adopted as follows: $2F = \int_0^L [(\partial^2 y(x, t)/\partial x^2)^2 + (\partial y(x, t)/\partial t)^2] dx$. Here, L is the length of the beam, and $y(x, t)$ is the deflection of the beam. Minimizing the time derivative of the function, vibrational control is achieved by bringing the system to equilibrium. Taking the time derivative of the function and substituting the governing equation of the beam yields the following (21): $\partial F/\partial t = \int_0^L \{ (1 - EI/\rho A)[\partial^3 y(x, t)/\partial t \partial x^2][\partial^2 y(x, t)/\partial x^2] dx + [cV(t)/\rho A][\partial^2 y(x, t)/\partial t \partial x] \}_{x=L}$. The variables E , I , ρ , and A represent the elastic modulus, inertia, density, and cross-sectional area of the beam, respectively. $V(t)$ is the control voltage, and c is a constant that implies the bending moment per volt. This constant is normally determined by the geometric and material properties of the structure. It is clear that the control voltage $V(t)$ should be chosen so that the second term of the time derivative equation is always as negative as possible. Therefore, two types of control laws are easily synthesized: (1) CAC: $V(t) = -K_1 \cdot \text{sign}(\dot{V}_f)$, and

(2) CGC: $V(t) = -K_2(\dot{V}_f)$. Here K_1 and K_2 are feedback gains. The variable $V_f(t)$ represents the output voltage produced from the piezoelectric sensor. The sensor voltage $V_f(t)$ is proportional to the sign of the angular displacement at the tip of the beam (16). It has been experimentally verified that the CAC is more effective than the CGC at the same maximum voltage (16). This is due to the fact that a square wave has more area than a sine wave of equal magnitude. However, from the practical point of view, the CAC causes undesirable residual oscillations in the settled phase that are attributed to the excessive control voltage from the time delay of the hardware system. This problem becomes more serious when small vibrational levels are considered at relatively high control voltages. On the other hand, the CGC also has some shortcomings under forced vibrational control. Due to insufficient control forces, the suppression efficiency is degraded. This problem becomes more serious when large vibrational levels are considered at relatively low control voltages. The multistep constant-amplitude controller (MCAC) has been also proposed to circumvent the drawbacks of the conventional CAC and CGC (21). The MCAC is given as follows: MCAC: $V(t) = -K_1 \cdot \text{sign}(\dot{V}_f)$, for $(V_f)_m > [(V_f)_{\max}/a_1]$; $V(t) = -K_2 \cdot \text{sign}(\dot{V}_f)$, for $[(V_f)_{\max}/a_2] < (V_f)_m \leq [(V_f)_{\max}/a_1]$; $V(t) = -K_3 \cdot \text{sign}(\dot{V}_f)[(V_f)_m/(V_f)_{\max}]$, for $(V_f)_m \leq [(V_f)_{\max}/a_2]$. The variable K_i is feedback gain, a_i is a switching constant to determine an appropriate voltage magnitude, $(V_f)_{\max}$ is the initial angular displacement in the absence of the control voltage, and $(V_f)_m$ is the controlled angular deflection at a certain time. The feedback gain K_i is chosen so that the maximum voltage amplitude does not exceed the voltage limit V_{\max} . This limit depends upon the breakdown voltage of the piezoactuator. The determination of the switching constant a_i is the key issue that makes the MCAC algorithm effective. The chattering magnitudes in the settled phase are normally experimentally evaluated with respect to imposed initial vibrational magnitudes and also to applied magnitudes of the control voltage in the CAC. And then, from this information, the switching constants are appropriately chosen so that undesirable chattering can be minimized for a certain initial magnitude and corresponding control voltage. The MCAC may be able to self-tune the voltage magnitude via the ratio $(V_f)_m/(V_f)_{\max}$. Implementing this type of controller provides a relatively large control force to suppress large oscillations at the beginning of the control action and subsequently a small control force to remove undesirable chattering in the settled phase. Figure 8 schematically presents the types of control input voltage for the CAC, the CGC and the MCAC, respectively.

Figure 9 presents the measured, transient vibrational control responses of a cantilevered beam that features a piezofilm actuator and sensor (21). The transient vibrational response characteristics were obtained by exciting the beam using the first-mode natural frequency and subsequently removing this excitation and feedback voltage applied. It is clearly observed that the CAC is more effective than the CGC but shows unwanted residual vibration (chattering) in the settled phase. The chattering phenomenon arises from the combined effect of the excessive supply of control voltage on the relatively small

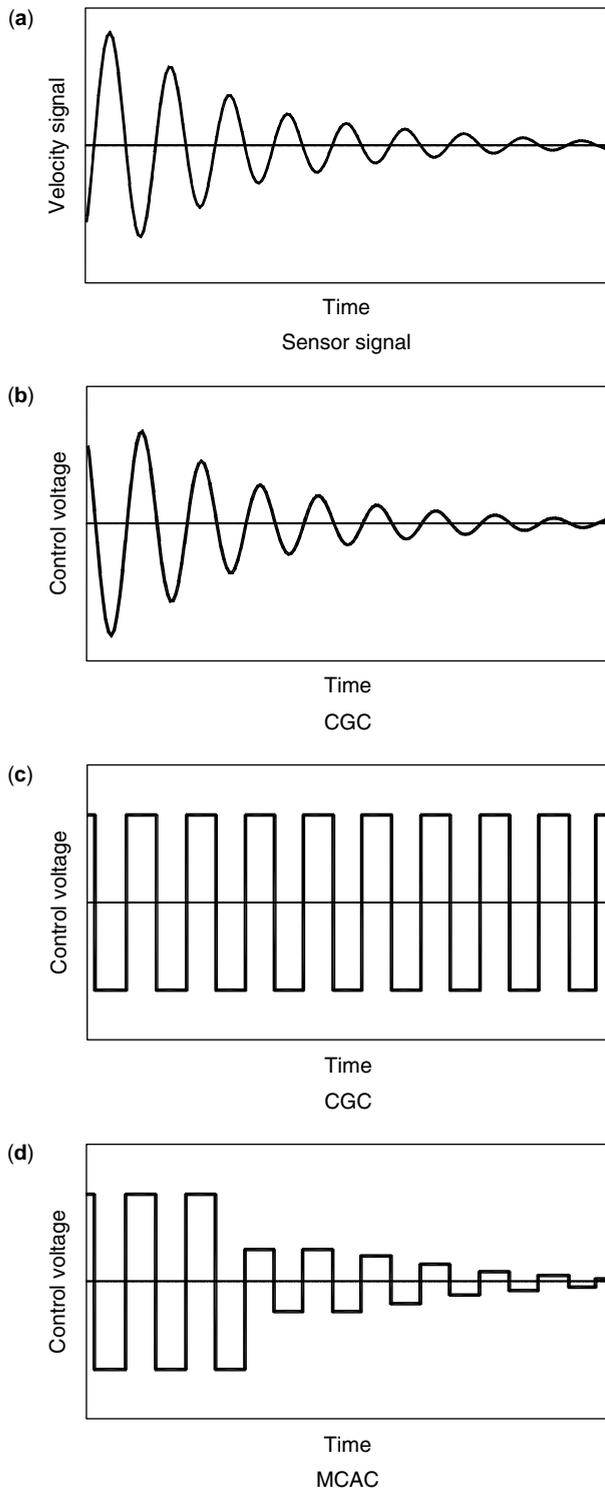


Figure 8. Vibrational control algorithms for a smart structure that features a piezoactuator and sensor.

oscillation and inevitable time delay of the hardware system. However, chattering was fairly well eliminated by employing the MCAC algorithm. This implies that the MCAC produces a relatively small adverse control force for the time delay in the settled phase. Note that the feedback signal from the piezofilm sensor represents the angular

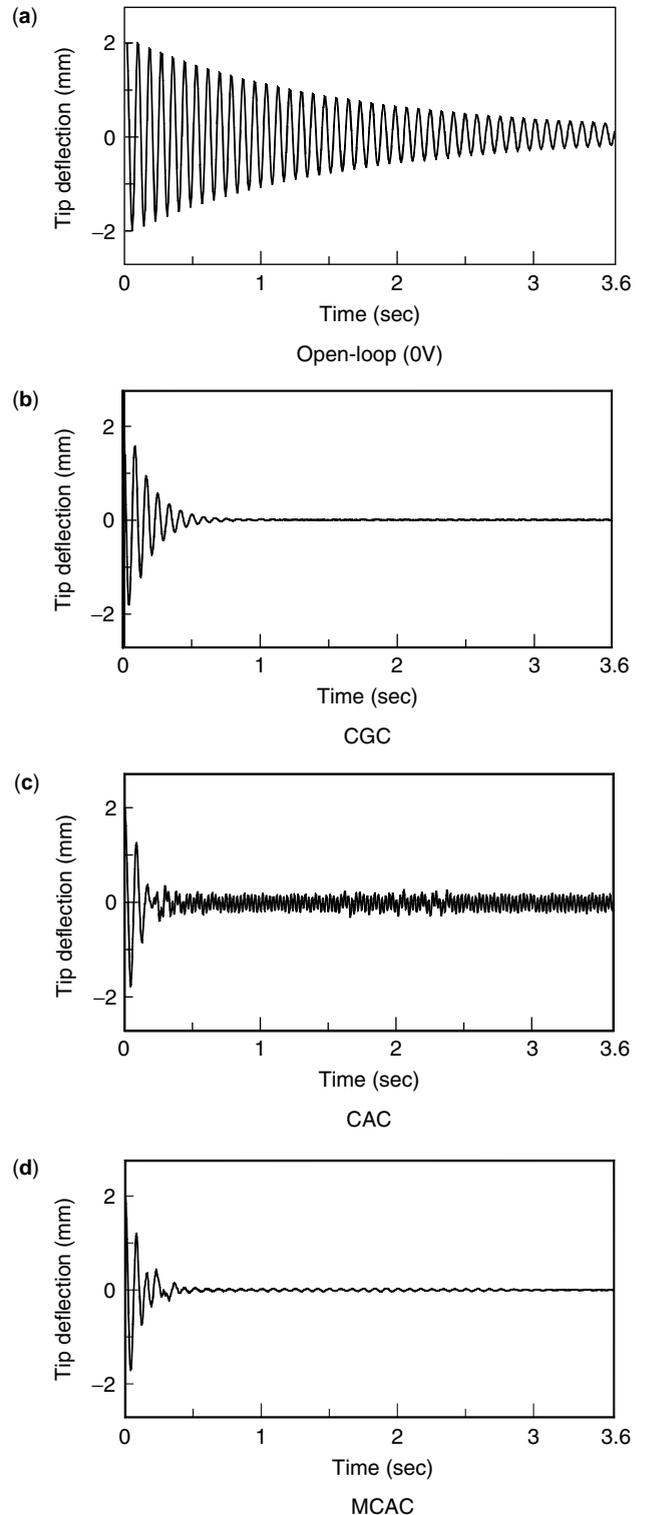


Figure 9. Transient vibrational responses of a cantilevered beam that features a piezoactuator and sensor.

displacement at the tip of the cantilevered beam. Thus, the distributed-parameter sensor catches the corresponding response caused by all of the vibrational modes. Furthermore, the CGC, the CAG, and the MCAC are derived without modal truncation of the plant model. These inherent

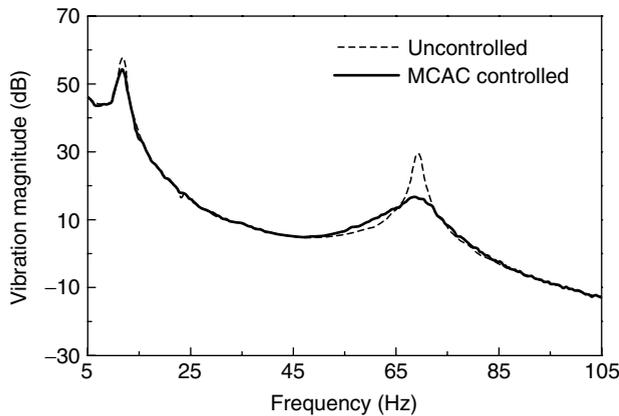


Figure 10. Simultaneous controllability of various vibrational modes in a piezofilm-based smart beam.

characteristics of the distributed sensor and control laws allow one the possibility of controlling all transverse vibrational modes at once, hence avoiding problems of spillover of uncontrolled vibrational modes (16). Figure 10 presents the measured transfer function, which is obtained from the ratio of the excitatory input force measured by the accelerometer to the tip deflection measured by the piezofilm sensor (21). It is clearly observed that both the first and second modes are effectively controlled by applying the MCAC algorithm without causing spillover problems.

The vibrational control of the piezoelectric-based smart beam structures can be extended without difficulties to vibrational control of the plate or shell structures. In the vibrational control of large structures, determining the optimal location for piezoactuators or/and sensors is very important to suppress effectively unwanted vibration caused by random disturbances which lead to exciting several mode shapes simultaneously. Furthermore, in practice, large flexible structures can easily be subjected to parameter variations such as natural frequency. Therefore, a robust control algorithm should be formulated for the piezoactuators to protect the robust vibrational control performance from these system uncertainties.

VIBRATIONAL CONTROL OF SMART SYSTEMS

Vehicle Suspension Using ER Damper

Recently, a great deal of attention was focussed on a damper design that significantly suppressed the vibration of a vehicle system (27). The vehicle vibration was to be attenuated for various road conditions. This is normally accomplished by employing a suspension system. So far, three types of suspensions were proposed and successfully implemented: passive, active, and semiactive. The passive suspension system that features an oil damper (or shock absorber) is simple to design and cost-effective. However, performance limitations are inevitable. On the other hand, the active suspension provides high control performance across a wide frequency range. However, the active suspension requires large power, sources, many sensors, servovalves, and sophisticated control logic. One way to resolve these requirements of the active suspension system

is to use a semiactive suspension system. The semiactive suspension system offers desirable performance that is generally enhanced in the active mode without large power sources and expensive hardware. Recently, a very attractive and effective semiactive suspension system featuring ER fluids was proposed by many investigators (28–34).

One of the salient properties of an ER fluid is its responds fast to an electric field, and hence it has a wide control bandwidth. This inherent feature has triggered tremendous research activities in the development of various engineering applications including dampers for controlling the vibration of vehicles. Sturk et al. (27) proposed a high-voltage supply unit that has an ER shock absorber and proved its effectiveness via a quarter-car suspension system. Nakano (28) constructed a quarter-car suspension model using an ER damper and proposed a proportional control algorithm to isolate vibration. Petek et al. (29) constructed a semiactive full suspension system that uses four ER dampers and evaluated the suspension performance by implementing a skyhook control algorithm that considers the heave, pitch, and roll motions of the car body. Gordaninejad et al. (30) proposed a cylindrical ER damper that has multielectrodes and proved its favorable capability for vibration control by implementing a bang-bang and a linear proportional controller. Sims et al. (31) proposed an ER valve-controlled vibrational damper, and obtained the linear behavior of the damping force with respect to the velocity by using a proportional feedback control gain. Peel et al. (32) proposed a long-stroke ER damper for effective vibrational control. Choi et al. (33) proposed a cylindrical ER damper for a passenger car, and proved its controllability of damping force by implementing a skyhook controller. Recently, Choi et al. (34) developed a sliding mode controller for a full car suspension in using four ER dampers. They constructed a full-car model, and evaluated its vibrational control performance via the hardware-in-the-loop simulation. The field test for the ER suspension system has also been undertaken (35).

In this article, a cylindrical ER damper shown in Fig. 11 is introduced to evaluate the vibrational control

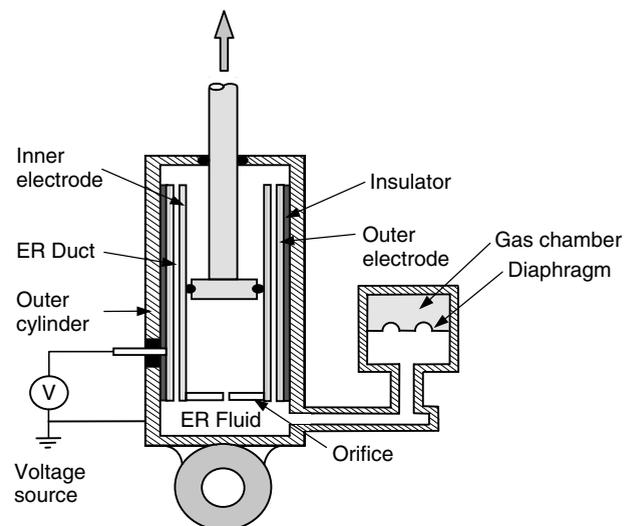


Figure 11. Schematic configuration of an ER damper.

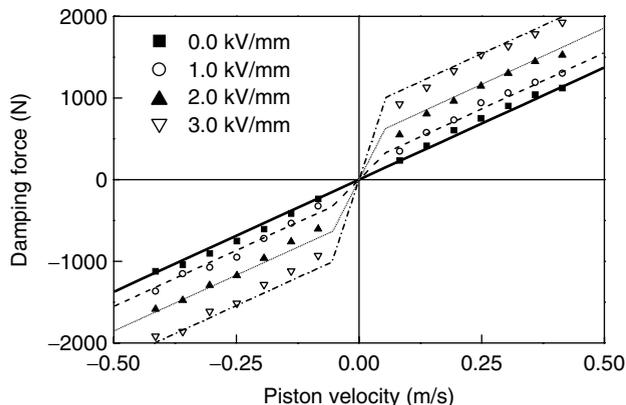


Figure 12. Field-dependent damping force of an ER damper.

performance in a passenger vehicle. The ER damper is divided into upper and lower chambers by a piston, that is filled with ER fluid. The ER fluid flows by the piston's motion through the duct between the inner and outer cylinders from one chamber to the other. A positive voltage is produced by a high-voltage supply unit connected to the inner cylinder, and the negative voltage is connected to the outer cylinder. The gas chamber located outside the lower chamber acts as an accumulator of the ER fluid induced by the piston's motion. If no electric field is applied, the ER damper produces a damping force caused only by fluid resistance. However, if a certain level of the electric field is supplied to the ER damper, the ER damper produces an additional damping force owing to the yield stress of the ER fluid. This damping force of the ER damper can be continuously tuned by controlling the intensity of the electric field. The damping force F of the ER damper shown in Fig. 11 can be obtained as follows (34): $F = k_e X_P + c_e \dot{X}_P + F_{ER}$. The variable k_e is the effective stiffness due to gas pressure, c_e is the effective damping due to the fluid viscosity, X_P is the excitation displacement, and F_{ER} is the field-dependent

damping force which is tunable as a function of applied electric field E . By adopting the Bingham model for the ER fluid, the controllable damping force F_{ER} can be expressed as $F_{ER} = (2L/h)(A_P - A_r)\alpha E^\beta \text{sign}(\dot{X}_P)$. A_P and A_r represent piston and piston rod area, respectively, L is the electrode length, h is the electrode gap, and α and β are intrinsic values of the ER fluid to be experimentally evaluated.

Figure 12 presents the measured damping force of a cylindrical ER damper for a passenger vehicle (35). As seen in the figure, the damping force increases as the electric field increases. For instance, the damping force is increased up to 1000 N at a piston velocity of 0.25 m/s by applying an electric field of 3 kV/mm. Note that the level of the damping force of a conventional passive oil damper is almost the same as this one at 0 kV/mm. Thus, we can expect improved suspension performance of the vehicle system by controlling the damping force. To evaluate the vibrational control performance of the vehicle system using the ER damper, we can construct a closed-loop control vehicle system, as shown in Fig. 13. A portable computer (microprocessor) equipped with a DSP (digital signal processor) board is normally positioned beside the driver's seat. Four pairs (one for the car body and the others for the wheels) of accelerometers are installed on four independent suspensions to measure the vertical motions of the vehicle. The signals from the accelerometers, gyroscope, and LDT (linear differential transformer) are fed back to the microprocessor, and depending upon the control algorithm employed, the required control input voltages are determined and applied to the four ER dampers through four high-voltage amplifiers positioned at four corners in the trunk. Among many controllers are candidates for the vehicle suspension, the skyhook control algorithm, which can be easily implemented, is frequently adopted and given as follows: $u_i = C_i |\dot{z}_{si}|$, for $\dot{z}_{si}(\dot{z}_{si} - \dot{z}_{usi}) > 0$; $u_i = 0$, for $\dot{z}_{si}(\dot{z}_{si} - \dot{z}_{usi}) < 0$. The variable u_i is the control damping force F_{ER} , \dot{z}_{si} denotes the vertical velocity of the car body, and \dot{z}_{usi} represents the

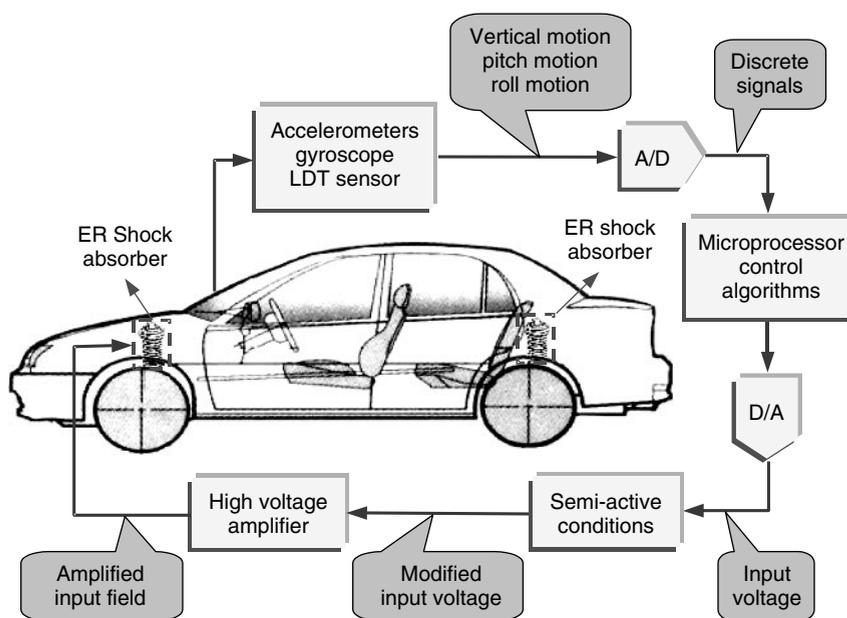


Figure 13. Vehicle configuration for an ER suspension test.

vertical velocity of the wheel. The control gain C_i needs to be determined depending upon the road excitation. In the final stage for practical use, the high-voltage amplifier should have short response time and should be integrated with an electronic control unit (ECU). Note that once the control input u_i is determined, the control electric field to be applied to the ER damper is obtained from the relationship between the electric field and the damping force.

The control characteristics for suppressing the vibration of the full-car suspension system are evaluated under two types of road excitation. The first excitation normally used to reveal the transient response characteristic is a bump. In bump excitation, the vehicle travels over the bump at a constant velocity of 3.08 km/h ($= 0.856$ m/s). The second type of road excitation normally used to evaluate the frequency response is a stationary random process. In random excitation, the values of road irregularity are chosen assuming that the vehicle travels on a paved road at a constant velocity of 72 km/h ($= 20$ m/s). Figure 14a presents the temporal responses of the ER suspension system to the bump excitation (34). It is generally known that vertical acceleration of the sprung mass and tire deflection are used to evaluate the ride comfort and the road holding of the vehicle, respectively. It is seen that both vertical acceleration of the sprung mass and tire deflection are substantially reduced by employing the control electric field. This implies that the ER suspension system can simultaneously provide both good ride comfort and driving safety to a driver by applying a control electric field to the ER dampers. Figure 14b presents frequency responses to random excitation (34). The frequency responses are obtained from the power spectral density (PSD) for the suspension travel and tire deflection. As expected, the power spectral densities for the suspension travel and tire deflection are substantially reduced in the neighborhood of body resonance (1–12 Hz). It is also observed that tire deflection is substantially reduced at wheel resonance (10–15 Hz). This indicates significant enhancement of the steering stability of the vehicle.

Note that most currently employed control algorithms for vibrational attenuation that use an ER fluid-based actuator are dubbed semiactive. The semiactive control system offers desirable performance generally enhanced in the active mode without requiring large power sources. One of the most popular control logics for the semiactive control system is the skyhook control algorithm because it is easy to formulate and implement in practice. Possible candidates for active controllers for the semiactive control system are the sliding mode control, neural network control, Lyapunov-based state feedback control, and optimal control. However, because the semiactive actuator cannot increase the mechanical energy of the control system, special attention (semiactive conditions in Fig. 13) should be given when these active control strategies are adopted. On the other hand, we can construct an active control system using an ER fluid by employing a hydraulic, closed-loop, ER valve–cylinder system. In this case, control logics adapted to conventional hydraulic servomechanism can be applied without any modification. The only difference is replacing the electromagnetic servovalve by the ER valve.

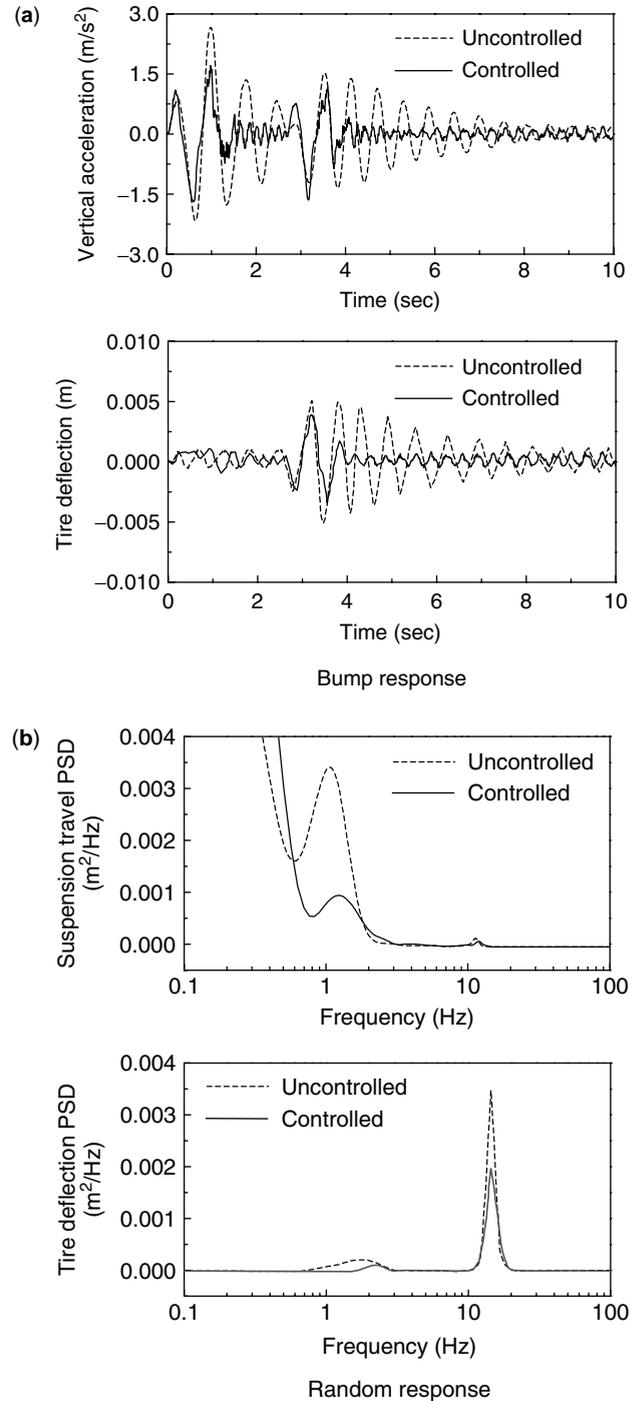


Figure 14. (a) Bump and (b) random responses of a passenger vehicle using ER dampers.

Flexible Manipulator That Features Piezoactuators

Though flexible robotic manipulators have some inherent advantages over conventional rigid robots, they have posed more stringent requirements on the control system design, such as accurate end-point sensing and fast suppression of transient vibration during rapid arm movements. Furthermore, model parameter variations such as natural frequencies and damping ratios may easily arise in practice due to a wide spectrum of various conditions in the design

and manufacturing process, dynamic modeling, and operating conditions. Numerous control strategies for flexible manipulators have been proposed in an attempt to find a successful and practical feedback control. Many of the previously proposed control strategies are based on optimal control theory (36,37). A few investigators strove to achieve effective control logics that accounted for the sensitivity of the control to parameter variations and extraneous disturbances. A robust control that guarantees stable system performance for all possible variations of the parameters was designed by employing the properties of the uniformly and ultimate uniformly boundedness of solving the system state equation (38). There were also several studies on sliding mode controllers (39,40) and an H_∞ controller (41) for the feedback control of flexible manipulators subjected to system uncertainties. The input torque of the motor in most of these control techniques for flexible manipulators is determined by simultaneously considering both the rigid body mode and finite elastic modes. The successful experimental realization of this type of torque may be very difficult under certain conditions due to hardware limitations such as saturation of the motor, computer speed, and signal noises from the motor and sensors. Furthermore, so-called spillover problems will occur because only some finite elastic modes are considered for controlling a distributed parameter system of infinite order. Other problems that plague existing conventional control methods include accurate estimation or measurement of state variables and the complexity of the control algorithm which makes on-line implementation infeasible.

Recently, a hybrid actuator control scheme that consists of two types of actuators, motors mounted at the hubs and piezoactuators bonded to the surface of the flexible links, has been proposed to resolve some of the existing problems and hence, to achieve accurate end-point position by suppressing unwanted vibration (42–44). In this article, this control technique is introduced by considering the elastodynamic flexural response in the horizontal plane (no gravity effect) of a two-link flexible manipulator that features surface-bonded piezoceramics (PZT) and piezofilms, as shown in Fig. 15. Piezoceramics on the right faces play the role of actuators, and piezofilms have the role of distributed sensors to measure elastic deflections caused by vibrational modes. The arm consists of two links connected by a revolute joint. Two links are normally modeled as a

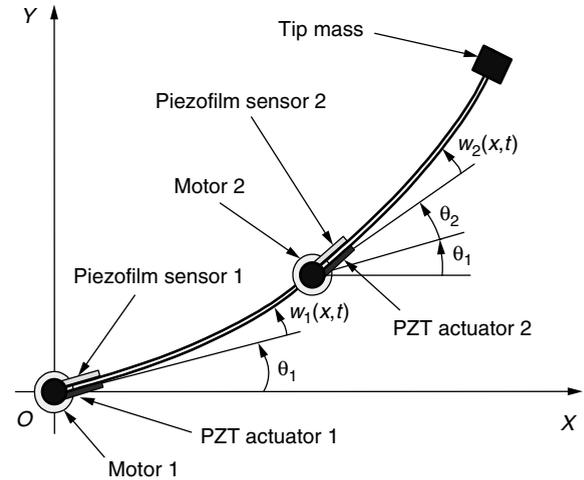


Figure 15. A two-link flexible manipulator that features piezoelectric actuators and sensors.

continuous and uniform beam. It is also generally assumed that the beams are flexible only in the direction transverse to their length in the plane of motion, so that there is no out-of-plane deflection or axial elongation of the links as the arms move. The first flexible link is clamped on the hub of the shoulder motor, and the second flexible link is clamped on the hub of the elbow motor at one end and has a concentrated tip mass at the other end. The motor torque that produces a desired angular position is obtained by employing the sliding mode controller on the rigid-link dynamics that have the same mass as that of the flexible link. Then, the torque is applied to the flexible manipulator to activate the commanded motion. However, during this control action, undesirable oscillations $w_i(x, t)$ occur due to the applied torques based on rigid-link dynamics. Subsequently, these vibrations are to be suppressed by applying the feedback voltage to the piezoceramic actuators. As a result, the desired tip motion is achieved favorably. Figure 16 presents a control block diagram that features hybrid actuators, motors and piezoceramics. In the figure, T_{fi} is the feed-forward term that compensates for the torque disturbance $[d_i(t)]$. The control torque T_i is determined so that the actual angular motion $\theta_i(t)$ tracks the desired motion $\theta_{id}(t)$ well. Because the design procedure for the sliding mode controller for the control torque T_i is the same as

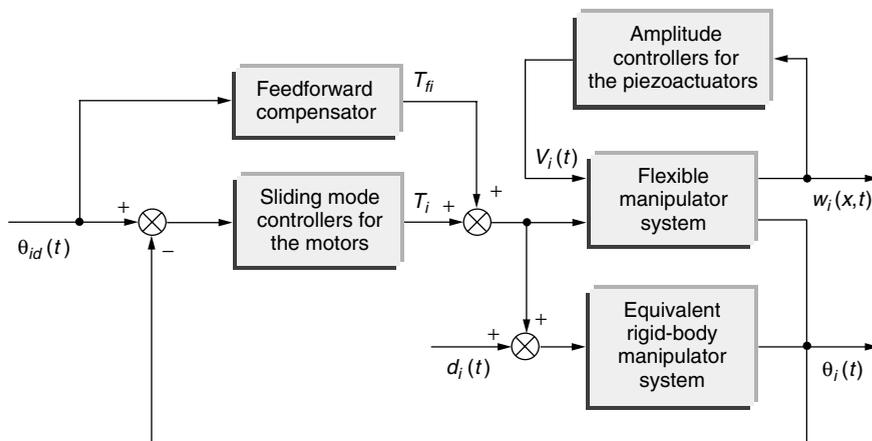


Figure 16. Control block diagram for a two-link flexible manipulator that features piezoactuators.

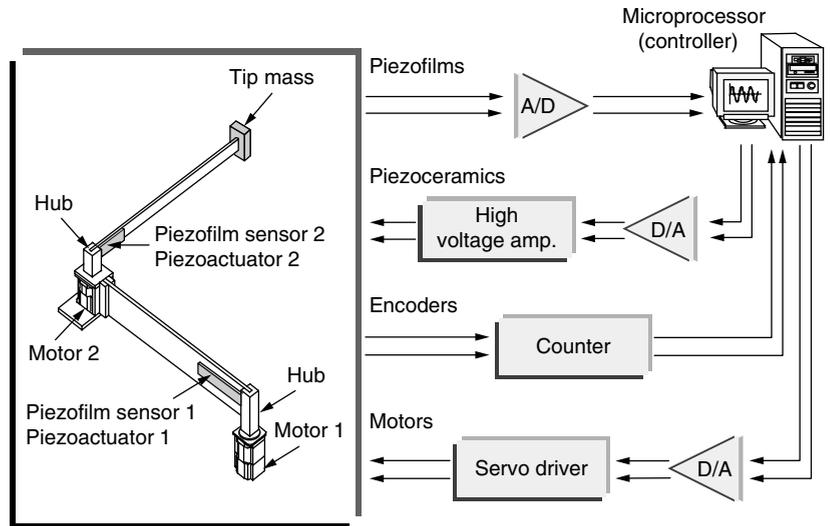


Figure 17. Schematic diagram of the experimental apparatus for end-point control.

that for the rigid-link robot, we can explain only the formulation of the controller for the piezoactuator. As mentioned earlier, one of the potential controller candidates for achieving favorable vibrational control of piezoelectric-based flexible structures is a constant-amplitude controller as follows: $V_i(t) = -K_i \cdot \text{sign}[V_{\dot{f}_i}(t)]$, $i = 1, 2$. Here, K_i is a feedback gain, and $V_{\dot{f}_i}(t)$ is the time derivative of the output signal voltage $V_{f_i}(t)$ from the piezofilm sensor bonded to the other surface of the flexible link. The output voltage produced by the piezofilm sensor is obtained by integrating the electric charge developed at a point on the piezofilm along the entire length of the film surface. The feedback gain K_i of the controller for the piezoactuator is chosen by considering the material property of the piezoceramic actuator as well as the geometry of the flexible link. Furthermore, the feedback gain should be determined so that the flexible manipulator system is stable. To investigate the stability of the system, we normally adopt a positive definite Lyapunov function that is basically a measure of the potential and kinetic energy of the system. The function is given as follows (43): $2F = \dot{\mathbf{z}}^T \mathbf{M} \dot{\mathbf{z}} + \mathbf{z}^T \mathbf{K} \mathbf{z}$. Here, \mathbf{M} is the system mass matrix, \mathbf{K} is the system stiffness matrix associated with the link elasticity, and \mathbf{z} is the generalized coordinate vector that consists of the angular displacement θ_i and the elastic deflection $w_i(x, t)$. We can guarantee the stability of a flexible manipulator system by choosing a proper feedback gain K_i that makes the time derivative of the Lyapunov function negative-definite. However, the stability of a flexible manipulator system can be violated by fast motions or by the acceleration phase of the hubs which in turn result in large oscillations of the flexible links. It is known that when the hub motions are in the deceleration phase, the stability of the system is satisfied by employing any positive feedback gain K_i . On the other hand, if the hub motions completely stop, the flexible links can be treated just as cantilever beams. Thus, in this case, Lyapunov stability is also satisfied by employing any positive feedback gain K_i . The proper determination of K_i normally depends on the magnitudes of the elastic vibration and angular velocity.

Noted that the controller for the piezoactuator is formulated on the basis of the distributed parameter model without truncating the vibrational mode. This allows one

the possibility of simultaneously controlling all of the vibrational modes. Therefore, the control spillover problem, which may occur in the truncated model from uncontrolled vibrational modes, can be avoided. We also note that the discontinuous property may cause undesirable chattering associated with time delay and hardware limitations in the actual implementation of constant-amplitude controllers. To remove the chattering effectively, we may use a so-called multistep amplitude controller that proportionally tunes the magnitude of the control voltage according to the output signal (21). In practice, we can measure the angular displacements by built-in optical encoders in the motors and the elastic deflections by the distributed piezofilm sensors. Therefore, we see that no state estimator, which may be inevitably necessary in most of the conventional control methods, is needed for implementing the hybrid actuator control scheme.

Figure 17 presents a typical experimental apparatus for implementing the hybrid actuator control scheme. The displacements of the motors are obtained from the optical encoders and sent to the microprocessor through the encoder board. The vibrational signals of the flexible links are measured by the piezofilm sensors and fed back to the microprocessor through the low-pass analog filter and A/D converter. Input torques determined from the sliding mode controller are applied to the motor through a D/A converter and a servodriver, and the input voltages determined from the constant-amplitude controller are supplied to the piezoceramic actuator through the D/A converter and high-voltage amplifier. Figure 18 presents the elastic deflections of the two-link flexible manipulator during the regulating control action (43). It is clear that the unwanted deflections are significantly reduced by applying feedback voltages to the piezoceramic actuators. This result directly indicates that the undesirable tip deflection of each flexible link can be effectively suppressed by employing the hybrid actuator control strategy that features servomotors and piezoactuators. Note that the deflection of each link could be reduced more by increasing the feedback gain K_i of the constant-amplitude controller. However, in this case, the breakdown voltage of the piezoactuator should be considered.

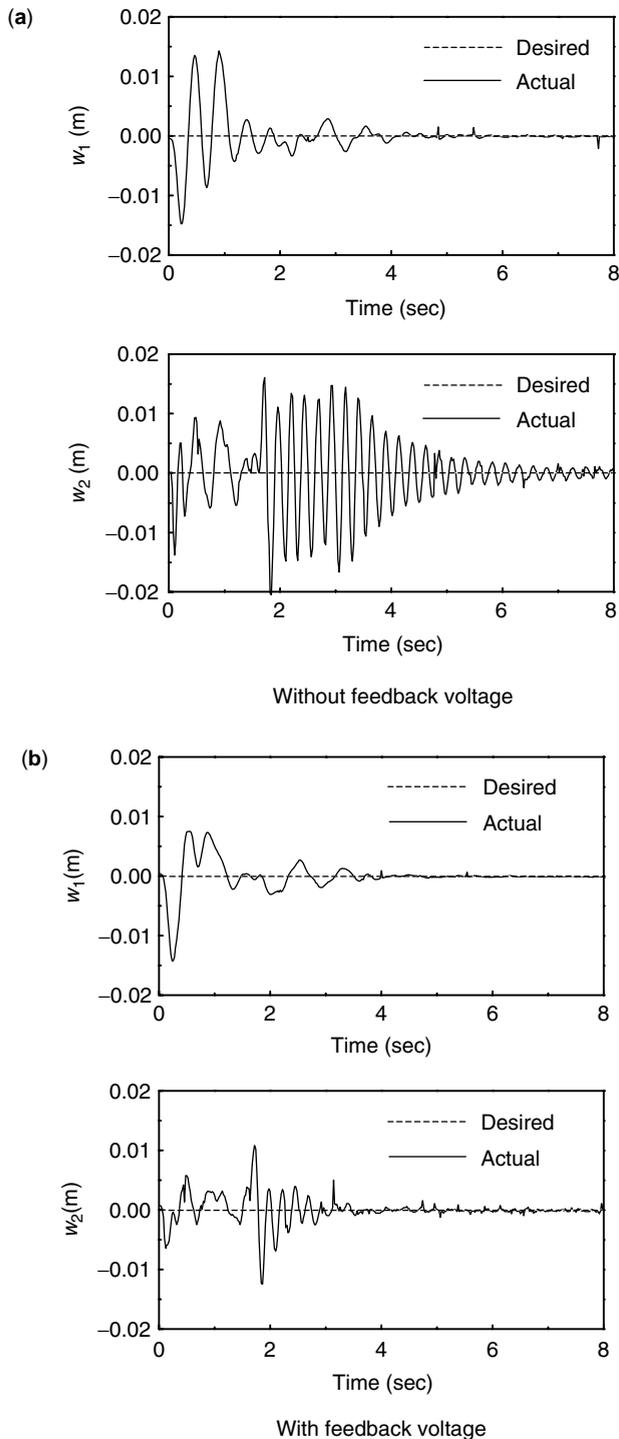


Figure 18. Vibrational magnitudes of a two-link flexible manipulator.

BIBLIOGRAPHY

1. M.V. Gandhi and B.S. Thompson, *Smart Materials and Structures*. Chapman & Hall, London, 1992.
2. S.B. Choi, B.S. Thompson, and M.V. Gandhi, *Proc. Damping '89 Conf.*, West Palm Beach, FL, Feb. 1988, **1**, pp. CAC.1–CAC.14.
3. M.V. Gandhi, B.S. Thompson, S.B. Choi, and S. Shakir,

4. *ASME J. Mech. Transmissions Autom. Design* **111**(3): 328–336 (1989).
5. S.B. Choi, Y.K. Park, and J.D. Kim, *Int. J. Mech. Sci.* **35**(9): 757–768 (1993).
6. Y. Choi, A.F. Sprecher, and H. Conrad, *J. Intelligent Mater. Syst. Struct.* **3**: 17–29 (1992).
7. J.P. Coulter and T.G. Duclos, *Proc. 2nd Int. Conf. ER Fluids*, Raleigh, NC, Aug. 1989, pp. 300–325.
8. C.D. Rahn and S. Joshi, *Modeling and Control of an Electrorheological Sandwich Beam, Active Control of Vibration and Noise*. ASME Publication **75**, New York, NY, pp. 159–167.
9. S.O. Oyadiji, *J. Intelligent Mater. Syst. Struct.* **7**: 541–549 (1996).
10. S.B. Choi, Y.K. Park, and C.C. Cheong, *J. Intelligent Mater. Syst. Struct.* **7**(4): 411–419 (1996).
11. H. Gong and M.K. Lim, *J. Intelligent Mater. Syst. Struct.* **8**: 401–413 (1997).
12. M. Yalcintas and J.P. Coulter, *Smart Mater. Struct.* **7**(1): 128–143 (1998).
13. S.B. Choi and Y.K. Park, *J. Sound Vib.* **172**(3): 428–432 (1994).
14. D.J. Mead and S. Markus, *J. Sound Vib.* **10**(2): 163–175 (1969).
15. H.S. Tzou and G.L. Anderson, *Intelligent Structural Systems*. Kluwer Academic, London, 1992.
16. S.B. Choi, Y.K. Park, and S.B. Jung, *J. Aircraft* **36**(2): 458–464 (1999).
17. T. Bailey and J.E. Hubbard, Jr., *J. Guidance, Control Dynamics* **8**(5): 605–611 (1985).
18. A. Baz and S. Poh, *J. Sound Vib.* **126**(2): 327–343 (1988).
19. H.S. Tzou and M. Gadre, *J. Sound Vib.* **136**(3): 477–490 (1990).
20. H.S. Tzou, *ASME J. Dynamic Syst. Meas. Control* **113**: 494–499 (1991).
21. A. Baz, S. Poh, and J. Fedor, *ASME J. Dynamic Syst. Meas. Control* **114**(1): 96–103 (1992).
22. S.B. Choi, C.C. Cheong, and S.H. Kim, *J. Intelligent Mater. Syst. Struct.* **6**(5): 430–435 (1995).
23. S.B. Choi and M.S. Kim, *J. Guidance Control Dynamics* **20**(5): 857–864 (1997).
24. S.M. Yang and G.S. Lee, *ASME J. Dynamic Syst. Meas. Control* **119**(1): 34–39 (1997).
25. J.L. Meyer, W.B. Harrington, B.N. Agrwal, and G. Song, *Smart Mater. Struct.* **7**: 95–104 (1998).
26. J. Tang, K.W. Wang, and M. Philen, *Proc. SPIE Conf. Smart Struct. Integrated Syst.* Newport Beach, CA, Mar. 1999, **3668**, pp. 543–554.
27. S.B. Choi, S.S. Cho, and Y.P. Park, *ASME J. Dynamic Syst. Meas. Control* **121**(1): 27–33 (1999).
28. M. Sturk, X.M. Wu, and J.Y. Wong, *Vehicle Syst. Dynamics* **24**: 101–121 (1995).
29. M. Nakano, *Proc. 5th Int. Conf. ER Fluids MR Suspensions Assoc. Technol.* (1996), Sheffield, UK, pp. 645–653.
30. N.K. Petek, D.J. Romstadt, M.B. Lizell, and T.R. Weyenberg, *SAE Technical Paper Series 950586*, Detroit, MI, 1995.
31. F. Gordaninejad, A. Ray, and H. Wang, *ASME J. Vib. Acoust.* **119**: 527–531 (1997).
32. N.D. Sims, R. Stanway, and S.B. Beck, *J. Intelligent Mater. Syst. Struct.* **8**: 426–433 (1997).
33. D.J. Peel, R. Stanway, and W.A. Bullough, *Proc. SPIE Smart Struct. Mater.* San Diego, 1998, pp. 416–427.
34. S.B. Choi, Y.T. Choi, E.G. Chang, S.J. Han, and C.S. Kim, *Mechatronics* **8**(2): 143–161 (1998).
35. S.B. Choi, Y.T. Choi, and D.W. Park, *ASME J. Dynamic Syst. Meas. Control* **122**(1): 114–121 (2000).

35. H.K. Lee, M.S. Thesis, Department of Mechanical Engineering, Inha University, Korea, 1999.
36. E. Schmitz, Ph. D. Thesis, Department of Aeronautics and Astronautics, Stanford University, 1985.
37. Y. Sakawa, F. Matsuo, and S. Fukushima, *J. Robotic Syst.* **2**(4): 453–472 (1985).
38. S.B. Choi, C.C. Cheong, B.S. Thompson, and M.V. Gandhi, *J. Sound Vib.* **162**(2): 195–208 (1993).
39. K.S. Yeung and Y.P. Chen, *Int. J. Control* **52**(1): 101–117 (1990).
40. S.B. Choi, C.C. Cheong, and H.C. Shin, *J. Sound Vib.* **179**(5): 737–748 (1995).
41. A. Tchernychev, A. Sideris, and J. Yu, *ASME J. Dynamic Syst. Meas. Control* **119**: 206–211 (1997).
42. S.B. Choi and H.C. Shin, *J. Robotic Syst.* **13**(6): 359–370 (1996).
43. H.C. Shin, Ph. D. Dissertation, Department of Mechanical Engineering, Inha University, Korea, 1998.
44. D. Sun and J.K. Mills, *Proc. IEEE Int. Conf. Robotics Autom.* Detroit, MI, 1999, pp. 849–854.

VIBRATION CONTROL IN SHIP STRUCTURES

T.S. KOKO

Martec Limited
Halifax, NS
Canada

U.O. AKPAN

Martec Limited
Halifax, NS
Canada

A. BERRY

GAUS, University de Sherbrooke
Sherbrooke, Quebec
Canada

P. MASSON

GAUS, University de Sherbrooke
Sherbrooke, Quebec
Canada

O. BESSLIN

GAUS, University de Sherbrooke
Sherbrooke, Quebec
Canada

J. SZABO

DREA, Dartmouth, NS
Canada

N. SPONAGLE

DREA, Dartmouth, NS
Canada

INTRODUCTION

The dominant sources of noise radiation in water are from ship engines and machinery—the propeller cavitation noise, the noise radiation from propeller blades, and the hydrodynamic pressure fluctuations induced by turbulent water flow along the ship's hull. At speeds below propeller cavitation inception, a ship's acoustic signature is generally dominated by structurally transmitted noise from onboard machinery. Reduction or control of ship noise

has traditionally been implemented by passive means, such as by the use of vibration isolation mounts, flexible pipe-work, and interior acoustic absorbing materials. However, these passive noise control techniques are effective mostly for attenuating high-frequency noise; they are generally ineffective for controlling low-frequency noise. There are, on the other hand, active noise control methods that have been proven to be effective in controlling low-frequency and tonal noise. These active control methods could be used instead of, or in combination with, passive techniques, for controlling or reducing ship noise.

Active noise control (ANC) involves the reduction or elimination of noise by modification of the dynamic properties of a system or by noise cancellation through linear superposition of a secondary noise field of equal but opposite strength. An active noise control system will typically consist of all or some of the following ingredients: sensors, actuators, and controllers.

FUNDAMENTAL CONCEPTS OF SHIP NOISE CONTROL

Noise Sources and Transmission Paths in Ship Structures

There are many sources of noise within a ship's structure. Among these are the propulsion systems, exhaust stacks, and various onboard equipment. The principal noise source is the engine system. A typical ship engine along with its mounting system is schematically depicted in Fig. 1. The figure shows the various vibroacoustic paths through which the engine vibration is transmitted to the ship structure and eventually radiated into the surrounding mediums. The various vibroacoustic paths transmit noise in different ways. For example:

- The noise from the exhaust stack and the fuel intake and the cooling systems can be viewed as duct and piping noise. In this mechanism the pressure wave in the duct is excited and transmitted as noise.
- The mounting systems, consisting of the engine cradle, isolation mount, raft, and foundation, are mechanical connections between the ship hull and the machine. Vibration is transmitted from the engine motion to the ship hull through these connections. The induced hull vibration is transmitted to the surrounding medium and is radiated as acoustic noise. This noise transmission mechanism is referred to as structural acoustic radiation.
- The engine vibration leads to airborne radiation within the enclosure, which may induce an acoustic load on the ship hull. This resulting excitation is radiated to the surrounding medium as acoustic noise.

The objective of ship noise control is the minimization of the acoustic radiation from the ducting and piping systems and from the ship's hull, and appendages to the surrounding water.

Passive and Active Ship Noise Control

In general, then, passive and active control methods are two distinct methods that can be used to reduce acoustic

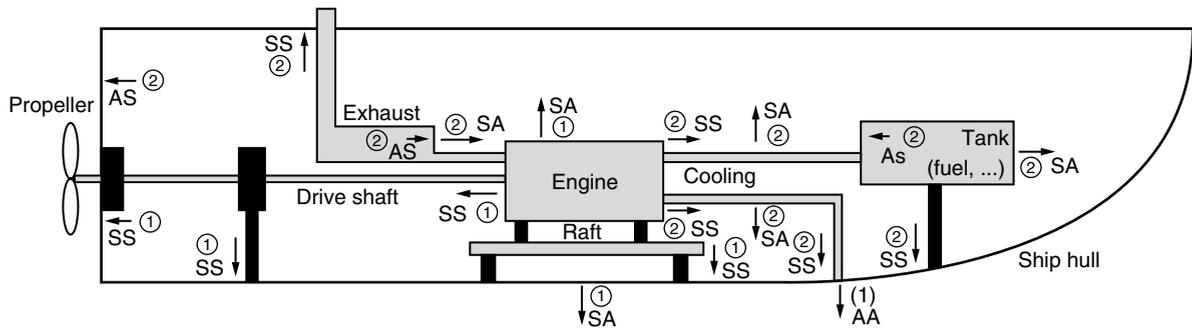


Figure 1. Typical marine diesel engine mounted on a ship hull. AA: acoustic to acoustic coupling, SS: structural to structural coupling, AS: acoustic to structural coupling, SA: structural to acoustic coupling. The relative importance of energy coupling for radiation into seawater is illustrated by a number (1) for more important and (2) for less important.

noise and radiation. Passive noise control essentially reduces unwanted noise by utilizing the absorption property of materials. In this approach, sound absorbent materials are mounted on or around the primary source of noise or along the acoustic paths between the source and the receivers of noise. At low frequencies, however, passive control techniques are not effective because the long acoustic wavelength of the noise requires large volumes of the passive absorbers (1)

Active noise control involves the injection of secondary sound by actuators, which by linear superposition is additive, to the primary sound field. It operates on the principle of superposing waveforms, by generating a canceling waveform whose amplitude and envelope match those of the unwanted noise, but whose phase is shifted by 180° (2). The main features of an active control system are illustrated in Fig. 2. The basic components are the physical system (this encompasses the plant, the sensors and the actuators) and the electronic control system (3). The main features are:

1. The primary source of noise/disturbance and the system to be controlled. This is usually referred to as the plant.
2. The input and error sensors. The input sensors are the electroacoustic (microphones) or electromechanical (accelerometers, tachometers) devices that measure the disturbance from the primary source and communicate it to the controller. They are often referred to as reference sensors. The error

sensor monitors the performance of the active controller.

3. The actuators. These are the electroacoustic or electromechanical devices that generate the secondary noise or anti-noise in order to reduce or cancel the primary noise. In some cases, the actuators modify the dynamic properties of the system in order to reduce their noise radiation efficiencies. Examples of actuators include speakers, piezoelectric material, and vibration shakers. The actuators, plant, and the sensors are collectively referred to as the physical system.
4. The active controller. This is the signal processor (usually a digital electronic system) that gives command to the actuators. The controller bases its output on sensor signals (primary noise sensor/error sensor) and usually on some knowledge of how the plant responds to the actuator.

The performance of an active controller depends on the physical arrangements of the control sources (actuators) and the sensors, causality, controllability, observability, and the stability of the control system. Active noise control methodologies (ANC) can be classified into two main categories, namely feedforward control (FFC) and feedback control (FBC). A summary of the description of the two methodologies is given by (4). The controllers that have been used in active noise control methodologies (FFC and FBC) have evolved over the years from analog to digital designs.

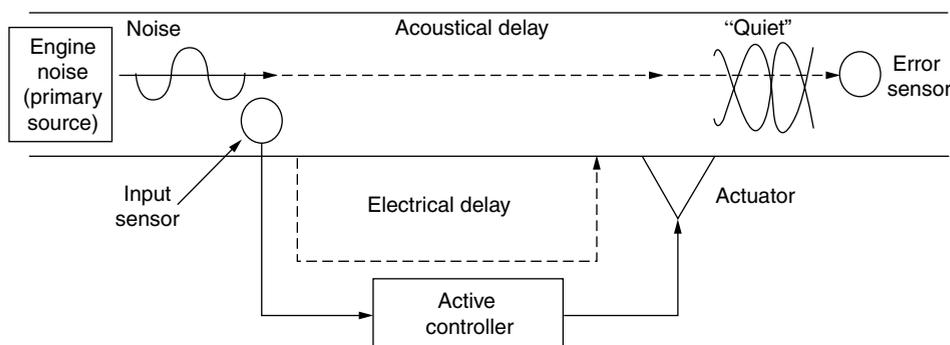


Figure 2. Main features of an active noise control system.

SENSORS AND ACTUATORS FOR ACTIVE NOISE AND VIBRATION CONTROL (ANVC)

Piezoelectric Materials

Piezoelectric materials are the oldest and most reliable materials used in high-speed sensor and actuator technologies. Piezoelectricity was discovered by the Curie brothers in 1880. When a piezoelectric material is subjected to a mechanical stimulus, an electrical charge or voltage is induced in the material. This is called the “direct piezoelectric effect,” which enables the material to be used as a sensor. On the other hand, when the piezoelectric material is subjected to an electrical charge or voltage, a mechanical force or strain is induced in the material. This is called the “converse piezoelectric effect,” which enables the material to be used as an actuator. The induced strain is directly proportional to the applied electric field and the linear piezoelectric constitutive equations are given by

$$T = cS - eE, \quad (1a)$$

$$D = e^T S + \varepsilon E, \quad (1b)$$

where T , E , S , and D are the vectors of stress, electric field, strain, and electric displacement (charge per unit area), c , e , and ε are the matrices of the elastic stiffness coefficients, piezoelectric stress constants, and dielectric coefficients, respectively. Equations (1a) and (1b) describe the direct and converse effects, respectively. Piezoelectric materials are also called soft ceramics because they are characterized by high dissipation factors (dielectric losses). As a result, they have high hysteresis in the displacement versus voltage curves (5).

The three most important piezoelectric materials are lithium niobate (LiNbO_3), polyvinylidene fluoride (PVDF or PVF2), and lead zirconate titanate (PZT) (6). LiNbO_3 is a crystal with a high electromechanical coupling and very low acoustical attenuation. Piezoelectricity is obtained from a strip of PVDF by stretching it under a high voltage. PVDF, originally discovered in 1969, is known for its flexibility, lightweight, durability, and relatively low acoustic impedance. PZT is by far the most commonly used piezoelectric material. This is a ferroelectric ceramic material with direct and converse piezoelectric properties. A wide variety of PZT formulations have been developed, with PZT-5 being one of the most widely used formulations for actuator applications (7–10). PZT can be used as sensors or actuators. For actuators, the device usually consists of a stack of many layers of the PZT, alternatively connected to the positive and negative terminals of a high voltage source (7).

The mechanical, dielectric and electromechanical coupling properties of some piezoelectric materials are shown in Table 1. Many studies, both theoretical and experimental, have been focused on the application of piezoelectric materials for vibration control of flexible structures (6,11–14). Of direct relevance to the noise control problem that is of interest in this study is the work by Sumali and Cudeny (15) who developed an actuator from a stack of layers of piezoelectric material in an actively controlled engine mount that was designed to reduce structural

vibrations. Most of the theoretical and experimental studies on the use of piezoelectric materials for active noise control have been directed at aircraft cabin noise control. For instance, Grewal et al. (16) have investigated the use of piezoceramic elements to reduce cabin noise in the Havillan Dash-8 series 100/200 aircraft. Their study shows that by judicious actuator and sensor design considerations systems using bonded piezoelectric actuators and vibration sensors alone are capable of simultaneously providing significant noise reduction as well as vibration suppression. Other studies include the works of Sutliff et al (17) on active noise control of low-speed fan rotor-stator modes; and Simonich (5) on the application of rainbow piezoceramic actuators (18) for active noise control of gas turbine engines.

Electrostrictive Materials

Electrostrictive materials are similar to piezoelectric materials. When a mechanical force or strain is applied to the material, an electric charge or voltage is induced; conversely, when an electric field is applied across an electrostrictive material, a mechanical strain is induced. Hence, electrostrictive material can also be used as sensors or actuators. However, there are several differences between electrostrictive and piezoelectric materials. In electrostrictive materials, the induced mechanical strain is proportional to the square of the electric field, whereas it is proportional to the electric field in piezoelectric materials. Thus electrostrictive materials always produce positive displacements regardless of the polarity. As a result, they are always in compression when doing work and avoid typical weakness of ceramics in tension (19). Electrostrictive materials exhibit microsecond recovery time upon withdrawal of the electric field, compared to milliseconds for piezoelectric materials. Electrostrictive materials have lower dissipation factors (and low displacements and hysteresis) compared to piezoelectric materials, and are regarded as hard ceramics (5). The most commonly used electrostrictive material is lead magnesium niobate (PMN) ceramic material.

Magnetostrictive Materials

Magnetostrictive materials are those materials that undergo an induced mechanical strain when subjected to a magnetic field. On the other hand, when a mechanical stress (or strain) is applied to the material, it undergoes domain changes that yield a magnetic field. These materials can thus be used as sensors and actuators due to the direct and converse effects. The most common magnetostrictive material is TERFENOL (consisting of TERbium, FE (iron) and dysprosium, which was developed by NOL, the Naval Ordnance Laboratory). The most commonly used formulation is TERFENOL-D. Magnetostrictive materials can exhibit strains of up to 0.2% at reasonably low magnetic field strength (20).

A detailed description of a magnetostrictive actuator is presented by Giugliuto et al. (7), who show the construction of a terfenol actuator. The actuator consists of a TERFENOL-D rod inside an electric coil that is enclosed

Table 1. Advantages and Disadvantages of Various Sensor and Actuator Technologies

Type of Sensor or Actuator	Advantages	Disadvantages
<p><i>Piezoelectric materials</i></p> <p>Examples: Lead zirconate Titanate (PZT) Polyvinylidene Fluoride(PVDF)</p>	<ul style="list-style-type: none"> • Used as sensors and actuators • Very large frequency range • Quick response time • Very high resolution and dynamic range • Possibility of integration in the structure for thin PZT actuators and PVDF sensors • Possibility of shaping PVDF sensors (spatial filtering) 	<ul style="list-style-type: none"> • Relatively low strain and low displacement capability (typically, less than 0.1% strain, and 1–100 microns displacement for stack actuators) • Actuators require relatively costly voltage amplifiers • Low recoverable strain (0.1%) • Piezoelectric ceramics are brittle • Cannot measure direct current • Susceptible to high hysteresis and creep when strained in direction of poling (e.g., stack actuators)
<p><i>Electrostrictive materials</i></p> <p>Example: Lead-magnesium niobate (PMN)</p>	<ul style="list-style-type: none"> • Used as sensor and actuators • Lower hysteresis and creep compared to piezoelectric • Potentially larger recoverable strain than piezoelectric 	<ul style="list-style-type: none"> • More sensitive to temperature variations than piezoelectrics
<p><i>Magnetostrictive materials</i></p> <p>Example: Terfenol-D</p>	<ul style="list-style-type: none"> • Higher force and strain capability than piezoceramics (typically, 1000 microstrain deformation) • Suited for high-precision applications • Suited for compressive load carrying components • Very durable 	<ul style="list-style-type: none"> • Low recoverable strain (0.15%) • Only for compression components • Nonlinear behavior
<p><i>Shape-memory alloys (SMA)</i></p> <p>Example: NITINOL</p>	<ul style="list-style-type: none"> • Large recoverable strain (8%) used largely for actuation due to large force generation • Low voltage requirements 	<ul style="list-style-type: none"> • Suited for low-frequency (0–10 Hz) and low-precision application • Slow response time • Complex constitutive behavior with large hysteresis
<p><i>Optical fibers</i></p> <p>Examples: Bragg grating, Fabry-Perot</p>	<ul style="list-style-type: none"> • Suited for remote sensing of structures • Corrosion resistant • Immune to electric interference • Small, light, and compatible with advanced composite 	<ul style="list-style-type: none"> • Used for sensing alone • Behavior is complicated by thermal strains
<p><i>Electrorheological fluids (ER)</i></p> <p>Example: Alumino-silicate in paraffin oil</p>	<ul style="list-style-type: none"> • Simple and quiet devices • Suitable for vibration control • Offers significant capability and flexibility for altering structural response • Low density 	<ul style="list-style-type: none"> • Low-frequency applications • Nonlinear behavior • Cannot tolerate impurities • Fluid and solid phases tend to separate • Not suitable for low temperature applications • High-voltage requirements (2–10 kV) • Higher η_p/τ_y^2 ratio than MR*
<p><i>Magnetorheological fluids (MR)</i></p>	<ul style="list-style-type: none"> • Simple and quiet devices • Quick response time • Suitable for vibration control • Offers significant capability and flexibility for altering structural response • Low voltage requirements • Behavior not affected by impurities • Suitable for wide range of temperatures • Lower η_p/τ_y^2 ratio than ER* 	<ul style="list-style-type: none"> • Nonlinear behavior • Higher density than ER
<p><i>Microphones</i></p>	<ul style="list-style-type: none"> • Low cost • Large dynamic range • Excellent linearity 	<ul style="list-style-type: none"> • Sensitive to turbulent flow • Need to achieve directionality in some active control systems (e.g., ducts) • Need protection to dust, moisture, high temperature

(cont.)

Table 1. (Continued.)

Type of Sensor or Actuator	Advantages	Disadvantages
<i>Displacement sensors</i> Example: Proximity probe, LVDT, LVIT	<ul style="list-style-type: none"> • Good low-frequency sensitivity (0–10 Hz) • Noncontacting measurement (proximity probe) • Well suited to measurement of relative displacement in active mounts 	<ul style="list-style-type: none"> • Low-frequency range (typically, below 100 Hz) • Low dynamic range (typically, 100 : 1) • Low resolution
<i>Velocity sensors</i> (magnetic)	<ul style="list-style-type: none"> • Noncontacting measurement • Well suited to measurement of relative velocity in active mounts 	<ul style="list-style-type: none"> • Low dynamic range (typically 100 : 1) • Low resolution • Heavy
<i>Accelerometers</i>	<ul style="list-style-type: none"> • Large dynamic range • Excellent linearity 	<ul style="list-style-type: none"> • Low sensitivity in low frequency (0–10 Hz) • Require relatively expensive charge amplifiers (piezoelectric accelerometers)
<i>Loudspeakers</i>	<ul style="list-style-type: none"> • Low cost 	<ul style="list-style-type: none"> • Nonlinear behavior if driven close to maximum power • Space requirement (backing enclosure) • Need protection to dust, moisture, high temperature, corrosive environment
<i>Electrodynamic and electromagnetic actuators</i>	<ul style="list-style-type: none"> • Relatively large force/large displacement capability • Excellent linearity • Extended frequency range 	<ul style="list-style-type: none"> • May need a large reaction mass to transmit large forces • Space requirement
<i>Hydraulic and pneumatic actuators</i>	<ul style="list-style-type: none"> • Large force/large displacement capability 	<ul style="list-style-type: none"> • Low-frequency range (0–10 Hz for pneumatic; 0–150 Hz for hydraulic) • Need for hydraulic or compressed air power supply • Nonlinear behavior • Space requirement

in an annular armature. When the coil is activated, the TERFENOL rod expands and produces a displacement. The TERFENOL-D bar, coil, and armature are assembled between two steel washers and put inside a protective wrapping to form the basic magnetoactive induced strain actuator unit (7). The main advantage of terfenol is its high-force capability at relatively low cost (21). It also has the advantage of small size and light weight, which makes it suitable for situations where no reactive mass is required such as in stiffened structures of aircraft and submarine hulls. The disadvantages of TERFENOL include its brittleness and low tensile strength (100 MPa) compared to compressive strength (780 MPa). Its low displacement capability is also a major disadvantage especially in the low-frequency range (less than 100 Hz). In addition, it also exhibits large hysteresis resulting in a highly nonlinear behavior that is difficult to model in practical applications (20,21). Tani et al. (20) have reviewed of studies on modeling the nonlinear behavior of TERFENOL-D as well as its application in smart structures. Ackermann et al. (22) developed a transduction model for magnetostrictive actuators through an impedance analysis of the electro-magneto-mechanical coupling of the actuator device. This model provided a tool for in-depth investigation of the frequency-dependent behavior of the magnetostrictive actuator, such as energy conversion, output stroke, and force. The feasibility of using embedded magnetostrictive mini actuators (MMA) for vibration suppression has been investigated by (20).

Shape-Memory Alloys (SMAs)

Shape-memory alloys (SMAs) are materials that undergo shape changes due to phase transformations associated with the application of a thermal field. When a SMA material is plastically deformed in its martensitic (low-temperature) condition, and the stress is removed, it regains (memory) its original shape by phase transformation to its austenite (high-temperature) condition, when heated. SMAs are considered as functional materials because of their ability to sense temperature and stress loading to produce large recovery deformations with force generation. TiNi (nitinol), which is an alloy comprising approximately 50% nickel and 50% titanium, is the most commonly used SMA material. Other SMA material including FeMnSi, CuZnAl, and CuAlNi alloys have also been investigated (20,23).

Typically, plastic strains of 6% to 8% can be completely recovered by heating nitinol beyond its transition temperature (of 45–55°C). According to Liang and Rogers (24) restraining the material from regaining its memory shape can yield stresses of up to 500 MPa for 8% plastic strain and a temperature of 180°C. By transformation from the martensite to austenite phase, the elastic modulus of nitinol increases threefold from 25 to 75 GPa, and its yield stress increases eightfold from 80 to 600 MPa (25).

SMAs can be used for sensing or actuation, although they are largely used for actuation due to their large force generation capabilities. They have very low voltage

requirements for operation and are very suited for low-frequency applications. However, their use is limited by their slow response time, which makes them suitable for low-precision applications only. Also, they exhibit complex constitutive behavior with large hystereses, which makes it difficult to understand their behavior in active structural systems. To provide a better understanding of the behavior of SMAs, several researchers have focused on the development of constitutive models for SMAs. Some of the most prominent and commonly used ones are those by Tanaka (26), Liang and Rogers (24), and Boyd and Lagoudas (27). These models are derived from phenomenological considerations of the thermomechanical behavior of the SMAs. Because of the numerous advantages they offer, several investigations on the application of SMAs have been carried out within the present decade. Reviews of these applications, focusing on fabrication of SMA hybrid composites, analytical and computational modeling, active shape control, and vibration control, are presented in (20,23).

Optical Fibers

For many applications, ideal sensors would have such attributes as low weight, small size, low power, environmental ruggedness, immunity to electromagnetic interference, good performance specifications, and low cost. The emergence of fiber-optic technology, which was largely driven by the telecommunication industry in the 1970s and 1980s, in combination with low-cost optoelectronic components, has enabled fiber-optic sensor technology to realize its potential for many applications (28–30). A wide variety of fiber-optic sensors are now being developed to measure strain, temperature, electric/magnetic fields, pressure, and other measurable quantities. Many physical principles are involved in these measurements, ranging from the Pockel, Kerr, and Raman effects to the photoelastic effect (31). These sensors use intensity, phase, frequency, or polarization modulation (32). In addition, multiplexing is largely used for many-sensor systems. Fiber-optic sensors can also be divided in discrete sensors and distributed sensors to perform spatial integration or differentiation (33). Three types of fiber-optic strain sensors are reviewed in the following: extrinsic interferometric sensors, Bragg gratings, and sensors based on the photoelastic effect.

The most widely used phase modulating fiber-optic sensors are the *extrinsic interferometric sensors*. Two fibers and directional couplers are generally used for these sensors. One of the fibers acts as a reference arm, not affected by the strain, while the other fiber acts as the sensing arm measuring the strain field. By combining the signals from both arms, an interference pattern is obtained from the optical path length difference. This interference pattern is used to evaluate the strain affecting the sensing arm (e.g., by fringe counting). These sensors have a high sensitivity and can simultaneously measure strain and temperature. One interferometer now being used in industrial applications is the Fabry-Perot interferometer, where a sensing cavity is used to measure the strain (34). This sensor uses a white-light source and a single multiple mode

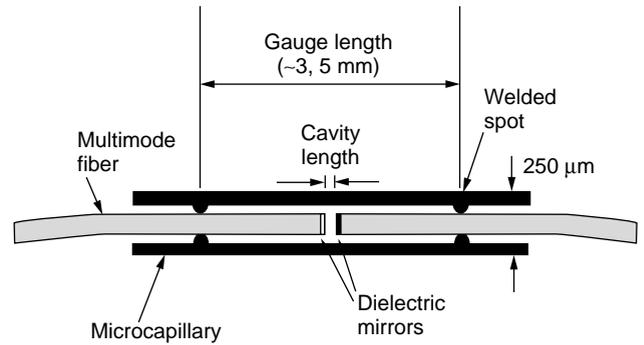


Figure 3. Fabry-Perot sensors used for ice impact monitoring and encapsulated version.

fiber, and provides absolute measurements. This extrinsic interferometer sensor is shown in Fig. 3.

Bragg grating reflectors can be written on an optical fiber using a holographic system or a phase mask to generate a periodic intensity profile (35). These sensors can be used as point or quasi-distributed sensors. The reflected signal from these sensors consist of frequency components directly related to the number of lines per millimeter of each grating reflector and, thus, to the strain experienced by the sensor. Fiber-optic sensors based on Bragg gratings are used to measure strain and temperature, either simultaneously or individually (36). The Bragg gratings are traditionally interrogated using a tunable Fabry-Perot or a Mach-Zender interferometer. Recently, long-period gratings have been used to interrogate Bragg sensing gratings (37). Bragg gratings have been used to measure vibrations either directly or through the development of novel accelerometers. A typical fiber Bragg grating (FBG) system is illustrated in Fig. 4.

The principle of operation of the sensors based on the *photoelastic effect* is a phase variation of the light passing through a material (fiber) that is undergoing a strain variation. This phase variation can be produced by two effects on the fiber: (1) the variation of the length produced by the strain; (2) the photo-elastic effect and the modal dispersion caused by the variation of the diameter of the fiber. These sensors are classified in modal interferometric sensors and polarimetric sensors. As it integrates the strain effect over its length, the modal interferometric sensor can act as a spatial filter if the propagation constant is given a spatial weighting (38).

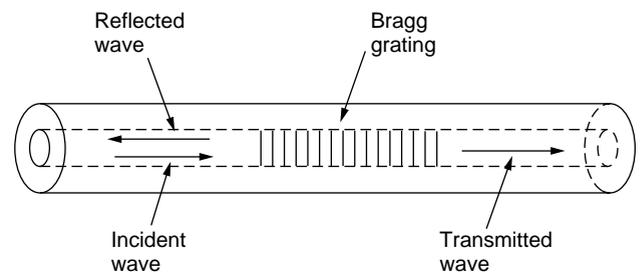


Figure 4. Bragg grating on an optical fiber.

Electrorheological Fluids (ER)

Electrorheological fluids (ER) are a class of controllable fluids that respond to an applied electric field with a dramatic change in rheological behavior. The essential characteristic of ER fluids is their ability to reversibly change from free-flowing linear viscous liquids to semisolids having controllable yield strength in milliseconds when exposed to an electric field (23). The ER fluids provide very simple, quiet and rapid response interfaces between electronic controls and mechanical systems. They are very suitable for vibration control because of the ease with which their damping and stiffness properties can be varied with the application of an electric field.

ER materials consist of a base fluid (usually a low viscosity liquid) mixed with nonconductive particles, typically in the range of 1 to 10 μm diameter. These particles become polarized on the application of an electric field, leading to solidification of the material mixture. Typical yield stresses in shear for ER materials are about 5 to 10 kPa. The most common type of ER material is the class of dielectric oils doped with semiconductor particle suspensions, such as aluminosilicate in paraffin oil. The material exhibits nonlinear behavior, which is still not completely understood by the research community. This lack of understanding has hindered efforts in developing optimal applications of ER materials. However, electrorheological fluids may be suitable for many devices, such as shock absorbers and engine mounts (23,25).

Magnetorheological Fluids (MR)

Magnetorheological fluids (MR) are similar to ER materials in that they are also controllable fluids. These materials respond to an applied magnetic field with a change in the rheological behavior. MR fluids, which are less known than ER materials, are typically noncolloidal suspensions of micron-sized paramagnetic particles. The key differences between MR and ER fluids are highlighted in Table 1. In general, MR fluids have maximum yield stresses that are 20 to 50 times higher than those of ER fluids, and they may be operated directly from low-voltage power supplies compared to ER fluids which require high-voltage (2–5 kV) power supplies. Furthermore, MR fluids are less sensitive to contaminants and temperature variations than are ER fluids. MR fluids also have lower ratios of η_p/τ_y^2 than ER materials, where η_p is the plastic viscosity and τ_y the maximum yield stress. This ratio is an important parameter in the design of controllable fluid device design, in which minimization of the ratio is always a desired objective. These factors make MR fluids the controllable choice for recent practical applications. Several MR fluid devices developed by Lord Corporation in North Carolina under the Rheo-etic trade name (23).

Microphones

Microphones are usually the preferred acoustic sensors in active noise control applications. Relatively inexpensive microphones (electret or piezoelectric microphones) can be used in most active noise control systems because the frequency response flatness of the microphones is not critical

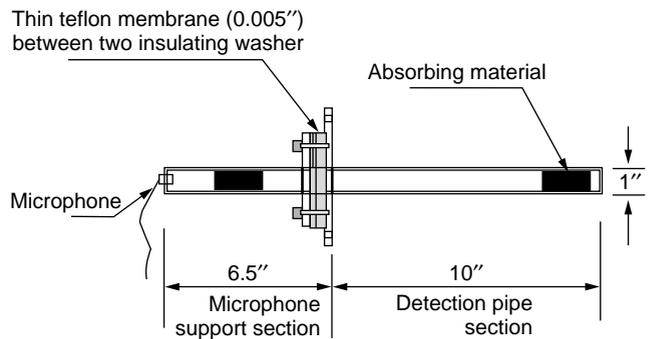


Figure 5. Sound pressure and particle velocity sensing.

in digital active control systems, as it is compensated in the identification of the control path. The most common types of microphones are omni-directional, directional, and probe microphones.

Whenever turbulent flow is present in the acoustic medium (e.g., a turbulent flow in a duct conveying a gas or a fluid), turbulent random pressure fluctuations are generated in the flow, adding to the disturbance pressure field. The most common way of reducing the influence of turbulent noise is to use a probe tube microphone consisting of a long, narrow tube with a standard microphone mounted at the end. The walls of the tube are porous or contain holes or an axial slit. The probe tube microphone must be oriented with the microphone facing the flow. Probe tube microphones are convenient as reference sensors in active control systems in ducts because they act as both directional sensors and turbulence filtering sensors. Details on the principle of operation can be found in (39). Low-cost microphone probes for hot corrosive industrial environments are also available from Soft dB Inc. Figure 5 shows a microphone adapted for such environments.

Displacement and Velocity Transducers

Although their dynamic range is usually much less than that for accelerometers, displacement and velocity transducers are often more practical for very low frequencies (0–10 Hz) where vibration amplitudes can be of the order of a millimeter or more for heavy structures whose corresponding accelerations are small. Also, in low-frequency active control systems, displacement or velocity rather than acceleration can be the preferred quantities to minimize. The displacement and velocity transducers are described below.

Proximity probes are the most common type of displacement transducers. There are two main types of proximity probes, the capacitance probe and the Eddy current probe. Proximity probes allow noncontact measurement of vibration displacements. They are well suited to vibration displacement measurements on rotating structures. The dynamic range of proximity probe is very small—typically 100 : 1 for low-frequency applications (<200 Hz). The resolution varies from 0.02 to 0.4 mm.

The *linear variable differential transformer (LVDT)* is a displacement transducer that consists of a single primary and two secondary coils wound around a cylindrical bobbin.

A movable nickel iron core is positioned inside the windings, and it is the movement of this core that is measured. The dynamic range of an LVDT is typically 100 : 1, with a resolution ranging from 0.01 to 1 mm. The frequency range is typically dc to 100 Hz. The total length of the sensor varies from 30 to 50 mm for short stroke transducers to about 300 mm for long stroke transducers.

The *linear variable inductance transformer* (LVIT) is a displacement transducer based on the measurement of inductance changes in a cylindrical coil. The coil is excited at about 100 kHz, and the inductance change is caused by the introduction of a highly conductive, nonferrous coaxial rod sliding along the coil axis. It is the movement of this coaxial rod that is measured. This type of transducer is particularly suited for measuring relative displacements in suspension systems. Transducer sizes vary from diameters of a few millimeters to tens of millimeters.

Often used among the velocity transducers is the *non-contacting magnetic type* consisting of a cylindrical permanent magnet on which is wound with an insulated coil. A voltage is produced by the varying reluctance between the transducer and the vibrating surface. This type of transducer is generally unsuitable for absolute measurements, but it is very useful for relative velocity measurement such as needed for active suspension systems. The frequency range of operation is 10 Hz to 1 kHz; the low-resonance frequency of the transducer makes it relatively heavy. Velocity transducers cover a dynamic range between 1 and 100 mms^{-1} . Low-impedance, inexpensive voltage amplifiers are suitable.

Accelerometers

Accelerometers are the most employed technology for vibration measurements. They provide a direct measurement of the acceleration, usually in the transverse direction of a vibrating object. The acceleration is a quantity well correlated to the sound field radiated by the vibrating object. Therefore, accelerometers can be a convenient alternative to microphones as error sensors for active structural acoustic control. Accelerometers usually have a much larger dynamic range than displacement or velocity sensors. A potential drawback of accelerometers, in low-frequency active noise control systems, is their low sensitivity at low frequency (typically 0–10 Hz).

Small accelerometers can measure higher frequencies, and they are less likely to affect the dynamics of the structure by mass loading it. However, small accelerometers have a lower sensitivity than bigger ones. Accelerometers range in weight from miniature 0.65 g for high-level vibration amplitudes up to 18 kHz on lightweight structures, to 500 g for low-level vibration amplitudes on heavy structures up to 1 kHz. Because of the three-dimensional sensitivity of piezoelectric crystals, piezoelectric accelerometers are sensitive to vibrations at right angle to their main axis. The transverse sensitivity should be less than 5% of the axial sensitivity. There are two main types of accelerometers: piezoelectric and piezoresistive.

A *piezoelectric accelerometer* consists of a small seismic mass attached to a piezoelectric crystal. When the accelerometer is attached to a vibrating body, the inertia force

due to the acceleration of the mass produces a mechanical stress in the piezoelectric crystal that is converted into an electric charge on the electrodes of the crystal. Provided that the piezoelectric crystal works in its linear regime, the electric charge is proportional to the acceleration of the seismic mass. The mass may be mounted to produce either compressive or tensile stress, or alternatively, shear stress in the crystal. A piezoelectric accelerometer should be used below the resonance of the seismic mass–piezoelectric crystal system. Since piezoelectric accelerometers essentially behave as electric charge generators, they must generally be used with high-impedance charge amplifiers. The cost of such amplifiers can represent a significant amount of the total cost of an active control system when a large number of accelerometers are used.

Piezoresistive accelerometers rely on the measurement of resistance change in a piezoresistive element usually mounted on a small beam and subjected to stress. Piezoresistive accelerometers are less sensitive than piezoelectric. They require a stable, external dc power supply to excite the piezoresistive elements. However, piezoresistive accelerometers have a better sensitivity at low frequency, and they require less expensive, low-impedance voltage amplifiers. The piezoresistive element is sometimes replaced by a piezoelectric polymer film (PVDF), and the electric charge across the electrodes of the PVDF is collected as the sensor output. Such a PVDF accelerometer has a sensitivity and frequency response similar to the piezoresistive accelerometer, and it is less expensive than the piezoelectric accelerometer.

Loudspeakers

The electrodynamic loudspeaker is the most commonly employed actuator technology for active noise control applications. When selecting a loudspeaker for an active noise control system, the important parameter is the cone volume velocity required to cancel the primary sound field (21).

For small systems, (small-duct, low-noise, domestic ventilation system), active acoustic noise control can be achieved with small commercial medium-quality speakers (radio-type speaker). However, for bigger systems, precautions have to be taken.

Electrodynamic loudspeakers exhibit a nonlinear behavior when they are driven close to maximum power or maximum membrane deflection. It can significantly degrade the performance of active control systems based on linear filtering techniques. It is thus important that loudspeakers should be driven at a fraction of the maximum power or maximum deflection specifications, especially in situations where single-frequency or harmonic noise has to be attenuated. For random noise, the peak cone velocity requirements for active control are likely to be four or five times the estimated rms velocity requirements (39).

In active control of single-frequency noise, it is desirable to design the loudspeaker so that its mechanical resonance lies close to the frequency of interest. This resonance frequency can be adjusted to suit a particular application either by adding mass to the cone (to reduce the frequency) or by adding a backing enclosure to the speaker (to increase the frequency).

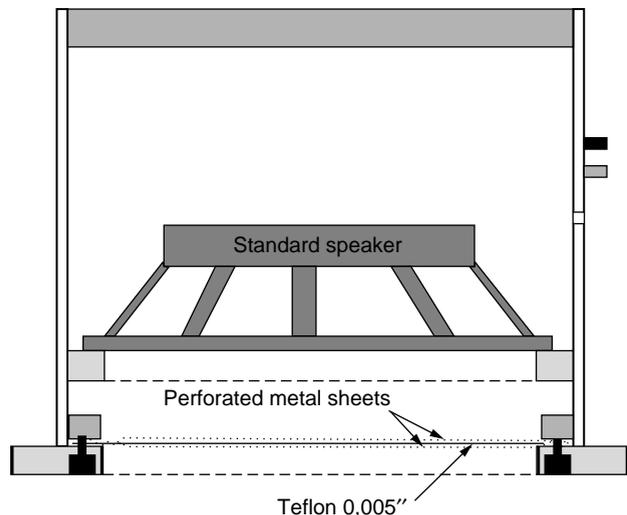


Figure 6. Protective system for loudspeaker membrane.

Operation in industrial environments requires considerable precautions. In high-humidity, high-temperature and corrosive environments, the loudspeaker cone must be protected with a heat shield. Soft dB used a Teflon membrane and a perforated metal sheet to protect the membrane of the speaker from corrosive gas (see Fig. 6).

Electromagnetic Actuators

For vibration control purposes, electromagnetic actuators can be classified into electrodynamic shakers and electrical motors. The latter can be used for low-frequency vibration control. Electrodynamic shakers are generally defined as devices having a central inertial core (usually a permanent magnet) surrounded by a winding. This type of inertial actuator applies a point force to a structure by reacting against the inertial mass. As in a loudspeaker, a time-varying voltage is applied to the coil in order to move the inertial mass and to force the movement of the structure onto which the shaker is attached.

Other inertial type actuators are available which use, for example, the piezoelectric effect, instead of a coil, to move the inertial mass. Proof-mass actuators (also called inertial actuators) are very similar in their operation to electrodynamic shakers. They usually consist of a mass that is moved by an alternating electromagnetic field. These devices can generate relatively large forces and displacements and can be good alternatives to costly electrodynamic shakers. The devices can excite very stiff structures such as electrical power transformers. Another advantage of proof-mass actuators is that their resonant frequency can be easily tuned for optimal efficiency at a given frequency.

Electrical Motors

The advent of new control strategies and digital controllers has revolutionized the way electrical motors can be used and now allows for the use of motor technologies that were previously difficult to implement in practical applications. Simple motor drives were traditionally designed with relatively inexpensive analog components that suffer from susceptibility to temperature variations and component aging. New digital control strategies now allow for the use of electrical motors in active vibration control applications. These efficient controls make it possible to reduce torque ripples and harmonics and to improve dynamic behavior in all speed ranges. The motor design is optimized due to lower vibrations and lower power losses such as harmonic losses in the rotor. Smooth waveforms allow an optimization of power elements and input filters. Overall, these improvements result in a reduction of system cost and better reliability.

Electrical motors can be divided into motors with a permanent magnet rotor (ac and dc motors) and motors with a coiled rotor. Figure 7 illustrates a detailed classification of the electrical motors. With the advent of new controllers, the tendency is to classify electrical motors under ac or dc according to the control strategy.

Due to its high reliability and high efficiency in a reduced volume, the *brushless motor* is actually the most

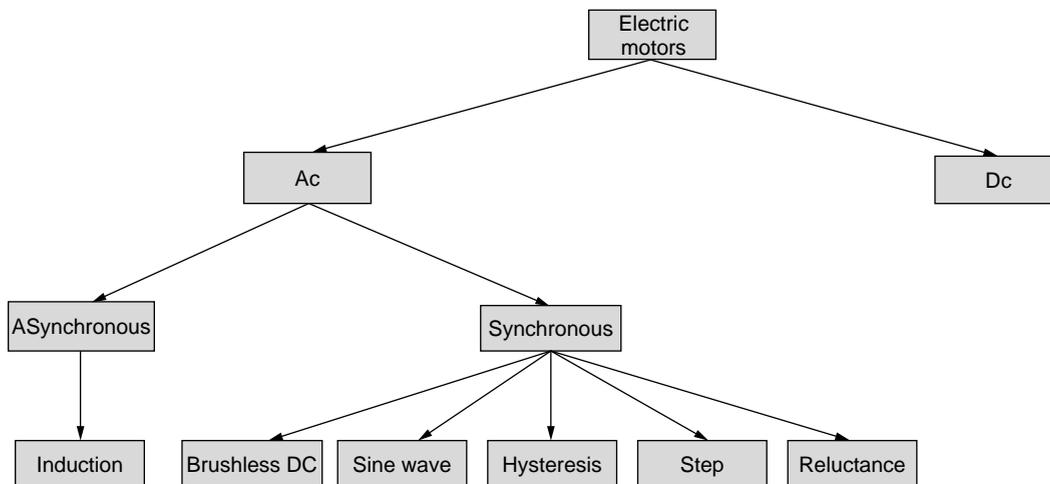


Figure 7. Classification of electric motors (42).

interesting motor for application to active vibration control (40). Although the brushless characteristic can be applied to several kinds of motors, the brushless dc motor is conventionally defined as a permanent magnet synchronous motor with a trapezoidal back EMF waveform shape, while the brushless ac motor is conventionally defined as a permanent magnet synchronous motor with a sinusoidal back EMF waveform shape. New brushless and coreless motors are now available which are very linear over a wide speed range (41). The brushless motor control consists of generating variable currents in the motor phases. The regulation of the current to a fixed 60° reference can be realized in two modes: *pulse width modulation* (PWM) or *hysteresis* mode. Shaft position sensors (incremental, Hall effect, resolvers) and current sensors are used for the control. Linear permanent magnet motors are also available that, in addition to the linear action, allow better magnetic dissipation in the core as it is distributed in space.

If volume is not a major concern, a second type of motor to be used in active vibration control is the *induction or ac motor* (41). As for the brushless motor, the performance of an ac motor is strongly dependent on its control. DSP controllers enable enhanced real time algorithms. There are several ways to control an induction motor in torque, speed, or position; they can be categorized in two groups: the *scalar* and the *vector control*. Scalar control means that variables are controlled only in magnitude, and the feedback and command signals are proportional to dc quantities. The vector control is referring to both the magnitude and phase of these variables. Pulse width modulation techniques are also used for the control of induction motors, and indirect current measurement (using a shunt or Hall effect sensor) is used as a feedback information for the controller.

The third electrical motor used for active vibration control is the *switched reluctance motor* (40). This motor is widely used mainly because of its simple mechanical construction and associated low cost and secondarily because of its efficiency, its torque/speed characteristic and its very low requirement for maintenance. This type of motor, however, requires a more complicated control strategy. The switched reluctance motor is a motor with salient poles on both the stator and the rotor. Only the stator carries windings. One stator phase consists of two series-connected windings on diametrically opposite poles. Torque is produced by the tendency of its movable part to move to a position where the inductance of the excited winding is maximized. There are two ways to control the switched reluctance motor in torque, speed and position. Torque can be controlled by the *current control method* or the *torque control method*. The pulse width modulation (PWM) strategy is used in both current and torque control approaches to drive each phase of the switched reluctance motor according to the controller signal.

Hydraulic and Pneumatic Actuators

Hydraulic and pneumatic actuators are good candidate technologies when low frequency, large force, and displacements are required. Hydraulic actuators consist of a hydraulic cylinder in which a piston is moved by the action of a high-pressure fluid. The main advantage of hydraulic

actuators is their large force and large displacement capability for a relatively small size. The disadvantages include the need for a hydraulic power supply (which can require space and generate noise), the high cost of servo-valves, the nonlinear relation between the servo-valve input voltage and the output force or displacement produced by the actuator, and the limited bandwidth of the actuator (0–150 Hz). Hydraulic actuators have been used in the design of active dynamic absorbers for ship structures (42,43).

The principle of operation of pneumatic actuators is very similar to hydraulic actuators, except that the hydraulic fluid is replaced by compressed air. Due to the higher compressibility of air, the bandwidth of pneumatic actuators is reduced (typically 0–10 Hz), which restricts the application to nonacoustic problems. Pneumatic actuators may be an attractive option when an existing air supply is already available

APPLICATIONS OF NOISE CONTROL IN SHIP STRUCTURES

A typical marine diesel engine mounted on a ship hull is schematically depicted in Fig. 1. The figure shows the various vibroacoustic paths through which the engine vibration is transmitted to the ship structure, and eventually radiated into seawater. In the figure the coupling between structural and acoustic energy is classified using the following symbols: AA: acoustic to acoustic coupling, SS: structural to structural coupling, AS: acoustic to structural coupling, SA: structural to acoustic coupling. The relative importance of energy coupling for radiation into seawater is illustrated by a number. As shown, there are five possible energy transmission paths, including (1) the mounting system, consisting of the engine cradle, isolation mounts, raft, and foundation; (2) the exhaust stack; (3) the fuel intake and cooling system; (4) the drive shaft; and (5) the airborne radiation of the engine. In this study, these five paths are grouped into four categories, corresponding to generic active control problems:

- Path 1: Active vibration isolation (mounting system).
- Path 2: Active control of noise in ducts and pipes (exhaust stack; fuel intake and cooling system).
- Path 3: Active control of vibration propagation in beam-type structures (drive shaft).
- Path 4: Active control of enclosed sound fields (airborne radiation of the engine).

Path 1: Active Vibration Isolation

Active vibration isolation involves the use of an active system to reduce the transmission of vibration from one body or structure to another (e.g., transmission of periodic vibration from a ship's engine to the ship's hull). Such an active isolation system will be used in practice to complement passive, elastomeric isolation mounts between the engine and supporting structure. An active isolation system is usually much more complex and expensive than its passive counterpart, but has the advantage of offering better low-frequency isolation performances, and can

be designed for a better static stability of the supported equipment.

The first class of system involves the control of system damping, and is often referred as a *semiactive* isolation system, Fig. 8(a). The damping modification is usually achieved by a hydraulic damper with varying orifice sizes. This system is often used for active suspensions in cars. Such a system involves control time constants significantly longer than the disturbance time constants, with the advantage of a simpler and less expensive implementation. However, low-frequency performance is much less than for fully active systems described in the following.

A second class of system involves an *active control actuator in parallel* with a passive system, with the actuator

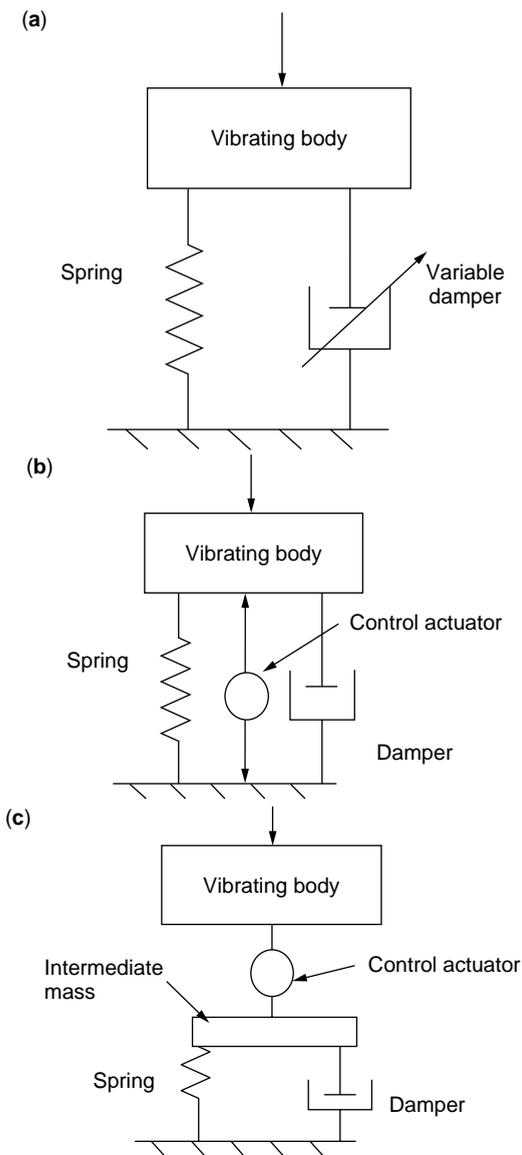


Figure 8. Active vibration isolation systems: (a) semiactive system with variable damper; (b) active system with control force applied to both vibrating body and base structure; (c) active system with control force in series with passive mount.

exerting a force on either the base structure or the rigid mass, Fig. 8(b). In this parallel configuration, the actuator is not required to withstand the weight of the machine; as compared to the configuration of Fig. 8(c), the required control force is smaller above the natural frequency of the system (44). The main disadvantage of this configuration is that at higher frequency (outside the frequency range where the actuator is effective), the actuator itself can become a transmission path. At low frequency, the large displacement/large force requirements for heavy structures preclude the use of piezoelectric, magnetostrictive actuators. Instead, hydraulic, pneumatic, or electromagnetic actuators (with their associated weight, space, and possibly fluid supplies problems) must be used. As far as practical application of active control is concerned, the use of an actuator in parallel with a passive isolation stage could have distinct advantages. In a given application, if an actuator can be found that provides a control force of the order of the primary force exciting the machine, then it may be possible to use of much higher mounted natural frequency associated with the passive isolation stage than would be otherwise possible. This in turn has advantages for the stability of the mounted machine.

A third configuration with the active system in series with the passive mount is shown on Fig. 8(c). Such a system has several advantages over the parallel configuration. The active system is now isolated from the dynamics of the receiving structure, which simplifies the control in the case of a flexible base structure, and the use of an intermediate mass creates a two-stage isolation system that offers better isolation performance in higher frequency.

Path 2: Active Control of Noise in Ducts and Pipes

The reduction of duct noise is the first-known application of active noise control. Active control systems for duct noise are now a mature technology, with several commercial systems available for ventilation systems, chimney stacks, or exhausts. All existing commercial systems are based on feedforward adaptive control systems. In the case of ducts containing air or a gas, loudspeakers are generally used as control sources, and microphones as error sensors.

Two important classes of systems must be distinguished, depending on the frequency and the cross-sectional dimension of the duct:

1. Systems for which only plane wave propagation exists in the duct. Such systems will necessitate a single-channel control system (one control source and one error sensor).
2. Systems for which higher-order acoustic modes propagate in the duct. Such systems will require a multi-channel control system.

The occurrence of higher-order modes in a duct depends on the value of the cut-on frequency. For a rectangular duct, the cut-on frequency is given by $f_c = c_0/2d$, where d is the largest cross-sectional dimension and c_0 is the speed of sound in free space. For a circular duct, $f_c = 0.586 c_0/d$, where d is the duct diameter. Higher-order modes will propagate at frequencies larger than f_c .

For active noise control (ANC) in the large duct, multichannel acoustical ANC systems are necessary, and M error sensors have to be used to control M modes for high-order propagation cases. The error sensors should not be located at the nodal lines (observability condition) (45). For a rectangular duct, the location of the error sensors is relatively simple because the nodal lines are fixed along the duct axis. However, in circular ducts, the location of the nodal lines changes along the duct axis, since the modes usually spin as a function of the frequency, temperature, and speed (46,47). Those variations of the nodal lines may explain why ANC of high-order modes in circular or irregular ducts appears to be difficult (48). Instead of using the modal approach (i.e., the shape of the modes to be controlled) to determine the error sensors location, an alternative strategy has recently been proposed by A. L'Espérance (49)—the error sensor plane concept. This concept calls for a quiet cross section to be created in the duct so that, based on the Huygen's principle, the noise from the primary source cannot propagate over this cross section. A multichannel ANC in a circular duct accords with this strategy (50).

The principles of active control of noise propagating in liquid-filled ducts are much the same as in air ducts (51). The higher speeds of sound in liquids means that plane wave propagation occurs in a larger frequency range than in air ducts. However, considerable care must be exercised to the possible transmission of energy via the flexible duct walls in this case, as a result of the strong coupling between the duct walls and the interior fluid.

Path 3: Active Control of Vibration Propagation in Beam-Type Structures

The active control of vibration in one-dimensional systems such as beams, rods, struts, and shafts can be approached from two different perspectives, depending on the description of the structural response. The response can be described in terms of vibration modes or in terms of waves propagating in the structure. The modal perspective is more appropriate to finite, or short, beams and to *global* reduction of the vibration. The description of the response in terms of structural waves is more appropriate to infinite, or long, beams and to reducing energy flow from one part of the beam to another (control of vibration transmission). The wave description is then more appropriate to the case of the transmission of vibration from a ship's engine via the drive shaft, since in this case the source of vibration is known and the objective is to block the vibration transmission along the shaft. The active control of vibration in beams is widely covered in the literature (21,44). The following presentation is mostly limited to feedforward control systems, since it is assumed that for the problem of vibration transmission along a marine drive shaft, an advanced signal correlated to the disturbance, or a measurement of the incoming disturbance, wave is possible.

Simultaneous Control of All Wave Types (Flexural, Longitudinal, Torsional). In a general adaptive feedforward controller used for the active control of multiple wave types

in a beam, sensor arrays (e.g., accelerometer arrays) are used to measure the different types of waves propagating upstream (detection array) or downstream (error array) of the control actuators, and an array of actuators is used to inject and control the various wave types in the beam (44).

Wave analyzers are necessary to extract the independent wave types (assumed uncoupled) from the sensor arrays, and wave synthesizers are necessary to generate the appropriate commands to the individual actuators. This approach has the advantage that independent control filters can be used to control the flexural, longitudinal, and torsional waves. However, it necessitates excellent phase matching of the sensors and a detailed knowledge of the structure in which the waves propagate. An experimental laboratory implementation of this approach has been conducted by (52), on a thin beam, for the control of two flexural wave components and one longitudinal wave using PZT actuators. Another, easier option avoids implementing wave analyzers and synthesizers by simply minimizing the sum of squared output of the error sensors to control the different wave types. This approach, however, requires a fully coupled multichannel control system. This approach has been tested for the control of two flexural waves and one longitudinal wave in a strut using three magnetostrictive actuators (53,54).

Control of Flexural Waves. The dispersive nature of flexural waves implies that a control force applied transversely to the beam generates propagative waves as well as evanescent waves localized close to the point of application of the force. If one transverse control force is applied at some location on the beam, it generates downstream and upstream propagating waves plus downstream and upstream evanescent waves. This actuator can minimize the total, transmitted downstream wave, but it generates a reflected wave toward the source and two evanescent components that may be undesirable. A total of four actuators will be necessary to control downstream and upstream, propagating, and evanescent components. Therefore, the control of flexural waves in beams will in general require *actuator arrays* (55). Combinations of *force* and *moment* actuators can also be used in the actuator array. The simplest feedforward control system uses only one control force and one error accelerometer, together with one reference accelerometer to measure the incoming wave. This system has been studied theoretically (56), and tested experimentally (57).

Physical limits of this system have been identified. The first limit is associated with the detection of the control actuator evanescent wave by the error sensor that puts a limit on the actuator-error sensor separation: in practice, the sensor should be at least 0.7λ from the control actuator (λ being the flexural wavelength). The second limit is related to the delay between detection and actuation that should be sufficient to allow the active control system to react at the control actuator location before the primary wave has propagated from the detection sensor to the control actuator. This puts a limit on the reference sensor-actuator separation, which depends on the characteristics of the control system.

Similarly to control actuator arrays, error sensor arrays need to be implemented for the control of flexural waves

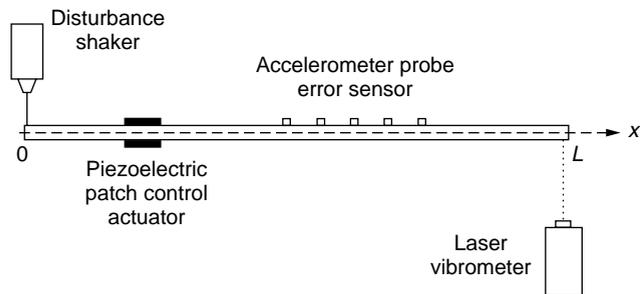


Figure 9. Typical experimental setup for the control of active structural intensity.

to distinguish between the various propagating waves and evanescent waves at the error sensor locations. This is particularly needed if the error sensors must be located at a short distance from the control actuators. In this case, an array of four accelerometers can discriminate between the two propagating waves and the two evanescent waves at the location of the error sensor array, and extract the components that need to be reduced (e.g., the downstream propagating wave).

Other sensing strategies have also been suggested, such as measuring and minimizing the structural intensity due to flexural waves (58,59). Structural intensity can be measured in practice using an array of four or more closely spaced accelerometers, as presented in Fig. 9.

Practical Implementations. There are a limited number of practical implementations of these principles to large, machinery structures. Semiactive or active devices have been used to attenuate the transmission of *longitudinal* vibration on a large tie-rod structure (60). The tie-rod is similar to that found in marine machinery to maintain the alignment of a machinery raft. A tunable pneumatic vibration absorber was used as the semiactive device, and an electrodynamic shaker or a magnetostrictive actuator was used as the active device. A load cell was used as the error sensor, such that the force applied by the tie-rod to a receiving bulkhead was minimized.

The suppression of vibration that is generated on rotating machinery with an overhung rotor has been presented (61). In this case, the vibration of the rotor-shaft system is controlled by active bearings. The active bearings consist of a bearing housing supported elastically by rubber springs and controlled actively by electromagnetic actuators. These actuators are controlled by displacement sensors at the pedestal and/or the roller and can apply an electromagnetic force that suppresses any vibration of the roller. The active vibration control (AVC) of rotating machinery utilizing piezoelectric actuators was also investigated (62). The AVC is shown to significantly suppress vibration through two critical speeds of the shaft line.

Path 4: Active Control of Enclosed Sound Fields

There exists a vast body of literature on the subject of active control of enclosed sound fields. Only the previous work relevant to the problem of canceling the sound field radiated

by a ship engine in its enclosed space will be reviewed here. More comprehensive presentations of the generic problem can be found in (3,21). Active control of enclosed sound fields has found applications essentially for automobile interior noise (63,64) and for aircraft interior noise (65–67), leading in some cases to commercial products.

There are two main categories of active control systems related to enclosed sound field minimization:

- Active control of sound transmission through elastic structures into an enclosure.
- Active control of sound field into rigid enclosures.

Only the second category will be reviewed here. The active control of sound transmission has been investigated using essentially modal approaches (68,69). The same type of analytical approach based on modes of the acoustic enclosure can be used to investigate the active control of sound field into rigid enclosures. It should be mentioned, however, that finite element approaches have also been used to study the active control of sound field into enclosures of complex geometries (70,71). Additionally, the objective of the *active* control in an enclosure can be to minimize the sound field *globally*, or *locally*. Only the approaches directed toward global attenuation of the sound field are reviewed here. In this respect, some important physical aspects of this problem are discussed in the following. These physical aspects depend primarily on the modal density of the enclosure.

Enclosures with a Low Modal Density. For enclosures with a low modal density (i.e., a small enclosure, or at low frequency), the active control will usually consist of placing a series of control loudspeakers in the enclosure; the loudspeakers are driven to minimize the sound pressure measured by discrete error microphones. In the case of an enclosed acoustic space, the performance metrics for the control should be the acoustic potential energy integrated over the volume of the enclosure,

$$E_p = \frac{1}{4\rho_0 c_0^2} \int_V |p(\mathbf{r})|^2 dv,$$

where $p(\mathbf{r})$ is the local sound pressure, ρ_0 is the density of the acoustic medium, and c_0 is the speed of sound. The active control scheme should aim at reducing the acoustic potential energy as much as possible.

It has been shown that active control of sound fields in lightly damped enclosures is most effective at the resonance of the acoustic modes (72). In these instances, the problem is essentially the control of a single mode. Significant attenuation of the acoustic potential energy is obtained using a single control source and a single error microphone (provided that neither the control source nor the error microphone is located on a nodal surface of the acoustic mode). For a multiple-mode (off-resonance) response of the cavity, the number of control sources and error microphones should be increased. However, the potential for attenuation is never as large as at a resonance frequency.

The number and placement of control sources and error sensors are critical for multiple-mode control. The corresponding optimization problem is nonlinear and usually involves many local minima. Optimization processes, such as multiple regression (21) or genetic algorithms (73), are used. As a general rule is that the number and locations of the control sources should be such that the secondary sound field matches as closely as possible the primary sound field in the enclosure.

Enclosures with a High Modal Density. As the frequency increases or the enclosure becomes larger, global attenuation of the sound field becomes more difficult to achieve using an active control system. To quantify these limitations, there are some approximate formulas, which are summarized here. These formulas are approximate, but they give useful expected performance of an active control system in a high modal density enclosure.

First, assuming a single primary point source and a single secondary point source in the enclosure, it is possible to derive the ratio of the minimized potential energy (after control) to the original potential energy (before control), (74):

$$\frac{E_{p,\min}}{E_{p,0}} = 1 - \left[1 + \frac{\pi}{2} M(\omega) \right]^{-2},$$

where $M(\omega)$ is the modal overlap of the cavity, which quantifies the likely number of resonance frequencies of other modes lying within the 3 dB bandwidth of a given modal resonance. For a rigid rectangular enclosure and for oblique acoustic modes, namely three-dimensional modes, such as the (1,1,1) mode,

$$M(\omega) = \frac{\zeta \omega^3 V}{\pi c_0^3},$$

where ζ is the damping ratio in the enclosure (assumed identical for all acoustic modes), ω is the angular frequency of the sound field, and V is volume of the enclosure.

If the modal density is low (at low frequency),

$$\frac{E_{p,\min}}{E_{p,0}} \approx \pi M(\omega),$$

which means that the achievable attenuation is dictated by the modal overlap (and hence the modal density and damping of the enclosure).

If the modal density is large (at high frequency),

$$\frac{E_{p,\min}}{E_{p,0}} \approx 1,$$

which means that no attenuation can be obtained after control. Another expression can be derived from the asymptotic expression of modal overlap in high frequency (75),

$$\frac{E_{p,\min}}{E_{p,0}} = 1 - \sin^2 kd,$$

where k is the acoustic wave number and d is the separation between the primary and control sources. Thus, as the control source becomes remote from the primary source, such that $kd \geq \pi$, any global attenuation of the sound field becomes impossible. This provides an explicit analytical demonstration that the global control of enclosed sound fields of high modal density is only possible with closely spaced compact noise sources. In other words, assuming an extended primary source such as a ship engine, the only viable solution in this case is to distribute control loudspeakers around the engine and in the close vicinity of it (within a fraction of the acoustic wavelength).

Advanced Sensing Strategies. Recently, alternatives to sensing and minimizing squared sound pressure have been suggested in active control of enclosed spaces. Sensing strategies based on total acoustic energy density minimization instead of sound pressure minimization have been suggested (76,77). The advantage of sensing the total energy density is that the control is less sensitive to the sensor locations, and in general, a superior attenuation is obtained. The energy density can be measured using combinations of microphones (2 to 6); in this case, finite differences between individual microphones are applied to obtain approximate measurements of the pressure gradient in several directions. Precise measurements of the pressure gradient requires an excellent phase matching of the individual microphones, which can result in more expensive microphones. Associated adaptation algorithms for the minimization of energy-based quantities have been derived (78).

RECOMMENDATIONS ON SENSORS AND ACTUATORS FOR ANVC OF MARINE STRUCTURES

Steps in Design of Active Control Systems in Marine Structures

The choice of the sensors and actuators in an active control system will depend on factors such as the frequency of the disturbance, operating environment, cost, expected performance, and magnitude of the vibratory or acoustic disturbance to be controlled along the various vibroacoustic paths. Figure 10 suggests a systematic approach that proceeds through the various steps of the active control system design. As a general recommendation, the first (and perhaps most important) phase of the design of every active control system is to acquire a thorough understanding of the vibroacoustic behavior of the system on which active control is to be applied. This involves carefully identifying and ranking the various paths along which vibroacoustic energy flows. This may imply addressing questions such as the transmission of moments or in-plane forces through the engine mounting, or the relative contribution of fluid-borne and structure-borne energy along pipes. This early phase is crucial in determining the active control strategy to be implemented. A number of experimental techniques and numerical simulation tools can be used to estimate the relative contribution of the various paths at a given receiving point (e.g., in water). Based on some contractors' previous experiences, a major transmission path appears to be the engine-mounting system.

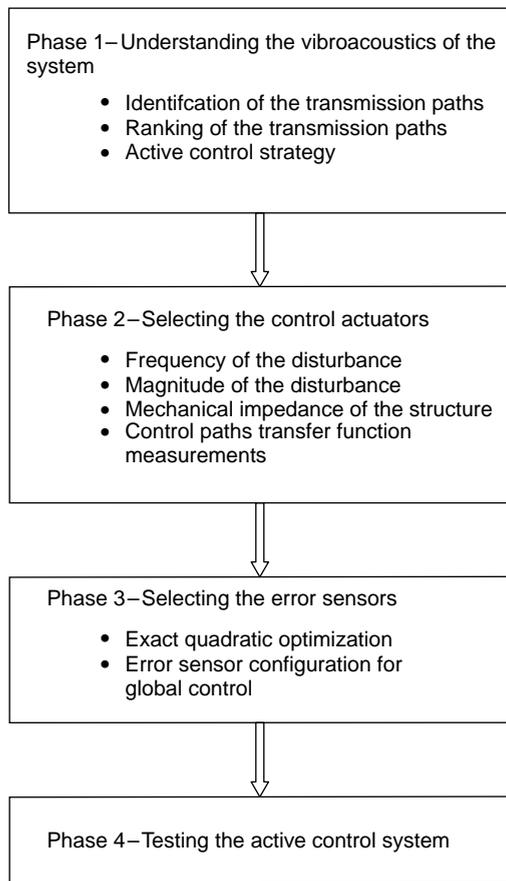


Figure 10. Suggested design steps of an active control system.

The second phase will determine the *type, number, and locations of the control actuators*. When global control is desirable (e.g., when attenuation of the sound field is desired at all positions in water), these parameters are determined by the requirement that the sound field generated by the control actuators should spatially match the primary sound field. The type of control actuators to be used will be based primarily on the frequency of the disturbance and the magnitude of the disturbance at the actuator location (for simplicity, the control actuators need to generate a secondary field with a magnitude equal to the disturbance at the actuator location). Once the type, number, and locations of the control actuators are known, extensive transfer function measurements need to be taken between individual actuators and field points (vibratory or acoustic), with the primary source turned off. Since this may involve a considerable experimental task, numerical simulations can be of a great help here.

The third phase will address the error sensors. Again, if global control is desirable, the *type, number, and locations of the error sensors* are dictated by the requirement that if the control actuators are driven to minimize the signal at the error sensors, then the resulting sound field is globally reduced. The measured transfer functions between individual actuators at field points and the magnitude of the primary disturbance at these field points are used, in conjunction with classical exact quadratic optimization

techniques, to calculate the optimal control variables (i.e., the required inputs of the control actuators) that minimize the error signals for a given error sensor arrangement. The final phase will be to test the active control with a real controller.

Recommended Sensor and Actuator Technologies for Various Ship Noise Paths

Path 1: Active Vibration Isolation. In selecting sensors and actuators for active vibration isolation of engine noise, due consideration has to be given to the size and weight of the structure (engine) being isolated. Since the engine is a heavy structure weighing over 6000 kg, it is necessary that the actuators are capable of delivering very high control forces. In addition, the nature of the noise through this path is nonacoustic, and hence nonacoustic sensors and actuators have to be used. Based on these considerations, the recommended sensors and actuators are (1) accelerometers and force transducers for sensing and (2) hydraulic and electrodynamic actuators for actuation. The recommendations are summarized in Table 3. For increased efficiency, the control systems must be designed to provide control forces in translational and rotational directions, since engine vibrations could take place in all directions. Furthermore, the active control systems should be used in conjunction with passive control systems, to reduce cost as well to provide fail-safe designs.

Path 2: Active Control of Noise in Ducts and Pipes. The feedforward algorithm has been recommended for the control of noise associated with a marine diesel engine where a reference signal is accessible (4). For ducts, generally associated with large cross-sectional dimensions, higher-order modes are more likely to exist, requiring a large number of sensors and actuators with an appropriate positioning strategy. For pipes, generally associated with small cross-sectional dimensions, it is expected that only plane wave propagation will exist, thereby limiting the number of elements needed to one sensor and one actuator. The following sensing configurations are possible: microphones, piezoelectric sensors, or accelerometers. For actuation, loudspeakers and inertial actuators are recommended.

Path 3: Active Control of Vibration Propagation in Beam-Type Structures. Feedforward control was recommended for the vibration control of a propeller shaft (4). The configuration of sensors and actuators to be used will depend on the excitation source and on the modal behavior of the shaft. For modal control, the sensors and actuators can be located either on the shaft itself or connected to it by a stationary mechanical link, such as by a bearing mounted on the shaft. Potential mounted actuators include curved piezoelectric actuators (PZT) and magnetostrictive actuators. For wave transmission control, sensors and actuators arrays are required to measure the downstream propagating and evanescent waves and to inject the control waves in the structure.

Mounted sensors to be used include piezoelectric (PVDF) sensors to measure the strain and accelerometers,

Table 2. Properties of Selected Piezoelectric Materials

Property	PZT-4	PZT-5H	BM532 Piezoceramic	Motorola 3203 HD PZT	PZT-G1195N
<i>Elastic properties</i>					
E_{11} (GPa)	81.3	170.0	71.4	1.77	63
E_{22} (GPa)	81.3	170.0	71.4	1.77	63
E_{33} (GPa)	64.5	158.0	5.0	1.77	63
G_{12} (GPa)	30.6	23.0	27.5	0.681	24.2
G_{23} (GPa)	25.6	23.0	27.5	0.681	24.2
G_{13} (GPa)	25.6	23.0	27.5	0.681	24.2
$<_{12}$	0.33	0.3	0.3	0.3	0.3
$<_{13}$		0.3	0.3	0.3	0.3
$<_{23}$		0.3	0.3	0.3	0.3
<i>Piezoelectric coefficients ($10^{12} m/V$)</i>					
d_{31}	-122		-200	-295	-254
d_{32}	-122		-200	-295	-254
d_{33}	285		580	569	374
d_{24}			0	560	584
<i>Piezoelectric stress constants (C/m^2)</i>					
e_{31}		-6.5			
e_{32}		-6.5			
e_{33}		23.3			
e_{23}		17.0			
<i>Electric permittivity</i>					
γ_{11}/γ_0	1480	1695	3250	2418	1729
γ_{22}/γ_0	1480	1695	3250	2418	1729
γ_{33}/γ_0	1300	1695	3250	3333	1695
Mass density Δ (kg/m^3)	7600	7500	7350	7600	

Note: $\gamma_0 = 8.85 \times 10^{-12}$ farad/m, electric permittivity of air.

if the rotation speed permits, for acceleration measurement. The mounted actuators include piezoelectric (PZT) actuators to induce strain in the structure. For robustness, it is recommended that the actuators be combined with passive control elements such as a viscoelastic layer bonded to the shaft.

Path 4: Active Control of Radiated Sound Fields. There are two types of radiated noise to be controlled for ship structures. These are the airborne engine noise into an enclosure, and the noise radiated by the noise into the sea. As stated in (4) both cases require the use of global control techniques that involve multiple input and multiple

Table 3. Recommended Sensors and Actuators for Ship Noise Control

Ship Noise Path	Recommended Sensors and Actuators	
	Sensors	Actuators
Path 1: Active vibration isolation	<ul style="list-style-type: none"> • Force transducers • Accelerometers 	<ul style="list-style-type: none"> • Hydraulic actuators • Electrodynamic actuators
Path 2: Active control of noise in ducts and pipes	<ul style="list-style-type: none"> • Microphones • Piezoelectric sensors • Accelerometers 	<ul style="list-style-type: none"> • Loudspeakers • Electric motors
Path 3: Active control of vibration propagation in beam-type structures	<ul style="list-style-type: none"> • Piezoelectric sensors • Accelerometers • Electrodynamic shakers • LVDT 	<ul style="list-style-type: none"> • Piezoelectric actuators • Magnetostrictive actuators • Electrodynamic shakers
Path 4: Active control of airborne engine noise		
Sound field into enclosures	<ul style="list-style-type: none"> • Combination microphones • Accelerometers 	<ul style="list-style-type: none"> • Loudspeakers
Radiated noise into sea	<ul style="list-style-type: none"> • Piezoelectric sensors • Accelerometers 	<ul style="list-style-type: none"> • Piezoelectric actuators • Magnetostrictive actuators

output transducers. Control of radiated noise can be achieved either by active noise cancellation (ANC) or by active structural acoustic control (ASAC) techniques. For active cancellation, the following sensors and actuators are recommended: (1) combination microphones and accelerators as sensors and (2) loudspeakers as actuators. For active structural acoustic control the following sensors and actuators are recommended: (1) piezoelectric sensors (shaped or not) and accelerometers as sensors, and (2) piezoelectric and magnetostrictive materials as actuators.

SUMMARY AND CONCLUSIONS

Among the wide range of sensor and actuator materials that could be used for active noise and vibration control in ship structures are piezoelectric and electrostrictive materials magnetostrictive materials, shape-memory alloys, optical fibers, electrorheological and magnetoheological fluids, microphones, loudspeakers, electrodynamic actuators, and hydraulic and pneumatic actuators. In making the selection, due consideration must be given to factors such as cost, frequency of the disturbance, operating (marine) environment, experience in other applications, ease of implementation, and the expected performance. In general, the following recommendations are made:

1. Nonacoustic sensors and actuators (e.g., accelerometers, force transducers, hydraulic actuators, piezoelectric materials, and electrodynamic actuators) are best for nonacoustic paths, namely for the engine-mounting system, the drive shafts, and mechanical couplings.
2. Acoustic sensors and actuators (e.g., microphones and loudspeakers) are best for acoustic paths, namely for the exhaust stacks and piping systems, and the air-borne noise.

It was also recommended that the active control strategies be combined with passive treatments whenever possible, to increase the robustness of the control system and to provide a fail-safe design.

BIBLIOGRAPHY

1. C.R. Fuller and A.H. von Flotow. *IEEE Cont. Sys. Mag.* **15**(6): 9–19 (1995).
2. R.R. Leitch. *IEEE Proc.* **134**(6): 525–546 (1987).
3. P.A. Nelson and S.J. Elliott. *Active Control of Sound*. Academic Press, San Deogo, CA, 1992.
4. U.O. Akpan, O. Beslin, D.P. Brennan, P. Masson, T.S. Koko, S. Renault, and N. Sponagle. *CanSmart-99, Workshop*, St.-Hubert, Quebec, 1999.
5. J.C. Simonich. *J. Aircraft.* **33**(6): 1174–1180 (1996).
6. S.S. Rao and M. Sonar. *Appl. Mech. Rev.* **47**(4): 113–123 (1994).
7. V. Giurgiutiu, C.A. Rogers, and Z. Chaudhry. *J. Intell. Mater. Sys. Struct.* **7**: 656–667 (1996).
8. L. Bowen, R. Gentilman, D. Fiore, H. Pham, W. Serwatke, C. Near, and B. Pazol. *Ferroelectr.* **187**: 109–120 (1996).
9. S. Sherrit, H.D. Wiederick, B.K. Mukherjee, and S.E. Prasad. *Ferroelectr.* **132**: 61–68 (1992).
10. Thunder™. Actuators and Sensors, FACE International Corporation Product Information. 1997.
11. E.F. Crawley and J. de Luis. *AIAA J.* **25**(10): 1373–1385 (1987).
12. V.V. Varadan, Y.-H. Lim, and V.K. Varadan. *Smart Mater. Struct.* **5**: 685–694 (1996).
13. H.S. Tzou and R. Ye. *J. Vib. Acous.* **116**: 489–495 (1994).
14. T.S. Koko, I.R. Orisamolu, M.J. Smith, and U.O. Akpan. *SPIE Conf. on Smart Structures and Materials*, Vol. 3039, pp. 125–134 (1997).
15. H. Sumali and H.H. Cudney. *AIAA-94-1406-CP*, pp. 1233–1241 (1994).
16. A. Grewal, D.G. Zimcik, and B. Leigh. *CanSmart Workshop on Smart Materials and Structures*, CSA, St.-Hubert, Quebec, Canada, 1998.
17. D.L. Sutliff, Z. Hu, F.G. Pla, and L.J. Heidelberg. *Proc. 3rd AIAA/CEAS Aeroacoustic Conf.*, Atlanta, GA, AIAA-97-1641, 1997.
24. C. Liang and C.A. Rogers. *J. Intell. Mater. Sys. Struct.* **1**: 207–234 (1990).
25. T.S. Koko, U.O. Akpan, L. Guertin, A. Berry, and P. Masson. DREA Contractor Report, in review (1999).
26. K. Tanaka. *Res. Mechanic* **18**: 251–263 (1986).
39. I.C. Shepherd, R.F. La Fontaine, and A. Cabelli. *J. Sound Vib.*, **130**: 125–135 (1989).
40. Texas Instruments. <http://www.ti.com/sc/docs/dsps/dcs/appnotes.htm> (1999).
41. Micro Mo Electronics. <http://www.micromo.com/index.html> (1999).
42. W.J. Hsueh and Y.J. Lee. *Trans. ASME* **116**(1): 43–48 (1994).
43. T. Kakinouchi, T. Asano, K. Tanida, and N. Takahashi. *Naval Eng. J.* **104**(3): 46–52 (1992).
44. C.R. Fuller, S.J. Elliott, and P.A. Nelson. *Active Control of Vibration*. Academic Press, New York, 1996.
45. S. Douglas and J. Olkin. *Proc. Recent Advances on Active Control*, 1993.
46. C.L. Morfey. *J. Sound Vib.* **1**: 60–87 (1964).
47. A. Bihhadi and Y. Gervais. *Acta-Acoustica* **2**: 343–357 (1994).
48. S. Laugesen. *J. Sound Vib.* **195**(1): 33–56 (1996).
49. A. L'Espérance, M. Bouchard, B. Paillard, C. Guigou, and A. Boudreau. *Appl. Acous.* **57**: 357–374 (1999).
50. A. L'Espérance, M. Bouchard, and B. Paillard. *Canadian Inst. Mining, Light Metal Sec. Metall. Soc.* **90**(1012): 94–99 (1997).
51. W.G. Culbreth, E.W. Hendricks, and R.J. Hansen. *J. Acous. Soc. Am.* **83**: 1306–1310 (1988).
52. R.L. Clark, J. Pan, and C.H. Hansen. *J. Vib. Acous.* **92**(2): 871–876 (1992).
53. M.J. Brennan, S.J. Elliott, and R.J. Pinnington. In *Proc. 1st Int. Conf. on Motion and Vibration Control*, 1992, pp. 605–609.
54. S.J. Elliott, T.J. Sutton, M.J. Brennan, and R.J. Pinnington. In *Proc. IUTAM Symp. on the Active Control of Vibration*, 1994, pp. 1–8.
55. R.B. Mace. *J. Sound Vib.* **114**: 253–270 (1987).
56. X. Pan and C.H. Hansen. *J. Sound Vib.* **165**: 497–510 (1993).
57. S.J. Elliott and L. Billet. *J. Sound Vib.* **163**: 295–310 (1993).
58. A.E. Schwenk, S.D. Sommerfeldt, and S.I. Hayek. *J. Acoust. Soc. Am.* **96**(5): 2826–2835 (1994).
59. P. Audrain, P. Masson, and A. Berry. *J. Acous. Soc. Am.* (1999) in press.

60. M.J. Brennan, M.J. Day, and R.J. Randall. *J. Vib. Acous.* **120**(1): 1–12 (1998).
61. K. Kato, T. Fujii, K. Mizutani, Y. Kurita. *Trans. Jpn. Soc. Mech. Eng.* **C63** (609): (1997).
62. A.B. Palazzolo, S. Jagannathan, A.F. Kascak, T. Griffin, and J. Giriunas. *ASME Int. Gas Turbine and Aero-engine Congress and Exposition* 1993, pp. 1–12.
63. R.J. Bernhard. In *Proc. Active 95*, 1995, pp. 21–32.
64. T.J. Sutton, S.J. Elliott, A.M. McDonald, and T.J. Saunders. *Noise Contr. Eng. J.* **42**: 137–147 (1994).
65. S.J. Elliott, P.A. Nelson, I.M. Stothers, and C.C. Boucher. *J. Sound Vib.* **140**: 219–238 (1990).
66. I.U. Borchers, U. Emborg, A. Sollo, E.H. Waterman, J. Paillard, P.N. Larsen, G. Venet, P. Goransson, and V. Martin. In *Proc. 4th NASA/SAE?DLR Aircraft Interior Noise Workshop* 1992.
67. U. Emborg and C.F. Ross. In *Proc. Recent Advances in Active Control of Sound and Vibration*. 1993, pp. 100–109.
68. C.R. Fuller. *J. Sound Vib.* **109**: 141–156 (1986).
69. J. Pan, C.H. Hansen, and D.A. Bies. *J. Acous. Soc. Am.* **87**: 2098–2108 (1990).
70. C.G. Mollo and R.J. Bernhard. *J. Vib. Acous.* **112**: 230–236 (1990).
71. K.A. Cunefare and G.H. Koopmann. *J. Acous. Soc. Am.* **90**: 365–373 (1991).
72. A.R.D. Curtis, P.A. Nelson, and S.J. Elliott. *J. Acous. Soc. Am.* **88**: 2265–2268 (1990).
73. K.H. Baek and S.J. Elliott. *J. Sound Vib.* **186**: 245–267 (1995).
74. S.J. Elliott. *ISVR Memorandum* 695 (1989).
75. P.A. Nelson, A.R.D. Curtis, S.J. Elliott, and A.J. Bullmore. *J. Sound Vib.* **117**: 1–13 (1987).
76. S.D. Sommerfeldt, J. Parkins, and Y. Park. In *Proc. Active 95*, 1995, pp. 477–488.
77. B.S. Cazzolato. Thesis. Department of Mechanical Engineering, University of Adelaide, 1998.
78. S.D. Sommerfeldt and P.J. Nashif. *J. Acous. Soc. Am.* **96**: 300–306 (1994).

VIBRATIONAL ANALYSIS

DAVE S. STEINBERG
Westlake Village, CA

INTRODUCTION

Vibration is present almost everywhere we travel in modern society. Vibrationally induced failures are very common in products such as television sets and computers that are shipped by trains and trucks. Vibrational failure in a television set may be just an inconvenience. However, vibrational failure in a large passenger airplane can lead to many deaths. Methods of vibrational analysis are available that are accurate and can reveal weak structural areas. Steps can then be taken either to repair or replace critical items. Vibrational analysis is a combination of science and art. The science uses sophisticated computers extensively to solve large complex problems. This method requires extensive training and often takes a long time to reach a satisfactory solution. The art uses approximations, short cuts, and test data to reduce the time needed to reach a

satisfactory solution. The approximations and short cuts can sharply reduce the time required for a solution, but it also reduces the accuracy of the analysis. Vibrational analysis can be used to make some materials work smarter by making small changes in their physical properties. These changes can often increase the fatigue life of critical structural members without a significant increase in the size, weight, cost, or impact on production and delivery schedules.

VIBRATIONAL REPRESENTATION

In a broad sense, vibration means an oscillating motion, where something moves back and forth. If the motion repeats itself, it is called periodic. If continuous motion never repeats itself, it is called random motion. Simple harmonic motion is the simplest form of periodic motion, and it is typically represented by a sine wave, as shown in Fig. 1. The reciprocal of the period is known as the frequency, and it is measured in cycles per second, or hertz (Hz). The maximum displacement is called the amplitude of the vibration.

DEGREES OF FREEDOM

A coordinate system is usually used to locate the positions of various elements in a system. When only one element is involved, it is restricted to moving along only one axis, and only one dimension is required to locate the position of the element at any instant, then it is called a single-degree-of-freedom system. The same is true for a torsional system. When one element is restricted to rotating about one axis so that only one dimension is required to locate the position of the element at any instant, it is a single-degree-of-freedom system. Two degrees of freedom requires two coordinates to locate the positions of the elements, and so on.

A single rigid body is usually considered to have six degrees of freedom, translation along each of the three orthogonal x , y , and z axes and rotation about each of the same three axes. Real structures are usually considered to have an infinite number of degrees of freedom.

VIBRATIONS OF SIMPLE STRUCTURES

The natural frequency (often called the resonant frequency) of a simple single-degree-of-freedom system can

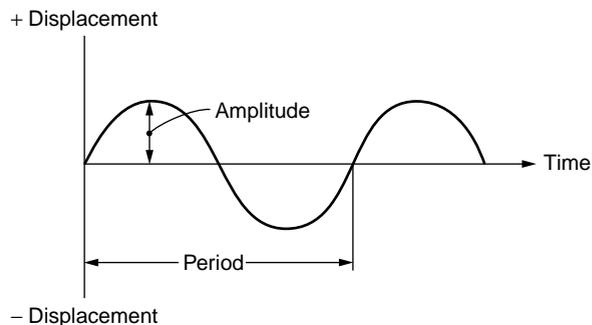


Figure 1. Simple harmonic motion: a sine wave.

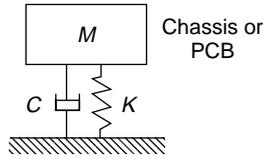


Figure 2. Single-degree-of-freedom spring-mass system.

often be obtained from the strain energy and the kinetic energy of the system. Consider the single spring and mass system shown in Fig. 2. When there is no damping in the system, then no energy is lost, and the strain energy must be equal to the kinetic energy. This results in the natural frequency equation (1),

$$f_n = \frac{1}{2\pi} \sqrt{\frac{g}{Y_{st}}} \quad (\text{Hz}), \quad (1)$$

where

$g = 9.80 \text{ m/s}^2$ (386 in/s^2), the acceleration of gravity and Y_{st} in meters (inch) is the static displacement.

Sample Problem: Natural Frequency of a Simple Structure

When the static displacement of a structure $Y_{st} = 1.27 \times 10^{-5} \text{ m}$ (0.00050 in), its natural frequency is 140 Hz .

The natural frequency is important because it is often considered the heart of a vibrating system. It influences the number of fatigue cycles and the displacement, which affect the fatigue life of a system. It also influences the damping, which affects the dynamic acceleration Q level, and the stress level, which also affects the fatigue life.

NATURAL FREQUENCIES OF UNIFORM BEAM STRUCTURES

Natural frequencies of uniform beam structures can be determined by equating the strain energy to the kinetic energy without damping. This method of analysis leads to simple solutions and very little error because beam types of structures normally have very little damping. The resulting equations for natural frequency apply to uniform beams that are forced to bend only in the vertical axis without bending in the horizontal axis and without torsion or twisting. The beam equation is (1)

$$f_n = \frac{a}{2\pi} \sqrt{\frac{EIg}{WL^3}} \quad (\text{Hz}), \quad (2)$$

where

$a = 3.52$ for a cantilevered beam,

$a = \pi^2 = 9.87$ for a beam that is supported (hinged) at each end,

$a = 22.4$ for a beam that is clamped (fixed) at both ends, E in newtons (N) m^2 (lb/s in^2) is the modulus of elasticity for beam material,

I in m^4 (in^4) is the area moment of inertia for a beam cross section,

$g = 9.80 \text{ m/s}^2$ (386 in/s^2), the acceleration of gravity,

W in Ns (N) (lb) is the total weight of the beam, and

L in m (in.) is the length of the beam between supports.

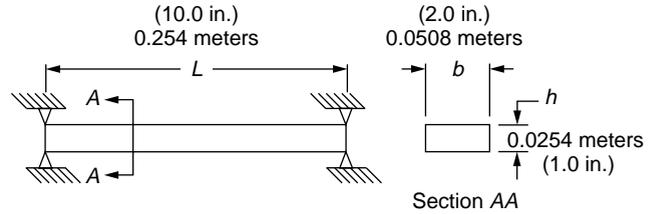


Figure 3. Uniform beam simply supported at each end.

Sample Problem: Natural Frequency of a Simply Supported Uniform Beam

For example, consider the simply supported (hinged) aluminum beam shown in Fig. 3, where $E = 6.894 \times 10^{10} \text{ N/m}^2$ ($10 \times 10^6 \text{ lb/in}^2$), $L = 0.254 \text{ m}$ (10.0 in), $I = 6.937 \times 10^{-8} \text{ m}^4$ (0.1667 in^4), and $W = 8.896 \text{ N}$ (2.0 lb). The resulting natural frequency is 890 Hz .

NATURAL FREQUENCIES OF UNIFORM PLATES AND CIRCUIT BOARDS

The natural frequencies of different types of flat, uniform plates that have different types of supports can often be obtained by using trigonometric or polynomial series (1). Again, when damping is ignored, the strain energy can be equated to the kinetic energy of the bending plate to obtain the natural frequency. A printed circuit board (PCB) that supports and electrically interconnects various electronic components can be analyzed as a flat rectangular plate, often simply supported (hinged) on all four sides, that has a uniformly distributed load across its surface. The natural frequency for this type of installation is (2).

$$f_n = \frac{\pi}{2} \left(\frac{D}{\rho} \right)^{1/2} \left(\frac{m^2}{a^2} + \frac{n^2}{b^2} \right) (\text{Hz}), \quad (3)$$

where

E in newtons/ m^2 (lbs/in^2) is the modulus of elasticity for plate material,

h in m (in) is the plate thickness,

μ is Poisson's ratio, dimensionless,

$D = \frac{Eh^3}{12(1-\mu^2)} \text{ N m}$ (lb in), the plate bending stiffness factor, (4)

$\rho = \frac{W}{gab} \text{ N s}^2/\text{m}^3$ ($\text{lb s}^2/\text{in}^3$), the mass per unit area, (5)

$g = 9.80 \text{ m/s}^2$ (386 in/s^2), the acceleration of gravity,

W in newtons (lb), is the total weight of the PCB,

a in m (in) is the length of the plate,

b in m (in) is the width of the plate, and

m and n are integers: first harmonic $m = 1, n = 1$;

second harmonic $m = 2, n = 1$;

third harmonic $m = 1, n = 2$; fourth harmonic $m = 2, n = 2$.

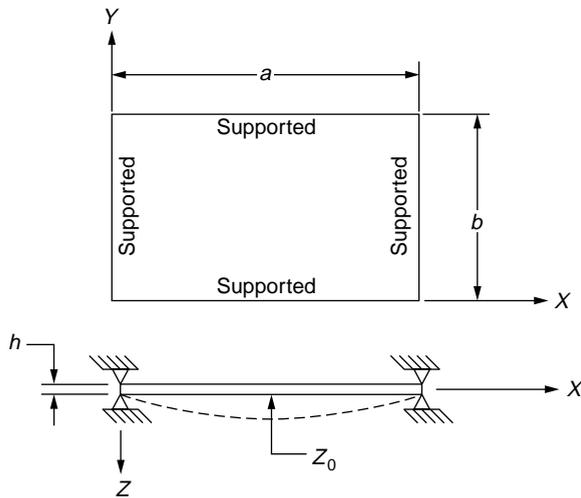


Figure 4. Uniform flat plate simply supported on four sides.

Sample Problem: Natural Frequency of a Rectangular PCB (see Fig. 4).

Consider a flat rectangular epoxy fiberglass PCB, supported (hinged) on four sides, where $E = 1.379 \times 10^{10}$ N/s m^2 , (2.0×10^6 lb/in 2), $h = 0.00157$ m (0.062 in), $\mu = 0.12$ dimensionless, $D = 4.53$ N (40.1 lb in), $W = 4.448$ N (1.0 lb), $a = 0.203$ m (8.0 in), $b = 0.178$ m (7.0 in), $\rho = 12.56$ Ns $^2/m^3$ (0.463×10^{-4} lb s $^2/in^3$). The resulting natural frequency for the first harmonic ($m = 1, n = 1$) is 52.6 Hz.

METHODS OF VIBRATIONAL ANALYSIS

Hand calculations are still being used extensively for simple sinusoidal and random vibrational analyses in small companies due to the high costs of the computers, the specialized computer software, and the skilled personnel to operate the computers. Many reference books are available that show how to perform simplified vibrational analyses on different types of simple structures. However, when large complex structures are involved, hand calculations are not adequate to ensure reasonable accuracy. Small companies often subcontract the work to outside consulting organizations that specialize in these areas. Sometimes it can be cheaper, faster, and more accurate to build a model of the structure, so it can be examined in a vibrational test laboratory.

Most large companies rely extensively on various types of computers and specially formulated finite element modeling (FEM) software programs for vibrational analyses. Their computers are usually networked together, so each has access to the wide variety of software analytical programs available on the network. The new desktop personal computers (PC) are very popular for vibrational analyses using FEM. They are more powerful and faster than the large main frame computers of a few years ago.

PROBLEMS OF VIBRATIONAL ANALYSIS

Almost all computers and computer software FEM programs for vibrational analysis agree within about 2% when

they are used to determine eigenvalues (resonant frequencies) and eigenvectors (mode shapes) for many types of complex structures. However, sample problems solved by using different FEM software programs have shown significant variations in their stress values. The stress values from four different FEM programs had a total variation of about 60%. This was 30% above the average stress value of the four programs and 30% below the average value for similar models of the same structure, subjected to the same type of vibrational excitation. Different computer FEM programs typically use different algorithms to define the building blocks for their various beam, plate, and brick elements. These algorithmic variations probably cause the variations in the stress values. Because the fatigue life of a structure is closely related to its stress value, significant variations in the calculated stress levels can result in dramatic changes in the calculated fatigue life of a structure. For example, the results of this investigation showed that the fatigue life at the critical point in the structure can be expected to vary across a wide range because of the variations in the calculated stress values. The fatigue life in the lead wires of PCB electronic component parts can be as much as five times greater than the average calculated fatigue life, or it can be as little as one-fifth of the average calculated fatigue life. [See Ref. 5, Chap. 12, Figs. 12.1–12.19 for more detailed information on finite element modeling.]

The results shown before may vary substantially. Only four different FEM software programs were involved in this investigation. At least several dozen new software programs are available now. When the different modeling techniques of different computer analysts are considered, these factors are expected to have a significant impact on the computer calculated stress values and the resulting calculated fatigue life.

PROBLEMS OF MATERIAL PROPERTIES

Material properties are often difficult to evaluate for vibrational environments. The life of any structure excited by vibration depends on the fatigue properties of the most critical materials used in fabricating and assembling the structure. When structural elements are forced to bend and twist back and forth, perhaps millions of times in severe vibrational environments, three very important factors have to be defined:

1. the very basic fatigue properties of the materials used in the structure
2. The manufacturing tolerances that will affect the physical dimensions of the parts
3. stress concentrations due to holes, notches, small radii, and rapid sectional changes.

First, consider the basic fatigue properties of the materials in the structure. Test data have shown that there are significant variations in the fatigue life of virtually identical parts that are machined to very close tolerances from the same forging, as shown in Fig. 5. These test data are plotted on log–log curves of stress (S) against the number of cycles (N) to failure. Only one average straight-line

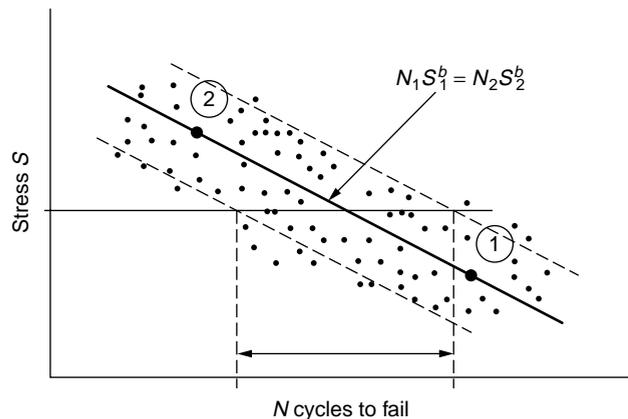


Figure 5. S - N fatigue curve showing large variations in the life test data.

usually represents the fatigue life properties of a material (1,3-5). When all of the failure test data points for all of the test samples are plotted, a wide variation in the fatigue life is revealed. Because these are log-log plots, the spread in the possible variation in fatigue life of virtually identical parts can be very great, sometimes reaching values of 10 to 1. Engineers involved in vibrational and fatigue life analysis do not like to reveal this type of data to upper management personnel. Personal experience with nontechnical upper management people is that they often expect mechanical designers and analysts to predict the fatigue life of their structures to within plus and minus 20%. This is an almost impossible task, when all of the possible variations are considered.

To compensate for these large variations in fatigue life of virtually identical structural elements, safety factors (sometimes called factors of ignorance) must be used when these structures are being designed and analyzed. Building models for vibrational life testing in a laboratory can be a great help in estimating the fatigue life of a structure. However, if no tests are run or if the number of samples tested is low, there is always the danger of erratic bursts of high failure rates in the production units because of the large scatter associated with fatigue.

Next, consider the effects of manufacturing tolerances on the physical dimensions of the structural elements in an assembly. Mass-produced products always show some variations in the physical dimensions of what appear to be identical parts. Even die cast parts that are made from the same mold have slightly different physical dimensions. Some manufactured devices, like the automatic transmission in an automobile, can have many precision gears, ground to very close tolerances. Holding very tight manufacturing tolerances can be very expensive. Therefore, looser tolerances are used in production parts that do not require tight tolerances for precision assembly work because they this reduce costs. When manufactured parts that have loose tolerances are exposed to severe vibration, the failure rates often go up and down erratically. Changes in the physical dimensions of load-carrying structural members can alter the load path through the structure, which can change the dynamic loads and stresses in it. It is too expensive to keep

track of manufactured parts that have extremes in their dimensional tolerances. These parts can be anywhere in large production programs. This means that failures which are difficult to predict and to control, may occur randomly in harsh environments.

To reduce costs, for example, the electronics industry tends to use very loose tolerances in the dimensions that control the external physical sizes of the length, width, and thickness of their printed circuit boards (PCBs) and electronic component parts. These large variations in tolerance of these parts further increase the difficulty in trying to predict the fatigue life accurately of electronic assemblies that are exposed to different vibrational environments.

RELATION OF DISPLACEMENT TO ACCELERATION AND FREQUENCY

Vibrational displacements are often very small, so they are difficult to observe during vibrational tests. Because these displacements are small, it does not mean that the resulting stresses are also small. Vibrational environments usually impose alternating displacements and alternating stresses on various structural load-carrying elements within a system. If the vibrating system experiences many thousands of stress reversals, fatigue failures can occur in critical structural members, even at relatively low displacements and stress levels. This is the nature of fatigue failures that occur at relatively low stress levels near small holes, small notches, and sharp bends. These geometric shapes are known as stress concentration factors, which can increase peak stress levels in these areas by a factor of 3 or 4 or more (4).

When vibrational tests are run in a laboratory, the normal procedure is to use small accelerometers to monitor the resulting acceleration values in different parts of the structure. When an electrodynamic shaker is used to generate a sinusoidal wave for the vibrational test, the electronic control system will show the frequency of the imposed wave in cycles per second, or hertz (Hz). With this type of setup, the test engineer will know the acceleration level and the frequency at any instant. This information is often incomplete without the resulting displacement at any instant. The resulting displacement at any instant can be obtained by considering a rotating vector that generates a sinusoidal wave based on the full relationship (1),

$$Y = Y_0 \sin \Omega t, \quad (6)$$

where

Y is the displacement at any time,

Y_0 is the maximum single amplitude displacement from zero to peak, and

$\Omega = 2\pi(f)$ rad/s, the frequency.

The acceleration a can be obtained from the second derivative of the displacement with respect to time from the preceding equation. The maximum acceleration occurs when the sine function is one. It is convenient to represent

the acceleration in terms of gravity units G :

$$G = \frac{a}{g} \quad (\text{gravity units, dimensionless}), \quad (7)$$

where

a in m/s^2 (in/s^2) is the acceleration level and
 $g = 9.80 \text{ m/s}^2$ (386 in/s^2), the acceleration of gravity.

The final results show the displacement Y_0 in terms of the frequency f in Hz and the number of dimensionless gravity units G :

$$Y_0 = \frac{AG}{f^2} \quad (\text{single amplitude displacement}), \quad (8)$$

where

$A = 0.248$ for meters displacement (9.8 for inch displacement)

G is the acceleration, in gravity units, dimensionless (same in English units), and

f is the frequency in cycles/s (Hz) (same in English units).

Sample Problem: Finding the Displacement from the Frequency and the G Level

For example, when the acceleration G level is 3.0 dimensionless gravity units and the frequency is 120 Hz, the single amplitude displacement is 0.0000517 m (0.00204 in).

This equation is probably the most important relationship in the entire field of dynamics. It shows that when any two of the parameters of Y_0 , G or f , are known, then the third parameter is automatically known. This equation can be used for sine vibration, random vibration, shock, and acoustics (1).

EFFECTS OF VIBRATION ON STRUCTURES

Vibrational environments can dramatically magnify the dynamic forces and stresses in different types of structures, when the structural natural frequencies are excited. Forces and stresses can be magnified and amplified by factors of 10, 30, and even 100 in many different types of structures for different types of vibrational excitation. The magnitude of the magnification, called the transmissibility Q , often depends on the amount of damping in the vibrating system. Figure 6 shows damping for a single-degree-of-freedom system. There are very few single-degree-of-freedom systems in the real world. For example, consider a two-degree-of-freedom system for an electronic assembly where the chassis is mass 1. The plug-in PCBs are attached to the chassis so they are mass 2. The response of mass 1 will be the input to mass 2. Testing experience, including different damping methods, has shown that the transmissibility Q of PCBs as mass 2 will depend far more on the dynamic coupling phase relation and frequency ratio between mass 1 and mass 2 than the damping in either mass 1 or mass 2 because the transmissibility Q 's between masses 1 and 2 do not add, they multiply.

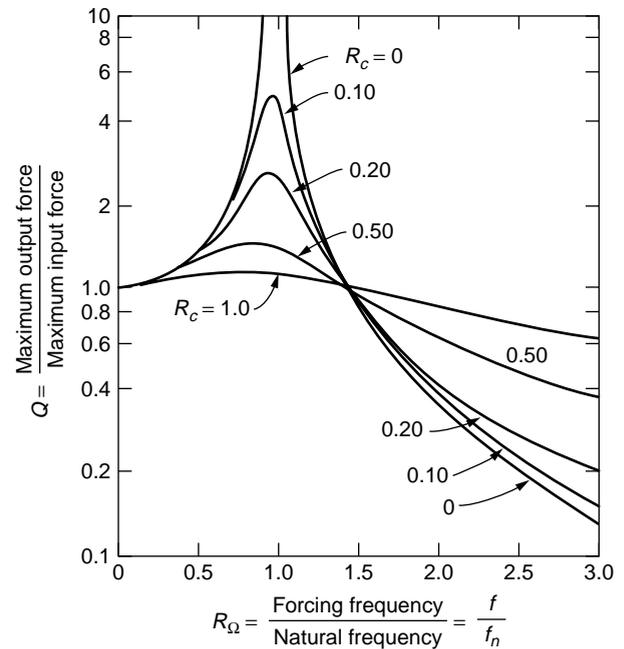


Figure 6. Effects of damping on the transmissibility Q plots.

The Q of a system is defined as the ratio of the output (or response of the system) divided by the input. The output and the input are usually defined in terms of the displacements, or the acceleration values. If the damping in a simple system is zero, the vibration theory states that the value of the transmissibility Q will be infinite. If the transmissibility Q is infinite, the resulting dynamic forces and stresses will also be infinite. However, because all real systems have some damping, Q can never be infinite. However, in lightly damped systems, Q can be very high. A high Q will result in high forces, displacements, and stresses, which can sharply reduce the fatigue life of the structure.

ESTIMATING THE TRANSMISSIBILITY Q IN DIFFERENT STRUCTURES

The transmissibility Q is strongly influenced by the damping in a vibrating structure. One form of damping is the conversion of kinetic energy into heat. This can be shown by rapidly bending a metal paper clip back and forth about 20 times through a large angle. Immediately place your finger on the paper clip in the bending area. This area will be quite warm. It may even be hot. The strain energy of bending has been converted into heat energy, which cannot be converted back into strain energy. It is lost energy. When heat energy is lost, it means there is also a loss of kinetic energy. Therefore, when damping is increased in a vibrating system, there is less energy available to convert into kinetic energy. Less kinetic energy means that there is less energy available to excite the structure at its natural frequency, so that the transmissibility Q is decreased. Conversely, when there is a decrease in the damping, this makes more kinetic energy available to excite the structure, so the transmissibility Q is increased.

In general, simple systems that have only a few structural elements have less damping than more complex systems that have many structural elements, when both systems are subjected to the same vibrational environment. Bolted joints usually have a lot of friction and damping at the bolted interfaces in vibrational environments, so structures that have many bolted joints usually have high damping. Therefore, a simple beam type of structure usually has less damping than a more complex plate type structure. Then, a beam structure should have a higher transmissibility Q than a plate structure in the same vibrational environment. The same thinking can be applied to a more complex box type of structure that has removable bolted covers to provide access to internal subassemblies. The box type of structure should have more damping than the plate structure for similar vibrational exposure because the box structure is much more complex than a plate structure. This means that the box structure should have a lower transmissibility Q than the plate structure for similar vibrational exposure. Extensive vibrational test data shows that this is the natural trend for damping in different types of structures.

Higher dynamic forces in a structure typically result in higher dynamic stresses and higher dynamic displacements. This results in higher damping, which reduces the dynamic transmissibility Q for that system. Therefore, higher acceleration G levels can be expected to result in lower transmissibility Q values.

Higher natural frequencies result in lower dynamic displacements, when the acceleration G level is held constant, as shown in Eq. (8). Lower displacements mean lower stresses. Lower stresses reduce damping. Lower damping increases the transmissibility Q value. Therefore, higher frequencies, at the same G level, increase the value of Q .

Vibrational test data from different types of structures can be used to estimate the transmissibility Q values expected for different types of common systems at the start of a preliminary vibrational analysis. The three most common types of structures in the order of their complexity are beams, plates, and enclosed boxes that have bolted covers. The approximate transmissibility Q for these three types of structures is

$$Q = J \left[\frac{f_n}{(G_{\text{in}})^{0.6}} \right]^{0.76}, \quad (9)$$

where

$J = 1.0$ for a beam type of structure (cantilever or restrained at each end),

$J = 0.50$ for a plate type of structure (supported around the perimeter),

$J = 0.25$ for a box type of structure,

f_n in Hz is the natural frequency of the structure, and G_{in} is the input acceleration G level in dimensionless gravity units.

Sample Problem: Finding the Approximate Q for Beams and Plates

For example, consider a beam structure whose natural frequency is 300 Hz and input acceleration level is 0.25 G in a sine vibrational test. The expected transmissibility Q is about 144. Now increase the input acceleration level to 5.0 G . The expected transmissibility Q will now drop to about 37. Next, consider the plate structure for the same conditions. For a 0.25 G input, the Q is about 72. For a 5.0 G input, the Q is about 18.3. Now take the square root of the 300-Hz resonant frequency for the plate structure, which is 17.3. This shows that a good approximation for the plate Q (frequently used for PCBs) is the square root of the natural frequency, when the input level is about 5 G (1,2);

$$\text{Good PCB approximation of } Q = \sqrt{f_n}. \quad (9a)$$

This demonstration should be taken as a warning. Performing vibrational tests at very low input acceleration G levels will result in very high transmissibility Q values. Very low input G levels are often used to prevent damage to prototype PCBs. These types of tests are not valid for evaluating PCBs that must operate at much higher G levels. Vibrational tests should be run on prototypes using the correct input G levels to verify the correct dynamic characteristics of the test specimen and future production models.

METHODS FOR EVALUATING VIBRATIONAL FAILURES

Vibration can cause failures in many different types of structures ranging from earthquakes and airplanes to electric knives and washing machines. Vibrational failures are often experienced during vibrational tests to evaluate the reliability of a product. Sinusoidal vibration is very useful in tests to diagnose the cause of specific structural vibrational failures, to determine the transmissibility Q of a structure, and to find the fatigue life of different types of structures. A very effective device that is often used in sinusoidal vibrational tests is the strobe light. This often allows the observer to see just how structures bend and twist during resonance. This information can be critical in determining why and where a structure will fail. Steps can then be taken to modify the structure to prevent future failures.

Finite element modeling (FEM) programs are available for use with new high speed small PCs that can generate models of very complex structural systems. When the computer models are generated by skilled engineers, the dynamic results from the model are often very similar to the actual vibrational test results. The problems most often encountered in these areas are the types of models that are generated by individuals who are familiar with FEM but do not have any real testing experience. The resulting structural models may look good, but their vibrational response will often have gross errors due to improper boundary conditions and to improper damping values. These parameters can be obtained only from extensive vibrational testing experience (1).

Vibrational failures are often difficult to trace. Sometimes the failures result from poor design, poor maintenance, or poor manufacturing processes. Very often the failures are a combination of all three. These failures are usually difficult to trace because it is often very difficult to get the information necessary to implement any corrective action.

DETERMINING DYNAMIC FORCES AND STRESSES IN STRUCTURES DUE TO SINE VIBRATION

Dynamic forces in a structure can be obtained from Newton's equation where force F is equal to mass m times acceleration a . When weight W is used with the acceleration of gravity g , and structural accelerations are in terms of dimensionless gravity units G , the following relationships are convenient:

$$F = ma, \quad (10)$$

where

$$m = \frac{W}{g}$$

and

$$G = \frac{a}{g}.$$

Substitute in this equation to obtain the dynamic force F (1).

Then,

$$F = WG. \quad (11)$$

When the dynamic transmissibility Q is included in this equation and the input acceleration level is shown in dimensionless gravity units G_{in} , then the maximum output (or response F_{out}) dynamic force due to sine vibration is obtained:

$$F_{out} = WG_{in}Q. \quad (12)$$

Sample Problem: Finding the Natural Frequency, Transmissibility Q , Dynamic Force, Displacement, and Stress in a Beam Excited by Sine Vibration

Consider the simply supported (hinged) weightless aluminum beam, shown in Fig. 7, that has a modulus of elasticity E of 7.238×10^{10} Nm² (10.5×10^6 lb/in²), a concentrated load W of 8.896 N (2.0 lb) acting at the center of the beam, a length L of 0.203 m (8.0 in), a cross-sectional width of 0.0305 m (1.2 in), a thickness of 0.0127 m (0.50 in), and an area moment of inertia of 5.206×10^{-9} m⁴ (0.0125 in⁴). Find the natural frequency, the transmissibility Q , the maximum expected dynamic force, the dynamic displacement, and the maximum expected bending stress in the beam due to a 5- G sine vibrational input.

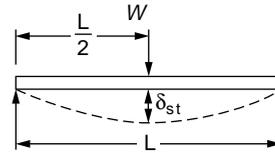


Figure 7. Simply supported beam that has a concentrated load at its center.

Beam Natural Frequency. The beam natural frequency can be obtained from the static displacement Y_{st} of a beam that has a concentrated load, using standard beam equations from a handbook (1):

$$Y_{st} = \frac{WL^3}{48EI} \quad (13)$$

Substituting the physical properties for the beam in the preceding equation results in a static displacement of 4.114×10^{-6} m (1.62×10^{-4} in). Substituting these numbers in Eq. (1) where the acceleration of gravity is 9.80 m/s² (386 in/s²), results in a natural frequency of 246 Hz.

Beam Transmissibility Q . The transmissibility Q for the beam in a 5- G input sine vibrational environment can be obtained from Eq. (9), where $J = 1.0$ and the natural frequency is 246 Hz. This results in a Q value of about 31.5.

Dynamic Output Force on Beam. The dynamic force acting on a beam can be obtained from Eq. (12); the given concentrated load is W , the sine input level is 5 G , and the transmissibility Q is 31.5. This results in an output force of 1401 N (315 lb).

Single Amplitude Dynamic Displacement of Beam. The single amplitude dynamic displacement at the center of the beam can be obtained by using Eq. (8) and adding the transmissibility Q for sine vibrational, as shown in Eq. (14). See Eq. (8) for values of A .

$$Y_0 = \frac{AG_{in}Q}{f_n^2} \quad (14)$$

For a 5- G sine input, a transmissibility Q of 31.5 and a natural frequency of 246 Hz, the single amplitude displacement is expected to be about 0.000645 m (0.0254 in).

Maximum Dynamic Bending Stress in Beam for Sine Vibration. Equation (15) gives the dynamic bending stress S_b . Stress occurs at the center of the beam, as shown in Fig. 8. A dimensionless geometric stress concentration factor (k) should be included when machined parts will be exposed to tens of thousands of stress reversals in vibrational environments. These types of fabricated parts usually have small defects in the form of cuts, scrapes, and scratches, which are known as stress risers or stress concentrations. These defects increase the magnitude of the local stresses which reduces the fatigue life of the structure. The stress concentration must be used *only once*. It can be used directly as shown in Eq. (15), or it can be used

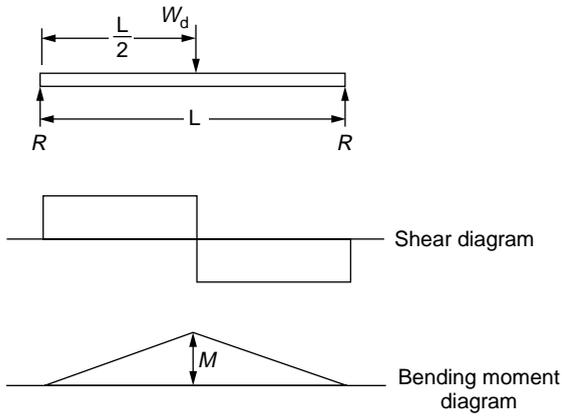


Figure 8. Shear and bending moment diagram for a beam that has a concentrated load.

to modify the slope of the fatigue curve shown in Eqs. (18) and (19), but not in both places.

$$S_b = \frac{kMc}{I} \tag{15}$$

A stress concentration factor of about 2 is a good place to start preliminary stress investigations. The dynamic bending moment M can be obtained from the geometry of the beam, using the reaction force R , as follows:

$$M = \frac{RL}{2} \tag{16}$$

Because of symmetry, the reaction R will be half of the dynamic load or 700.5 N (157.5 lb). Using a length L of 0.203 m (8.0 in) results in a bending moment of 71.1 N m (630 lb in). The c distance is half the beam thickness, or 0.00635 m (0.25 in). Using the value of $5.20 \times 10^{-9} \text{ m}^4$ (0.0125 in⁴) for the moment of inertia from Eq. (13) and substituting it in Eq. (15) results in a dynamic bending stress of $1.737 \times 10^8 \text{ N/m}^2$ (25,200 lb/in²).

DETERMINING THE FATIGUE LIFE IN A SINE VIBRATIONAL ENVIRONMENT

Accurate fatigue properties of materials that have varying stress concentrations are very difficult to obtain. Figure 5 shows there is a great deal of scatter in typical fatigue data. The normal method for calculating the approximate fatigue life from a known stress value is to use the $S-N$ (stress versus number of cycles to failure) curve for the particular material involved in the investigation. If the fatigue properties of the materials are unknown, then the fatigue life cannot be calculated. Tests should be run on prototypes or on structural members to establish their fatigue properties. If the fatigue properties of the materials are not known, then there is a very great risk of many fatigue failures in production units that will be exposed to vibrational environments.

When the fatigue properties of the materials are known, these properties are often plotted using a sloped line on a log-log curve. The typical fatigue curve for the aluminum

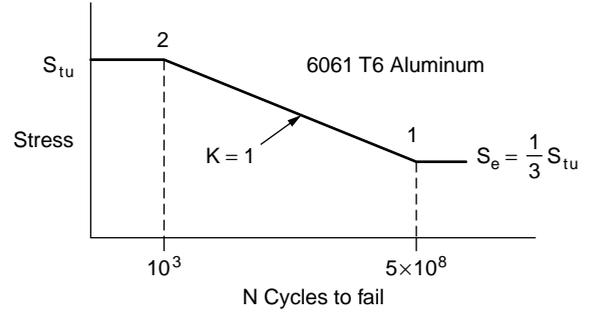


Figure 9. $S-N$ fatigue curve for a smooth specimen of 6061 T6 aluminum.

alloy 6061 T6 shown in Fig. 9 will be used in the sample problem following.

Sample Problem: Finding the Fatigue Life of a Beam Excited by Sine Vibration

The approximate fatigue life of the beam in the previous sample problem can be obtained from Fig. 9 along with the calculated bending stress value of $1.737 \times 10^8 \text{ N/m}^2$ (25,200 lb/in²), as obtained from Eq. (15). The following fatigue damage equations can be used in several different ways to obtain the fatigue life of the structure; t is time and Z is displacement:

$$N_1 S_1^b = N_2 S_2^b,$$

or

$$t_1 G_1^b = t_2 G_2^b,$$

or

$$Z_1 G_1^b = Z_2 G_2^b. \tag{17}$$

The slope b of the fatigue line can be obtained by modifying Eq. (17) and using the Fig. 9 reference break points in the following equation:

$$\frac{N_1}{N_2} = \left(\frac{S_2}{S_1} \right)^b \tag{18}$$

Let N_1 be 5×10^8 , N_2 be 10^3 , S_2 be $3.102 \times 10^8 \text{ N/m}^2$ (45,000 lb/in²), and S_1 be $1.034 \times 10^8 \text{ N/m}^2$ (15,000 lb/in²). Substitute in the preceding equation for the slope:

$$\frac{5 \times 10^8}{10^3} = \left(\frac{3.102 \times 10^8}{1.034 \times 10^8} \right)^b,$$

or

$$5 \times 10^5 = 3^b,$$

or

$$\log 5 \times 10^5 = b \log 3,$$

$$b = \frac{5.699}{0.477} = 11.95 \text{ (slope of fatigue line)}. \tag{19}$$

The b exponent was obtained *without* any stress concentration or safety factor in the material stress. A stress

concentration or safety factor should always be used in vibration. Sometimes, it is more convenient to use a stress concentration (typically 2) directly in the stress value, as in Eq. (15). Sometimes it is more convenient to use the stress concentration in the S - N fatigue curve. Either method can be used, as long as the safety factor is not *used twice*. If a safety factor of 2 is used in the S - N fatigue curve, the value of the b exponent is typically 6.4 for nonferrous alloys and 8.3 for ferrous alloys.

The approximate fatigue life of the vibrating beam can be obtained from the bending stress level $1.737 \times 10^8 \text{ N/m}^2$ (25,200 lb/in²) which was obtained from Eq. (15). Use reference points N_1 at 5×10^8 cycles and S_1 at $1.034 \times 10^8 \text{ N/m}^2$ (15,000 lb/in²) at the right break point in Fig. 9. This results in the number of cycles to failure:

$$N_2 = N_1 \left(\frac{S_1}{S_2} \right)^b = (5 \times 10^8) \left(\frac{1.034 \times 10^8}{1.737 \times 10^8} \right)^{11.95} \quad (20)$$

$$= 1.015 \times 10^6 \text{ cycles to failure}$$

The fatigue life expected at a natural frequency of 246 Hz for a sine resonant dwell condition can be obtained as

$$\text{Life} = \frac{1.015 \times 10^6 \text{ cycles}}{\frac{246 \text{ cycles}}{\text{s}} \times \frac{3600 \text{ s}}{\text{h}}} = 1.146 \text{ h to failure} \quad (21)$$

EFFECTS OF HIGH VIBRATIONAL ACCELERATION LEVELS

High vibrational acceleration levels can result in many different types of failures in different types of systems. High vibrational acceleration levels can be generated by earthquakes, explosions, aircraft buffeting, gunfire, unbalanced rotating devices, rough roads, and rough tracks, to name a few. Vibrational isolators are often used in the foundations of large buildings to protect them from earthquakes. Vibrational isolators are often used on military naval ships and submarines to protect sensitive equipment such as electronics from explosions. When vibrational isolation systems cannot be used, then brute force methods must be used to reinforce structural elements to keep them from failing. The method of reinforcing often results in very large, heavy, and expensive products.

High vibrational acceleration levels can be caused by high input acceleration levels, by severe coupling between adjacent structural elements, and by very low damping in the structure. High vibrational acceleration levels are often caused by careless structural designs, where the natural frequencies of closely linked structural members are very close together. When this happens, the transmissibilities of the adjacent structural elements *multiply*, they do not add. This can cause very rapid failure in almost any structure.

High accelerations in electronic systems can result in large PCB deflections, which can cause impacting between PCBs, high stresses, and rapid failures in the electrical lead wires and solder joints of the components mounted on the PCBs, when the PCBs are forced to bend back and forth thousands of times. High PCB displacements can break pins on electrical connectors and cause electrical short

circuits and cracked components. High acceleration levels can cause relays to chatter, crystal oscillators to malfunction, potentiometers slugs to slip, electrical failures, and cracked castings. Cables and harnesses can whip around causing wires and connections to fail.

MAKING STRUCTURAL ELEMENTS WORK SMARTER IN VIBRATION

One of the biggest problems in structures exposed to vibration is severe coupling between adjacent structural members. PCBs mounted within a chassis or an enclosure are a good example. When the enclosure has an input vibrational acceleration level of 10 G s and a Q of 10 and the PCBs have a Q of 10 and a natural frequency close to the enclosure, the PCBs experience acceleration levels of $10 \times 10 \times 10$ or 1000 G s. Acceleration levels this high cause electronic failures in just a few seconds.

Using the Octave Rule to Improve Vibrational Fatigue Life

One way that structures can be made to work smarter is to design them to follow the octave rule. Octave means to double. When adjacent structural members have natural frequencies that are separated by an octave, or by a factor of 2 to 1, they cannot experience severe coupling.

It does not matter if the natural frequency of each PCB is two (or more) times greater than the natural frequency of the outer housing or if the natural frequency of the outer housing is two (or more) times greater than the natural frequency of each PCB. As long as the natural frequencies of these adjacent structural members are separated by a ratio of 2 (or more), there will be a large reduction in the coupling between them, *as long as the weight of the PCB is very small compared to the weight of the housing*. If high shock levels are also expected, then it is best to use the *reverse octave rule*. The *reverse octave rule* applies when the natural frequency of the outer housing is two (or more) times greater than the natural frequency of any PCB (1).

The *reverse octave rule* works only in dynamic systems where the weight of each PCB is much smaller than the weight of the outer housing (or enclosure). Much smaller means by a factor of 10 or more. In other words, the weight of the enclosure must be more than ten times greater than the weight of any one PCB in that enclosure. If this ratio is not followed, severe dynamic coupling can occur and cause problems.

There are never any problems using the *forward octave rule*, where the natural frequency of each PCB is two or more times greater than the natural frequency of the outer enclosure. This works well no matter what the weight ratio is between the PCB and the enclosure. Each PCB can weigh four times more than the enclosure, or the enclosure can weigh four times more than any PCB. Using the *forward octave rule*, there is never a severe coupling problem. When the weight of any one PCBs is less than about one-tenth the weight of the chassis enclosure, the *reverse octave rule* works a little better in high shock environments.

The octave rule can be very effective in reducing vibrational and shock dynamic coupling acceleration levels

in plug-in types of PCBs installed in a chassis enclosure. When properly used, the octave rule is almost always more effective than damping in reducing the acceleration G levels transferred from the chassis to the internal PCBs. The dynamic acceleration G response of the chassis, which is usually the first degree of freedom, will be the dynamic input to the PCBs, which is usually the second degree of freedom. Transmissibility Q values that are transferred from the chassis to the PCBs do not add, they multiply.

Vibrational test data and computer-generated dynamic analyses have shown that the octave rule can reduce the acceleration G levels transferred from the chassis to PCBs by as much as 75%. When the natural frequencies of the chassis and the internal PCBs are close together, a good constrained layer damping system will reduce the acceleration G levels transferred to the PCBs by only about 15 to 20%. (See (1), Figs. 7.2–7.5 and Fig. 7.8.)

When a constrained layer damping system is added to a plug-in type PCB, some electronic components have to be removed to make room for the damper. When a stiffening rib must be added to a plug-in type of PCB to increase its natural frequency so that it follows the octave rule, some electronic components may have to be removed to make room for the stiffening rib. A stiffening rib will take up much less room on a PCB than a good constrained layer damper. Test data and past experience in damping and stiffening for PCBs to increase their vibrational reliability and fatigue life has shown that increasing the PCB natural frequency has almost always been the better choice.

Equation (14) shows that dynamic displacements are inversely related to the square of the natural frequency. This is a general relationship that applies to almost every type of structure exposed to dynamic vibration, shock, and acoustic environments. Consider the case where the input acceleration G level is held constant and the transmissibility Q value is approximated by Eq. (9a) as $\sqrt{f_n}$. When the PCB natural frequency is doubled, the resulting dynamic displacement of the PCB will be reduced:

$$Z = \frac{(f_n)^{\frac{1}{2}}}{(f_n)^2} = \frac{1}{(f_n)^{\frac{3}{2}}} = \frac{1}{(2)^{1.5}} = \frac{1}{2.83} \text{ (displacement ratio)} \quad (21a)$$

The fatigue life of the structure will increase because the displacement is reduced, which reduces the stress in the same proportion for a linear system. The fatigue life is strongly related to the b fatigue exponent slope of the $S-N$ fatigue curve shown in Fig. 5 and in Eq. (17).

For a smooth polished structure that has no stress concentrations where $k = 1$, Eq. (19) shows that the exponent b for materials used in electronic assemblies has a value of about 11.95. However, real structures almost always have some type of stress riser or stress concentration. A typical stress concentration value k for electronic structures is about 2. This results in a value for the b fatigue exponent slope of about 6.4. This means that the vibrational fatigue lives of typical electric components, their electrical lead wires, solder joints, fasteners, electrical connectors, and circuit traces on plug-in type of PCBs are increased when the natural frequency is doubled. However, doubling the PCB natural frequency uses up the fatigue life twice

as fast. This must be considered when the fatigue life improvement is evaluated;

$$\text{Fatigue life improvement} = \frac{(2.83)^{6.4}}{2} = 389 \text{ times} \quad (21b)$$

Various damping techniques have been applied very successfully in reducing the displacement amplitudes and stresses in tall buildings and long suspension bridges subjected to earthquakes and high winds. Damping is also used extensively to reduce noise levels in air ducts, automobile panels, washing machines, fan-cooled electronic systems, and aircraft jet engines. Damping, however, has not been used extensively to increase the dynamic fatigue life of plug-in types of PCBs because of cooling problems, repair costs, and changes in material damping properties at high temperatures.

A large midwest electronics company won a large contract to supply an electronic system that was required to operate in a severe vibrational environment. A decision was made to use viscoelastic damping materials for plug-in PCBs to reduce the vibrational acceleration G levels acting on the PCBs. Each plug-in PCB module consisted of two circuit boards bonded together, back to back, using the viscoelastic damping material. Vibrational tests were run on prototypes to verify the reliability and fatigue life of the proposed design. The tests were very successful, so the company went into full production using the viscoelastic damped plug-in PCB modules. One of the production electronic assemblies was selected for the vibrational qualification test required by the contract. The qualification test required the electronic system to be operating so that any electrical failures could be observed immediately. The vibrational qualification test was a disaster. Electronic component parts were breaking loose and flying off the PCBs. The engineers were stunned. The vibrational tests on the prototype viscoelastic damped PCBs were very successful. What happened? The engineers went back to their prototype test modules and repeated their previous vibrational tests. Their tests were successful once again. One of the engineers noted that the vibrational tests on their prototypes were run at room temperature. They decided to repeat the vibrational tests on their prototype models at an elevated temperature that simulated the temperatures experienced by the electrically operating production assembly. The elevated temperature vibrational tests on the prototype models were a disaster. Electronic components were breaking loose and flying off the PCBs. The elevated temperatures had sharply reduced the damping properties of their viscoelastic material, so that their design failed. The company was still under contract to deliver production electronic systems that could pass the vibrational requirements while they were operating electrically. The company had to redesign the electronic system and rerun prototypes at elevated temperatures to prove the new design integrity. They had to scrap the old production systems, retool for the new systems, and go into production to fulfill their contract requirements without change in the contract price. The company lost a substantial amount of money on that contract, and several engineers had to look for new jobs.

The octave rule can increase the vibrational fatigue life of a system and reduce the size, weight, and cost of the equipment at the same time. Once this rule is followed, it becomes possible to make PCB materials work smarter. This can be accomplished by understanding the relationships between the dynamic forces, displacements, stresses, and fatigue life of the materials used in fabricating and assembling electronic systems. A great deal of information related to eigenvalues (natural frequencies) and eigenvectors (mode shapes) can be obtained by using one of the many finite element analysis (FEA) programs available. Accurate dynamic stress levels in various structural members and approximate fatigue life in different environments are much more difficult to obtain from any FEA program. This cannot be done by analysis alone. An extensive amount of vibrational test data is required, where several similar electronic assemblies are vibrated until they fail. The failed parts are then examined closely to evaluate the physics of the failures. The test data are then combined with the FEA dynamic analysis and fatigue theory. This method of evaluation can result in equations that show what *minimum* natural frequency a rectangular plug-in PCB should have to achieve a fatigue life of about 10 million stress cycles in a sinusoidal vibrational environment. The physical properties of the circuit board materials, electronic component materials, solder joint materials, and the effects of surface-mounted and through-hole assembly practices must be understood to make these materials work smarter.

Finding the Maximum Allowable PCB Dynamic Displacement for Sine Vibration

Test data for sine vibrational environments are used to establish *maximum* allowable dynamic displacements Z_{\max} for PCBs to achieve a 10-million cycle life. This is based on the size of the PCB, the types and sizes of the electronic components, and the location of these components on the PCB. The following equation includes an added safety factor of 1.3 to ensure the effective fatigue life of each PCB further (1).

$$Z_{\max} = \frac{UB}{Chr\sqrt{L}} \text{ maximum allowable displacement, m (in)} \quad (22)$$

where

$$U = 8.90 \times 10^{-7} \text{ metric (} 2.20 \times 10^{-4} \text{ English)}$$

B in m (in), the length of the PCB parallel to the length of a component

h in m (in), thickness of the PCB

L in m (in), the length of an electronic component

C is a component type constant, metric and English

1.0 for a standard dual inline package (DIP)

1.26 for a DIP using side brazed lead wires

1.26 for a pin grid array (PGA) or hybrid that has two parallel rows of wires extending from the bottom surface of the component

1.0 for a PGA that has wires around the perimeter extending from the bottom surface of the component

2.25 for a leadless ceramic chip carrier (LCCC)

1.0 for a leadless chip carrier that has J leads or gull wing leads

0.75 for axial leaded devices such as resistors and capacitors

1.75 for ball grid array (BGA) components

r is a relative positional factor for components mounted on a rectangular PCB supported around its perimeter, metric and English

1.0 when the component is mounted at the center of the PCB at $x = a/2$ and $y = b/2$

0.707 when the component is mounted off the center of the PCB at $x = a/2$ and $y = b/4$

0.50 when the component is mounted at the one-quarter mounting points off the center of the PCB at $x = a/4$ and $y = b/4$

Extensive vibrational test data on different types of PCBs have shown that the electronic component lead wires fail far more often than their related solder joints in through-hole mounted devices. Surface-mounted components experience a small increase in the solder joint failures, but again, the greatest number of failures occur in the lead wires (1,2).

Combining Eq. (22) with Eqs. (8) and (9a) results in the following equation for the *minimum* desired natural frequency f_d that a PCB must have to provide a fatigue life of about 10 million stress cycles for the most critical components and their lead wires in a sine vibrational environment (1):

$$f_d = \left[\frac{AG_{\text{in}}Chr\sqrt{L}}{UB} \right]^{\frac{2}{3}} \text{ (minimum desired PCB natural frequency)} \quad (23)$$

where

$$A = 0.248 \text{ metric (} 9.80 \text{ English)}$$

$$U = 8.90 \times 10^{-7} \text{ metric (} 2.20 \times 10^{-4} \text{ English)}$$

G_{in} = peak input acceleration, dimensionless gravity units, metric and English

Sample Problem—Determining the Minimum Desired PCB Natural Frequency

When the octave rule is followed, it is possible to design PCBs without considering the vibrational response of the outer enclosure that supports the PCBs because there is a large reduction in the dynamic coupling between these two structures.

Find the minimum desired PCB natural frequency and the approximate fatigue life for a 0.00157 m (0.062 in) thick PCB that has a hybrid 0.0508 m (2.0 inch) long made of two parallel rows of pins extending from the bottom surface ($C = 1.26$). The hybrid is mounted at the center of a 0.152 m (6.0 in) \times 0.203 m (8.0 in) rectangular plug-in PCB, parallel to the 0.203 m (8.0 in) edge. The PCB must operate in a 5.0- G peak sine vibrational environment. Substituting in the preceding equation for metric or English units results in a minimum desired natural frequency of 210 Hz.

The approximate fatigue life for a resonant dwell condition can be obtained from the expected 10 million cycle

fatigue life for sine vibration, as follows:

$$\text{Life} = \frac{10 \times 10^6 \text{ cycles}}{\frac{210 \text{ cycles}}{\text{s}} \times \frac{3600 \text{ s}}{\text{h}}} = 13.23 \text{ h to failure} \quad (24)$$

Using Snubbers to Decrease Dynamic Displacements and Stresses (6)

High vibrational and shock accelerations can produce high dynamic displacements in structural members, which can cause very rapid failures. These high displacements can often be reduced by using snubbers. Snubbers are small devices that can be added to adjacent structural elements to make them act smarter by limiting their dynamic displacements. When these snubbers are properly placed, they leave only small clearances between the adjacent snubbing members, so that the snubbers strike each other. The striking action between the snubbers reduces the dynamic displacements, forces, and stresses in these structures and results in increased fatigue life. [See (1), Figs. 7.6 and 7.7.] Snubbers can be made from different grades and shapes of rubber, nylon, aluminum, or epoxy fiberglass. Rubber snubbers work well on PCBs that have low natural frequencies below about 50 Hz. When the PCB natural frequency is above about 100 Hz, the resulting dynamic displacements are so small that soft rubber does not work well. A harder material, similar to epoxy fiberglass, works much better. Good results have been obtained using 0.0063-m (0.25-in) diameter epoxy fiberglass snubbers epoxy bonded directly to the surface of the PCB near the center. Snubbers work well even if they cannot be mounted at the center of each PCB. For a new design, it is usually possible to leave a small amount of room at the center of each PCB for mounting snubbers. For existing PCBs, there may not be any room near the center, so the snubbers may have to be bonded to almost any convenient space between the components. The snubbers should not be directly bonded to the electronic components themselves. The snubbers should not be allowed to impact any components or any protruding lead wires or solder joints on adjacent PCBs because impact during vibration may cause failures in these areas (6).

Increasing PCB Stiffness to Decrease Dynamic Displacements

PCBs are often considered the heart of an electronic system because they hold most of the important electronic components that control the system. Large dynamic displacements on PCBs must be avoided in vibrational environments because they can result in high stresses and rapid fatigue failures. PCBs can be made to act smarter by increasing their natural frequency. This rapidly reduces the dynamic displacements and stresses and substantially increases the fatigue life. For new PCB designs, it is often very easy simply to increase the basic thickness of the circuit board. A 10% increase in the basic circuit board thickness can increase the PCBs natural frequency by 15%, which can increase the fatigue life of the components by a factor of about 6.

When the basic circuit board thickness cannot be increased, the PCBs natural frequency for a new design can often be increased to make the PCB act smarter by adding more copper planes to the multilayer board assembly. Many circuit boards already have at least two full one-ounce copper planes, 0.0000356 m (0.00140 in) thick for ground and for voltage. The natural frequency can be increased by simply doubling the thickness of the copper and adding another two copper planes (which can be called heat sinks), totaling four copper planes. The high copper modulus of elasticity can increase the natural frequency of the PCB by about 12%, which can increase the component fatigue life four times.

Stiffening ribs are often added to new designs of PCBs to increase the stiffness and the natural frequency (1). Stiffening ribs can significantly increase the natural frequency, but ribs take up room on the PCB, so that fewer components can be mounted on it. The probability of adding effective stiffening ribs to existing PCBs is very low. When improved vibrational or shock performance is required on existing hardware, a better choice is using snubbers.

Adding Doublers to Increase Local Stiffness in Critical PCB Areas

When existing PCBs experience component vibrational fatigue failures, it is often possible to make these PCBs act smarter by increasing the stiffness of the PCB in local areas of the critical components. This can often be done by simply bonding shims (sometimes called doublers) to the critical areas on the PCB. The shims should be fabricated from the same material as the circuit board. Strips about half the thickness of the board should be bonded to both sides of the PCB, where possible, under the component on the top side of the PCB and on the back side of the PCB just under the component. Metal shims are sometimes used instead of plastic shims. Metals have a higher modulus of elasticity than plastics, so they should work better than plastics. The problem is that metal shims often do not work as well as plastic shims. Metals are usually very smooth and hard. Epoxy type adhesives do not adhere well to hard smooth surfaces, so metal shims often fall off. Think of a mountain climber. It is very difficult, almost impossible, to climb a steep mountain without punching holes in the mountain-side to get a better grip. If there is no grip, the climber will slip and fall. The same holds for metal shims. If there is no grip, they will slip. A large shim may require at least eight small, well separated holes to be punched (or drilled) through the shims to allow the adhesive to flow through the holes and form adhesive rivets. Tests have shown that these plastic rivets will hold metal shims very securely to PCBs in severe vibrational, shock, and thermal cycling environments.

Making Component Lead Wires Work Smarter by Changing Their Form

Stiff component electrical lead wires have been known to cause lead wire and solder joint failures in vibrational, shock, and thermal cycling environments (2). It is often

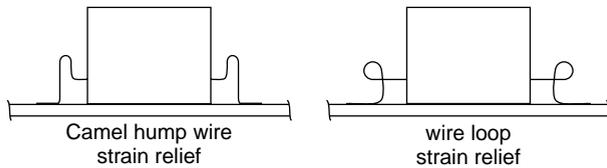


Figure 10. Electrical lead wire camel hump and wire loop strain relief.

possible to increase the fatigue life of lead wires and solder joints by *reducing* the stiffness of the lead wires. This is the theory behind this observation.

The spring rate K of a linear elastic system is defined as the force P divided by the displacement Y . Then, the force in the system is the product of the spring rate and the displacement:

$$P = KY. \quad (25)$$

When a system has a fixed (or constant) displacement Y , this equation shows that the best way to reduce the force P is to reduce the spring rate K . A constant displacement system occurs in vibration for an existing piece of hardware, when the natural frequency is known and the acceleration G level is known. Equation (8) shows that the displacement is fixed (or constant) when the natural frequency and the acceleration G level are defined.

Most lead wire and solder joint failures in vibration can be related to the bending action of the PCB, when its natural frequency is excited. The bending action of the PCB forces the component lead wires to bend as well. The spring rate of a wire in bending can be obtained by treating the wire as a beam, as shown in Eq. (13). Again defining the spring rate as the load divided by the displacement and ignoring the parameters that define the beam end restraints, the spring rate of the bending wire is:

$$K = \frac{EI}{L^3}. \quad (26)$$

An examination of this equation shows that the easiest parameter to change is the length L of the wire. Because this is a cubic function, a small change in the length of the wire produces a large reduction in the spring rate of the wire. It is often very easy to increase the length of the wire by looping it, adding camel humps in it, or simply by adding a small kink in the wire, as shown in Fig. 10. Another popular method for reducing the spring rate of the wire is to coin the wire. In this process, the wire is squeezed from a round cross section to a flat cross section. This reduces the area moment of inertia I which reduces the spring rate but not the area.

HOW STRUCTURES RESPOND TO RANDOM VIBRATION

Random vibration contains many different frequencies simultaneously across a broad frequency range. This means that all of the natural frequencies in a structure that are within the bandwidth of the random vibration will be

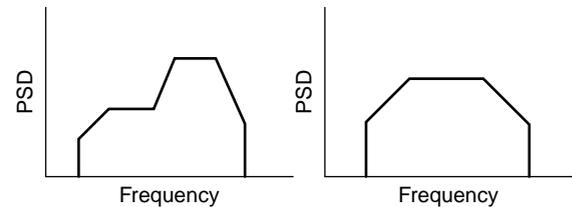


Figure 11. Typical power spectral density curves for random vibration.

excited at the same time. This contrasts with sine vibration, where only one natural frequency is excited at any given time during a sine sweep across a broad frequency bandwidth. This means that failures can occur in random vibration that cannot be duplicated in sine vibration (1).

Random vibrational profiles are plotted on log-log curves where power spectral density (psd) in terms of G^2/Hz is plotted on the y axis and frequency in Hz is plotted on the x axis, as shown in Fig. 11. The square root of the area under the curve represents the root-mean-square (rms) of the acceleration level G_{rms} . (1):

$$\sqrt{\text{Area}} = \sqrt{\frac{G^2}{\text{Hz}} \times \text{Hz}} = G_{\text{rms}} \quad (27)$$

Probability distribution functions have to be used to solve random vibrational problems because many different frequencies are present at any instant. The Gaussian distribution is used by many engineers involved with electronic equipment. This function represents the probability of the value of *instantaneous* accelerations at any time. The Rayleigh distribution, which is the probability of the distributions of *peak* accelerations, is also used. The argument here is that *peak* forces and stresses cause failures. A combination of these two functions, which is known as the three-band technique (1), is a Gaussian skewed toward a Rayleigh. This method is convenient for obtaining quick and relatively accurate solutions to random vibrational problems without using a computer. In the three-band technique, the rms represents the one-sigma (1σ) acceleration G level. The two-sigma (2σ) acceleration G levels are two times greater than the rms G level. The three-sigma (3σ) acceleration G levels, which are the maximum levels expected for the Gaussian distribution, are three times greater than the rms G level. The percentage of time that these levels occur in the three-band technique are as follows:

$$\begin{aligned} 1\sigma \text{ values occur } 68.3\% \text{ of the time;} \\ 2\sigma \text{ values occur } 27.1\% \text{ of the time;} \text{ and} \\ 3\sigma \text{ values occur } 4.33\% \text{ of the time.} \end{aligned} \quad (28)$$

Response of PCBs in Random Vibrational Environments

PCBs operating in random vibrational environments can be evaluated accurately as single-degree-of-freedom systems (1). The G_{rms} response for the PCB can then be obtained from the input psd value in G^2/Hz , the natural

Table 1. Fatigue Cycle Ratio n/N

Environment	Sine Vibration	Random Vibration	Thermal Cycles
1. Power on			$n_1/N_1 = 0.035$
2. Taxi		$n_2/N_2 = 0.001$	
3. Take-off		$n_3/N_3 = 0.203$	$n_3/N_3 = 0.047$
4. Climb		$n_4/N_4 = 0.097$	$n_4/N_4 = 0.022$
5. Attack	$n_5/N_5 = 0.091$	$n_5/N_5 = 0.191$	$n_5/N_5 = 0.033$
6. Return home		$n_6/N_6 = 0.009$	
7. Power off			$n_7/N_7 = 0.045$
$\sum \frac{n_i}{N_i}$	0.091	0.501	0.182

frequency f_n , and the Q value:

$$G_{rms} = \sqrt{\frac{\pi}{2} \text{psd } f_n Q} \tag{29}$$

The maximum allowable dynamic displacement Z_{max} for the PCB can still be shown by Eq. (22) for random vibration. The PCB fatigue life for random vibration is now expected to be about 20 million stress cycles. Because the maximum random displacement is based upon the $3\text{-}\sigma$ value, Eq. (8) must be multiplied by 3. Now Eqs. (8) and (22) can be combined with Eqs. (9a) and (29) to obtain the *minimum* desired PCB natural frequency for a component fatigue life of about 20 million stress cycles:

$$f_d = \left[\frac{VChr \sqrt{\frac{\pi}{2} (\text{psd}) L}}{UB} \right]^{0.80} \tag{30}$$

(*minimum*; desired PCB natural frequency),

where

- $V = 0.744$ metric (29.4 English)
- $U = 8.90 \times 10^{-7}$ metric (2.20×10^{-4} English)
- $\text{psd} = G^2/\text{Hz}$, power spectral density input, metric (English)

Sample Problem—Finding the Minimum Desired PCB Natural Frequency

Use the same physical dimensions for the PCB, as shown in the sine vibrational sample problem following Eq. (23), except use a PCB length of 0.228 m (9.0 in), parallel to the component length, and a random vibrational psd input of 0.10 G^2/Hz in the area of the PCB natural frequency. This results in a minimum desired natural frequency of about 178 Hz.

Determine the Approximate Fatigue Life of the Component on the PCB

$$\text{Life} = \frac{20 \times 10^6 \text{ cycles}}{\frac{178 \text{ cycles}}{\text{s}} \times \frac{3600 \text{ s}}{\text{h}}} = 31.2 \text{ h to failure} \tag{31}$$

Determine the Response of the PCB to the Previous Random Vibration

Substitute the previous information into Eq. (29) as follows:

$$G_{rms} = \sqrt{\frac{\pi}{2} (0.10)(178)\sqrt{178}} = 19.3 \text{ rms} \tag{32}$$

Find the Maximum Expected Acceleration G Level for the PCB

The maximum acceleration level expected is three times the rms, which will be

$$\text{Max } G = 3(19.3) = 57.9 \text{ G} \tag{33}$$

Find the Maximum Expected Displacement for the PCB

The maximum displacement expected for the PCB can be obtained from Eq. (8), which results in a dynamic displacement of 4.54×10^{-4} m (0.0179 in).

MINER'S CUMULATIVE DAMAGE FOR ESTIMATING FATIGUE LIFE

Miner's cumulative damage theory states that every time a structure experiences a stress cycle, part of its life is used up. This is shown as a series of ratios where the actual number of stress cycles (n) is divided by the number of cycles required to produce a failure (N) for many different stress environments. When the total of all of the ratios equals one, all of the life will be used up, so the structure will fail. Miner's method can be used to add up all of the damage accumulated in sine vibrational, random vibrational, shock, acoustic noise, and thermal cycling environments. Miner's damage ratio R_n is (1,2)

$$R_n = \frac{n_1}{N_1} + \frac{n_2}{N_2} + \frac{n_3}{N_3} + \frac{n_4}{N_4} + \frac{n_5}{N_5} + \dots = 1.0 \text{ for failure} \tag{34}$$

Sample Problem: Estimating the Fatigue Life of the Lead Wires on an Aircraft Electronic Component Subjected to Gunfire Sine Vibration, Flight Random Vibration, and Flight Thermal Cycling Environments

The attack uses the Gatling gun which generates a sine type of vibration. Anytime the engines are running, the system experiences random vibration. Anytime there is a

significant power change there is a significant temperature change, which is considered a thermal cycle.

Fatigue failures are difficult to predict because of the typically wide scatter in the fatigue life test data available. To ensure the reliability of the electronic systems in a military aircraft, it is good policy to include a scatter factor, or safety factor, in the design and analysis of these systems. Therefore, a scatter factor of 2.0 will be used to evaluate the fatigue life for the electronics. The normal design life for a military aircraft is about 10 years or 10,000 flying hours. Using a scatter factor of 2.0, the fatigue life of this electronic system will be designed for an operational time of 20,000 hours.

$$R_n = 0.091 + 0.501 + 0.182 = 0.774 \quad (35)$$

The cumulative damage ratio is less than 1.0, so the electronic design is acceptable.

BIBLIOGRAPHY

1. D.S. Steinberg, *Vibration Analysis For Electronic Equipment*. 3e, J Wiley, NY 2000.
2. D.S. Steinberg, *Cooling Techniques For Electronic Equipment*. 2e, J Wiley, NY 1991.
3. MIL-Handbook-5A, *Metallic Materials And Elements For Aerospace Vehicle Structures*, Department of Defense, Washington, DC
4. R.E. Peterson, *Stress Concentration Design Factors*. J Wiley, NY 1959.
5. D.S. Steinberg, *Mach. Design Mag.*, May 25, 1989.
6. D.S. Steinberg, *Mach. Design Mag.*, March 24, 1977.

VIBRATIONAL DAMPING, DESIGN CONSIDERATIONS

MICHAEL DRAKE
University of Dayton Research
Dayton, OH

INTRODUCTION

This article discusses the phases of a damping design effort. The basic steps in a damping design effort are identical for either a passive or active damping concept. The steps in this general approach are summarized in Table 1. The first four steps in this design process ensure that the designer completely defines the problem to be solved. During these steps, the designer verifies that the problem is the result of a resonant vibration, defines the vibrational characteristics of the structure under consideration, defines the environmental conditions in which the structure operates, and defines the level of damping required to solve the problem. These parameters are obviously needed because an effective design for the candidate damping concept (either active or passive) depends on the level of damping required, the frequency and mode shape, and the operational temperature range of the vibrating system. Based on these parameters, the problem is completely defined, and the

designer can make a logical choice of appropriate damping concepts to be evaluated that lead to the final design (1–3).

Although this design approach identifies individual steps, these steps are not independent. A successful damping design project is a systems engineering problem and must be solved using a concurrent engineering process. Most unsuccessfully damping designs are the direct result of not addressing all of the critical issues in a timely manner. The following paragraphs detail the critical tasks contained in each of the process steps and present examples from actual case histories of the way various issues were addressed.

DYNAMIC PROBLEM IDENTIFICATION

The proper initial step in solving any problem always is first completely defining the problem. This fact holds true when attempting to solve a vibration-induced problem by using damping technology. Therefore, the first step in this damping design approach is to substantiate that the problem to be solved results from structural resonance. If the vibration problem is not due to structural resonance, then a damping design will not be effective.

In a new structural system design, the designer must obtain the anticipated force input or excitatory environment for the system and correlate the frequency content of this information with the results of a natural frequency analysis of the structure (4). If there are natural frequencies within the frequency band of the excitation expected, the designer has identified the potential for dynamics problems. These dynamic problems must be evaluated to determine if the vibrations will keep the system from fulfilling the intended purpose.

If a problem develops in an existing part, the designer might choose one of the following approaches to identify the problem.

When a component cracked, a crack analysis should be run to verify that the crack is a high cycle fatigue failure. An instrumented operational test of the component should also be run to identify the frequencies and vibrational levels of the problem. Operational deflection patterns also can be defined to support the identification of the mode shapes encountered. This operational test can be run using strain gauges or accelerometers for measurements. The use of thermocouples enables one to obtain peak vibrational levels and corresponding temperature data, as well as maximum operating temperature data.

If the problem under consideration is a high-noise radiative problem, an operating evaluation should be done to determine both the frequencies and magnitudes of the noise being radiated and the source of radiation (5). An unacceptable vibrational level environment problem should be attacked in the same basic manner as the noise problem.

As a result of these investigations, the designer has determined the operating dynamic cause of the problem and the resonant frequencies that are developing the high dynamic response.

Table 1. General Approach Steps

Step Number	Description
1	Verify that the problem is resonant vibration induced
2	Complete dynamic analysis of the system determining resonant frequencies, mode shapes, and system damping
3	Define the environmental conditions in which the system operates
4	Define the system damping required to eliminate the problem
5	Select the appropriate damping concept and basic damping configuration
6	Develop the required design from the data collected
7	Prototype the design and complete laboratory verification tests
8	Develop tooling and manufacturing methods and complete field validation tests

DYNAMIC CHARACTERISTICS

A successful damping design is developed from a complete understanding of the dynamic behavior of the structural system and the component to be damped. Generally, a frequency range across which dynamic information is needed is defined from the analysis completed during the first step. The dynamic range can be defined from operational testing or can be determined from knowledge of the system under consideration. For example, problems such as a component where the excitation forces are known to be engine-order related, low frequency excitation from road roughness to the suspension system, or acoustic excitation to aircraft fuselage components due to jet engine exhaust all enable rapid determination of the excitatory frequency range. Once a frequency range is defined for the problem, a complete dynamic structural characterization must be completed. One must accurately determine all of the resonant frequencies, corresponding structural mode shapes, and inherent modal damping values that occur in the required frequency range. This data can be obtained analytically or experimentally.

If a prototype is available in the early structural design stages, the optimum solution for data acquisition is experimental analysis of the prototype structure used to refine analytical models. These models are then used for further component damping design evaluation (6).

Often, when a damping application is used as a redesign approach, the necessary dynamic characterization can be acquired efficiently by using modern experimental methods. Experimental methods can quickly determine the data needed for a highly complex structural system; however, measurements on operating systems can often be extremely difficult and costly.

The Fourier analyzer is a powerful experimental tool available to do the experimental work; however, holographic methods for determining mode shapes and standard sine sweep methods for resonant frequencies and modal damping values are extremely useful (7–9). The designer must choose the most expedient method for developing the required data.

ENVIRONMENTAL DEFINITION

A major data point needed in the quest for an optimally designed damping application is the operational environment that the design will see. This data, at first thought,

might seem to be a rather simple task, but the importance of accurate information cannot be overstressed. This data must cover the entire life of the damping concept from fabrication through operation.

A broad-brush approach to temperature, such as the standard temperature range of -65 to 250°F for operation of many aircraft components, is not the answer. This range may be the maximum range seen by the component; however, it will not generally be necessary to provide high damping across this entire range. The engineer must determine the specific temperature range across which the damage is occurring and design his application for that range while maintaining a total awareness of the required survivability temperature range. Time-related recordings of vibration and temperature data from operating tests are used to determine the temperature range across which damaging vibrational levels occur. Operating tests can also supply the necessary maximum temperature limits to be used in the design. If temperature data from a large number of different operating tests are available, a statistical study of the data will reveal the temperature range in which the majority of operating time is spent. An example of this type of data is shown in Fig. 1 illustrating minimum and maximum temperatures along with the percentage of total operating time spent in each temperature range. It is easy to see the value of this type of data, particularly if vibrational level and temperature data cannot be simultaneously obtained for operating conditions.

In the early stages of system design, complete operating temperature data may not be available. Then, data from

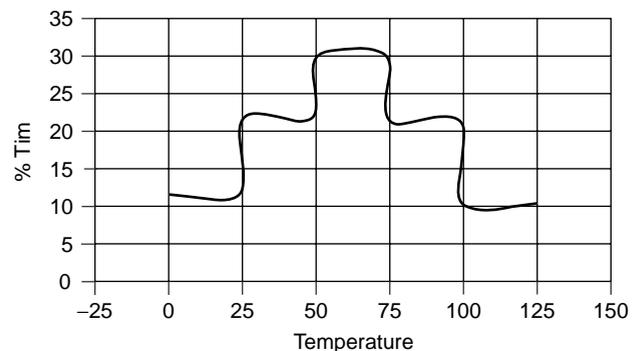


Figure 1. Percentage of flight time spent in various temperature ranges during 500 flight hours.

similar systems should be reviewed, and the best estimates of temperature should be developed and used in the design procedure.

Temperature is not the only environmental factor that must be considered. The engineer must know if the application will come in contact with contaminants such as salt water, gasoline, jet engine fuel, hydraulic fluid, or any other substance that might affect the performance or longevity of the candidate damping concepts (10,11).

When the damping concept will be integrated into the structural system and installed during the structural system manufacturing process, the damping concept components must be able to survive manufacturing processing, including temperatures, pressures, and processing.

REQUIRED DAMPING INCREASE

The remaining question to be answered before a damping design can be started is "How much damping is needed to eliminate the problem?" In the "fix-it" damping business, the general approach found in the literature is to design an optimum damping system and test it in service. If the failures are eliminated, the problem is solved. In reality, the designer wants the minimum value of system damping that will eliminate the vibrational problem. If the damped design accomplishes just the minimum required damping using an optimum damping system, the design also should be optimized from the standpoints of weight, size, and cost.

The method for determining the minimum required system damping depends on the problem to be solved. The inherent system damping has been determined from the dynamic characterization. The corresponding vibrational problem (high dynamic stress, noise level radiated, high dynamic amplitude response, etc.) is directly related to the inherent damping. Quick calculations can be made to determine the required increase in system damping to eliminate the vibrational problem. Basically, if a 20% decrease of system response is needed, then the system damping needs to be increased 20%. If an analytical model has been developed, an analysis can be conducted to verify the value of system damping needed to eliminate the vibrational problem.

Table 2 presents values of typical system and material damping for various structural systems and structural materials (4).

DAMPING CONCEPT SELECTION AND APPLICATION DESIGN

Until this point, the primary function of the designer has been to develop an accurate and complete definition of the resonant vibrational problem. Now, it becomes a simple matter to determine which resonant modes of the component are creating the vibrational problem. This information defines the frequencies that need to be damped, the corresponding mode shapes, and the undamped modal loss factors. The required level of damping and the environmental conditions complete the data required to start defining appropriate damping concepts and analyzing the effectiveness of the concepts.

Table 2. Typical Damping Values at Room Temperature

Systems/Materials	Loss Factor
Welded metal structure	0.001 to 0.0001
Bolted metal structure	0.01 to 0.001
Aluminum	0.0001
Brass, bronze	0.001
Beryllium	0.002
Copper	0.002
Glass	0.002 (0.3 to 1.5 at elevated temperatures)
Lead	0.5–0.002
Magnesium	0.0001
Plexiglas™	0.03
Sand	0.6–0.1
Steel	0.0001
Iron	0.0006
Tin	0.002
Wood, fiberboard	0.02
Wood, oak, fir	0.008–0.01
Wood, plywood	0.01
Zinc	0.0003

Often, the temperature range for effective damping and the survivability temperature limits are evaluated first. If the survival temperature is above 400°F, most organic passive damping materials and many of the piezoelectric materials are eliminated from consideration. Therefore, if you had a requirement for damping in the 100°–300°F range and a survival temperature of 600°F, most constrained layer and free layer damping materials and piezoelectric materials would be ruled out.

The next consideration is the mode shapes of the resonant frequencies that must be damped. Free-layer and constrained-layer passive damping concepts and induced strain actuation active damping concepts are effective for damping "plate-like" modes that have large areas of bending deformation. Highly localized strain distribution will negate the effectiveness of these damping concepts. An example of the localized strain condition is discussed in (12). The high-cycle fatigue cracks initiated in the corner areas on the antenna are shown in Fig. 2. For the mode generating the failure, all of the strain was concentrated in the corners, and the rest of the cone area was moving in a rigid body motion. As a result, layered passive concepts and piezoelectric induced strain active concepts were not applicable. The displacements of the mount were such that a displacement-sensitive concept such as a passive tuned damper or an active reaction mass concept could eliminate the problem. In this case, a passive design was used.

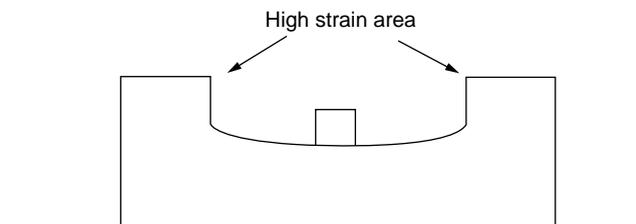


Figure 2. Cross section of the antenna base showing high strain areas.

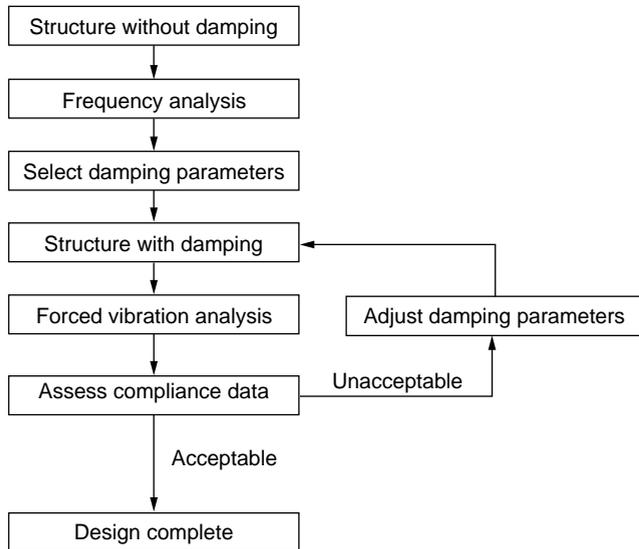


Figure 3. Basic damping design flow chart.

From the environmental conditions and the dynamic characteristics, the designer can choose the appropriate classes of damping concepts for the starting point to design the specific application for the structure under investigation. The basic principles of passive and active damping concepts and analysis methods are given in (13–18).

Various design analysis methods are often appropriate for problems; however, a successful design can be developed only after all of the basic information discussed previously is obtained. A design flow chart appropriate for any of the design analysis techniques is given in Fig. 3. The dynamic and temperature data are the inputs, and the output is the structural loss factor. The process loops through the design analysis step until the proper loss factor is achieved.

Note also that any successful design must use concurrent engineering principles and consider, from the beginning, the manufacturing methods that will be used to integrate the damping concept into the structural system and the maintenance processes for the damping concept.

PROTOTYPE FABRICATION AND LABORATORY VERIFICATION

Once the design is complete, the next step is to fabricate the design prototype and verify the design in the laboratory. The fabrication processes used for the prototype should be scalable to production processes, whenever possible. These processes must not degrade the function of the damping concept in any way.

Generally, the laboratory test setup is some scaled version of the total structural system. The primary consideration in the laboratory test is that the test article has the same dynamic characteristics as the full structural system and that the laboratory test environment simulates the critical operating conditions.

The laboratory validation test results should verify the analytical method used to develop the design. Any variation in the comparison of the test results and the model results should be evaluated and the test or the model, whichever is found in error, modified. After satisfactory comparisons are obtained, critical analysis for the field evaluation test should be conducted.

PRODUCTION TOOLING AND FIELD VALIDATION

Production tooling and processes should be refined from the process used to fabricate and install the prototype damping system. All lessons learned from the laboratory testing should be applied to the field test effort.

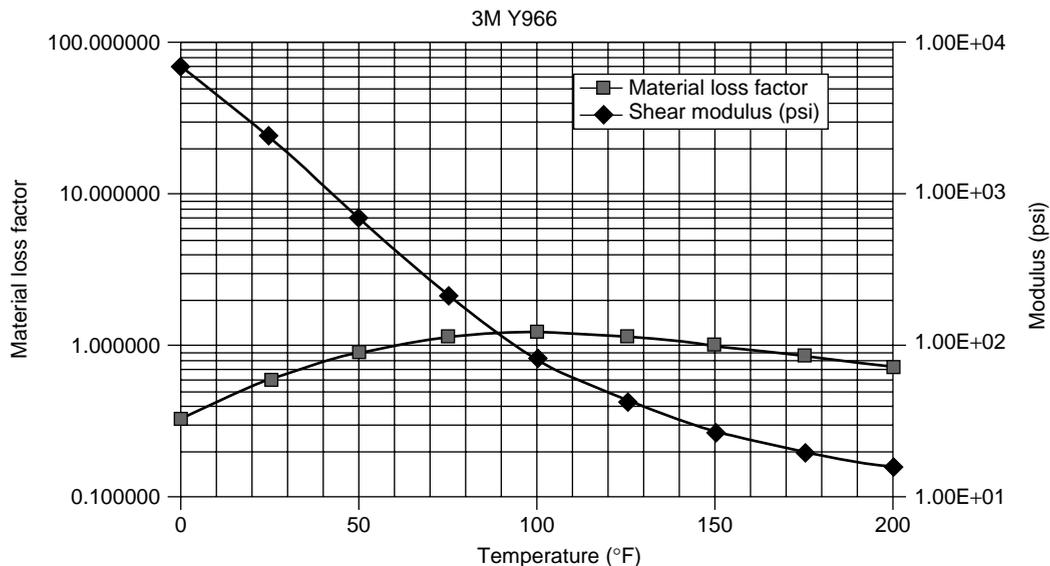


Figure 4. Y966 damping material properties as a function of temperature.

SUMMARY

Restating the importance of the problem definition is appropriate at this point. Inaccurate temperature range data will eliminate any beneficial effects of the damping concept. For passive systems, this effect can be seen in Fig. 4 (dynamic properties of 3M Y-966) where a temperature shift of 100°F causes a significant reduction in the material loss factor. If the survival temperature limits are incorrect, the damping concept may well provide the necessary reduction in the vibration levels, but the concept will be destroyed by the first overtemperature condition (19). Guesses at temperature data will invariably lead to failure of the damping design.

The other major area where accurate data are necessary is the dynamic characteristics of the structural system under investigation. Placing a strain-sensitive damping concept on a portion of the structure which is not undergoing significant strain for a particular mode is as ineffective as placing a displacement-sensitive concept on a node line of the mode that you wish to control.

As in any design project, successful results require accurate information upon which to base the design. Operational data and dynamic characteristics are the two prime factors that must be meticulously defined to obtain good damping design results.

BIBLIOGRAPHY

1. S.E. Olson, et al., *SPIE Smart Struct. Mater.* 1994, Vol. 2190.
2. A.J. Bronowicki, et al., *SPIE Smart Struct. Mater.* 1994, Vol. 2190.
3. M.L. Drake, et al. 1999 USAF Aircraft Struct. Integrity Program Conf.
4. J. Soovere and M.L. Drake, *Aerospace Structures Technology Damping Design Guide, Vol. 2-Design, Guide*, AFWAL-TR-84-3089, December 1985.
5. B.J. Smith, *Acoustics and Noise Control*. 2ed, Addison-Wesley, Reading MA, 1996.
6. M.F. Kluesener and R.A. Brockman, *Shock and Vib. Bull.* **60**: 1990.
7. Society for Experimental Mechanics, Inc., Course 101-Modal Analysis Theory & Application, February 2000.
8. K.A. Ramsey, "Effective Measurements for Structural Dynamics Testing," *Sound Vib.* November 24–35 (1975).
9. M.L. Drake and J.P. Henderson, "An Investigation of the Response of a Damped Structure Using Digital Techniques," *Shock Vib. Bull.* **45** (Part 5): 1975.
10. Flora et al., "Dynamic Analysis and Testing of Damped Inter-module Plates for the Sigma Laser Device," ASIAC Report No. 1182.1A, November 1982.
11. M.L. Drake, ed., University of Dayton Vibration Damping Short Course, Section 4.
12. D.I.G. Jones, J.P. Henderson, 1/Lt. G.H. Burns, *13th Ann. Air Force Sci. Eng. Symp.*, Arnold Air Force Station, Tennessee, September 27–29, 1966.
13. C.T. Sun and Y.P. Lu, *Vibration Damping of Structural Elements*. Prentice Hall, (1995).
14. J. Soovere and M.L. Drake, *Aerospace Structures Technology Damping Design Guide Vol. 1, 2, and 3*, AFWAL-TR-84-3089, December 1985.
15. A.D. Nashif et al., *Vibration Damping*. J. Wiley, NY, 1985.
16. C.R. Fuller et al., eds. *Active Control of Vibration*. Academic Press, 1996.
17. Colin S. Hansen et al., *Active Control of Noise and Vibration*. Routledge, 1996.
18. G. Rosenhouse, *Active Noise Control: Technologies and Design Concepts in Active Noise and Vibration*. Computational Mechanics, Inc, WIT Press, 2000.
19. M.L. Drake, ed., University of Dayton Vibration Damping Short Course, Section 11.

WINDOWS

J.M. BELL

I.L. SKRYABIN

J.P. MATTHEWS

Queensland University of Technology
Brisbane Qld

J.P. MATTHEWS

Sustainable Technologies
Queanbeyan, Australia

INTRODUCTION

The term Smart Window was introduced in the mid 1980s by Claes Granqvist to describe optically switchable electrochromic glazings. These devices exemplify the fundamental characteristic of all “smart windows”: controllable variation in the optical transmittance of the window. The variation in optical transmittance of electrochromic smart windows occurs through the simultaneous injection of electrons and ions (usually H^+ or Li^+ ions) into an electrochromic material such as WO_3 . In WO_3 , this leads to the development of either a broad absorption band or a reflection edge (depending on the details of the material preparation) that leads to a low transmittance state. The process is (ideally!) reversible; extraction of the ions and electrons returns the material to a transparent state. The control of the process is accomplished by applying a small voltage (1–2 V) (or passing a small current) that controls ion injection and extraction. Since the initial discovery of electrochromism in thin film WO_3 by Deb (1) in 1969, an enormous amount of research has been directed toward the goal of large area switchable windows for architectural applications. This initially focused on electrochromic systems that covered a wide range of materials and device structures. However, during the 1980s and subsequently, several alternate optical switching systems were developed that fall into the general category of smart windows. These include other electrically activated systems such as suspended particle devices (2) and phase dispersed liquid crystals (3), temperature controlled switchable devices such as thermochromic (4) and thermotropic devices (5), and recently developed gasochromic devices that are controlled by using reducing or oxidizing gas mixtures in a window unit (5).

ARCHITECTURAL GLAZING APPLICATIONS FOR SMART WINDOWS

The potential reductions in heating, cooling, and lighting energy use that can result from using switchable glazings in buildings has provided the impetus for the majority of this research. More than 675 U.S. patents have been filed in this field since 1976. It is now well established that reduced energy consumption of up to 50% is possible from the use of electrochromic glazings in commercial buildings, and

savings of 20–30% are obtainable in most climatic conditions. This is also the primary application for thermochromic and thermotropic windows that change transmittance as a function of temperature and therefore can reduce transmittance of infrared radiation (i.e., heat) when the temperature is high. However, some of the other “smart” windows, such as the polymer dispersed liquid crystal (PDLC) devices, do not provide significant energy savings but are used as privacy screens. To understand the role of switching in a smart window, it is necessary to understand the nature of glazings and the way the optical properties of a window affect its performance and utility.

Physics of Windows

Common to all applications of glazing materials is the need for transmission of light, and in most applications, reflectivity is also very important. To quantify the light transmission properties, the transmittance and reflectance spectra, $T(\lambda)$ and $R(\lambda)$ can be measured and used to define several different average transmittance and reflectance quantities. The most important of these are the solar and visible transmittance and reflectance:

$$T_{\text{sol}} = \frac{\int_0^{\infty} T(\lambda)\varphi_{\text{sol}}(\lambda)d\lambda}{\int_0^{\infty} \varphi_{\text{sol}}(\lambda)d\lambda}, \quad (1)$$

$$R_{\text{sol}} = \frac{\int_0^{\infty} R(\lambda)\varphi_{\text{sol}}(\lambda)d\lambda}{\int_0^{\infty} \varphi_{\text{sol}}(\lambda)d\lambda},$$

$$T_{\text{vis}} = \frac{\int_{370}^{770} T(\lambda)\varphi_{\text{vis}}(\lambda)d\lambda}{\int_{370}^{770} \varphi_{\text{vis}}(\lambda)d\lambda}, \quad (2)$$

$$R_{\text{vis}} = \frac{\int_{370}^{770} R(\lambda)\varphi_{\text{vis}}(\lambda)d\lambda}{\int_{370}^{770} \varphi_{\text{vis}}(\lambda)d\lambda},$$

where φ_{sol} and φ_{vis} are the solar spectrum and visible response of the human eye, respectively. Usually, the air mass 1.5 solar spectrum is used to define the solar averages, although it is not necessarily the most appropriate at all locations, and there can be significant differences between different solar spectra (6). The difference in the spectral ranges for φ_{sol} , φ_{vis} , illustrated in Fig. 1, immediately gives rise to the concept of spectral selectivity, which is central to many advanced window glazing systems. This refers to the ability of a window to transmit, for example, visible radiation (high T_{vis}) but to reflect heat (low T_{sol} and high R_{sol}). This is fully explained in Granqvist (7). Therefore,

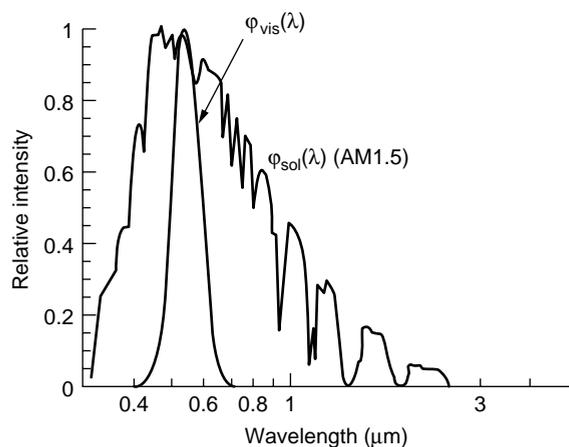


Figure 1. The three components related to the ambient radiation environment which need to be considered for designing windows. The three distinct spectral regions correspond to the wavelengths $0.3 < \lambda < 2.5 \mu\text{m}$ [solar radiation— $\varphi_{\text{sol}}(\lambda)$], $\lambda > 3 \mu\text{m}$ (thermal IR), not shown here, and $0.37 < \lambda < 0.77 \mu\text{m}$ corresponding to the visible response of the human eye, $\varphi_{\text{vis}}(\lambda)$.

the selectivity $T_{\text{vis}}/T_{\text{sol}}$, which has a maximum value of approximately two, can be used as a figure of merit for glazings. Another important optical quantity is the solar absorption, $A_{\text{sol}} = 1 - T_{\text{sol}} - R_{\text{sol}}$. A window that absorbs solar radiation strongly will heat up considerably, and this heat can be transmitted into the building, negating the reduction in insolation that results from low solar transmittance (8).

Although these are the basic quantities of interest, there are several derived quantities, such as the total solar energy transmittance (6) that are commonly used to classify window glazings. The key parameters of interest in switchable glazings are the change in the visible and solar transmittance between the two states of the window. The values usually quoted are the normal-normal transmittance and reflectance, that is, the fraction of an incident beam normally incident on the window that is specularly transmitted. The spectral response typical of different types of smart windows is illustrated in Fig. 2.

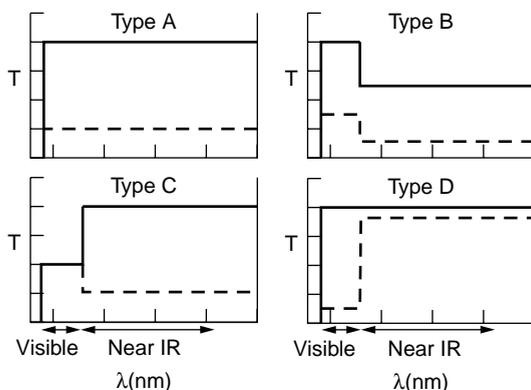


Figure 2. Schematic illustration of the different types of switching of transmitted radiation that are possible using different switchable window systems.

Most windows are designed to have low diffuse scattering, but many of the smart windows available or in development switch between a specularly transmitting and reflecting state and a diffusely transmitting and reflecting state. In this case, the normal-hemispherical or sometimes the hemispherical-hemispherical transmittance and reflectance are quoted.

Even for the purposes of minimizing energy use in buildings, there is no single figure of merit for smart windows that can be used to identify the optimum smart window. In general, the important characteristics and those quoted in this work where they are available, are the selectivity, defined before, and the contrast ratio, or amount of switching, in both the visible and solar spectral ranges: $T_{\text{vis}}^{\text{clear}}/T_{\text{vis}}^{\text{colored}}$, and $T_{\text{sol}}^{\text{clear}}/T_{\text{sol}}^{\text{colored}}$, respectively. The contrast ratio is frequently quoted as the change in optical density, $\Delta\text{OD} = \log_{10}(T_{\text{vis}}^{\text{clear}}/T_{\text{vis}}^{\text{colored}})$. A complete description of window behavior involves the variation of the selectivity or contrast function of the switching parameter, schematically illustrated for several different devices in Fig. 3. Some authors prefer to use the g factor or total solar energy transmittance (TSET) rather than the solar transmittance because this includes the effects of thermal conduction through the window and secondary heat gains that arise from absorption in the glazing (6). A good single number that is frequently used to rate a smart window is the ratio of the maximum visible transmittance to the minimum value of total solar energy transmittance ($T_{\text{vis}}^{\text{clear}}/T_{\text{sol}}^{\text{colored}}$). Another approach to defining the performance of smart windows in buildings is the energy cost for a building when different types of windows are used. Selkowitz and Lampert (9) introduced this concept through a graph that compares the cooling and lighting costs (the costs most

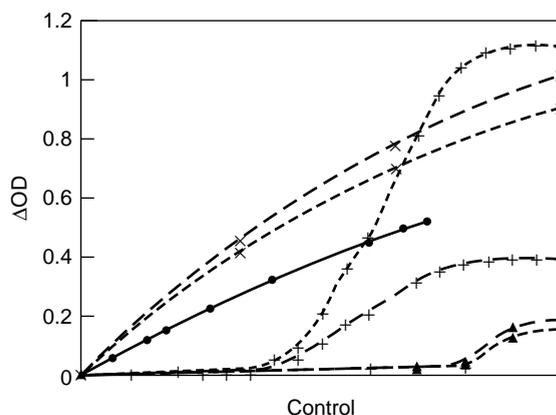


Figure 3. The variation in optical density for four differently electrically activated smart window devices. •—a sol-gel deposited device manufactured by Sustainable Technologies Australia Ltd, and based on WO_3 ; ×—a sputtered device manufactured by Asahi Glass Co based on a WO_3 -NiO complementary device; +—an organic electrochromic from Gentex Corporation, based on viologens; ▲—a polymer-dispersed liquid crystal device manufactured by 3M Corporation. The different curves represent different spectra used in calculating the optical density: — broadband transmittance; --- visible transmittance T_{vis} ; - - - solar transmittance T_{sol} . The divergence in the optical density for visible and solar transmittance for the Gentex viologen-based device reflects the type of spectral change (Type D in Fig. 2).

Table 1. Characteristics of Smart Window Systems

System	Mechanism of Color Change	Operational Switching Parameter	Control Parameter	Switching Behavior	Spectral Response (See Fig. 2)	Memory	Selectivity	Contrast	Ref.
							Colored State Clear State	Visible Solar	
Electrochromic (organic)	Reduction of organic species—e.g., viologen	Electrical control—voltage/current (approx. 2–3 Vdc; 50–1000 $\mu\text{A}/\text{cm}^2$)	Voltage	Threshold voltage, then approx. linear response to saturation	Type D	No	0.2 1.3 Specular	13 2.5	14
Electrochromic (inorganic)	Reduction of electrochromic (e.g., WO_3) by electrochemical ion insertion		Charge injected into electrochromic layer (mC/cm^2).	Approx. linear change in optical density with charge	Type A and B	Yes	1.4–1.9 1.3 Specular	3–10 3–11 Depends on system	14
Thermochromic	Metal–insulator transition	Temperature (T_{switch} depends on material)	Glazing temperature ($^{\circ}\text{C}$)	Rapid change at T_c (transition point)	Type C	N/A	1.0–1.2 <1 Specular	~1 1.1–1.6	4,15
Thermotropic	Phase separation leading to multiple scattering	Temperature (T_{switch} depends on material)	Glazing temperature ($^{\circ}\text{C}$)	Gradual transition (in hydrogel, from 30°C to 65°C)	Type A	N/A	1.5–1.8 1.3–1.7 (hydrogel) 1.1 1.2 (cloud gel) Scattering/specular	4.6 5 (hydrogel) 15 16 (cloud gel)	5,16
Polymer dispersed liquid crystal	Alignment of liquid crystal particles in electric field	Electrical control—(~ 100 Vac—voltage depends on system)	Applied voltage	Highly nonlinear response	Type D	No	0.9 1.0 Scattering/specular (with haze)	1.6 1.5	3,14
Suspended particle	Alignment of suspended particles in electric field	Electrical control (50–200 Vac) Power 2–10 W/m^2	Applied voltage	Linear response 0–50 V, saturation by ~ 100 V	Type D	No	N/A Specular	14 –	2
Gasochromic	Reduction of gasochromic (e.g., WO_3) by catalytic ion insertion	Gas control concentration of H_2 in window cavity	Concentration of H_2 in window cavity	Nonlinear response to gas concentration	Type A and B	Yes	1.2–1.7 1.0–1.5 Specular	4.2 4.9	5,17,18

affected by advanced glazing in commercial buildings) for different types of switchable glazings.

Other important aspects of windows and the materials used to make windows are durability (10), cost (11), color (12), and haze, or the amount of radiation which is diffusely transmitted (13).

SURVEY OF SMART WINDOWS

The principal smart window systems are summarized in Table 1 (2–5, 14–18). This table shows the switching mechanism, the method by which this mechanism is operated, whether the device exhibits memory, and the type of spectral response (selectivity and contrast) that occurs. The table also includes a control parameter that in most cases is the same as the operational switching parameter. However, in the context used here, the control parameter is one that can be measured and provides a degree of control over the optical state of the device. This is extremely important for practical applications of smart windows where it is essential to control the state of a device and to be able to switch the device. As can be seen in Fig. 3, of the electrically activated devices, the inorganic electrochromic devices provide the most linear response with respect to the control parameter, which in this case is injected charge density.

Electrochromic Smart Windows

Electrochromic materials undergo a change in their optical properties when an electrical potential difference is applied. A wide range of both organic and inorganic materials exhibit electrochromism. Inorganic electrochromic materials, most commonly transition-metal oxides, change from a clear to a colored state when ions and electrons are inserted into their lattice, under the application of a small dc voltage, that leads to optical absorption. Extraction of these ions and electrons is effected by a reversal of the electrical field and the material returns to the clear state. Organic electrochromic materials involve stoichiometric redox couples, where the oxidized and reduced forms have different colors. Therefore, the application of an electrical potential sufficient to change the oxidation state of the material causes a color change.

Inorganic Electrochromic Smart Windows. Many inorganic electrochromic materials may be produced in thin films on transparent electrically conducting substrates, thereby allowing their use in window fabrication. Tungsten trioxide (WO_3) the most commonly used inorganic electrochromic material is the compound in which electrochromism was first observed. Tungsten oxide can be colored by injecting electrons and small ions, including H^+ , Na^+ , K^+ and Li^+ , as schematically illustrated in the reaction scheme in Eq. (3a). Insertion of lithium and electrons into the tungsten trioxide structure (intercalation) forms the tungsten bronze Li_xWO_3 [reversible for $0 < x < 0.4$ (19)] which has a deep blue color. To fabricate an electrochromic device, a source of ions and electrons is required. The most common device structure used for inorganic electrochromic devices is illustrated in Fig. 4. There are a number of variants of this device structure that have different components, and this is discussed

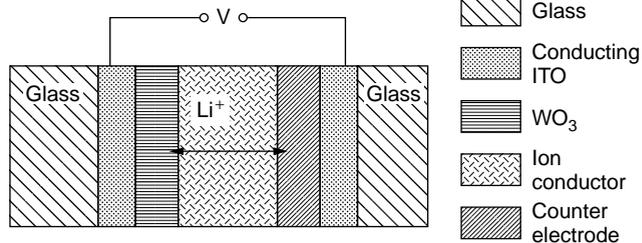
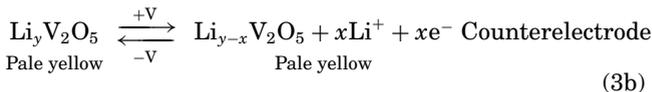
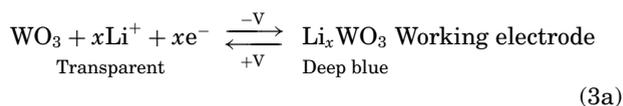


Figure 4. A schematic illustration of the structure of a typical inorganic electrochromic device that shows the five-layer structure.

later. The most widely used counterelectrode materials are cerium–titanium oxide ($\text{CeO}_2\text{--TiO}_2$) (20), vanadium oxide (V_2O_5) (21), and derivatives. The complementary reaction at the counterelectrode in a $\text{WO}_3\text{--V}_2\text{O}_5$ device is illustrated in Eq. (3b), where Li^+ is the mobile ionic species (22):



The change in optical density is usually proportional to the charge density injected into the electrode (23,24), and the coloration efficiency, $\text{CE} = \Delta\text{QD}/Q_n$, is frequently used to characterize the performance of a film or device. A typical coloration efficiency for WO_3 -based electrochromics is between 55 and 70 mC/cm^2 at $\lambda = 700 \text{ nm}$ (25).

The electrochromic material usually changes its optical constants by developing an absorption band (26) or by increasing the effective free electron density in the material (27). These give rise to optical changes known as “absorption modulation” and “reflectance modulation” (7), respectively, according to whether the change in transmittance results from an increase in optical absorption or reflectance. Reflectance modulation is the preferable optical change for glazing materials because these materials do not absorb radiation significantly in the colored state, minimizing secondary heat gain and reducing the TSET value of the window. The specific behavior depends on the microstructure of the material; crystalline WO_3 films exhibit reflectance modulation, and amorphous WO_3 films show absorption modulation. The coloration efficiency spectra in Fig. 5 illustrate the change in the absorption characteristics of sputtered WO_3 as the deposition temperature is increased. Despite the better performance that can be obtained from reflectance modulation, the majority of prototype devices produced to date use absorption modulation.

Devices based on inorganic electrochromics are usually spectrally selective to some degree (Type B behavior in Fig. 2), and it is possible to modulate the solar transmittance of the window strongly and still maintain moderate visible transmittance, thus allowing significant daylighting in a building. This makes WO_3 an excellent candidate for smart windows where low solar transmittance is a major requirement and daylighting is a direct advantage.

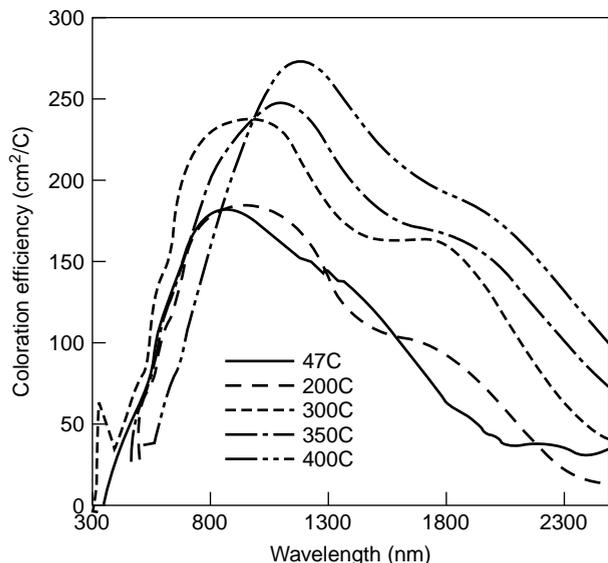


Figure 5. The spectral coloration efficiency that is directly proportional to change in optical density in sputtered WO_3 . This shows the development of an absorption band in WO_3 during ion injection. Note the shift in the absorption band peak position as the substrate temperature increases and the film becomes more crystalline, representing the change from absorption to reflection modulation.

Organic Electrochromic Smart Windows. The absorption coefficients of organic electrochromic materials are generally extremely high compared to those of inorganic materials, so it is possible to produce devices that undergo drastic optical changes that involve very intense colors. The majority of organic electrochromic materials are conducting polymers. Some of these organic materials, such as polyaniline, have different colors for various oxidation states, so it is possible to achieve multiple coloration by stepwise oxidation or reduction (28). The transmittance of organic electrochromic materials depends on the specific compound; however, most of these materials switch primarily in the visible region and show little modulation of solar transmittance compared to inorganic systems (Type D spectral behavior). This makes organic electrochromics less suitable than inorganic electrochromics for smart window applications.

The most widely studied family of organic electrochromic materials is the viologens, which are currently used in a commercial, variable reflectance mirror for automobile rear view mirrors (Gentex Corporation). In this application, it is free from many of the problems associated with solar/UV degradation that occur in architectural glazing applications of smart windows. Polyaniline is a promising candidate for organic electrochromic smart windows due to its high stability and ability to undergo multiple color changes (28,29).

Thermochromic Devices

Thermochromic materials change their optical properties when heated and return to the original state when cooled to the initial temperature (30). The most widely studied inorganic thermochromic material and the best

candidate for large area thermochromic smart window applications is VO_2 and various doped forms of VO_2 (15). This material undergoes a metal-insulator phase transition at 68°C and exhibits metallic properties at high temperature and associated infrared reflectivity, as shown in Fig. 6a. The visible part of the spectrum is largely unaffected by the transition (Type C spectral behavior), so the principal change that occurs at switching is in solar transmittance. The visible transmittance decreases slightly, but there is a moderately strong absorption band in the visible spectrum in the low-temperature insulating state which means that visible transmittance is relatively low in both states.

The major area of research in developing large area smart windows that are useful for energy control in buildings is developing materials that have higher visible transmittance and, most importantly, a transition temperature close to room temperature. This means that windows will switch to a low transmittance state to prevent overheating in the building. Both of these areas have been the subject of considerable research, and numerous

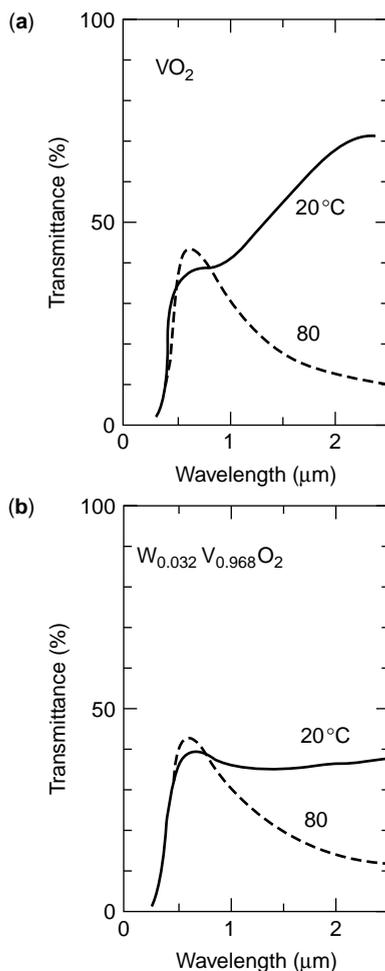


Figure 6. The transmittance of (a) VO_2 and (b) W-doped VO_2 at temperatures above (80°C) and below (20°C) the thermochromic transition. The spectra are type C in both cases (see Fig. 2). Reprinted from Sobhan et al. (15), with permission from Elsevier Science.

efforts have been made to dope VO_2 with various elements to achieve these goals. Fluorination of VO_2 (4) increases visible transmittance from 27 to 35% (in the low-temperature state) but also reduces the effect of switching; a change in solar transmittance from 44 to 40% occurs at the transition. However, the efforts to reduce the transition temperature have been much more successful; significant reduction in the thermochromic transition temperature from 68°C to approximately 10°C has been achieved for $\text{W}_{0.032}\text{V}_{0.968}\text{O}_2$ (15,31). This is illustrated in Fig. 6b.

Apart from the relatively poor spectral selectivity and contrast ratio of thermochromic devices, the inability to control the state of thermochromic devices independently makes it difficult to envisage wide ranging applications as smart windows.

Thermotropic Devices

Thermotropic devices, originally classed with thermochromic devices in the literature (32,33), also change their optical properties as temperature increases. However, the mechanism of the change is sufficiently different from typical inorganic thermochromic materials that they are now recognized as a different class of material. Thermotropic materials change between a specularly transmitting transparent state to a white, diffusely reflecting state as their temperature rises. The change is reversible. This change occurs through a phase separation that occurs in the materials at elevated temperatures that transforms the material from a homogeneous optical material to a heterogeneous, two-phase material at high temperature. The size of the particles of the two phases leads principally to backscattering of incident radiation (16).

There are two main types of thermotropic materials: hydrogels and polymer blends. Both systems have a similar device structure, as illustrated in Fig. 7. In both cases, the materials form homogeneous transparent mixtures at low temperatures, in which the two components are bound

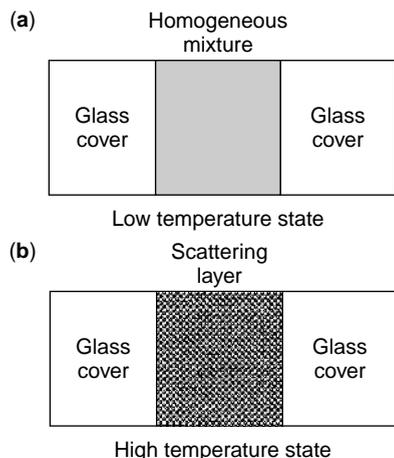


Figure 7. The structure of a thermotropic laminate. In the low-temperature state (a), the device is fully transparent, but in the high-temperature state (b), the thermotropic state separates into discrete particles, and the layer becomes scattering.

together at a molecular or molecular cluster (macromolecular) level. The bonding is usually hydrogen bonding, and these bonds can be broken as the temperature increases from room temperature. Once the bonds break, the materials separate into distinct domains of different refractive index and cause scattering. If the particle size is appropriate (close to the wavelength of light across the solar spectral range), then strong scattering occurs.

Hydrogels. This is one class of thermotropic materials; the thermotropic material CloudGel, commercially available from Suntek, is of this type. As the name suggests, a hydrogel is a water/polymer mixture. Another example of these materials is the TALD hydrogel, which is a mixture of colloidal particles of a polyether/water mixture, embedded in a carboxyvinyl polymer and water gel (16). The colloidal particles are soluble in the gel at room temperature and below, but as the temperature increases, water is expelled from the colloidal particles, and the polyether particles become denser and increase their refractive index.

The switching temperature of hydrogels can be varied between 5 and 60°C by changing the proportions of the components. The materials switch across a reasonably wide temperature range, as illustrated in Fig. 8. The size of the particles in the phase-separated state and the thickness of the hydrogel layer (typically 1 mm) ensure that the majority of the radiation is backscattered by multiple scattering in the high-temperature state (see Fig. 8). This means that these thermotropic materials have good potential as energy efficient windows.

Polymer Blends. Polymer blends of thermotropic materials have the advantage of not containing liquid water. An example of a polymer blend hydrogel is a high refractive index (1.585) cross-linked matrix of polystyrol and polyhydroxyethyl methacrylate that has a second, low refractive index (1.45) phase of polypropylene oxide interpenetrating the matrix. At low temperatures, the lower refractive index

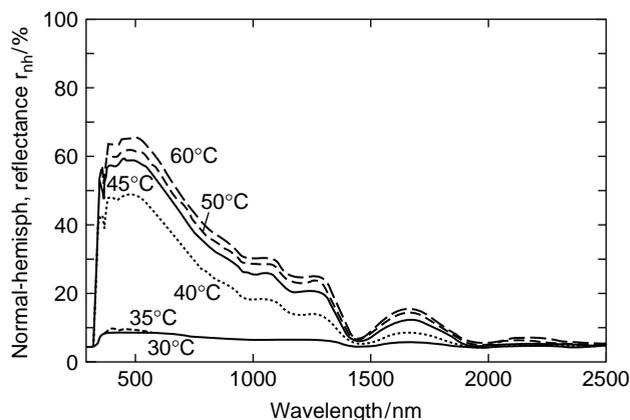


Figure 8. The normal-hemispherical reflectance of a thermotropic hydrogel laminate at a range of temperatures below (30°C) and above (>35°C) the thermotropic transition temperature. The spectra are type B and demonstrate a high degree of selectivity.

component is hydrogen bonded to the matrix. At higher temperatures, these bonds break, and the polypropylene oxide component forms small inclusions in the matrix that lead to scattering. The size of the scattering particles and the transition temperature depend on the proportions of the two components and the degree of cross-linking of the matrix.

Applications. The principal factors that constrain the application of thermotropic devices are the inability to control the device transmittance independently and the scattering state at high temperatures. The latter means that thermotropic devices are unlikely to be used in view windows, but they have applications in skylights and for overheating protection on brick facades. These applications have been extensively investigated (16), and large area windows can be routinely produced.

Polymer Dispersed Liquid Crystal Devices

There are many classes of liquid crystals. Most liquid crystals consist of thin, needle-like or rod-shaped organic molecules that exhibit significant ordering properties. There are several classes of electro-optic devices that can be formed using liquid crystals, but only two classes—guest–host systems and polymer dispersed liquid crystals (PDLC)—have been extensively investigated for use in smart windows. Lampert (34) provides an excellent introduction to the physics of liquid crystals, and Basturk and Grupp (35) and Montgomery (3) provide good reviews of the basics of both types of liquid crystal devices.

The principal liquid crystal devices currently available are PDLC devices, which are commercially available from 3M under the name “Privacy Film.” This device consists of a PDLC film sandwiched between ITO-coated sheets of glass, as illustrated in Fig. 9. The PDLC film consists of spherical liquid crystal droplets embedded in a polymer matrix. The typical droplet diameter is about $1\ \mu\text{m}$ or less, of the same order of magnitude as light wavelengths. When a voltage is applied to the ITO electrodes, the liquid crystal molecules are oriented in the field direction, resulting in index matching between the droplets and the matrix, and thus high transmittance for light propagating parallel to the applied field. When the field is switched off, the liquid crystal molecules reorient randomly; the (unmatched) value of the extraordinary refractive index also becomes effective and causes scattering at the droplet surfaces, and the film appears milky.

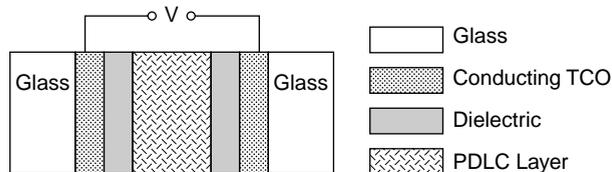


Figure 9. The structure of a PDLC device. Unlike an electrochromic device, it is not a conducting device, and switching depends on the electrical field across the PDLC layer.

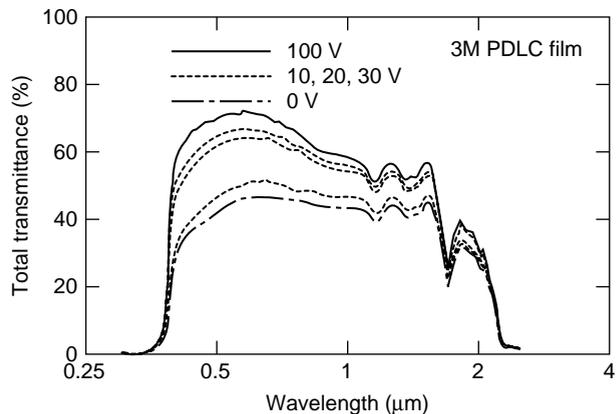


Figure 10. Normal-hemispherical transmittance spectra for a PDLC device at applied voltages of 0, 10, 20, 30 and 100 V. The device is opaque (in the sense that images cannot be perceived through the device in the off state (0V), but it still transmits significant energy. Reproduced with permission of Dr. Arne Roos.

Although PDLC devices act very effectively as privacy screens and have essentially zero specular transmittance in the zero voltage state and good transmittance but some diffuse scattering in the “on” state, the devices are not particularly effective for energy conservation because the majority of the scattered radiation is forward scattered. This is illustrated in Fig. 10, which shows total transmittance (including diffuse transmittance). The device is totally opaque in the “off” state (0V), but the total transmittance is still substantial. Another drawback of PDLC devices and all liquid crystal systems is the lack of memory in the devices. This is intrinsic to liquid crystal systems because the alignment of liquid crystal molecules is caused by dipolar polarization in the applied field. The curves in Fig. 10 also show the highly nonlinear behavior of switching as voltage is applied, which is the control parameter in Fig. 10 for the PDLC device.

Suspended Particle Devices

Suspended particle devices, also termed “light valves,” consist of a suspension of fine particles that are strongly absorbing in the “off” state (no applied voltage), but which can be aligned by applying an electrical field, leading to a transparent “on” state. The device requires constraining the suspension between conducting electrodes; in the large area devices currently being commercialized by Hankuk Glass, these are deposited on a PET film, which is then laminated to glass (2). The suspended particles are usually needle-shaped particles of polyiodides or parathite approximately $1\ \mu\text{m}$ long that are suspended in an organic liquid (32). In current devices, methods have been developed to encapsulate microdroplets that contain the suspended particles in a polymer matrix. The device structure is essentially identical to the PDLC device (Fig. 9), and the suspended particle layer replaces the PDLC layer.

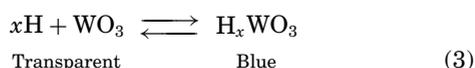
In the absence of an applied electrical field, the particles are randomly arranged and absorb incident radiation. The absorption in all states (from zero applied voltage to the

maximum applied voltage) depends on the concentration of particles in the film, and the color can also be altered by the choice of particles. Increasing the applied voltage increases the transmittance approximately linearly to about 50 V, and then the transmittance starts to saturate and reaches saturation at about 100 V. The major drawback of the devices for use as smart windows is that they do not provide significant spectral selectivity (figures for solar transmittance are not available) or significant reductions in solar transmittance. Similarly to polymer dispersed liquid crystal devices, they are principally privacy and visual comfort devices; however, these devices have a high viewing angle unlike liquid crystal devices.

Gasochromic Devices

Gasochromic smart windows are devices that color by a complex process involving the adsorption and dissociation of a gas onto a catalyst, followed by a “spill over” (36) of the dissociated gas molecules into the gasochromic layer that causes coloration (5,17,18,37–40). Gasochromic devices are similar to electrochromic devices; they both color when another species is inserted into the atomic network of the host material. Tungsten trioxide is a good material for both electrochromic and gasochromic devices. The major differences between electrochromic and gasochromic devices lie in the different mechanisms of coloration and the relative structural simplicity of the gasochromic devices.

Simple gasochromic films have a laminar structure that involves a substrate (usually glass), the gasochromic layer, and a catalyst layer. These films are fabricated into gasochromic smart windows by adding another pane of glass a short distance from the catalyst surface, so that a small cavity is formed inside the device, as shown in Fig. 11. Various gas mixtures can then be introduced into this interior cavity to color and bleach the window. Gasochromic tungsten oxide colors according to the following simplified redox reaction, when hydrogen gas is used to induce coloration:



This reaction is essentially identical to the reaction involved in the coloration of electrochromic WO_3 . The principal difference is that the catalyst layer allows the dissociation of H_2 gas into atomic hydrogen, which can diffuse into and react with WO_3 . In an electrochromic system, the positive and negative charges are introduced via an electrical circuit, and the polymer electrolyte and

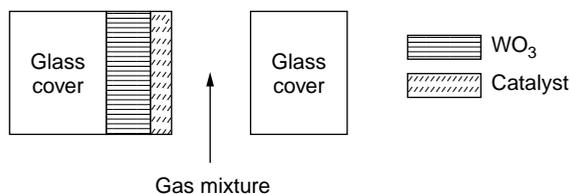


Figure 11. A typical gasochromic device structure.

counterelectrode are used to separate the charge and store cations, respectively.

Pure hydrogen is not required to color gasochromic films, and very dilute, nonexplosive mixtures can be used. Gasochromic films are usually bleached by filling the window cavity with a dilute oxygen mixture; however bleaching can also be achieved by using argon or a vacuum (18). Gasochromic devices can be partially colored or bleached by replacing the gas in the window cavity with an inert gas such as argon, once a suitable level of coloration has been attained.

Gasochromic Technology. Gasochromic technology has stemmed largely from research into electrochromism; tungsten oxide is the most commonly reported gasochromic material (5,17,18,37). Platinum is frequently used as the catalyst layer due to its ability to dissociate hydrogen and oxygen gas molecules. Some new systems whose gasochromic layer is based on alloys of lanthanides have been recently reported (38–40). Catalyst-coated lanthanum hydrides (LnH_x) undergo extreme color changes when x is varied from two (film reflective in blue region/transmitting in red) to three (yellow/transparent) (38). Exposure of catalyst-coated LnH_2 films to hydrogen gas quickly results in the formation of the transparent LnH_3 species. When exposed to air or a vacuum, the film reverts to the LnH_2 form. The function of the catalyst is to allow dissociation of gas molecules, but it also prevents the underlying film from rapidly oxidizing in air (40). The addition of magnesium reduces absorption in the transparent state and transmission in the opaque state, but it increases the opaque state reflectivity and also has the advantage of producing more color-neutral windows (38,40). A particularly impressive system involves a gadolinium–magnesium alloy $\text{Gd}_{40}\text{Mg}_{60}$ that reportedly switches from a silver white mirror that has almost no transmission and high reflection to a fully transparent color-neutral state (38).

Applications. The function of the glass pane used opposite the gasochromic layer is to contain the gas used for switching, and ideally it has high visible transmittance. Ordinary float glass is suitable; however, glass that has a low-emittance coating can be used to improve thermal performance. Due to the small number of layers required for gasochromic smart windows, the dynamic transmittance range is very high compared to electrochromics and other chromogenic glazings. Gasochromic smart windows reportedly have visible transmittances ranging from 76% (bleached state) to 18% (colored state). The advantages of the relative structural simplicity of gasochromic smart windows are somewhat offset by the complexity of the infrastructure needed to hold and transfer the gas mixtures required to operate them. The gas system must be capable of storing gas, preparing and filtering gas mixtures, and delivering these to a bank of multiple windows. Georg et al. (5) offer some innovative solutions to the problems of gas support.

As for all smart windows, lifetime is a critical issue for gasochromic devices. The principal degradation mechanisms for gasochromics are catalyst poisoning and varying water content of the films (37). Recent results are

promising, however; one gasochromic smart window 1.1×0.6 m has been cycled 10,000 times, and there was no significant reduction in its performance (18).

ELECTROCHROMIC SMART WINDOWS

As noted in the second section one of the most promising candidates for a future commercial smart windows is an electrochromic glazing based on inorganic electrochromic materials. Inorganic electrochromic smart windows are also the closest to commercial production for large area architectural applications and have substantial potential for reducing building energy use. Hence, the discussion following is based on windows made from inorganic materials, in particular, tungsten oxide.

The method of window control is also very important. To optimize energy performance, it is essential to be able to set specific optical states independent of temperature and irradiance incident on the window. The lifetime of electrochromic films and entire devices can be reduced by using improper control methods (41).

Various different types of electrochromic material and window testing and control (42) have been used: constant voltage control, trapezoidal voltage waveform (10), cyclic voltammetry (triangular voltage waveform) (43), and constant current control (44). There have also been many theoretical analyses of the charge injection and extraction processes. It is widely recognized that the coloration process is based on double injection of ions and electrons that form tungsten bronze M_xWO_3 (22). There are various theories regarding the mechanisms that limit the coloration current. They include the presence of a barrier at the ion injecting interface (electrolyte/ WO_3 interface) (22,45), diffusion limited motion of ions within the electrochromic films (46,47), and the series resistance of the cell (48,49). Impedance spectroscopy has been used by a number of authors recently to study ion injection kinetics in electrochromics (50,51), and this holds some promise as a useful tool for analyzing the degradation of smart windows in operation.

Electrochromic Smart Window Structures

The basic multilayer electrochromic device structure is shown in Fig. 4. There are two electrodes in all electrochromic devices to enable electron flow, an active electrochromic layer, and an ion storage and/or ion conducting layer to allow ions to move into and out of the active layer. It is possible to classify various electrochromic devices according to the structure and materials used to achieve ion injection and extraction in a device. There are three principal generic device types, which are all essentially the same as that shown in Fig. 4, but use different approaches to the ion transport and ion storage layers. The materials used in these devices are discussed in more detail later.

Type 1—Ion Conducting Layer and Passive Counterelectrode. The most widely used device structure has an ion transport layer (also called the electrolyte layer) between the working electrode and an ion storage

electrode (which is also called the counterelectrode). The fundamental requirement of the ion storage layer is charge capacity—the amount of charge (number of ions) that can reversibly be stored in the layer. The charge capacity is normalized to the electrode area and is measured in mC/cm^2 . The charge capacity of the ion storage electrode should be in excess of the charge required for sufficient coloration of the working electrode. Typical values are from 15–30 mC/cm^2 . It is also essential to achieve high transparency of the counterelectrode, which often conflicts with the requirement for high charge capacity. Another requirement for the ion storage layer in this structure is low coloration efficiency. A good ion storage electrode should have very large charge capacity and be optically passive. Two examples of devices of this type are the prototype commercial window produced by Sustainable Technologies Australia (STA) and the commercially available window produced by Flabeg (Pilkington). These devices incorporate a modified WO_3 electrochromic electrode, a V_2O_5 -based counterelectrode, and a lithium-doped cross-linked polymer electrolyte. STA uses sol-gel deposition of the electrodes, whereas Flabeg uses vacuum deposition. Spectra of the colored and bleached states of the STA device are shown in Fig. 12.

Type 2—Combined Ion Conducting Layer and Counterelectrode. In the second approach to device design, charge delivery and charge storage are realized in one layer. This is the ion conducting layer, which contains a dispersed redox couples; thus in the deintercalated state, ions are stored in the electrolyte. An example of this type of device is based on WO_3 as the working electrode and a solid polymer electrolyte that contains lithium polyorganodisulfide redox salts (52). These devices have the advantage of simplicity of manufacture and require deposition of one less layer than Type 1 devices. In addition, the redox salts are generally optically passive, so the charge capacity of the counter electrode is not an issue. However, the redox reaction occurs only at the surface of the layer, and

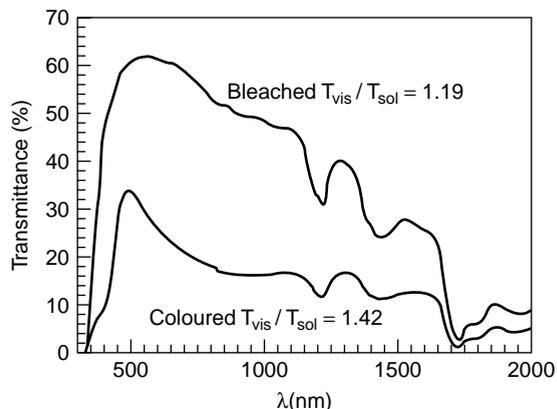


Figure 12. Transmittance spectra for colored and bleached states of an STA electrochromic device. The selectivity in both states is shown on the graph, indicating the increase in selectivity in the colored state. The dynamic range for the device is $T_{sol} = 51.5\%$ to $T_{sol} = 20.6\%$, and the injected charge is $10 mC/cm^2$.

transport of the electrons across the polymer layer from the TEC on the counterelectrode side of the device can limit the device.

Type 3—Ion Transport Layer and Complementary Counterelectrode. The third device structure is very similar to the first type because there are separate ion transport and counterelectrode layers. However, the counterelectrode is an electrochromic material which is complementary to the working electrode, that is, it is an anodically coloring electrochromic material that darkens upon extraction of ions from the counterelectrode and bleaches upon injection of ions (53). In this way, both electrodes contribute to the coloration of the electrochromic device and enhance the overall performance of the device. This type of device has the advantage that it requires less overall charge transport to achieve the same level of coloration as a type 1 device. Hence, the charge capacity requirements of the electrodes are lower than those of type 1 devices. A device of this type has been developed by the Asahi Glass Company (Japan). This device has the following structure: glass/TEC (200 nm)/NiO_x (500 nm)/Ta₂O₅ (500 nm)/WO₃ (500 nm)/TEC (500 nm)/adhesive film (250 μm)/glass. The spectral response of this device is shown in Fig. 13 at four different levels of charge injection. Note that the change in optical density per unit charge is significantly larger for this device than for the STA device of type 1 (see Figs. 3 and 12) due to the coloration of both the working electrode (WO₃) and the counterelectrode (NiO_x).

Materials Used in Electrochromic Devices

Electrochromic Materials. Tungsten oxide is currently the most commonly used inorganic electrochromic material. Coloration during ion insertion (H⁺, Li⁺, Na⁺, and K⁺) and bleaching on ion extraction is termed cathodic coloration. In an electrochromic smart window, thin film electrodes must be in contact with the working and counterelectrodes, and an electrolyte is required for ion transport. Use of electrochromic materials in window fabrication necessitates the transparency of all components, and

these constraints limit the number of materials suitable for electrochromic device fabrication.

Numerous metal oxides are suitable for use as the electrochromic layer in smart windows; a detailed discussion of electrochromism in these materials is given in Granqvist (54). The most widely studied electrochromic materials are WO₃ and doped WO₃, where the dopant is another transition-metal oxide such as TiO₂ or MoO₃. Tungsten oxide films are commonly prepared by thermal evaporation and sputtering, but CVD (chemical vapor deposition), spray deposition, anodization, and sol-gel deposition have also been used. A detailed discussion of these deposition techniques and their effects on the properties of the deposited materials is given in Granqvist (54), Bell and Matthews (55), and references therein. On the basis of the amount of charge which can be inserted, the coloration efficiency and the transmittance change of the film, the electrochromic performance of WO₃ deposited by most techniques is similar; the most important effect is increased crystallinity in films sputtered onto heated substrates (above approximately 350°C) or in sol-gel deposited films heat treated at temperatures of 350–400°C after deposition. This increase in crystallinity is usually accompanied by an increase in reflectance modulation in films.

Other inorganic electrochromic systems that have received significant attention are NiO, TiO₂, IrO₂, Nb₂O₅, and more recently, SnO₂ and Pr₂O₃. A variety of techniques has been used to deposit all of these coatings, but none of these materials shows the electrochromic efficiency of WO₃.

Counterelectrode Materials. NiO and V₂O₅ have been extensively studied as potential counterelectrode materials for electrochromic devices (54); CeO₂-TiO₂, and more recently CeO₂-ZrO₂, were extensively studied in the past few years (20). NiO is particularly attractive because it colors anodically (on the extraction of ions) and hence complements both the coloration and bleaching of the cathodically coloring WO₃ layer (53). (see previous discussion). V₂O₅ has excellent charge capacity and can be produced relatively easily using sol-gel deposition (21), but it has the disadvantage of an absorption edge that encroaches into the visible region of the spectrum and results in a yellowish coloration in the films (56). Initially (up to the early 1990s), sputtering and evaporation dominated the deposition of electrochromics, but sol-gel deposition is becoming very widespread, except for NiO, although there has been one report of sol-gel deposited NiO (57).

Ion Transport Layer. The function of the electrolyte layer is to allow ions to travel between the working and counterelectrodes; hence, it is ionically conducting. Low electronic conductivity of the electrolyte layer is advantageous because it reduces the internal leakage current in a device and allows devices to stay colored for long periods of time without the need for an external power source.

Electrolytes used in electrochromic smart windows can be classified as either polymeric or superionic conductors. Polymeric electrolytes are prepared as liquids that contain some dissolved lithium salt, and are then used to

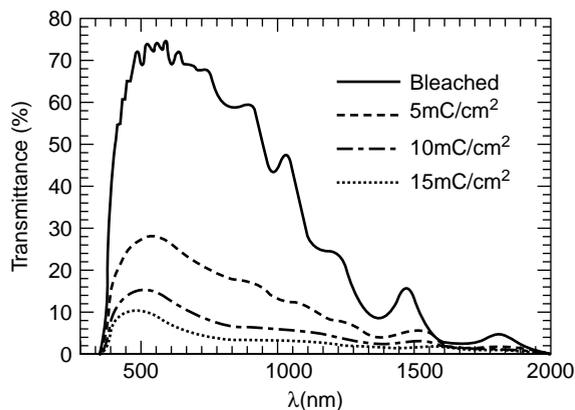


Figure 13. Transmittance spectra of an Asahi electrochromic device in the bleached state and for three colored states, where 5 mC/cm², 10 mC/cm² and 15 mC/cm² are injected into the WO₃ layer at a constant potential of 1.5 V.

lamine the working and counterelectrode layers together. Therefore, the polymer allows ionic conduction and gives the device its mechanical strength. Superionic conductors such as M- β -Alumina (where M is a monovalent metal such as Li, Na, or K) allow ions to move through the host lattice of alumina. These electrolytes are advantageous because they are not prone to damage by factors such as UV irradiation and can be deposited directly onto the electrochromic layer; however, they may have nonzero electrical conductivity which leads to reduced memory in the colored state. The Asahi device which uses Ta₂O₅ as an ion conducting layer has poor memory compared to devices that use polymer-based electrolytes (e.g., the STA device discussed previously) (14).

Transparent Electronic Conductors. The transparent electrical conductors in most electrochromic devices are thin films of metal oxides such as indium-doped tin oxide (SnO₂:In or ITO) and fluorine-doped tin oxide (SnO₂:F or FTO). These materials are commercially available precoated onto glass substrates by spray pyrolysis methods (58). Some electrochromic devices use conducting polymeric layers such as ITO-PET (ITO coated onto polyethylene terephthalate) as the transparent electrical conductors, which allows fabrication of polymeric solid-state electrochromic devices (59).

Control of Electrochromic Smart Windows

The application of electrochromic smart windows in architectural glazings requires that the devices perform tens of thousands of switching cycles during a period of 20–30 years and undergo minimal degradation in optical performance. The lifetime of a smart window is measured in the number of coloration/bleaching cycles, when the transmittance in each cycle is modulated between specified maximum and minimum values. However, the transmittance of the window, particularly of an installed window, is not easily accessible. As illustrated in Fig. 3, the optical density of an electrochromic window has a one-to-one (and very nearly linear) relationship to injected charge density. This relationship is also almost independent of temperature (Fig. 14), so a good switching algorithm for electrochromic smart windows can be based on charge control. When considered as an electrochemical system, an electrochromic device is very similar to a rechargeable battery. The energy in the battery is stored and released by reversible redox reactions. In electrochromic devices, these reactions result in optical changes. The durability of batteries has been investigated extensively during the last 50 years, and several empirical relationships between the lifetime and operational parameters have been found (60). In particular, it is found that battery lifetime is inversely proportional to the operational current, so the switching current must be small enough to provide a long lifetime.

In the design of a control algorithm for electrochromic windows, two fundamental limits that restrict fast switching of the window must be considered:

- Thermodynamic limit (overvoltage)—the voltage must be maintained below the potential at which destructive side reaction(s) occur; and

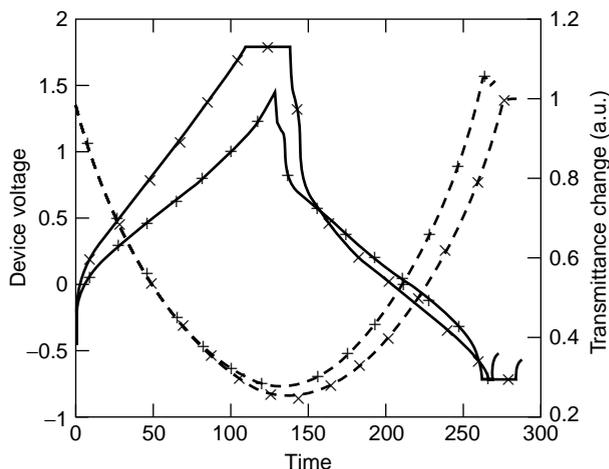


Figure 14. Electrochromic device performance at 50°C (+) and 18°C (x). The solid lines represent the device voltage, and the dashed lines the relative transmittance of the device. The operational parameters in both cycles are current density = 0.1 mA/cm², $Q_{in} = 13$ mC/cm²; $Q_{out} = -13$ mC/cm²; preset switching voltages are $V_{max} = 1.8$ V, $V_{min} = -0.7$ V. Note that at 50°C, the voltage does not reach V_{max} . The high-temperature cycle period is shorter and lower voltages are required; however, the transmittance change in the two cycles is almost identical.

- Kinetic limit (excessive charging current)—large currents cause irreversible cation trapping on the surface of the film (44) due to charge “pile-up” at the surface. This is usually caused by slow cation diffusion into the electrochromic layer.

Of course, if the voltage is too small, the desirable electrochromic reaction can never occur, and if the charging current is too small, the switching time of the window is unacceptably long. Therefore, both current and voltage need to be maintained within fixed limits to achieve optimum device lifetime, and charge is used as an independent parameter to control the optical state of the device.

Both thermodynamic (redox potentials) and kinetic (resistance, diffusion coefficients) properties of electrochromic window are temperature dependent; thus, they generally require adjustments in switching parameters for changing environmental conditions. For example, a constant applied voltage will not provide a constant contrast ratio at different temperatures due to the dependence of the film potential on temperature. At high temperatures, a voltage suitable for low temperature could exceed the potential of side reactions (thermodynamic limit) and will lead to higher currents and possible kinetic device failure.

Typical cycles of an electrochromic window at 50 and 18°C are presented in Fig. 14. The charge intercalated into the working electrode is the same in both cycles; therefore, the changes in optical transmittance are almost identical. The charging current density is also identical at the two temperatures; however at 50°C, the window required significantly lower coloring voltage because diffusion coefficients and electrolyte conductance increase with

temperature. The control algorithm implemented to produce these results includes constant current charging until the injected charge exceeds a given value Q_{in} . If the voltage exceeds a predetermined safe value V_{max} , the system automatically switches to a constant voltage mode. In the bleaching part of a cycle the extracted charge shall not exceed Q_{out} , and voltage shall be lower in magnitude than a predetermined maximum bleaching voltage V_{min} . Methods for controlling electrochromic devices by using these techniques are described in patents (61,62), and the theoretical background is outlined in Bell et al. (41,44).

FUTURE DIRECTIONS

Commercially available smart windows enabling energy control in buildings are now available (63). There are still significant questions about the durability of such windows, and the major research efforts in the immediate future will be to improve window durability during cycling. The advantage of electrochromic windows in terms of independent control of the optical state (using injected charge density) is actually a disadvantage in this context because it allows overdriving the device, whereas gasochromic and thermally switched devices are intrinsically self-limiting in switching the windows. However, other durability issues exist for these devices.

Two issues need to be addressed to improve the performance of smart windows in architectural glazing applications: the development of truly reflective smart window devices (like crystalline electrochromic materials, thermotropic glazings, and some of the recently discovered gasochromic materials) that are fully specular and (relatively) color-neutral; and control of the thermal emittance of glazings. The latter will broaden the range of climates in which smart windows can be useful and will also open up a new range of nonarchitectural applications.

BIBLIOGRAPHY

1. S.K. Deb, A novel electrophotographic system. *Appl. Opt., Suppl.* **3**, 192–195 (1969).
2. B.-S. Yu, E.-S. Kim, and Y.-W. Lee, Developments in suspended particle devices (SPD). *Proc. SPIE* **3138**, 217–225 (1997).
3. G.P. Montgomery, Polymer-dispersed and encapsulated liquid crystal films. In *Large Area Chromogenics: Materials and Devices for Transmittance Control* (C.M. Lampert and C.G. Granqvist, eds.), p. 577. SPIE, Washington, DC, 1998.
4. K.A. Khan and C.G. Granqvist, Thermochromism of Sputter deposited vanadium oxyfluoride coatings. In *Large Area Chromogenics: Materials and Devices for Transmittance Control* (C.M. Lampert and C.G. Granqvist, eds.), p. 160. SPIE, Washington, DC, 1998.
5. A. Georg, W. Graf, D. Schweiger, V. Wittwer, P. Nitz, and H. Wilson, Switchable glazing with a large dynamic range in Total Solar Energy Transmittance (TSET). *Sol. Energy* **62**(3), 215–228 (1998).
6. W.J. Platzer, Determination of key glazing parameters. Part I. Total solar energy transmittance. *Proc. Windows Innovations Conf. WIC '95*, Toronto, Canada, 1995, pp. 22–31.
7. C.G. Granqvist, Introduction to materials science for solar energy conversion systems. In *Materials Science for Solar Energy Conversion Systems* (A.A.M. Sayigh, ed.), p. 1. Pergamon, New York, 1991.
8. *ASHRAE Handbook of Fundamentals*, Chapter 27. American Society of Heating, Refrigeration and Air Conditioning Engineers, Atlanta, GA, 1981.
9. S.E. Selkowitz, and C.M. Lampert, Applications of large area chromogenics to architectural glazing. In *Large Area Chromogenics: Materials and Devices for Transmittance Control*, p. 22. (C.M. Lampert and C.G. Granqvist, eds.), SPIE, Washington, DC, 1998.
10. A.W. Czanderna, J.-G. Zhang, C.E. Tracy, D.K. Benson, and S.K. Deb, Accelerated life testing of large area electrochromic devices for window application. *Proc. SPIE* **3138**, 68–77 (1997).
11. R. Sullivan, S. Selkowitz, P. Lyons, P.C. Thomas, I. Heimonen, I. Andresen, O. Aschehoug, H. Simmler, P. Eggmann, and T. Frank, Energy simulation studies in IEA/SHC task 18 advanced glazings and associated materials for solar and building applications. *Proc. Windows Innovations Conf. WIC '95*, Toronto, Canada, 1995, pp. 635–662.
12. P.K. Nair, M.T.S. Nair, A. Fernandez, and M. Ocampo, Prospects of chemically deposited chalcogenide thin films for solar control applications. *J. Phys. D* **22**(6), 829–836 (1989).
13. J. Ferber and W. Platzer, Correct optical measurement of scattering samples. *Proc. SPIE* **2255**, 708–717 (1994).
14. M.D. Rubin, *IEA Task 18, Project B3: Chromogenic Glazing*, Final report. 1997.
15. M.A. Sobhan, R.T. Kivaisi, B. Stjerna, and C.-G. Granqvist, Thermochromism of sputter deposited $W_xV_{1-x}O_2$ films. *Sol. Energy Mater. Sol. Cells* **44**(4), 451–455 (1996).
16. H.R. Wilson, Potential of thermotropic layers to prevent overheating—A review. *Proc. SPIE* **2255**, 214–225 (1994).
17. D. Schweiger, A. Georg, W. Graf, and V. Wittwer, Examination of the kinetics and performance of a catalytically switching (gasochromic) device. *Sol. Energy Mater. Sol. Cells* **54**, 99–108 (1998).
18. A. Georg, W. Graf, R. Neumann, and V. Wittwer, Stability of gasochromic WO_3 films. *Sol. Energy Mater. Sol. Cells* **63**, 165–176 (2000).
19. J.-G. Zhang, D.K. Benson, C.E. Tracy, and S.K. Deb, The influence of microstructure on the electrochromic properties of Li_xWO_3 films. Part II. Limiting mechanisms in coloring and bleaching processes. *J. Mater. Res.* **8**(10), 2657–2667 (1993).
20. C.G. Granqvist, A. Azens, A. Hjelm, L. Kullman, G.A. Niklasson, D. Ronnow, M. Stromme Mattson, M. Veszelei, and G. Vairars, Recent advances in electrochromics for smart windows applications. *Sol. Energy* **64**(4), 199–216 (1998).
21. G.E. Tulloch, I.L. Skryabin, G. Evans, and J.M. Bell, Operation of electrochromic devices prepared by sol-gel methods. *Proc. SPIE* **3136**, 426–432 (1997).
22. B.W. Faughnan and R.S. Crandall, Electrochromic display based on WO_3 . *Top. Appl. Phys.* **40** (1980).
23. T. Kase, T. Miyamoto, T. Yoshimoto, Y. Ohsawa, H. Inaba, and K. Nakase, Performance of tungsten oxide/prussian blue device. In *Large Area Chromogenics: Materials and Devices for Transmittance Control* (C.M. Lampert and C.G. Granqvist, eds.), p. 504. SPIE, Washington, DC, 1998.
24. J.M. Bell, J. Barczynska, L.A. Evans, K.A. MacDonald, J. Wang, D.C. Green, and G.B. Smith, Electrochromism in sol-gel deposited TiO_2 films. *Proc. SPIE* **2255**, 324–331 (1994).
25. A. Agrawal, J.P. Cronin, and R. Zhang, Review of solid state electrochromic coatings produced using sol-gel techniques. *Sol. Energy Mater. Sol. Cells* **31**(1), 9–22 (1993).

Index Terms

Links

A

Absorption modulation, in smart windows	1136		
Accelerometers			
giant magnetostrictive material application	510		
with piezoelectric double amplifier			
smart skin	1032		
for ship noise control	1104		
truss structures	1067		
Acenaphthene, triboluminescence	1054		
<i>N</i> -Acetyl anthranilic acid,			
triboluminescence	1054		
Acid-base indicators	174		
Ac motors, for ship noise control	1105		
Acoustic emission sensors, smart paints for	754	757	
Acoustic noise reduction, aircraft	45		
Acoustic telemetry, fish	436		
Acoustic temperature sensors	919		
Acousto-optic tunable filter	411		
Acrylates, for enzyme immobilization	102		
Acrylic resins, as binder in smart			
paint composites	755		
Acryloyl chloride, for enzyme			
immobilization	101		
Activators, electrorheological materials	377	380	
Active Aeroelastic Wing project	48	51	
Active catheter robotic systems	635		
Active constrained layer (ACL) damping	31		
Active devices	860		
Active Flexible Wing project	20	48	51
Active flow control actuators	51		
Active flutter suppression, in aircraft	45		
Actively smart materials	337		
Active noise cancellation	1113		
Active noise control (ships)	1097		
applications	1106		
sensors and actuators for	1099	1110	
Active property tuning	553		
Active skins	1028		
See <i>also</i> Smart skins			
Active strain energy tuning	553		
Active structural acoustic control	1113		
for smart skins	1028		
Active vibration control	1128		
truss structures	1065		
Active vibration isolation	1106		
Active wing tip control	51		
Actuators. See <i>also</i> Piezoelectric actuators			
for active noise control in ships	1098	1099	1110
for adaptive systems	16	40	
analogy to muscles	218		

Index Terms	Links						
Actuators (cont.)							
conductive polymer applications	285						
defined	148						
electromechanical	338						
electrostrictive	340						
giant magnetostrictive material							
application	513						
macroworld/microworld association	503						
magnetostrictive materials for	600	611					
magnets in	591						
materials for	339						
for microrobotics	620						
piezoelectric ceramic, functional gradient	1						
power industry applications	882						
for smart highways	546						
for smart skin sound control	1028						
truss structures	1066	1071	1072	1079			
Adaptation	223						
Adaptive aircraft control	44						
rotary wing applications	28						
Adaptive all-movable aerodynamic surfaces	53						
Adaptive composite systems							
modeling and applications	16						
Adaptive materials	860						
Adaptive systems, rotary wing applications	28						
Adhesion promoters, flip-chips for underfill	443						
Advanced High-Resolution Radiation							
(AVHRR) sensors	984						
Aeroelastic divergence	49						
Aeroelasticity	47						
Aeroelastic tailoring	46	50					
Affinity precipitation							
chitosan-based gel applications	187						
smart polymers for	839						
Agarose, for enzyme immobilization	102						
Agonistic behavior studies	423						
Aileron reversal	48						
Aircraft. <i>See also</i> Helicopters							
adaptive composite system applications	21						
biological function emulation	715	716					
design	44						
gunfire sine vibration	1127						
health monitoring	45	302	303	313	420	718	736
sensing locations	904						
sensor array technology	903						
for ship monitoring	984						
smart foam skin	1029						
smart structure applications	42						
smart wings	928						
trade study chart	905						

Index Terms	Links				
Aircraft structures	45				
aeroelasticity	47				
smart structure applications	42				
Air/gas propagation transducers	693				
Alcohol dehydrogenase	100				
Alcohol oxidase	100	102			
Alcohols, Langmuir-Blodgett films	585				
Alginate, with chitosan-based gels	184				
Alkali halides, triboluminescence	190	1055	1056		
Alkaline battery	71				
Alkaline-earth aluminates					
triboluminescence of rare-earth doped	191	1059			
Alkyl-alkyl interactions	471				
Alkylamines, Langmuir-Blodgett films	585				
Alkyl- π interactions	471				
Allophane-rich soil	1027				
Alumina					
in composites with MEH-PPV	795				
for enzyme immobilization	102				
microtube fabrication substrate	645				
Aluminosilicate/poly(dimethylsiloxane)					
electrorheological fluid	364	368			
Aluminosilicates					
electrorheological fluid dispersed phase	379	380			
negative thermal expansion	1049				
Aluminum					
noncontact ultrasound examination	697	699	700	707	711
shape-memory alloy fiber/metal					
matrix composites	552				
shape-memory alloy particulate/aluminum	554				
matrix composites					
steel replacement	244				
Aluminum nitride, thin films					
Aluminum oxide. <i>See</i> Alumina	154				
Aluminum snubbers	1125				
American Ceramic Society	581				
Amides					
Langmuir-Blodgett films	585				
organogelators	475				
Amino acid organogelators	487				
<p>-Aminohippurat, pulsatile delivery</p> systems for	321				
Amorphous polymers, piezoelectricity in	867	871			
AMPLATZER septal occluder	932				
Amplitude, of vibration	1114				
α -Amylase, immobilization	102				
Anhydrous electrorheological fluids	362	363	378	380	
Aniline, polymerization in					
Langmuir-Blodgett films	586				

Index Terms	Links	
Animals. <i>See also</i> Fish aquatic studies		
behavior studies	423	
biomimetic electromagnetic devices	112	
flight smart structures	43	
as smart structures	415	
spectral sensitivity	112	
vision	112	
Anisotropic conductive adhesives	331	
Anisotropy determination		
noncontact ultrasound application	700	
ultrasound application	691	
Annealing, in shape-memory alloys	621	
local	640	
Anodes	68	
design considerations	71	
galvanic cells	357	
lithium-ion batteries	73	80
ANSYS modeling package	172	
Antagonistic design, of shapememory		
alloy actuators	626	640
Anthracene derivative gelators	471	475
Anthranlyl-cholesterol		
organogelators	484	
Anthraquinone derivative gelators	471	475
Antiarrhythmics	319	
Antibodies		
for affinity precipitation	839	
immobilization with conductive polymers	281	
molecularly imprinted polymers as mimics	678	
Antibody interaction drug delivery systems	324	
Antibody reaction biosensors	96	97
enzyme immobilization	103	
Anticancer drugs	319	
chitosan-based gel delivery systems	185	497
Anticounterfeiting indicator devices	179	
Antiferromagnets	591	
Anti-Invar effect	1044	
Antimony chalcogenide films	740	
Antireflective coatings, insect eyes	114	
Antiscald MemrySafe valve	930	
Ants, vision	112	
Aperiodic gratings	400	
Apodization	400	
Arachidic (n-eicosanoic) acid		
Langmuir-Blodgett films	585	
Archer fish, operant conditioning	425	
Architectural smart materials		
applications	61	
future design approaches	65	
relevant material characteristics	60	62

Index Terms	Links	
Architectural smart materials (cont.)		
for smart windows	1133	
types of	60	
Architecture	59	
Argon fluoride lasers, for Bragg grating writing	400	
Arsenic chalcogenide films	740	
Artificial extracellular matrices		
chitosan-based gel applications	188	
Artificial intelligence	682	
ultrasound application	691	
Artificial neural networks. <i>See</i> Neural network		
Artificial pancreas	99	
Artificial sphincter	90	
Artificial urethral valve	88	
ART programs	684	
Ascorbate oxidase	100	
Ascorbic acid biosensors	99	
Asymmetrical spring-mass element	520	
Atlantic salmon, radio telemetry	434	
Atomic force microscopy		
Langmuir-Blodgett films	588	
precision positioner for	339	
Atomic polarization	363	364
Atopic dermatitis, detecting with skin sensor	93	
Aurivillius phases	1000	
Austenite-martensite phase		
transformation	267	921
low-temperature martensite reorientation	965	977
magnetic-field induced	936	956
and material behavior	964	
modeling	967	
and shape-memory alloy loading modes	524	
and shape-memory effect	620	966
transformation temperature influences	952	
various types of shape-memory alloys	951	
Autoclave molding	293	
Auto exhaust sensor	347	
Autolysis	221	
Automated processing of materials	559	560
Automobile applications		
advanced automobiles	548	
electric vehicle batteries	68	
ESF controllers	453	
highways	545	
magnetorheological fluids	600	
shape-memory alloys	933	
suspension, electrorheological damper	389	1091
Autonomous diagnosis. <i>See</i> Self-diagnosis		
Averaging fiber-optic sensors	718	
Aviation. <i>See also</i> Aircraft	42	

Index Terms	Links				
Azo-cholesterol organogelators	486				
B					
Backpropagation neural networks	682				
Bacteria, infrared detection	115	116			
Bakelite	215				
Ballasts	63				
Ball grid array components, flip-chip technology application	439				
Balloon angioplasty	87				
Bamboo	245				
Barbiturate-2,6-diaminopyridine organogelators	483				
Barbiturate-melamine organogelators	480				
Barely visible impact damage	903	904			
Barium hexacelsians triboluminescence	1057				
Barium titanate	151	343			
composites with smart paints	754	755			
intelligent synthesis	573	578	579		
ionic polarization	386				
piezoelectric coefficients	150				
piezoelectricity in	143	162	781	788	1550
positive temperature coefficient ceramics	344	347			
positive temperature coefficient of resistivity	394				
properties	151	168			
in shape-memory alloy/piezoelectric heterostructures	557				
smart perovskite	992				
Barium zirconate	343				
Barret effect	609				
Basilica of St. Francis shape-memory alloy application for structural elements	933				
BathGard valve	930				
Batteries. <i>See also</i> Lithium-ion batteries					
design considerations	71				
performance parameters	70				
processing and applications	68				
smart	72				
Battleships, gun position control with adaptable fluid power	449				
Beam element					
damage detection	535				
nonuniform	529				
uniform	524				
Beam-type structures					
active vibration control in ships	1108	1111			
natural frequency	1115				
vibrational analysis	1120				

Index Terms	Links				
Beddington Trail Bridge, fiber Bragg grating sensor application	412				
Beers, electronic nose differentiation	284				
Bees, vision	112	114			
Beetles, infrared detection	115	116	118		
Behenic (n-docosanoic) acid Langmuir-Blodgett films	585				
Bellows microtubes	651	652	653		
Belousov-Zhabotinsky reaction	392				
Bending loading mode shape-memory actuators	624				
Benzoylchitosan membrane	187				
<i>N</i> -Benzyloxycarbonyl-L-alanine-4-hexadecanoyl-2-nitrophenyl ester organogelator	487				
Beryl	1051				
β -Blockers	319				
Biased martensite	965				
Bidirectional associative memory	684				
Bilirubin biosensors	107				
Bimorph actuators	1	159	622		
Binding assays, molecularly imprinted polymers application	678				
Bingham plastics electrorheological fluids	382	1092			
magnetorheological fluids	597				
smart fluids for flexible machines	451				
Biocatalysts, reversibly soluble	846				
Biochemical probes Langmuir-Blodgett film applications	584				
Biocompatibility chitosan	82				
drug delivery systems	182				
piezoelectric polymers	319				
polymer gel coatings for biosensors	871				
Bioerodible polymers	101				
Biological infrared detection	319	320			
Biologically-inspired creativity	115				
Bioluminescence	217				
Biomaterial intelligence	103				
Biomedical applications and Forum for Intelligent Materials	220				
hydrogels	82				
microrobotic	392				
noncontact ultrasound	491				
piezoelectric applications	634	635			
shape-memory alloy applications	712				
Biomedical diagnostics	83	91			
Biomedical sensing	83	88	925	926	931
Biomimetic electromagnetic devices	95				
	95				
	112				

Index Terms	Links	
Biomimetics	112	217
Biopolymers	835	
piezoelectricity	866	
Biopsy forceps	82	
shape-memory alloy application	86	
Biosensors		
in architectural smart materials	62	
hydrogels application	491	
immobilization techniques	100	
Langmuir-Blodgett film applications	584	
molecularly imprinted polymers		
application	674	
operations	103	
piezoelectric polymers for	871	
polymer coatings	99	
porous silicon	121	
Bioseparation		
affinity precipitation	187	839
chitosan-based gel applications	186	
partitioning	840	
Biotin, binding to streptavidin		
detection using porous silicon sensors	132	
Birds	43	
spectral sensitivity	112	
Birkhauzen effect	507	
Birth control drugs	319	
2,2-bis(hydroxymethyl)butanol		
trimethacrylate-based		
molecularly imprinted polymers	671	
Bismuth chalcogenides	745	
Bis-thiophene organogelator	476	
Bis-urea organogelators	478	
Blackbody fiber-optic sensor	416	
Black marlin, ultrasonic telemetry	432	
Bleomycin, chitosan-based gel		
encapsulation	185	
β -Blockers	319	
Blocks	1019	
Blood gas biosensors	107	
Boas (snakes), infrared detection	116	
Bolt breakers, electric power industry	884	
Bone plates	82	83
shape-memory alloy application	85	
Bone staples, shape-memory alloy		
application	85	
Boron carbide nanotubes	646	
Boron nitride nanotubes	646	
Bovine serum albumin, pulsatile		
delivery systems for	320	321
Bracket formulations	272	

Index Terms

Links

Bragg grating sensors. <i>See</i> Fiber			
Bragg grating sensors			
Bragg wavelength	397	399	408
Brassiere underwires, shape-memory			
alloy application	930		
Brevity, technology direction	582		
Brewster angle microscopy			
Langmuir-Blodgett films	587		
Bricks	1019		
Bridges			
damage self-diagnosis	891		
earthquake damage monitoring	582		
fiber Bragg grating sensors for	412	732	733
fiber-optic sensors for	421		
health monitoring	546		
magnetorheological fluids application	600		
Brillouin scattering sensors	418		
for health monitoring	303		
Bromocresol green	175		
Bromocresol purple	175		
Bromothymol blue	175		
Bronze Age	214	215	
Brownmillerite structure			
anion-deficiency induced	1006	1008	
Brown trout, radio telemetry	433		
Brushless motors, for ship noise control	1105		
Buckytube	646		
Buffeting, aircraft	50		
Buildings. <i>See also</i> Architectural			
smart materials; Health monitoring			
damage self-diagnosis	891		
earthquake damage monitoring	582		
fiber-optic sensors for	421		
smart materials for envelope heat transfer	61	62	
Bulk molding compounds	293		
Bundled microtubes	653		
BURPS process, for composite manufacture	145		
Butterfly loop	165		
Butterfly wings	114		
Flexinol for artificial	923		
Butyl methacrylate, chemical indicator	175		
C			
Cadmium sulfide, nanotubes	646		
Calcium manganite	1002		
Calcium silicate hydrates	1019		
Calcium titanate, intelligent synthesis	573	578	579
Camcorders, high energy density			
batteries required	68		
Camera components, electroceramic			
applications	339		

Index Terms	Links		
Camera focusing system	351		
CAMP, molecularly imprinted			
polymer-based sensor	676		
Cancer chemotherapy	319		
Canted antiferromagnets	591		
Capillary electrophoresis, molecularly			
imprinted polymers application	672		
Carbon anodes, lithium-ion batteries	80		
Carbon black, in composites with			
large positive temperature coefficients	274		
Carbon dioxide, contaminant in			
smart perovskites	577		
Carbon dioxide detection			
biosensors for	99		
material characteristics for			
architectural smart materials	62	64	
Carbon fiber-carbon matrix			
Composites, health monitoring			
via electrical resistance	311	313	
Carbon-fiber-reinforced concrete			
damage self-diagnosis	897		
health monitoring via electrical			
resistance	311		
for intrinsically smart structures	224		
smart highways	546	547	
Carbon-fiber-reinforced plastics	250		
for aircraft health monitoring	314		
for civil structure health monitoring	315		
cure monitoring	293	298	
damage self-diagnosis	891		
damage self-monitoring	392		
health monitoring via electrical			
resistance	311		
intelligent processing	561		
for intrinsically smart structures	233		
and ken materials	581	582	583
noncontact ultrasound examination	701	707	
Carbon steel, noncontact ultrasound			
propagation in	698		
Carboxymethylcellulose, for smart pills	839		
Cargo, dangers to ships	982		
Carotenes, Langmuir-Blodgett films	586		
Carpet, soil-ceramics contrasted	1025		
Cars. See Automobile applications			
Catalase	100	106	
Catalysts			
for epoxy resins in flip-chips	442		
molecularly imprinted polymers application	677		
reversibly soluble biocatalysts	846		
three-way	347		

Index Terms	Links		
Cathodes	68		
design considerations	71		
galvanic cells	357		
lithium-ion batteries	73		
Cavitation			
erosion mitigation	883	887	925
for pest control	762		
and ship noise control	1097		
Cell detachment, smart surfaces for	841		
Cellophane	215		
Cellular phones	438		
high energy density batteries required	68	72	
Cellular tissue-based biosensors	96	98	
Celluloid	215		
Cellulose	245		
for enzyme immobilization	102		
Langmuir-Blodgett films	586		
Cellulose acetate phthalate, for smart pills	839		
Cellulose copolymers, chemical indicators	175		
Cellulose nitrate, dielectric properties	368		
Cement matrix composites	223		
for damage sensing	226		
for strain sensing	224		
for temperature sensing	226		
for thermal control	228		
for vibration reduction	232		
Cephalins, Langmuir-Blodgett films	585		
Ceramic actuators	1		
Ceramic matrix composites	216	246	
for aircraft health monitoring	314		
damage self-diagnosis	891		
Ceramic microtubes	647		
Ceramic particulate/shape-memory			
alloy matrix composites	555		
Ceramics	245		
<i>See also</i> Electroceramics; Piezoelectric			
ceramics; Soil-ceramics			
damage detection in	891		
electrostrictive	139		
intelligent synthesis	568		
noncontact ultrasound examination	701		
piezoelectric and electrostrictive	139		
transducers	148		
Ceramic Society of Japan	581		
Ceramic tubing	645		
Cetyltrimethylammonium ions	475		
CFRP. <i>See</i> Carbon-fiber-reinforced			
plastics			
Chabazite, negative thermal expansion	1050		

Index Terms	Links	
Chalcogenide compound films		
optical storage in	749	
phase transition	738	742
photoinduced effects	745	
properties of	739	
Characterization		
ferroelectric liquid crystalline elastomers	851	
Langmuir-Blodgett films	588	
noncontact ultrasound application	707	
piezoelectric ceramic materials	162	
piezoelectric polymers	870	
of polymer films with surface		
acoustic wave sensors	914	
porous silicon	122	
smart perovskites	1010	
Charge transfer	221	
Chelation-based drug delivery systems	324	
Chemical indicating devices	173	
selecting	174	
Chemically induced triboluminescence	1055	
Chemically ordered covalent		
network, chalcogenide compounds	740	
Chemical organogelators	474	
Chemical reaction-sensitive		
hydrogels	495	
applications	500	
Chemical sensors		
Langmuir-Blodgett film applications	584	589
material characteristics for		
architectural smart materials	62	64
porous semiconductors	393	
porous silicon	129	
power industry applications	881	
Chemical valves	844	
Chemoluminescents, in architectural		
smart materials	63	
Chemoreceptors, in biosensors	96	
Chinook salmon, ultrasonic telemetry	432	
Chip fabrication	644	
<i>See also</i> Printed circuit boards		
Chip scale package components		
flip-chip technology application	439	
Chiral liquid crystals, ferroelectric		
behavior in	850	
Chirped gratings	400	
Chitin	182	
Chitosan	182	
Chitosan-based gels	182	
applications	185	497
Chitosan-based implants	182	

Index Terms	Links	
Chitosan-poly(ethylene oxide) microspheres	497	
Chloride biosensors	95	107
Chlorophenol red	175	
Cholesterol		
home test kits for	98	
Langmuir-Blodgett films	585	
organogelators based on derivatives	483	
Cholesterol oxidase	100	
Chromatography		
molecularly imprinted polymers application	672	
surface acoustic wave sensors application	912	
temperature-controlled	842	
Chromoproteins	112	
Chymotrypsin, immobilization	102	
Circuit boards. See Printed circuit boards		
Circuit breakers		
electric power industry	880	884
shape-memory alloy application	930	
Civil structures		
fiber-optic sensors for	421	
health monitoring	303	315
power industry	886	
smart paint application	754	
Classical lamination theory	260	
Clays, electrorheological material		
dispersed phase	377	
Climate control, by porous bodies	1016	
Closed-loop drug delivery systems	319	320
CloudGel, in smart windows	1138	
Clutch fan (automotive)		
shape-memory application	932	
Coal-fired power plants	874	
Coatings	190	
for corrosion detection in aircraft	905	
electrochromic sol-gel	356	
enteric, for smart pills	838	
Langmuir-Blodgett film applications	589	
for surface acoustic wave sensors		
for gas detection	913	
Cobalt, magnetostriction	602	
Cobalt-chromium alloys, biocompatibility	83	
Cobalt nitride anodes, lithium-ion batteries	81	
Cobalt oxide layered cathodes		
lithium-ion batteries	74	
Cobalt(tetracyanoethylene) ₂ -based magnets	595	
Coercive field	994	
Coho salmon		
operant conditioning	424	
radio telemetry	434	
Coiled microtubes	650	

Index Terms	Links	
ColdMark indicator	176	
ColdSNAP indicator	176	
Collagen	183	184
for enzyme immobilization	102	
Colloids. <i>See also</i> Gels; Hydrogels		
inherently conductive polymers	286	
memory effects	267	
Color centers, cooperative breeding of	773	
Colossal magnetoresistance	202	998
Colossal magnetoresistive materials	202	
<i>See also</i> Perovskite		
manganites		
chemical factors	204	
origin of effect	203	
Combustion process sensors	878	
Communications debriefing, for		
aircraft sensor array	910	
Compact Hybrid Actuators program	47	54
Complex fluids, memory effects	267	
Complexing hydrogels	495	
applications	499	
Composite actuators	2	
Composites	216	
<i>See also</i> Cement matrix composites;		
Ceramic matrix composites; Cure		
monitoring; Fiber-reinforced composites;		
Metal matrix composites;		
Polymer matrix composites		
adaptive	16	
in architectural materials	60	
classical lamination theory	260	
conductive polymer, with large		
temperature coefficients	274	
damage self-diagnosis	891	
functionally graded polymer blends	826	
fused deposition of	145	
future concepts	214	
hybrid	551	
intelligent processing	560	
intrinsically smart structures	223	
laminar failure	263	
macroscale behavior	258	
magnetostrictive/ferromagnetic tagged	310	
mesoscale behavior	250	
microscale behavior	246	
orientation and stacking effects	261	
particulate reinforcement	255	
with poly(p-phenylenevinylene) (PPV)	795	
strength theories	255	
survey	243	

Index Terms	Links		
Compressed air devices	215		
Computational fluid dynamics	265		
Computational techniques	265		
continuum and molecular descriptions	268		
nonequilibrium thermodynamics	271		
Concavalin A, in glucose-responsive insulin delivery system	325		
Concrete, fiber-reinforced. See Fiber-reinforced concrete			
Conductive adhesives	331		
Conductive polymer composites	274		
conductive filler effects	276		
with large temperature coefficients	274		
polymer matrix effects	277		
Conductive polymers	279		
applications	285		
colloidal	286		
in glucose biosensors	105		
microtube fabrication	645		
ring-substituted	286		
in sensor array for aircraft	903	904	
shape-memory alloys compared	623		
in smart windows	1137		
Confederation Bridge, fiber Bragg grating sensor application	413	733	
Conformal antennas	909		
Connectors			
electric power industry	884		
ZIF connector	935		
Conservation laws	265		
Constant-amplitude controller, for vibration control	1088		
Constant-gain controller, for vibration control	1088		
Constant-rate drug delivery systems	319		
Constitutive equations			
for adaptive composites	19		
for memory fluids	265		
piezoelectricity	165	617	862
Construction documents (CDs)	59		
Construction Specification Institute (CSI) classification	59		
Consumption-based society	1014		
Contact lenses, hydrogels application	491		
Continuous fiber reinforcement	251		
Continuous structural elements	522		
nonuniform structures	529		
Continuum descriptions, of smart materials	268		
Continuum thermodynamics	272		
Controlled collapse chip connection	438		

Index Terms	Links		
Controlled drug delivery	319		
Controlled porosity	843		
Controllers			
for ship noise control	1097	1098	
for smart skin sound control	1035		
for smart truss structures	1074		
for vibration control	1088		
Copolymers, Langmuir-Blodgett films	586		
Copper, thermal properties compared to Ni-Ti shape-memory alloys	629		
Copper-aluminum-nickel			
shape-memory alloys	622	922	952
applications	930		
for ship noise control	1101		
Copper-aluminum shape-memory alloys	952		
Copper-based shape-memory alloys	922	951	952
recovery stresses	961		
two-way shape-memory effect	957	958	
Copper ferric sulfide, negative thermal expansion	1041		
Copper halides, in photochromic glasses	770	773	
Copper nitride anodes, lithium-ion batteries	81		
Copper tin anodes, lithium-ion batteries	81		
Copper tubing	645		
Copper-zinc-aluminum			
shape-memory alloys	622	922	952
applications	930	931	
designing with	926		
for ship noise control	1101		
Copper-zinc shape-memory alloys	952		
Cordierite, negative thermal expansion	1051		
Corneal nipples	114		
Corn starch, electrorheological material dispersed phase	377		
Corrosion			
conductive adhesives	333		
marine environments	984		
power industry	886		
sensor array technology for			
detection in aircraft	903	904	905
ultrasonic measurement	691		
Corrosion cell	906		
Couplings, electric power industry	884		
Covalent molecularly imprinted polymers	669		
Crabs, smart materials for studying	423		
Crack development			
monitoring in structural concrete	64		
Schrodinger's equation-based simulation	271		
self-diagnosis in fiber-reinforced composites	891		

Index Terms	Links		
Cradle-through-grave approach	220		
<i>See also</i> Health monitoring			
Creatine biosensors	107		
<i>o</i> -Cresol red	175		
Critical magnetic field	001		
Critical temperature indicators	175		
Critical transition temperature	001		
Cross-linking, in hydrogels	491	493	
Crown ethers	471		
Cruise control	549		
Cryocon electrical connector	935		
Cryofit hydraulic pipe couplings	928	929	
Crystallization, ultrasound application	691		
Cure monitoring	292		
frequency-dependent			
electromagnetic sensing for	456	460	
smart automated cure control	463		
with surface acoustic wave sensors	915		
Current measurement, power			
industry applications	881		
<i>trans</i> -Cyclohexane-1,2-diamide			
organogelator	476		
<i>cis</i> -1,3,5-Cyclohexanetricarboxamide			
derivative organogelators	476		
Cymbal (piezoelectric actuator)	1	147	
D			
Dacron, for enzyme immobilization	102		
Damage detection			
carbon-fiber reinforced plastics			
self-monitoring	392		
cement matrix composites for	226		
in ceramics and large-scale structures	891		
earthquakes, self-diagnosis	582		
fiber-optic sensors for	307	415	419
polymer matrix composites for	236		
sensor array technology for aircraft	903	904	907
smart paints for	754		
triboluminescence applications	190	1063	
using dereverberated transfer function	520	533	
Dampers			
electrorheological fluids application	388	1091	
lockable damper mechanism	449		
magnetorheological fluids application	600		
space system	928		
Dams			
health monitoring	546		
structural integrity	882	887	
Data storage, optical. <i>See</i> Optical recording media			

Index Terms	Links					
Daylight sensing, material characteristics						
for architectural smart materials	62					
Dc motors, for ship noise control	1105					
Debye polarization, and						
electrorheological effect	363	364				
Decamethylferrocenium, tetracyanothanide						
ferromagnetism	592					
Defect detection						
noncontact ultrasound for	694	700	701	705	706	707
ultrasound for	690	691				
Deformation luminescence	190					
Deformation sensors						
magnetostrictives for	614					
triboluminescence applications	1064					
Defrost temperature indicator	176					
Degrees of freedom, of vibrating systems	1114					
DekDyne microsensor	457					
Delamination detection, ultrasound for	691					
delta E effect	507	609				
Demand controlled ventilation	64					
Denisyuk hologram	180					
Density determination, ultrasound for	690					
Dental applications	82					
hydrogels application	491					
shape-memory alloy application	85	931				
Dentures, shape-memory alloy application	86					
Dereverberated transfer function, in						
structural health monitoring	520					
damage detection approach using	533					
nonuniform structures	525					
Dermatitis, detecting with skin sensor	93					
Dew-point hygrometers	912	917				
Dew point measurement, surface						
acoustic wave sensors for	916					
Dextrans, for enzyme immobilization	102					
Diabetes monitoring	99	104				
1,2-diaminocyclohexane derivative						
organogelators	476					
2,6-diaminopyridine-barbiturate						
organogelators	483					
Diaphragm-type pressure						
transducers, for ship health monitoring	985					
Diazepam, molecularly imprinted						
polymer-based assay	680					
Dichromism	775					
Dielectrics						
property-performance relationship	458					
ultralow dielectric constant	438					

Index Terms	Links	
Dielectric sensors		
for composite manufacturing	562	
for cure monitoring	294	298
frequency-dependent		
electromagnetic sensing	456	
Diethylaminoethyl methacrylate		
chemical indicator	175	
Differential Global Positioning Systems	550	
Digital chronometer	448	
Dimensional analysis		
noncontact ultrasound application	713	
ultrasound application	691	
2,4-dinitrotoluene (DNT) detection		
porous silicon sensors	130	
Direct chip attach technique	439	
Direct fiber-optic sensors	719	
Discrete fiber-optic sensors	719	
Discrete spring-mass elements	520	
nonuniform structures	525	
Dispersions, memory effects	267	
Dispersion-strengthened composites	246	
Displacement transducers, for ship		
noise control	1103	
Displayed motion	923	
Dissolution-diffusion method, for preparing		
functionally graded polymer blends	826	827
Distributed-effect fiber-optic sensors	718	
Distributed fiber-optic sensors	418	718
Distributed fiber-optic temperature sensor	879	
Distributed smart material systems	873	
Ditriphenylphosphine oxide		
manganese bromide triboluminescence	1054	
Diverging tail	53	
DNA-biochips, conductive polymer		
application	283	
DNA-chitosan complexes	186	
<i>n</i> -docosanoic (behenic) acid		
Langmuir-Blodgett films	585	
Domain boundaries	994	
Donor-acceptor complexes	474	
Doppler vibrometers, for bridge monitoring	546	
Dot-matrix printer head		
piezoelectric actuator application	159	
Double exchange	202	998
Dragonflies	44	
vision	112	
Drainage systems, smart materials for	60	
<i>Drosophila melanogaster</i> , vision	112	113

Index Terms	Links		
Drug delivery systems	95	319	
chitosan-based gels applications	185		
glycoproteins for binding	99		
hydrogels applications	322	491	495
Dry (intrinsic) electrorheological materials	379	381	
Ducts, active noise control in ships	1107	1111	
Dyes			
as chemical indicators	174		
Langmuir-Blodgett films	589		
as pH indicators for power			
industry applications	879		
Dynamic aeroelasticity	49		
Dysprosium. <i>See also</i> Terfenol-D			
magnetostriction	602		
E			
Earth ceramics. <i>See</i> Soil-ceramics			
Earthenware	581		
Earthquakes			
damage monitoring	582		
fiber-optic sensors for detecting	421		
shape-memory alloy application for			
structural elements	933		
triboluminescence during	1055		
Eastern Pacific bonito, ultrasonic telemetry	435		
Eco-carat soil-ceramic material	1027		
Ecosystems	219		
Edrophonium chloride, pulsatile			
delivery systems for	321		
Eggo superelastic headphones (Sony)	930		
<i>n</i> -eicosanoic (arachidic) acid			
Langmuir-Blodgett films	585		
Elasticoluminescence	190		
Elastic properties determination			
noncontact ultrasound application	700		
ultrasound application	691		
Elastomers	245		
electrorheological materials	376		
epoxy resin toughening agent in			
flip-chips	443		
ferroelectric liquid crystalline	850		
magnetorheological materials	598		
Electrically-conductive adhesives	331		
Electrically-induced triboluminescence	1054		
Electrically-stimulated pulsatile			
drug delivery systems	321		
Electrical power industry. <i>See</i> Power			
industry applications			
Electric motors, for ship noise control	1105		
Electric vehicles, high energy density			
batteries required	68		

Index Terms	Links					
Electroceraic actuators	1	2				
Electroceraics	337					
electromechanical actuators	338					
electrostriction	342					
optical materials	351					
smart chemical systems	347					
thermal systems	344					
Electrochemical cells	68					
light-emitting, PPV application	799	804				
Electrochemical micromachining	647					
Electrochemical transduction, porous						
silicon sensors	131					
Electrochromics						
in architectural smart materials	60	62	63			
defined	356					
Electrochromic smart windows	352	357	360	1133	1135	1141
Electrochromic sol-gel coatings	356					
Electrochromism	356					
Electrodischarge machining	634					
Electroluminescent devices						
poly(p-phenylenevinylene) for	793	799				
Electroluminescents						
in architectural smart materials	62	63				
Langmuir-Blodgett film applications	589					
Electrolumiscence	1054					
Electromagnetic actuators, for ship						
noise control	1105					
Electromagnetic devices, biomimetic	112					
Electromagnetic interference						
Shielding, conductive polymer						
composites application	274					
Electromagnetic sensing, frequency						
dependent (FDEMS)	456					
Electromechanical actuators	338					
Electromyogram (EMG) assessment						
of fish using telemetry	434					
Electron energy-loss spectroscopy						
for smart perovskite characterization	1010					
Electronic nose	284					
Electronic packaging	438					
Electronic polarization, and						
electrorheological effect	363	364	387			
Electronic Route Guidance Systems	550					
Electronics						
Langmuir-Blodgett film applications	584					
shape-memory alloy applications	934					
Electron microscopy, precision						
positioner for	339					
Electrophoretic mobility						
electrorheological fluids	379					

Index Terms	Links		
Electrorheological fluids	362	376	
for adaptive systems	17	28	30
dielectric properties	363		
ESF controllers	451		
and fluid machines	450		
mechanism of effect	371	385	
memory effects	267		
for ship noise control	1100	1103	
for vibration control	62	1084	1091
Electrorheological materials	218	376	598
applications	388		
mechanical properties	382		
in oscillatory shearing	383		
in steady-state flow	383		
Electrorheological valves	390		
Electrostatic dissipation, conductive			
polymer composites application	274		
Electrostatic intermolecular interactions	471	472	
Electrostriction	139	218	342
measurement	141		
smart perovskites	997		
Electrostrictive actuators	1	340	
Electrostrictive ceramics	139		
<i>See also</i> Piezoelectric ceramics			
applications	146		
future trends	147		
Electrostrictive materials	860		
for adaptive systems	17	28	
for ship noise control	1099	1100	
Electrostrictive single crystals	143		
Elevated houses	1016		
Elinvar	1044		
Emulsions, memory effects	267		
Enclosed sound fields	1109	1112	
Energy supply systems, smart			
materials for	62	63	
Engineering materials	214	243	
Enhanced active constrained layer			
(EACL) damping	31		
Enlargement, technology direction	582		
Ensemble, of molecules in memory fluid	269		
Enteric coatings, smart pills	838		
Environmental and people applications	392		
Environmental design issues	219		
Environmentally-responsive			
self-regulated drug delivery systems	322		
Environmental sensors	61	64	
Enzymatic biosensors	95	97	98
enzymes used	100		
immobilization techniques	100		

Index Terms	Links			
Enzyme-based drug delivery systems	324			
Enzyme immobilization	101			
chitosan-based gel application	188			
conductive polymer application	281			
Enzyme reaction-sensitive hydrogels	495			
applications	500			
Epoxide-terminated polyurethane				
conductive adhesive	334			
Epoxy fiberglass snubbers	1125			
Epoxy molding compounds, flip-chips				
for underfills	446			
Epoxy resins				
as binder with PZT in smart paint				
composites	755			
in flip-chips	441			
Erasable/rewritable digital storage	738	751		
ERL-4221	444			
<i>Escherichia coli</i> , infrared detection	116			
ESF controllers	452			
Esters, Langmuir-Blodgett films	585			
<i>N,N</i> -1,2-Ethanediybis(2-methyl-2-propenamide)-based molecularly imprinted polymers	671			
Ethanol biosensor	98			
Ethinyl estradiol, controlled release	324			
Ethyleneglycol				
dimethylmethacrylate-based				
molecularly imprinted polymers	671	672	675	677
β -Eucryptite, negative thermal expansion	1050			
Euler buckling, material				
characteristics for architectural				
smart materials	62			
<i>Eurema</i> butterfly wings	115			
Evolution	217			
Exhaust sensor, automotive	347			
Extended irreversible thermodynamics	272			
Exterior radiation sensors, for louvers	62			
Externally-regulated drug delivery systems	319	320		
Extrinsic electrorheological materials	379			
Extrinsic Fabry-Perot interferometer sensors	724	725		
Extrinsic fiber-optic sensors	416	719		
Extrinsic interferometric sensors	1102			
Eyeglasses				
photochromic	1036			
superelastic frames	84	923	929	
F				
Fab fragments	840			
Fabry-Perot etalons	417			

Index Terms	Links					
Fabry-Perot fiber-optic sensors	719	727				
for cure monitoring	297					
for health monitoring	303					
for ship health monitoring	987					
Fabry-Perot interferometers	298	417	724	725	726	
power industry application	878					
Factors of ignorance	1117					
Fail-safe design, of aircraft	313					
Faraday-rotation transducers	881					
Fast ion conductors	349	359				
Fast polarization	364					
Fatigue						
shape-memory actuators	629					
shape-memory alloys	926					
Fatigue life						
octave rule for improving	1122					
sine wave vibration	1121					
Fatty acids						
gelators	475					
Langmuir-Blodgett films	585	589				
Faujasite, negative thermal expansion	1050					
Federal power agencies	874					
Feldspar	1020					
Ferrimagnets	591					
Ferrite, for composite tagging	310					
Ferroelectric ceramics in shape-memory	148	150				
alloy/piezoelectric heterostructures	557					
Ferroelectric hysteresis	341					
Ferroelectricity	860					
piezoelectric ceramics	162	164				
semicrystalline polymers	863					
smart perovskites	993					
Ferroelectric liquid crystalline elastomers	850					
Ferroelectric liquid crystals	850					
Ferroelectric relaxors	144	151	153	344	575	
Ferro fluids, for gas leakage prevention	455					
Ferromagnetic shape-memory alloys	936					
Ferromagnetic tagged composites	310					
Ferromagnets	591					
magnetostriction in	601					
Ferrous alloys	215					
Fiber Bragg gratings	395	728				
in photothermorefractive glasses	778					
reflection in	395					
Fiber Bragg grating sensors	395	719	724	725	726	728
applications	412					
for civil structure health monitoring	316					
composite manufacturing applications	562					
for cure monitoring	297					
β -eucryptite application	1051					

Index Terms**Links**

Fiber Bragg grating sensors (cont.)			
for health monitoring	303		
matched	411		
power industry applications	878		
sensor demodulation	408		
sensor multiplexing	401		
for ship health monitoring	988		
for ship noise control	1102		
Fiber Fabry-Perot filter	409	410	
<i>See also</i> Fabry-Perot fiber-optic sensors			
Fiber gratings	417	418	
Fiber optic Fresnel reflectometers			
composite manufacturing applications	562		
Fiber-optic sensors	62		
<i>See also</i> Fabry-Perot fiber-optic sensors;			
Fiber Bragg grating sensors			
for adaptive composite systems	17		
advantages	415		
applications	730		
in architectural smart materials	62	63	64
benefits and concerns	731		
for civil structure health monitoring	315		
classification	718		
comparison	725		
composite manufacturing applications	562	564	
for cure monitoring	294	295	
for data transmission in ship			
health monitoring	989		
effect of physical parameters on			
optical modulation	722		
for health monitoring	302	415	
for highway smart structure monitoring	546	547	
introduction, evaluation, and application	715		
for power industry chemical sensing	881		
sensing mechanisms	719	720	
for ship health monitoring	987		
for ship noise control	1102		
for smart structures	415		
theory and applications	415		
Fiber-optic transducers, power			
industry applications	876		
Fiber-reinforced composites	246		
continuous fiber reinforcement	251		
fiber-matrix interface	248		
randomly oriented fiber reinforcement	254		
textile architectures	253		
Fiber-reinforced concrete			
damage self-diagnosis	897		
health monitoring via electrical resistance	311		
smart highways	546	547	

Index Terms	Links		
Fiber-reinforced plastics	246		
See also Carbon-fiber reinforced plastics; Composites; Glass-fiber reinforced plastics			
cure monitoring	292		
damage self-diagnosis	891		
fiber Bragg grating sensors	405		
Fictitious control surface concept	50		
Field-responsive chitosan-based gels	188		
Filament winding	293		
Films, Langmuir-Blodgett. See Langmuir-Blodgett films			
Finite differences method	265		
Finite elements methods	265		
adaptive composite system modeling	19		
for aircraft structure design	50	57	
cavitative pest control calculation	763		
memory fluid calculation	271		
truss structure analysis	1067		
for vibrational analysis	1116	1119	1124
Finite volumes method	265		
Firechek valve	930		
Fish aquatic studies	119	423	
metabolic rate determination	107		
operant conditioning applications	423	424	
telemetry applications	431	432	
Fish kills	887		
Flaperon	32		
Flexible image carrier	715		
Flexible machines	448		
Flexible manipulator	1094		
Flexible structures			
large flexible space structures	1065		
microrobotic	641		
Flexinol	923	927	
Flexon optical frames	929		
Flexensional sonar transducers	611		
Flint	214		
Flip-chip applications			
anisotropic conductive adhesives	331		
molded (tablet) underfills	446		
reworkable underfills	444		
underfill materials	438		
Flip-chip in package techniques	439		
Flip-chip on board technique	439		
Floating monolayers	584		
Flow control actuators	51		
Flow-Gard valve	930		
Flow metering, ultrasound for	691		
Flow sensors, miniaturized	345		

Index Terms	Links	
Fluidic applications		
chip fabrication	644	
microrobotics	637	
microtubes	654	
Fluidlastic technology	30	
Fluid machines	448	
Fluids. <i>See also</i> Electrorheological fluids; Magnetorheological fluids; Memory fluids; Newtonian fluids; Non-Newtonian fluids		
memory effects in complex	267	
Fluorescence scattering sensors	418	
Fluorescent probes	417	
Fluorimetry, for cure monitoring with fiber-optic sensors	295	296
Fluorine-doped tin oxides, in smart windows	1143	
Fluorites, triboluminescence	1056	
Fluorohydrocarbons, Langmuir-Blodgett films	585	
Flutter, aircraft	45	49
Flux, for epoxy resins in flip-chips	443	
Foams		
magnetorheological materials	598	
smart foam skin	1029	
Food quality losses, reactions leading to	177	
Foraging behavior studies	423	
Force-sensing resistor	907	
Force sensors		
giant magnetostrictives for	511	
magnetostrictives for	616	
Formed-in-place membranes		
chitosan-based gel applications	187	
Forum for Intelligent Materials	392	
Fossil-fueled boilers	874	
Foundry technology	214	
Fractoluminescence	190	
Fracture	190	1055
<i>See also</i> Damage detection		
triboluminescence during		
Frangibolt release bolts	928	
Free-free beam element	529	
damage detection	535	
Free recovery	923	
Free stand zone melting, for giant magnetostrictive material manufacture	508	
Free system model	1070	
Frequency, of vibration	1114	
and displacement	1117	

Index Terms	Links		
Frequency dependent			
electromagnetic sensing (FDEMS)	456		
cure monitoring application	456	460	
life monitoring application	467		
molecular mobility sensing	459		
Frits	1051		
Frontier Ceramics Project	392	582	
Fructose biosensors	99		
Fuel cells	999		
Functional gradient piezoelectric			
ceramic actuators	1		
applications	10		
Functionally graded polymer blends	826		
Fused deposition of composites	145		
Fuselage noise attenuation	23		
Fuses, electric power industry	884		
Fuzzy ART	684		
Fuzzy associative memory	684		
G			
Gadolinium-magnesium alloys, in			
smart windows	1140		
Galactose biosensors	99	107	
Gallium arsenide	216		
Gallium chalcogenide films	745		
Galloping control, power			
transmission lines	883	888	
Galvanic cells	357		
Gas igniters, piezoelectric ceramic			
application	155		
Gasochromic smart windows	1133	1135	1140
Gas sensors			
Langmuir-Blodgett film applications	589		
selective	348		
surface acoustic wave	913		
Gastric acid inhibitors	319		
Gatling gun	1127		
Gelatin	487		
with chitosan-based gels	183		
Gelation	474		
Gelators, organic	471		
Gels	490		
See also Chitosan-based gels;			
Hydrogels; Sol-gels			
electrorheological materials	376		
magnetorheological materials	598		
memory effects	267		
for pulsatile drug delivery systems	322		
Gel transition	561		
Generators, electric power industry	880		
GENERIC framework	272		

Index Terms	Links				
Gene therapy, chitosan-based gel delivery	186				
Germanium chalcogenide films	740	741	745	751	
GFRP. See Glass-fiber-reinforced plastics					
Giant magnetoresistance	202				
Giant magnetostriction	601				
Giant magnetostrictive materials	503				
applications	510				
in magnetic					
particulate/shape-memory					
alloy matrix composites	555				
manufacturing	508				
origins of effect	504				
shape-memory alloys compared	623				
Ginzburg-Landau formulation					
relaxation equations	272				
Glass/epoxy composites, shapememory					
alloy fiber/polymer matrix composites	553				
Glasses	245	770			
for enzyme immobilization	102				
negative thermal expansion	1051				
photochromic and photothermorefractive	770				
photosensitivity	770				
triboluminescence	1055	1056			
Glass-fiber-reinforced concrete					
damage self-diagnosis	897				
smart highways	546				
Glass-fiber-reinforced plastics					
for aircraft health monitoring	314				
cure monitoring	298	300			
damage self-diagnosis	891				
damage self-monitoring	392				
and ken materials	581	582	583		
tagging with ferromagnetics	310				
Glass microtubes	647				
Glassy carbon, in anodes for					
lithium-ion batteries	80				
Glazing materials	60	62			
for smart windows	1133				
Global Positioning System (GPS)					
with advanced automobiles	548	549	550		
for fish telemetry studies	432				
for ship health monitoring	987				
Glucose biosensors	95	99	100	104	
home test kits	98				
Glucose oxidase	99	100	105	106	325
in chemical valve	844				
Glucose-6-phosphate dehydrogenase	100				
Glucose-responsive hydrogels	500				
Glucose-responsive insulin delivery	106	325			

Index Terms	Links					
Glyceraldehyde-3-phosphate						
Dehydrogenase, affinity precipitation	840					
Glycol chitosan	185					
Goldfish, operant conditioning	424	428				
Goldsmith tolerance factor	204					
Golf clubs, shape-memory alloy application	930					
GPS. <i>See</i> Global Positioning System						
Graphite, noncontact ultrasound						
examination	702					
Graphite/epoxy composites, shapememory						
alloy fiber/polymer matrix composites	553	554				
Graphite-fiber-reinforced plastics.						
<i>See</i> Carbon-fiber-reinforced plastics						
Graphitic carbon, in anodes for						
lithium-ion batteries	80					
Grazing incidence X-ray diffraction						
(GIXD), Langmuir-Blodgett films	587	588				
Greases						
electrorheological fluid phase	377					
magnetorheological materials	598					
Great white sharks, acoustic telemetry	436					
Green alumina, noncontact						
ultrasound examination	706					
Greenhouse window opener						
shape-memory alloy application	930					
Green porcelain, noncontact						
ultrasound examination	707					
Green tungsten carbide, noncontact						
ultrasound examination	706					
Gruneisen parameter	1039	1040				
Guidewire cores, shape-memory application	931					
H						
Hand lay-up methods	293					
Haptic sensor	92					
Hawkmoth, wing as antireflective device	114					
Health monitoring	219	301				
adaptive systems	39					
aircraft and space structures	45	302	303	313	420	718
civil structures	303	315				
damage detection approach	533					
dereverberated transfer functions						
of nonuniform structures	525					
dereverberated transfer functions						
of structural elements	520					
highway structures	546					
life monitoring	467					
ships	981					
smart paints for	754					
using wave dynamics	520					

Index Terms	Links		
Heat capacity			
extrapolation for polymers at very low			
temperatures, neural network application	684		
material characteristics for			
architectural smart materials	62		
Heat engines, shape-memory alloy			
applications	934		
Heat-shock proteins	116	120	
Heave, of ships	987	988	
Helicopters			
adaptive systems applications	28		
rotor control system, piezoelectric			
actuator application	160		
Heparin, hydrogel-based delivery systems	498		
Heterogeneity determination			
ultrasound application	691		
Heterogeneous photochromic glasses	774		
Heterogeneous systems, using smart			
polymers	838		
Hexacelsians, triboluminescence	1057		
Hexacyanometallate magnets	595		
Hexaphenylcarbodiphosphorane			
triboluminescence	1054		
Hexcel 8552, intelligent processing	563		
Hexsil process	644		
High density microvia printed wiring boards	438		
High density polyethylene, in composites			
with large positive temperature coefficients	274	277	278
High energy density batteries. See			
Lithium-ion batteries			
High performance liquid			
chromatography (HPLC)			
molecularly imprinted polymers application	672		
High pressure environments			
steam plants	874		
ultrasound application	691		
High radiation environments			
ultrasound application	691		
High temperature environments			
noncontact ultrasound application	713		
steam plants	874		
ultrasound application	691		
High-temperature indicating crayons	176		
High-temperature shape-memory alloys	955	956	
High-temperature superconductors			
smart perovskites	992	1001	1006
Highways	545		
advanced automobiles	548		
damage self-diagnosis	891		
monitoring earthquake damage	582		

Index Terms	Links				
Hill gratings	399				
Hirudin, chitosan-based gel encapsulation	185				
Hollow fiber membranes					
chitosan-based gel applications	187				
Holographic anticounterfeiting					
indicator devices	180				
Homeostasis	219	222			
Homing studies	423				
Hook-and-loop fastener	217				
Hopfield nets	684				
Hormone replacement therapy	319				
Hot isostatic pressing, intelligent processing application	567				
Hot-press molding	293				
Houses, made of soil	1025				
Hubble Space Telescope	1066				
Hull girder stress	981				
Hull girder whipping	982				
Hull Response Monitoring Systems	990				
Hull vibration	1097				
Humidity sensing					
material characteristics for					
architectural smart materials	62				
surface acoustic wave sensors for	916				
Hund's rule	998				
HVAC systems, smart materials for	60	62	64	65	67
Hyaluronic acid gels, for pulsatile drug delivery systems	322	324			
Hybrid composites	551				
Hybrid magnetostrictive/piezoelectric devices	613				
Hydraulic actuators, for ship noise control	1106				
Hydraulic circuit control					
electrorheological material applications	390				
Hydraulic devices	215				
Hydraulic servo-spool valve	449				
Hydrocarbon sensors					
for power industry applications	881				
surface acoustic wave	913				
Hydrogarnet	1019				
Hydrogels					
applications	497				
as chemical indicators	174	175			
defined	490				
for drug delivery systems	322				
enzyme immobilization in	101				
in smart windows	1138				
structure and properties	494				
types	494				

Index Terms	Links		
Hydrogen bonding	471	472	
chemical indicators	174		
hydrogels	491		
Hydrogen loading, with fiber Bragg gratings	400		
Hydrothermal processing, of			
soil-ceramics	1019		
Hydrothermal synthesis, of smart			
perovskites	568		
Hydroxypropylmethylcellulose			
acetate succinate, for smart pills	839		
Hydroxypropylmethylcellulose			
Phthalate, for smart pills	839		
8-hydroxypyrene-1,3,6-trisulphonic			
acid, as pH indicator	879		
Hygrometers, in architectural smart			
materials	62		
Hysteresis			
biosensors	97	98	
electrorheological materials	384		
ferroelectric	341		
ferroelectric liquid crystalline elastomers	850	852	
ferromagnetic shape-memory alloys	936		
giant magnetostrictive materials	505	508	
magnetostrictive materials	604	617	
organic magnets	592	593	
piezoelectric ceramics	164	170	
PVDF and copolymers	807		
shape-memory alloys	927	959	970
and shape-memory alloys in			
biomedical applications	84		
I			
Ice, dangers to ships	982		
IC fabrication	438		
Illuminance measurements, material			
characteristics for architectural			
smart materials	62		
Ilmenites, spontaneous polarization	994		
Immobilization. See Enzyme immobilization			
Immune responses, to drug delivery	319		
Immunization	319		
Immunoassays, smart polymers for	846		
Immunochemical biosensors	96	97	
Impact damage, sensor array			
technology for aircraft	903	904	907
Impact test, for smart truss structures	1077		
Implants	82		
See also Drug delivery systems			
chitosan-based	182		
Imprint lithography	645		
Indirect fiber-optic sensors	719		

Index Terms	Links		
Indium chalcogenide films	744	751	
Indium tin oxide (ITO)			
with PPV in photovoltaics and			
light-emitting devices	799		
in smart windows	1143		
Indium tungstate, negative thermal expansion	1049		
Indomethacin, pulsatile delivery systems for	321		
Induction motors, for ship noise control	1106		
Industrial Revolution	215		
Inertial damping, material			
characteristics for architectural			
smart materials	62		
Inflammation	319		
Inflammation-responsive drug			
delivery systems	324		
Inflatoplane	53		
Information integration	222		
Infrared detection, biological	115		
Inherently conducting polymers	279		
<i>See also</i> Conductive polymers			
ring-substituted	286		
Injection molding	293		
Injection molding mandrels			
shape-memory alloy application	934		
In-line fiber etalon sensor	729		
Innovative Control Effector program	51		
Inorganic electrochromic smart windows	1136		
Insect vision	112		
Insertion compounds, lithium	72	73	
Insulin	319		
glucose-responsive delivery	106	325	
hydrogel-based delivery systems	497	498	500
pulsatile delivery systems	320	321	
Insulin pumps	105		
Integral passives	438		
Integrating fiber-optic sensors	718		
Integration, technology direction	582		
Intelligent biomaterials	392		
Intelligent ceramics	392		
Intelligent components, in architecture	60		
Intelligent environments	60		
Intelligent fibers	392		
Intelligent hydraulics	449		
Intelligent materials	218	860	
<i>See also</i> Smart materials			
Intelligent polymer systems	279		
<i>See also</i> Conductive polymers; Smart			
polymers			
Intelligent processing of materials (IPM)	559		
Intelligent synthesis, of ceramic materials	568		

Index Terms	Links		
Intelligent transportation systems	546	548	551
Intensimetric optical fiber sensors	720	723	
Intensity-based sensors, for health monitoring	303	305	
Interaortic balloon pump	86	87	
Interdigital transducers	911		
Interface nucleation and Propagation, shape-memory material modeling	968		
Interfacial polarization, and electrorheological effect	363	378	386
Interferometric based demodulation	411		
Interferometric optical fiber sensors for health monitoring	723		
	303		
Interior daylight sensors, for louvers	62		
Interior heat generation, smart materials for controlling	62		
Intermolecular interactions	471	472	
Interpenetrating polymer network	183		
Intrinsically-distributed fiber-optic sensors	718		
Intrinsically-smart structures	223		
cement matrix composites for	223		
polymer matrix composites for	233		
Intrinsic electrorheological materials	379	381	
Intrinsic Fabry-Perot interferometer sensor	724	725	
Intrinsic fiber-optic sensors	416	719	
Intrinsic sensors	399		
Invar effect	1044		
Invars			
magnetostriction in	603		
magnetovolume effect	609		
negative thermal expansion	1043	1046	
Inverse Joule effect	506	608	
Inverse Wertheim effect	507		
Inverse Wiedemann effect	609		
Investor-owned utilities	874		
Ionic conductivity, smart perovskites	999		
Ionic displacement polarization, and electrorheological effect	363	364	386
Ionic hydrogels	492		
Ion transfer	221		
I-Point (Vitsab) time-temperature indicator	179		
Iron	215		
See also Invars; Terfenol-D			
magnetostrictive coefficient	607		
suspended particles as magnetorheological fluid	598		
Iron Age	214	215	

Index Terms	Links		
Iron-based shape-memory alloys	951		
magnetically activated ferromagnetic	936		
magnetic-field-induced martensitic transformation	956		
Iron-based spin-crossover materials	1036		
Iron-chromium-nickel alloys biocompatibility	83		
Iron-manganese-silicon			
shape-memory alloys	951		
for ship noise control	1101		
Iron oxide cathodes, lithium-ion batteries	80		
Iron-palladium shape-memory alloys			
magnetic field-induced strain	938	947	
Iron powder compact, noncontact ultrasound examination	711		
Iron-rhenium alloys, for optothermo magnetic motors	392		
Iron-terbium-dysprosium magnetostrictive materials. See Terfenol-D			
Iron(tetracyanoethylene) ₂ -based magnets	595		
Irreversible digital storage	738	745	751
Irreversible temperature labels	176		
Isocitrate dehydrogenase	100		
<i>N</i> -Isopropyl carbazole triboluminescence	1054		
Isotropic conductive adhesives	331		
Isotropic spontaneous magnetostriction	606		
ITO. See Indium tin oxide			
J			
Jahn-Teller distortion			
in cathodes for lithium-ion batteries	79		
in colossal magnetoresistive materials	202	207	
Joule annealing, of shape-memory alloys	640		
Joule effect	506		
Joule magnetostriction	601	606	609
<i>Journal of Intelligent Material Systems and Structures</i>	218		
<i>Journal of Smart Material Systems and Structures</i>	218		
K			
Kaolinite, hydrothermal processing	1019		
Keatite	1051		
Ken materials	581		
Ken Materials Research Consortium	392	581	
Kerosene, electrorheological fluid phase	377		
Kerr effect	353		
Kevlar 49	257		
Kevlaro, in shape-memory alloy			
fiber/polymer matrix composites	553		
Kiesewetter motor	612		
Knee prosthesis	449	451	

Index Terms	Links	
Kohonen self-organizing maps	684	
Koi, operant conditioning	426	
Krypton fluoride lasers, for Bragg grating writing	400	
Kunster marrow needle	84	
L		
LabView, fish studies application	429	
Lactate biosensor	98	
Lactate dehydrogenase	100	
Lactate oxidase	100	
Lactose biosensors	95	107
Lagrangian methods	272	
Lake trout, radio telemetry	435	
Lamellar composites	246	
Laminate failure	263	
Lamination theory	260	
Lamnoid sharks, acoustic telemetry	436	
Langmuir-Blodgett films	584	
characterization of transferred	588	
for microtube formation	645	
PVDF copolymers	815	
smart material applications	589	
Langmuir-Blodgett-Kuhn films	584	
Langmuir-Blodgett-Shaefer method	587	
Langmuir-Blodgett trough	584	
Langmuir film balance	584	585
Langmuir monolayers	584	585
Langmuir trough	584	
Lanthanide manganite perovskites		
colossal magnetoresistance in	202	
Lanthanum hydrides, in smart windows	1140	
Lanthanum manganite perovskites	1002	
colossal magnetoresistance in	202	
ionic conductivity	999	
Lanthanum strontium cobalt oxides	1009	
Laptop computers, high energy		
density batteries required	68	72
Large flexible space structures	1065	
Large-scale structures		
damage self-diagnosis	891	
flexible space structures	1065	
Laser annealing, of shape-memory alloys	640	
Laser-assisted chemical vapor deposition	645	
Laser-induced ultrasound	693	
Laser machining	632	
Lasers		
poly(p-phenylenevinylene) application	804	
for writing on optical recording media	738	746
Lattice cell mechanics		
shape-memory material modeling	967	

Index Terms	Links			
Laws relay	450			
Layered cobalt oxide cathodes				
lithium-ion batteries	74			
Layered iron oxide cathodes				
lithium-ion batteries	80			
Layered nickel oxide cathodes				
lithium-ion batteries	75			
Layered vanadium oxide cathodes				
lithium-ion batteries	80			
Lead-acid batteries	331			
lithium-ion batteries compared	72			
Lead lanthanum zirconate (PLZT), in shape-memory alloy/piezoelectric heterostructures	557			
Lead magnesium niobate	144	151		
ferroelectric relaxor	144	153		
intelligent synthesis	569	575		
Lead magnesium niobate-lead titanate ceramics, ferroelectric liquid crystalline elastomers contrasted	858			
Lead magnesium niobate (PMN)				
applications	159	160		
for ship noise control	1099	1100		
smart perovskite	992	997		
Lead meta-niobates, in dry coupling ultrasonic transducers	693			
Lead niobate				
intelligent synthesis	575			
properties	168			
Lead titanate	151	152		
composites with piezoelectric polymers	557			
composites with smart paints	755			
intelligent synthesis	569	572	576	578
negative thermal expansion	1043			
properties	143	151	168	
in shape-memory alloy/ piezoelectric heterostructures	557			
spontaneous polarization	994			
Lead-titanium hydrothermal system	571			
Lead zinc niobate, intelligent synthesis	569	578		
Lead zirconate-lead titanate	152			
discovery of piezoelectricity in	150			
in dry coupling ultrasonic transducers	693			
Lead zirconate titanate (PZT)	337	343		
for adaptive composite systems	17	18	31	
applications	156	157	160	
cation doping	999			
characterization	162			
composites with piezoelectric polymers	144	557		
composites with smart paints	754	755		

Index Terms**Links**

Lead zirconate titanate (PZT) (cont.)						
for cure monitoring	300					
for flexible manipulator	1094					
for health monitoring	308	309				
intelligent synthesis	569	571	573	576	578	
in magnetometers	616					
mixed-oxide route for preparing	163					
piezoelectric coefficients	150					
in piezoelectric double amplifier smart skin	1032					
piezoelectricity in	143	150	162	781	788	861
properties	151	168				
in sensors for transformer monitoring	881					
in shape-memory alloy/ piezoelectric heterostructures	557					
for ship noise control	1099	1100	1111			
smart perovskite	992	995				
in smart truss structures	1071					
thin films	2	147	151	154		
for vibration control in smart structures	1089					
Learning						
animal studies	423					
self-learning	219	220	222			
Lecithins, Langmuir-Blodgett films	585					
Leclanche battery	71					
LEDs (light-emitting diodes), in architectural smart materials	62	63				
Lens polishing machine	455					
Level sensing						
noncontact ultrasound for	694	699	713			
ultrasound for	690					
Lever	215					
Lidocaine, pulsatile delivery systems for	321					
Lifeline Fresh-check/Fresh-Scan time-temperature indicators	178					
Life monitoring. <i>See also</i> Health monitoring						
frequency dependent electromagnetic sensing for	467					
Lift-off fabrication process, of microdevices	632	633				
LIGA process	644					
Light-emitting devices						
in architectural smart materials	62	63				
poly(p-phenylenevinylene) application	799					
Light-emitting electrochemical cells poly(p-phenylenevinylene) application	799	804				
Lighting systems, smart materials for	60	61	62	63		
Light valves. <i>See</i> Suspended particle panels						
Lime, in soil-ceramics	1019					
Linear displacement transducers, for ship health monitoring	986					
Linear magnetostriction	606					

Index Terms	Links		
Linear (micro)actuator	636		
Linear motors, magnetostrictives for	612		
Linear potentiometer, for ship health monitoring	986		
Linear quadratic Gaussian control			
for truss structure vibration control	1065		
for vibration control	1089		
Linear variable differential transformers			
for ship health monitoring	986		
for ship noise control	1103		
Linear variable inductance transformer, for ship noise control	1103		
Lipid templated tubes	645		
Liposomes	838	844	
with temperature-sensitive hydrogels	498		
Liquid-condensed phases	586		
Liquid crystal display labels	176		
Liquid crystal displays			
in smart windows	1133	1135	1139
and spin-crossover materials	1036		
Liquid crystalline materials			
in architectural smart materials	62	63	
as chemical indicator	174		
composites with PPV	797		
ferroelectric liquid crystalline elastomers	850		
memory effects	267		
piezoelectricity	864	866	
PPV films	798		
Liquid-expanded phases	586		
Liquid molding	293		
Liquid-phase sintering conductive adhesive	335		
Lithium aluminosilicate gels	359		
Lithium-cobalt nitride anodes			
lithium-ion batteries	81		
Lithium-cobalt oxide cathodes, in			
lithium-ion batteries	74		
Lithium-copper nitride anodes			
lithium-ion batteries	81		
Lithium-copper-tin alloy anodes			
lithium-ion batteries	81		
Lithium insertion compounds	72	73	
Lithium-ion batteries	72		
carbon anodes	80		
conventional batteries compared	72		
layered cobalt oxide cathodes	74		
layered nickel oxide cathodes	75		
other anodes	81		
other cathodes	79		
spinel manganese oxide cathodes	78		

Index Terms	Links			
Lithium-ion cells	73			
Lithium-iron nitride anodes				
lithium-ion batteries	81			
Lithium-iron oxide cathodes				
lithium-ion batteries	80			
Lithium-manganese nitride anodes				
lithium-ion batteries	81			
Lithium-manganese oxide battery	71			
Lithium-manganese oxide cathodes				
lithium-ion batteries	78			
Lithium-nickel oxide cathodes				
lithium-ion batteries	75			
Lithium niobate	344			
piezoelectricity in	781			
for ship noise control	1099	1100		
single crystals	151			
for surface acoustic wave devices	157			
Lithium-sulfur oxide battery	71			
Lithium tantalate	344			
single crystals	151			
for surface acoustic wave devices	157			
Lithium-titanium oxide anodes				
lithium-ion batteries	81			
Lithium-vanadium oxide cathodes				
lithium-ion batteries	80			
Lithography	645			
Local transformation strain	965			
Lockable damper mechanism	449			
Long-period grating based sensors				
for health monitoring	303	304		
Loudspeakers, for ship noise control	1104			
Louver control systems				
shape-memory alloy application	930			
smart materials for	62			
Low density polyethylene, in				
composites with large positive				
temperature coefficients	276			
Low-temperature martensite				
reorientation	965			
modeling	977	978		
Luciferin	103			
Lumen/watt energy conversion ratio				
material characteristics for				
architectural smart materials	62	63		
M				
Mach-Zender interferometer	419	724	725	726
Macroporous hydrogels	491			
Macroscopic transformation strain	965			
Magnesium battery	71			
Magnesium biosensors	107			

Index Terms	Links	
Magnesium hydroxide/ poly(methylphenylsiloxane) electrorheological fluid	363	
Magnetically-activated ferromagnetic shape memory alloys	936	
Magnetically-responsive hydrogels	496	
Magnetically-stimulated pulsatile drug delivery systems	320	
Magnetic anisotropy magnetostrictive materials	603	
Magnetic domain processes	604	
Magnetic-field-induced martensitic transformation	936	956
Magnetic-field-induced twin rearrangement	940	
Magnetic films, Langmuir-Blodgett film applications	589	
Magnetic moment	591	
Magnetic moment jumping	603	
Magnetic particulate/shape-memory alloy matrix composites	555	
Magnetic powder brakes	598	
Magnetic transitions, and negative thermal expansion	1043	
Magnetism	591	
Magnetization models	601	604
nickel-manganese-gallium shape-memory alloys	939	
Magnetomechanical effect	608	
Magnetometers, magnetostrictives for	615	
Magnetoresistance	202	
smart perovskites	998	
Magnetorheological fluids	597	
for adaptive systems	17	30
ESF controllers	451	
and fluid machines	450	
memory effects	267	
for ship noise control	1100	1103
for structural vibration control	62	
Magnetorheological materials	218	597
Magnetostriction	218	504 601
field-induced strain contrasted	949	
isotropic spontaneous	606	
linear	606	
and microrobotics	642	
origin of	603	
saturation	606	607
Magnetostrictive/ferromagnetic tagged composites	310	

Index Terms	Links		
Magnetostrictive fluids	600		
<i>See also</i> Giant magnetostrictive materials			
Magnetostrictive materials	216	860	
<i>See also</i> Giant magnetostrictive materials; Terfenol-D			
for adaptive systems	16	28	
hybrid magnetostrictive/ piezoelectric devices	613		
shape-memory alloy/Terfenol-D heterostructures	557		
for ship noise control	1099	1100	1101
Magnetostrictive transducers	609		
Magnetovolume effect	609		
Magnets, organic/polymer	591		
Maleic anhydride copolymers			
Langmuir-Blodgett films	586		
Mammalok	931		
Manganese nitride anodes			
lithium-ion batteries	81		
Manganese oxide spinel cathodes			
lithium-ion batteries	78		
Manganese(tetracyanoethylene)2- based magnets	595		
Manifolded microtubes	652	654	
Manual processing of materials	559	560	
Manufacturing process monitoring	292		
Marrow needles	82		
shape-memory alloy application	84		
Mars Sojourner Rover actuator			
shape-memory alloys application	928		
Martensite. <i>See also</i> Austenitemartensite			
phase transformation			
aging and order state of	953		
biased and unbiased	965		
classification of nonferrous	951		
low-temperature reorientation	965	977	978
stress-strain behavior	939		
Mass loading, surface acoustic wave sensors	911		
Material processing, intelligent	559		
Materials evolution	214	243	
Materials science	214		
Mathematical modeling. <i>See</i>			
Computational techniques;			
Modeling			
Matrix models	272		
Matsushiro earthquake swarm			
(1965), triboluminescence during	1055		
Matteucci effect	609		

Index Terms	Links	
Maxwell-Wagner-Sillars interfacial polarization	378	387
Mechanically-induced triboluminescence	1056	
Mechanical properties determination		
noncontact ultrasound application	700	
ultrasound for	690	691
Mechanoluminescence	190	
alkaline aluminates doped with rare-earth ions	191	
devices for measurement	190	
zinc sulfide doped with transitionmetal ions	191	195
Medical applications. See Biomedical applications		
Medical checkup devices	82	
Medium-range order, chalcogenide compounds	740	
Megamouth shark, ultrasonic telemetry	432	
MEH-PPV. See Poly[2-((2-ethylhexyl)oxy-5-methoxy-p-phenylene)vinylene]		
Melamine-barbiturate organogelators	480	
<i>Melanophila acuminata</i> , infrared detection	118	
Membranes		
chitosan-based gel applications	186	
microtube fabrication in nanoporous	645	
molecularly imprinted polymers application	672	
Memory effects, computational techniques	266	
Memory fluids		
constitutive equations	265	
finite element calculation	271	
Memory scattering	353	
Mercury battery	71	
Mesophase-pitch-based carbon fibers in composites for intrinsically smart structures	233	
Metakaolin	1019	
Metal matrix composites	216	246
intelligent processing application	567	
Metal microtubes	647	
Metal processing, intelligent		
processing application	567	
Metal welding, intelligent material processing	559	
Metglas		
magnetoelastic properties	605	
magnetostriction in	602	
in sensors	614	
Methacrylic acid, chemical indicator	175	
Methotrexate, hydrogel-based delivery systems	497	
Methyl cellulose, dielectric properties	368	

Index Terms	Links				
Methyl ethyl ketone (MEK) detection					
with surface acoustic wave sensors	913				
Methyl methacrylate, chemical indicator	175				
Methyl red	175				
Michelson interferometer	724	725	726		
Microactuators	935				
Microbalances, surface acoustic wave sensors	912				
Microbend fiber-optic sensor	416				
Microcomputers					
macroworld/microworld association	503				
in smart monitoring systems	292				
for vibrational analysis	1116				
Microdevices	630				
micromachining and fabrication	631				
smart shape-memory alloy	637				
Microelectromechanical systems	337	355	556	630	644
shape-memory alloy applications	935				
surface acoustic wave sensors as thin films for	912				
	147	151			
Microendoscope	635				
Microgrippers	635	638			
locally annealed	641	642			
Microolithography, Langmuir-Blodgett film applications	584	589			
Micromachining					
for microrobotics	631	644			
for microtube formation	631	644			
Micro/macro computational methods	273				
Micromagnetics	941				
Microphones, for ship noise control	1103				
Microporous hydrogels	491				
Microrobotics, shape-memory alloys for	620				
Microstructure determination (of metals)					
noncontact ultrasound application	699	701	705	706	708
ultrasound application	691				
Microsystems	644				
Microtaggant anticounterfeiting					
indicator devices	180				
Microtube check valve	656				
Microtube flow limiter	656				
Microtube flow restricter	656				
Microtube internal bearings	665				
Microtube pressure/flow regulator	657				
Microtube pressure sensor	658				
Microtubes	644				
applications	648				
devices based on surface tension and wettability	654				
Microtube shutter mechanism	663				

Index Terms	Links
Migration studies	423
Milk products, noncontact ultrasound examination	711
Mineral oil, electrorheological fluid phase	377
Minerals, triboluminescence	190
Miner's cumulative damage	1127
Miniaturization	582
electronic packaging	438
and microrobotics	620
microtubes	644
Miniaturized flow sensor	345
Minirobotics	636
Mini-to microgrippers	635
Mission Adaptive Wing	45
Mitek Homer Mammalok	931
Mitek suture anchors	931
Mitosis	221
Modal domain optical fiber sensors	723
Modeling. <i>See also</i> Computational techniques	
adaptive composite systems	16
shape memory alloys	964
Modified Bridgman method, for giant magnetostrictive material manufacture	508
Molded (tablet) underfills, flip-chip applications	446
Mold release compounds, with flip-chips for underfills	444
Molecular actuators	285
Molecular complexity, and memory effects	266
Molecular descriptions, of smart materials	268
Molecularly-imprinted polymers	667
applications	672
for biosensors	109
Molecularly imprinting	667
Molecular mobility sensing	
frequency dependent	
electromagnetic sensing for	459
Molecular recognition	471
Molecular Sieve	
3A/poly(dimethylsiloxane)	
electrorheological fluid	366
Molecule-based magnets	595
Monoamine oxidase	100
Monoclonal antibodies, binding to	
Myoglobin, detection using porous silicon sensors	130

Index Terms**Links**

Monomers. <i>See also</i> Polymers polymerization in			
Langmuir-Blodgett films	586		
Monsoon regions, soil-ceramic applicability	1014	1016	
Monte Carlo methods	273		
neural network modeling of simulated polymer properties	686		
Moonie (piezoelectric actuator)	1	147	159
Morphine triggered naltrexone delivery systems	325		
Morphing Program	51		
<i>Morpho</i> butterfly wings	114		
Morphotropic phase boundary	996		
Mosquitoes, ultrasonic pest control	761		
Mossybacked atopic dermatitis detecting with skin sensor	93		
Motion sensors			
pest control applications	761		
for ship health monitoring	987		
Motors, electric power industry	880		
Mucoadhesive drug delivery systems	186		
Multifunctional systems	39		
Multifunction integrated film	635		
Multimode graded index optical fibers	716		
Multimode stepped index optical fibers	716		
Multistep constant-amplitude controller, for vibration control	1089		
Multiwire tension device shape-memory alloy application	934		
Myoglobin, binding to monoclonal antibodies, detection using porous silicon sensors	130		