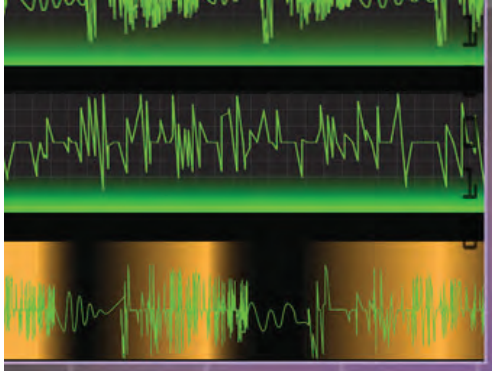
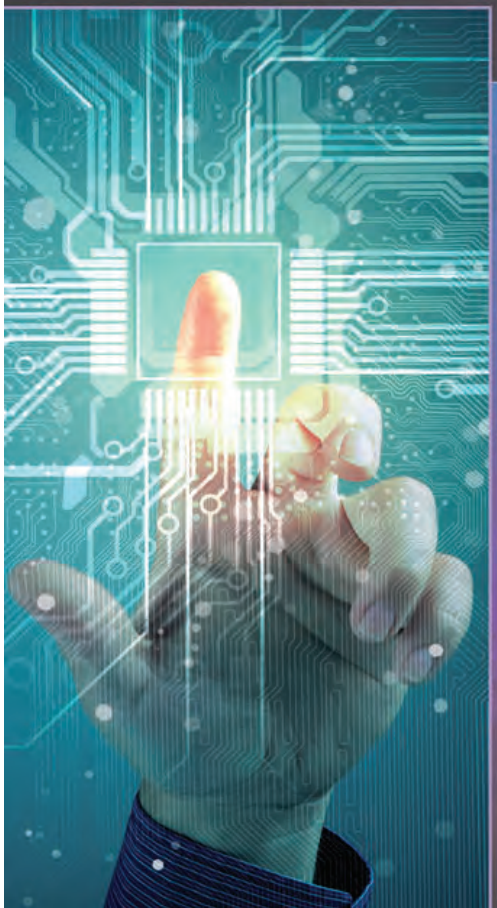








SECOND EDITION

Measurement, Instrumentation, and Sensors Handbook

Spatial, Mechanical, Thermal, and Radiation Measurement

EDITED BY
John G. Webster
Halit Eren



CRC Press
Taylor & Francis Group

Valve Exhibit 1046
Valve v. Immersion

SECOND EDITION

Measurement, Instrumentation, and Sensors Handbook

Spatial, Mechanical, Thermal, and Radiation Measurement



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

SECOND EDITION

Measurement, Instrumentation, and Sensors Handbook

Spatial, Mechanical, Thermal, and Radiation Measurement

EDITED BY
John G. Webster
Halit Eren


CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

MATLAB® is a trademark of The MathWorks, Inc. and is used with permission. The MathWorks does not warrant the accuracy of the text or exercises in this book. This book's use or discussion of MATLAB® software or related products does not constitute endorsement or sponsorship by The MathWorks of a particular pedagogical approach or particular use of the MATLAB® software.

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2014 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works

Printed in Canada on acid-free paper
Version Date: 20140716

International Standard Book Number-13: 978-1-4398-4888-3 (Hardback)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (http://www.copyright.com/) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Library of Congress Cataloging-in-Publication Data

Measurement, instrumentation, and sensors handbook : spatial, mechanical, thermal, and radiation measurement / editors, John G. Webster, Halit Eren. -- Second edition.
pages cm
Includes bibliographical references and index.
ISBN 978-1-4398-4888-3 (hardcover : alk. paper)
1. Physical measurements--Handbooks, manuals, etc. 2. Measurement--Handbooks, manuals, etc. 3. Scientific apparatus and instruments--Handbooks, manuals, etc. I. Webster, John G., 1932- editor of compilation. II. Eren, Halit (Professor of electrical engineering), editor of compilation.

QC39.M393 2014
530.8--dc23 2013025789

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

Preface xiii
Acknowledgments xv
Editors xvii
Contributors xix

PART I Instrumentation and Measurement Concepts

1 Measurements, Instrumentation, and Sensors **1**-1
Halit Eren

2 Characteristics of Instrumentation **2**-1
John R. Hansman, Jr.

3 Operational Modes of Instrumentation **3**-1
Richard S. Figliola

4 Static and Dynamic Characteristics of Instrumentation **4**-1
Peter H. Sydenham

5 Measurement Accuracy **5**-1
Ronald H. Dieck

6 Development of Standards **6**-1
Halit Eren

7 Measurement Standards **7**-1
DeWayne B. Sharp

8 Calibrations in Instrumentation and Measurements **8**-1
Halit Eren

9 Intelligent Sensors and Instruments **9**-1
Halit Eren

10 Virtual Instruments **10**-1
David Potter and Halit Eren

11 Fail-Safe Instruments and Devices **11**-1
Davide Quatrini, Giuseppe Fazio, Mauro Giaconi, and Adelio Salsano

12 Dynamic Error Measurements of Force Sensors **12**-1
Akihiro Takita, Jin Tao, and Yusaku Fujii

PART II Spatial Variables

13 Thickness Measurement **13**-1
John C. Brasunas, G. Mark Cushman, and Brook Lakew

14 Distance Measurement **14**-1
W. John Ballantyne

15 Altitude Measurement **15**-1
Dimitris E. Manolakis

16 Attitude Measurement **16**-1
Mark A. Stedham, Partha P. Banerjee, Seiji Nishifuji, and Shogo Tanaka

17 Inertial Navigation **17**-1
Halit Eren

18 Level Measurement **18**-1
Detlef Brumbi

19 Area Measurement **19**-1
Charles B. Coulbourn and Wolfgang P. Buerner

20 Volume Measurement **20**-1
René G. Aarnink and Hessel Wijkstra

21 Tilt Measurement **21**-1
Adam Chrzanowski and James M. Secord

22 Proximity Sensing for Robotics **22**-1
Ricardo E. Saad, Ben Benhabib, A. Bonen, and K.C. Smith

PART III Displacement

23 Resistive Displacement Sensors **23**-1
Keith Antonelli, James Ko, and Shyan Ku

24 Inductive Displacement Sensors **24**-1
Halit Eren

25 Capacitive Sensors: Displacement, Humidity, Force **25**-1
Halit Eren

26 Piezoelectric Sensors and Transducers **26**-1
Ahmad Safari, Victor F. Janas, Amit Bandyopadhyay, and Andrei Kholkine

27 Laser Interferometer Displacement Sensor **27**-1
Bernhard Günther Zagar

28 Bore-Gaging Displacement Sensors **28**-1
Viktor P. Astakhov

29 Ultrasonic Displacement Sensors **29**-1
Nils Karlsson and Ole Pedersen

30 Optical Encoder Displacement Sensors **30**-1
J.R. René Mayer

31 Magnetic Displacement Sensors **31**-1
David S. Nyce

32 Synchro/Resolver Displacement Sensors **32**-1
Robert M. Hyatt, Jr. and David Dayton

33 Optical Fiber Displacement Sensors **33**-1
Richard O. Claus, Vikram Bhatia, and Anbo Wang

34 Optical Beam Deflection Sensors **34**-1
Grover C. Wetsel

35 Velocity Measurement **35**-1
Charles P. Pinney and William E. Baker

PART IV Mechanical Variables

36 Acceleration, Vibration, and Shock Measurement **36**-1
Halit Eren

37 Strain Measurement **37**-1
Christopher S. Lynch

38 Tactile Sensing **38**-1
Ricardo E. Saad, A. Bonen, K.C. Smith, and Ben Benhabib

39 Pressure Measurement **39**-1
Kevin H.L. Chau

40 Vacuum Measurement **40**-1
Ron Goehner, Emil Drubetsky, Howard M. Brady, and William H. Bayles, Jr.

41 Force Measurement **41**-1
M.A. Elbestawi

42 Angle Measurement **42**-1
Robert J. Sandberg

43 Mass, Weights, and Instrumentation **43**-1
Emil Hazarian

44 Torque and Power Measurement **44**-1
Ivan J. Garshelis

45 Density Measurement **45**-1
Halit Eren

46 Fluid Viscosity Measurement **46**-1
R.A. Secco, M. Kostic, and J.R. deBruyn

47 Surface Tension Measurement **47**-1
David B. Thiessen and Kin F. Man

PART V Acoustics

48 Acoustic Measurement **48**-1
Per Rasmussen

49 Ultrasound Measurement **49**-1
Peder C. Pedersen

PART VI Flow and Spot Velocity

50 Capillary-Type Mass Flow Meter **50-1**
Reza Pakdaman Zangabad and Manouchehr Bahrami

51 Differential Pressure Flowmeters **51**-1
Richard Thorn

52 Variable Area Flowmeters **52**-1
Adrian Melling, Herbert Köchner, and Reinhard Haak

53 Positive Displacement Flowmeters **53**-1
Zaki D. Husain and Donald J. Wass

54 Turbine and Vane Flowmeters **54**-1
David Wadlow

55 Impeller Flowmeters **55**-1
Harold M. Miller

56 Electromagnetic Flowmeters **56**-1
Halit Eren

57 Ultrasonic Flowmeters **57**-1
Hans-Peter Vaterlaus, Thomas Hossle, Paolo Giordano, and Christophe Bruttin

58 Vortex-Shedding Flowmeters **58**-1
Wade M. Mattar and James H. Vignos

59 Thermal Anemometry **59**-1
Jugal K. Agarwal and John G. Olin

60 Coriolis Effect Mass Flowmeters **60**-1
Jesse Yoder

61 Drag Force Flowmeters **61**-1
Rekha Philip-Chandy, Roger Morgan, and Patricia J. Scully

62 Pitot Probe Anemometer **62**-1
John A. Kleppe

63 Thermal Dispersion Mass Flow Meters **63**-1
John G. Olin

64 Laser Anemometry **64**-1
Rajan K. Menon

PART VII Thermal and Temperature Measurement

65 Temperature Measurements, Scales, and Calibrations **65**-1
Franco Pavese

66 Thermal Conductivity Measurement **66**-1
William A. Wakeham and Marc J. Assael

67 Heat Flux **67**-1
Thomas E. Diller

68 Resistive Thermometers **68**-1
Jim Burns

69 Thermistor Thermometers **69**-1
Rod White and Meyer Sapoff

70 Thermocouple Thermometry **70**-1
R.P. Reed

71 Semiconductor Junction Thermometers **71**-1
Randy Frank

72 Noncontact Thermometers **72**-1
Jacob Fraden

73 Pyroelectric Detectors **73**-1
Jacob Fraden

74 Liquid-in-Glass Thermometers **74**-1
Rod White and J.V. Nicholas

75 Manometric Thermometers **75**-1
Franco Pavese

76 Temperature Indicators **76**-1
Jan Stasiek, Tolestyn Madaj, and Jaroslaw Mikielewicz

77 Fiber-Optic Thermometers **77**-1
Brian Culshaw

78 Thermal Imaging **78**-1
Herbert M. Runciman

79 Calorimetry Measurement **79**-1
Sander van Herwaarden and Elina Iervolino

PART VIII Radiation

80 Radioactivity Measurement **80**-1
Bert M. Coursey

81 Radioactivity Detectors **81**-1
Larry A. Franks, Ralph B. James, and Larry S. Darken

82 Charged-Particle Measurement **82**-1
John C. Armitage, Madhu S. Dixit, Jacques Dubeau, Hans Mes, and F. Gerald Oakham

83 Neutron Measurement **83**-1
Steven M. Grimes

84 Dosimetry Measurement **84**-1
Brian L. Justus, Mark A. Miller, and Alan L. Huston

PART IX Wireless Instrumentation

85 Wireless Instrumentation **85**-1
J.P. Carmo and J.H. Correria

86 Wireless Sensor Node Hardware **86**-1
Michael Healy, Thomas Newe, and Elfed Lewis

87 Mobile Instrumentation with Wireless Design and Implementation **87**-1
Frederick Fortson and Kenneth Johnson

88 Powering Autonomous Sensors **88**-1
Manel Gasulla, Maria Teresa Penella, and Oscar Lopez-Lapeña

89 Wireless Sensing Technology **89**-1
Gregory C. Willden, Ben A. Abbott, and Ronald T. Green

90 Telemetry **90**-1
Albert Lozano-Nieto

PART X Control and Human Factors

91 PID Control **91**-1
F. Greg Shinskey

92 Optimal Control and the Software **92**-1
Halit Eren

93 Electropneumatic and Electrohydraulic Instruments: Modeling **93**-1
M. Pachter and C.H. Houpis

94 Explosion-Proof Instruments **94**-1
Sam S. Khalilieh

95 Measurement and Identification of DC Brush and Brushless Stepping Motors **95**-1
Stuart Schweid, Robert Lofthus, and John McInroy

96 Human Factors in Displays **96**-1
Jeffrey D. Onken, Barrett S. Caldwell, and Steven A. Murray

Appendix: Units and Conversions **Appendix**-1
B.W. Petley

Index **Index**-1



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Preface

Introduction

The second edition of *The Measurement, Instrumentation, and Sensors Handbook* comes in two volumes. This edition reflects the current state of the art in measurement, instrumentation, and sensors. In this two-volume set, all chapters have been updated, and 40 new chapters have been included to provide the finest possible reference that is both concise and useful for engineers practicing in industry, scientists and engineers involved in R&D, designers, college and university personnel and students, as well as managers, together with many others who are involved in instrumentation and measurement design and applications.

The handbook covers an extensive range of topics that comprise the subject of measurement, instrumentation, and sensors. It describes the use of instruments and techniques for practical measurements required in engineering, physics, chemistry, environmental science, and the life sciences. It also explains sensors, techniques, hardware, and the associated software. The handbook includes information processing systems, automatic data acquisition, reduction and analysis, operation characteristics, accuracy, errors, calibrations, standards, and their incorporation for control purposes. Emphasis is given on modern intelligent instruments and techniques, wireless network operations, human factors, and modern display methods, as well as virtual instruments.

The chapters include descriptive information for professionals, students, and workers interested in measurement. They include equations to assist engineers and scientists who seek to discover applications and solve problems that arise in fields not in their specialty. They also include specialized information needed by informed specialists who seek to learn advanced applications of the subject, evaluative opinions, and possible areas for future study. Thus, the handbook serves the reference needs of the broadest group of users—from the advanced high-school science student to industrial and university professionals.

Organization

In this edition, the first volume has 10 parts, each having several chapters, for a total of 96 chapters written by experts in their areas. It concentrates on concepts in instrumentation and measurements, spatial variable measurement, displacement measurement, mechanical variable measurement, acoustics, flow and spot velocity, thermal and temperature measurement, and radiation. It reflects recent trends in instrumentation and measurements with the addition of a new part on wireless instrumentation. Concepts in control systems and human factors are given as a separate part.

The second volume has 10 parts, each having several chapters, for a total of 96 chapters written by experts in their areas as in volume 1. It concentrates on sensors and sensor technology, electric variable measurement, electromagnetic variables, time and frequency, optical measurement, chemical variables,

medical, biomedical and health, and environmental measurement. Signal processing, and displays and recorders constitute the last two parts of this volume.

Locating Your Topic

To find out how to measure a given variable, skim the table of contents, turn to that section, and find the chapters that describe different methods of making the measurement. Consider the alternative methods of making the measurement and each of their advantages and disadvantages. Select a method, sensor, and signal processing method. Many chapters list a number of vendors to contact for more information. You can also visit http://www.globalspec.com/ to obtain a list of vendors.

For more detailed information, consult the index, since certain principles of measurement may appear in more than one chapter.

MATLAB® is a registered trademark of The MathWorks, Inc. For product information, please contact:

The MathWorks, Inc.
3 Apple Hill Drive
Natick, MA 01760-2098, USA
Tel.: 508-647-7000
Fax: 508-647-7001
E-mail: info@mathworks.com
Web: www.mathworks.com

John G. Webster and Halit Eren
Co-Editors

Acknowledgments

We would like to thank all the authors for their valuable contribution toward this two-volume set book. We appreciate the time and effort devoted by all our new authors and those authors who went an extra mile to revise and update their chapters. We are grateful to the CRC Press team for their encouragement to prepare this second edition. The publication of this book would not have been possible without their tireless dedication in putting it together. Last but not least, we would like to thank all our readers in selecting this book for advancing their knowledge and technical skills.

John G. Webster and Halit Eren
Co-Editors



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Editors

John G. Webster received his BEE from Cornell University, Ithaca, New York, in 1953, and his MSEE and PhD from the University of Rochester, Rochester, New York, in 1965 and 1967, respectively.

He is professor emeritus of biomedical engineering at the University of Wisconsin–Madison. He is a highly cited researcher at King Abdulaziz University, Jeddah, Saudi Arabia. In the field of medical instrumentation, he teaches undergraduate and graduate courses and does research on intracranial pressure monitors, ECG dry electrodes, and tactile vibrators.

Dr. Webster is the author of *Transducers and Sensors, an IEEE/EAB Individual Learning Program* (Piscataway, NJ: IEEE, 1989). He is the coauthor, with B. Jacobson, of *Medicine and Clinical Engineering* (Englewood Cliffs, NJ: Prentice-Hall, 1977) and, with R. Pallas-Areny, of *Sensors and Signal Conditioning*, Second Edition (New York: Wiley, 2001), and *Analog Signal Conditioning* (New York: Wiley, 1999). He is the editor of *Encyclopedia of Medical Devices and Instrumentation*, Second Edition (New York: Wiley, 2006), *Tactile Sensors for Robotics and Medicine* (New York: Wiley, 1988), *Electrical Impedance Tomography* (Bristol, U.K.: Adam Hilger, 1990), *Teaching Design in Electrical Engineering* (Piscataway, NJ: Educational Activities Board, IEEE, 1990), *Prevention of Pressure Sores: Engineering and Clinical Aspects* (Bristol, U.K.: Adam Hilger, 1991), *Design of Cardiac Pacemakers* (Piscataway, NJ: IEEE Press, 1995), *Design of Pulse Oximeters* (Bristol, U.K.: IOP Publishing, 1997), *Medical Instrumentation: Application and Design*, Fourth Edition (Hoboken NJ: Wiley, 2010), *Encyclopedia of Electrical and Electronics Engineering* (New York, Wiley, 1999), *Minimally Invasive Medical Technology* (Bristol, U.K.: IOP Publishing, 2001), and *Bioinstrumentation* (Hoboken NJ: Wiley, 2004). He is the coeditor, with A. M. Cook, of *Clinical Engineering: Principles and Practices* (Englewood Cliffs, NJ: Prentice-Hall, 1979) and *Therapeutic Medical Devices: Application and Design* (Englewood Cliffs, NJ: Prentice-Hall, 1982), with W. J. Tompkins of *Design of Microcomputer-Based Medical Instrumentation* (Englewood Cliffs, NJ: Prentice-Hall, 1981) and *Interfacing Sensors to the IBM PC* (Englewood Cliffs, NJ: Prentice Hall, 1988), and, with A. M. Cook, W J. Tompkins, and G. C. Vanderheiden, of *Electronic Devices for Rehabilitation* (London, U.K.: Chapman & Hall, 1985).

Dr. Webster has been a member of the IEEE-EMBS Administrative Committee and the NIH Surgery and Bioengineering Study Section. He is a fellow of the Institute of Electrical and Electronics Engineers, the Instrument Society of America, the American Institute of Medical and Biological Engineering, the Biomedical Engineering Society, and the Institute of Physics. He is the recipient of the IEEE EMBS Career Achievement Award.

Halit Eren received his BEng in 1973, MEng in 1975, and PhD in 1978 from the University of Sheffield, United Kingdom. He obtained an MBA from Curtin University in 1999.

After his graduation, Dr. Eren worked in Etibank (a mining and metallurgy company in Turkey) as an instrumentation engineer for two years. He was an assistant professor at Hacettepe University in 1980–1981 and Middle East Technical University in 1982. He has been at Curtin University since

1983, researching and teaching primarily in the areas of control systems, instrumentation, and engineering management.

Dr. Eren was appointed as visiting associate professor at the Polytechnic University in Hong Kong in 2004 and is currently a visiting professor at the University of Wisconsin, USA. He is a senior member of IEEE, taking roles in Region 10 activities, various committees for organizing conferences, and as a member of editorship in transactions. Dr. Eren has over 180 publications in conference proceedings, books, and transactions. He is the author of *Electronic Portable Instruments—Design and Applications* (Boca Raton, FL: CRC Press, 2004) and *Wireless Sensors and Instruments—Networks, Design and Applications* (Boca Raton, FL: CRC Press, 2006). He has coedited, with Bela Liptak, *Instruments Engineers' Handbook—Process Software and Digital Networks*, Vol. 3, edn. 4 (Boca Raton, FL: CRC Press, 2011). He is involved in writing a number of books in the field of instrumentation and measurement. Dr. Eren is active in researching and publishing on intelligent sensors, wireless instrumentation, wireless sensor networks, automation and control systems, and very large control systems.

Contributors

René G. Aarnink
University Hospital Nijmegen
Nijmegen, the Netherlands

Ben A. Abbott
Southwest Research Institute
San Antonio, Texas

Jugal K. Agarwal
TSI, Inc.
Shoreview, Minnesota

Keith Antonelli
Kinetic Sciences, Inc.
Vancouver, British Colombia, Canada

John C. Armitage
Ottawa–Carleton Institute for Physics
Carleton University
Ottawa, Ontario, Canada

Marc J. Assael
Faculty of Chemical Engineering
Department of Chemical Engineering
Aristotle University of Thessaloniki
Thessaloniki, Greece

Viktor P. Astakhov
Astakhov Tool Service
Rochester Hills, Michigan

Manouchehr Bahrami
University of Tabriz
Tabriz, Iran

William E. Baker
Department of Mechanical Engineering
University of New Mexico
Albuquerque, New Mexico

W. John Ballantyne
Department of Systems Engineering
Spar Aerospace Ltd.
Brampton, Ontario, Canada

Amit Bandyopadhyay
Department of Ceramic Science and Engineering
Rutgers University
Piscataway, New Jersey

Partha P. Banerjee
Department of Electrical and Computer Engineering
University of Alabama at Huntsville
Huntsville, Alabama

William H. Bayles, Jr.
The Fredericks Company
Huntington Valley, Pennsylvania

Ben Benhabib
Department of Mechanical and Industrial Engineering
University of Toronto
Toronto, Ontario, Canada

Vikram Bhatia
Virginia Tech
Blacksburg, Virginia

A. Bonen
University of Toronto
Toronto, Ontario, Canada

Howard M. Brady
The Fredericks Company
Huntington Valley, Pennsylvania

John C. Brasunas
Goddard Space Flight Center
National Aeronautics and Space Administration
Greenbelt, Maryland

Detlef Brumbi
Krohne Messtechnik GmbH
Duisburg, Germany

Christophe Bruttin
Rittmeyer Ltd.
Zug, Switzerland

Wolfgang P. Buerner
Los Angeles Scientific Instrumentation Co.
Los Angeles, California

Jim Burns
Burns Engineering, Inc.
Minnetonka, Minnesota

Barrett S. Caldwell
Purdue University
West Lafayette, Indiana

J.P. Carmo
Department of Industrial Electronics
University of Minho
Guimarães, Portugal

Kevin H.L. Chau
Micromachined Products Division
Analog Devices, Inc.
Cambridge, Massachusetts

Adam Chrzanowski
University of New Brunswick
Fredericton, New Brunswick, Canada

Richard O. Claus
Bradley Department of Electrical Engineering
Virginia Tech
Blacksburg, Virginia

J.H. Correria
Department of Industrial Electronics
University of Minho
Guimarães, Portugal

Charles B. Coulbourn
Los Angeles Scientific Instrumentation Co.
Los Angeles, California

Bert M. Coursey
Ionizing Radiation Division
Physics Laboratory
National Institute of Standards and Technology
Gaithersburg, Maryland

Brian Culshaw
Department of Electronic and Electrical Engineering
University of Strathclyde
Glasgow, England

G. Mark Cushman
Goddard Space Flight Center
National Aeronautics and Space Administration
Greenbelt, Maryland

Larry S. Darken
Oxford Instruments, Inc.
Oak Ridge, Tennessee

David Dayton
ILC Data Device Corp.
Bohemia, New York

J.R. deBruyn
University of Western Ontario
DeKalb, Illinois

Ronald H. Dieck
Pratt & Whitney
and
Ron Dieck Associates, Inc.
Palm Beach Gardens, Florida

Thomas E. Diller
Virginia Tech
Blacksburg, Virginia

Madhu S. Dixit
Centre for Research in Particle Physics
Carleton University
Ottawa, Ontario, Canada

Emil Drubetsky
The Fredericks Company
Huntington Valley, Pennsylvania

Jacques Dubeau
Centre for Research in Particle Physics
Carleton University
Ottawa, Ontario, Canada

M.A. Elbestawi
Department of Mechanical Engineering
McMaster University
Hamilton, Ontario, Canada

Halit Eren
Department of Electrical and Computer Engineering
Curtin University
Perth, Western Australia, Australia

Giuseppe Fazio
Department of Electronic Engineering
University of Rome "Tor Vergata"
Rome, Italy

and

Isatel Ltd.
Moscow, Russia

Richard S. Figliola
Department of Mechanical Engineering
Clemson University
Clemson, South Carolina

Frederick Fortson
Solidica, Inc.
Ann Arbor, Michigan

Jacob Fraden
Fraden Corporation
San Diego, California

Randy Frank
Randy Frank & Associates, Ltd
Scottsdale, Arizona

and

Motorola, Inc.
Phoenix, Arizona

Larry A. Franks
Sandia National Laboratories
Livermore, California

Yusaku Fujii
Department of Electronic Engineering
Gunma University
Kiryu, Japan

Ivan J. Garshelis
Magnova, Inc.
Pittsfield, Massachusetts

Manel Gasulla
Universitat Politècnica de Catalunya
Barcelona, Catalonia, Spain

Mauro Giaconi
Department of Electronic Engineering
University of Rome "Tor Vergata"
Rome, Italy

Paolo Giordano
Rittmeyer Ltd.
Zug, Switzerland

Ron Goehner
The Fredericks Company
Huntington Valley, Pennsylvania

Ronald T. Green
Southwest Research Institute
San Antonio, Texas

Steven M. Grimes
Department of Physics and Astronomy
Ohio University
Athens, Ohio

Reinhard Haak
University of Erlangen–Nuernberg
Erlangen, Germany

John R. Hansman, Jr.
Department of Aeronautics and Astronautics
Massachusetts Institute of Technology
Cambridge, Massachusetts

Emil Hazarian
Master of Science Quality Assurance Program
and
Bachelor of Science Quality Assurance Program
California State University, Dominguez Hills
Carson, California

and

ANSI-ASQ National Accreditation Board
ACLASS
Alexandria, Virginia

Michael Healy
University of Limerick
Limerick, Ireland

Sander van Herwaarden
Xensor Integration
Delft, the Netherlands

Thomas Hossle
Rittmeyer Ltd.
Zug, Switzerland

C.H. Houpis
Air Force Institute of Technology
Wright–Patterson Air Force Base, Ohio

Zaki D. Husain
Daniel Flow Products, Inc.
Bellaire, Texas

Alan L. Huston
Naval Research Laboratory
Washington, District of Columbia

Robert M. Hyatt, Jr.
Howell Electric Motors
Plainfield, New Jersey

Elina Iervolino
Xensor Integration

Ralph B. James
Sandia National Laboratories
Livermore, California

Victor F. Janas
Center for Ceramic Research
Rutgers University
Piscataway, New Jersey

Kenneth Johnson
Solidica, Inc.
Ann Arbor, Michigan

Brian L. Justus
Department of Optical Science
Naval Research Laboratory
Washington, District of Columbia

Nils Karlsson
Department of Sensor and EW Systems
Swedish Defence Research Agency
Linköping, Sweden

Sam S. Khalilieh
Department of Electrical Engineering
Earth Tech
Grand Rapids, Michigan

Andrei Kholkine
Rutgers University
Piscataway, New Jersey

John A. Kleppe
Department of Electrical and Biomedical Engineering
University of Nevada, Reno
Reno, Nevada

James Ko
Kinetic Sciences, Inc.
Vancouver, British Columbia, Canada

Herbert Köchner
University of Erlangen–Nuernberg
Erlangen, Germany

M. Kostic
Northern Illinois University
DeKalb, Illinois

Shyan Ku
Kinetic Sciences, Inc.
Vancouver, British Columbia, Canada

Brook Lakew
Goddard Space Flight Center
National Aeronautics and Space Administration
Greenbelt, Maryland

Elfed Lewis
University of Limerick
Limerick, Ireland

Robert Lofthus
Xerox Corporation
Rochester, New York

Oscar Lopez-Lapeña
Universitat Politècnica de Catalunya
Barcelona, Catalonia, Spain

Albert Lozano-Nieto
Penn State University
Lehman, Pennsylvania

Christopher S. Lynch
Department of Mechanical Engineering
Georgia Tech
Atlanta, Georgia

Tolestyn Madaj
Technical University of Gdansk
Gdansk, Poland

Kin F. Man
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Dimitris E. Manolakis
Alexander Technological Education Institute
of Thessaloniki
Thessaloniki, Greece

Wade M. Mattar
The Foxboro Company
Foxboro, Massachusetts

J.R. René Mayer
Department of Mechanical Engineering
Ecole Polytechnique de Montreal
Montreal, Quebec, Canada

John McInroy
Department of Electrical Engineering
University of Wyoming
Laramie, Wyoming

Adrian Melling
University of Erlangen–Nuernberg
Erlangen, Germany

Rajan K. Menon
Laser Velocimetry Products
TSI Inc.
St. Paul, Minnesota

Hans Mes
Centre for Research in Particle Physics
Carleton University
Ottawa, Ontario, Canada

Jaroslaw Mikielewicz
Institute of Fluid Flow Machinery
Gdansk, Poland

Harold M. Miller
Data Industrial Corporation
Mattapoisett, Massachusetts

Mark A. Miller
Naval Research Laboratory
Washington, District of Columbia

Roger Morgan
School of Engineering
Liverpool John Moores University
Liverpool, England

Steven A. Murray
University of San Diego
San Diego, California

Thomas Newe
University of Limerick
Limerick, Ireland

J.V. Nicholas
Temperature Standards Section
Measurement Standards Laboratory of
New Zealand
Lower Hutt, New Zealand

Seiji Nishifuji
Department of Electrical and Electronic
Engineering
Yamaguchi University
Ube, Japan

David S. Nyce
Revolution Sensor Company
Cary, North Carolina

F. Gerald Oakham
Centre for Research in Particle Physics
Carleton University
Ottawa, Ontario, Canada

John G. Olin
Sierra Instruments, Inc.
Monterey, California

Jeffrey D. Onken
Purdue University
West Lafayette, Indiana

M. Pachter
Air Force Institute of Technology
Wright–Patterson Air Force Base, Ohio

Reza Pakdaman Zangabad
Faculty of Engineering and Natural Sciences
Sabanci University
Istanbul, Turkey

Franco Pavese
Thermodynamic Division
National Institute of Meterology Research
Turin, Italy

Ole Pedersen
Department of Science and Technology
Linköping University
Linköping, Sweden

Peder C. Pedersen
Department of Electrical and Computer Engineering
Worcester Polytechnic Institute
Worcester, Massachusetts

Maria Teresa Penella
Urbiotica S.L.
Barcelona, Catalonia, Spain

B.W. Petley
Centre for Basic, Thermal and Length Metrology
National Physical Laboratory
London, England

Rekha Philip-Chandy
School of Engineering
Liverpool John Moores University
Liverpool, England

Charles P. Pinney
Pinney Technologies, Inc.
Albuquerque, New Mexico

David Potter
National Instruments
Austin, Texas

Davide Quatrini
Italcertifer, Inc.
and
Department of Electronic Engineering
University of Rome "Tor Vergata"
Rome, Italy

Per Rasmussen
G.R.A.S. Sound and Vibration
Vedback, Denmark

R.P. Reed
Proteun Services
Albuquerque, New Mexico

Herbert M. Runciman
Pilkington Optronics
Scotland, United Kingdom

Ricardo E. Saad
Harmonics Lightwaves
Sunnyvale, California

Ahmad Safari
Department of Ceramic Science and Engineering
Rutgers University
Piscataway, New Jersey

Adelio Salsano
Italcertifer, Inc.
and
Department of Electronic Engineering
University of Rome "Tor Vergata"
Rome, Italy

Robert J. Sandberg
Department of Mechanical Engineering
University of Wisconsin–Madison
Madison, Wisconsin

Meyer Sapoff
MS Consultants
Princeton, New Jersey

Stuart Schweid
Xerox Corporation
Rochester, New York

Patricia J. Scully
School of Engineering
Liverpool John Moores University
Liverpool, England

R.A. Secco
University of Western Ontario
London, Ontario, Canada

James M. Secord
Department of Geodesy and Geomatics Engineering
University of New Brunswick
Fredericton, New Brunswick, Canada

DeWayne B. Sharp
Shape of Things
San Luis Obispo, California

F. Greg Shinskey
Process Control Consultant
North Sandwich, New Hampshire

K.C. Smith
University of Toronto
Toronto, Ontario, Canada

Jan Stasiek
Department of Mechanical Engineering
Technical University of Gdansk
Gdansk, Poland

Mark A. Stedham
Department of Electrical and Computer Engineering
University of Alabama at Huntsville
Huntsville, Alabama

Peter H. Sydenham
University of South Australia
Mawsons Lakes, South Australia, Australia

Akihiro Takita
Department of Electronic Engineering
Gunma University
Kiryu, Japan

Shogo Tanaka
Department of Electrical and Electronic Engineering
Yamaguchi University
Ube, Japan

Jin Tao
Department of Electronic Engineering
Gunma University
Kiryu, Japan

David B. Thiessen
California Institute of Technology
Pasadena, California

Richard Thorn
School of Engineering
University of Derby
Derby, United Kingdom

Hans-Peter Vaterlaus
Department of Instrument
Rittmeyer Ltd.
Zug, Switzerland

James H. Vignos
The Foxboro Company
Foxboro, Massachusetts

David Wadlow
Consultant
Basking Ridge, New Jersey

William A. Wakeham
Department of Chemical Engineering
Imperial College, London
London, England

Anbo Wang
Bradley Department of Electrical Engineering
Virginia Tech
Blacksburg, Virginia

Donald J. Wass
Daniel Flow Products, Inc.
Houston, Texas

Grover C. Wetsel
Erik Jonsson School of Engineering and Computer Science
University of Texas at Dallas
Richardson, Texas

Rod White
Temperature Standards Section
Callaghan Innovation
Measurement Standards Laboratory of New Zealand
Lower Hutt, New Zealand

Hessel Wijkstra
University Hospital Nijmegen
Nijmegen, the Netherlands

Gregory C. Willden
Southwest Research Institute
San Antonio, Texas

Jesse Yoder
Flow Research, Inc.
Wakefield, Massachusetts

and

Automation Research Corporation
Dedham, Massachusetts

Bernhard Günther Zagar
Department of Electrical Engineering
Technical University of Graz
Graz, Austria

I

Instrumentation and Measurement Concepts

1 Measurements, Instrumentation, and Sensors *Halit Eren*.....**1**-1
Introduction • Measurements • 1.3 Instruments • Analog and Digital Instruments • Sensors and Transducers • Instrumentation and Networks • Software Support for Sensors and Instruments • Application Examples • Bibliography • Partial List of Vendors and Suppliers

2 Characteristics of Instrumentation *John R. Hansman, Jr.* **2**-1
Simple Instrument Model • Bibliography

3 Operational Modes of Instrumentation *Richard S. Figliola*..... **3**-1
Null Instrument • Deflection Instrument • Analog and Digital Sensors • Analog and Digital Readout Instruments • Input Impedance • Defining Terms • Further Information

4 Static and Dynamic Characteristics of Instrumentation *Peter H. Sydenham* **4**-1
Static Characteristics of Instrument Systems • Dynamic Characteristics of Instrument Systems • Calibration of Measurements • References

5 Measurement Accuracy *Ronald H. Dieck*..... **5**-1
Error: The Normal Distribution and the Uniform Distribution • Measurement Uncertainty Model • Calculation of Expanded Uncertainty • Summary • Defining Terms • References • Further Information

6 Development of Standards *Halit Eren*..... **6**-1
Introduction • Standard Organizations • Development of Standards • Obtaining and Using Standards • Examples of Standards • Software Standards • Bibliography

7 Measurement Standards *DeWayne B. Sharp* **7**-1
Historical Perspective • What Are Standards? • Conceptual Basis of Measurements • Need for Standards • Types of Standards • Numbers, Dimensions, and Units • Multiplication Factors • Conversions of Units • Examples of Defining Terms • Bibliography

8 Calibrations in Instrumentation and Measurements *Halit Eren* **8**-1
Introduction • Errors and Uncertainties in Calibrations • Benefits of Calibrations • Calibration Procedure and Personnel • Calibration Methods • Laboratories and Institutions • Calibration Software Support • Cost of Calibrations • Trends in Calibrations • Calibration Examples • Bibliography • Partial List of Calibration Service and Software Providers

9 Intelligent Sensors and Instruments *Halit Eren* **9**-1
Introduction • Intelligent Sensors • The IEEE 1451 Standard • Communications of Intelligent Sensors • Plug-and-Play Capability • Applications and Examples of Intelligent Sensors • Intelligent Instruments • Examples of Intelligent Instruments • References • Bibliography • Partial List of Manufacturers/Suppliers

10 Virtual Instruments *David Potter and Halit Eren* **10**-1
Introduction • Virtual Instrument Architecture • Virtual Instrument Software • Supervisory Control and Data Acquisition • Conclusions • Bibliography • Partial List of Suppliers

11 Fail-Safe Instruments and Devices *Davide Quatrini, Giuseppe Fazio, Mauro Giaconi, and Adelio Salsano* **11**-1
Introduction • Fail-Safe Realization: Relevant Features • Fail-Safe Realization: In-Depth Example • Case Study: Level 1 ERTMS • Summary • References

12 Dynamic Error Measurements of Force Sensors *Akihiro Takita, Jin Tao, and Yusaku Fujii* **12**-1
Introduction • Problems and Solutions of Dynamic Calibrations • Evaluation of Dynamic Characteristics of Force Sensors • Impulse Response • Oscillation Response • Conclusions • References

1

Measurements, Instrumentation, and Sensors

1.1 Introduction **1**-1
1.2 Measurements **1**-2
1.3 Instruments **1**-2
Design of Instruments • Testing and Use of Instruments • Response and Drift • Accuracy and Errors • Error Reduction • Calibration of Instruments
1.4 Analog and Digital Instruments **1**-6
Analog Instruments • Digital Instruments
1.5 Sensors and Transducers **1**-7
Smart Sensors • Wireless and Autonomous Sensors and Instruments
1.6 Instrumentation and Networks **1**-10
1.7 Software Support for Sensors and Instruments **1**-11
1.8 Application Examples **1**-12
Bibliography **1**-13
Partial List of Vendors and Suppliers **1**-14

Halit Eren
Curtin University

1.1 Introduction

Measurement is a process of gathering information from a physical world and comparing this information with agreed standards. As highlighted in this chapter and discussed in detail in this book, measurements are essential activities for observing and testing scientific and technological investigations.

Measurements are carried out by using instruments, which are designed and manufactured to fulfill specific tasks. Sensors are used as the primary elements in instruments to respond to the physical variable under investigation. In this book, a diverse range of sensors and instruments are discussed; the advances and the recent developments in measurements, instrumentation, and sensors are introduced in the proceeding sections and chapters.

In many applications, many sensors and instruments are used to collect information about the process under investigation. These instruments are connected together using wired, optical, or wireless networks. The details of the supporting hardware and software technologies behind these networks are discussed extensively.

1.2 Measurements

If the behavior of the physical variable is known, its performance can be monitored and assessed by means of suitable methods of sensing, signal conditioning, and termination. The applications of instruments range from laboratory conditions to arduous environments such as inside nuclear reactors or remote locations such as satellite systems and spaceships.

The sensor-output signals of the majority of modern instruments are in electric form. This is due to electric signals being easy to process, display, store, and transmit. Once converted to electric forms, the relation between the sensor signals and the physical variations can be expressed in the form of *transfer functions*. The transfer function is a mathematical model between the sensor signal and the physical variable. In a continuous system, the transfer function may be linear or nonlinear. A linear relationship may be expressed by the following equation:

$$y = a + bx \tag{1.1}$$

where

y is the electric signal
x is the physical stimulus
a is the intercept on the y-axis, which gives the output signal for a zero input
b is the slope that is also known as the sensitivity

In ideal cases, this relationship should be consistent without any errors, nonlinearity, or deviation.

However, in measurements, there may be many sources of errors; therefore, it is important to identify these sources and draw up an error budget by considering factors, such as

- Imperfections in electric and mechanical components (e.g., high tolerances and noise or offset voltages)
- Changes in component performances (e.g., shift in gains, changes in chemistry, aging, and drifts in offsets)
- External and ambient influences (e.g., temperature, pressure, and humidity)
- Inherent physical fundamental laws (e.g., thermal and other electric noises, Brownian motion in materials, and radiation)

Once the measurements are made, various forms of data analysis may be used to extract useful information about the measurements and draw conclusions. Data analysis may include statistical methods, curve fitting, selecting or discarding subsets of data, or many other techniques. For example, a typical data analysis tool is the data mining, which aims to discover unforeseen patterns hidden in the data. There is an extensive range of software available for the analysis (e.g., MATLAB®) to suit specific needs of information obtained from experimental or test results. Further information on measurements and analysis can be found throughout this book.

1.3 Instruments

Instruments are man-made devices for determining the value of the quantity/variable. They are designed to maintain prescribed relationships between the parameters being measured and the physical variables under investigation. Instruments can be designed and constructed to be analog, digital, or hybrid.

The construction of an instrument can be broken into smaller elements, as illustrated in Figure 1.1. Typically, an instrument will have a sensor or transducer stage, a signal-conditioning stage, and an output or termination stage. All instruments have some or all of these functional blocks.

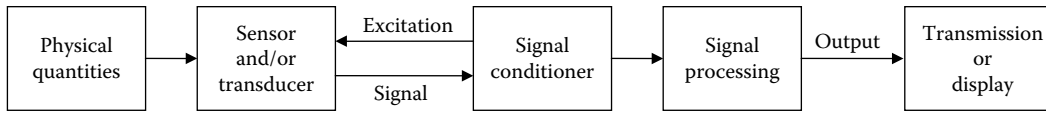


FIGURE 1.1 Construction structure of a typical instrument.

A diverse range of sensors and transducers may be available to meet the measurement requirements of a physical system. The sensors can be categorized in a number of ways depending on the energy input and output, input variables, sensing elements, and electric or physical principles.

In recent years, the rapid growth of the integrated circuit (IC) electronics and the availability of cost-effective processors have led to an impressive progress in instrumentation and measurements in all fields. This coupled with the improvement of mathematical methods, the extensive applications of digital techniques, and the addition of new application areas enabled modern instruments to excel to new heights as discussed in detail in this two-volume *Measurements, Instrumentation, and Sensors Handbook*.

1.3.1 Design of Instruments

Instruments are designed on the basis of existing knowledge, which is gained either from the experiences of people about the physical process or from structured understanding of the process. In any case, ideas conceived about an instrument are translated into hardware and software that can perform well within the expected standards and easily be accepted by the end users.

Usually, the design of instruments requires many multidisciplinary activities. In the wake of a rapidly changing technology, instruments are upgraded often to meet the demands of the marketplace. Depending on the complexity of the proposed instrument, it may take many years to produce an instrument for a relatively short commercial lifetime. In the design and production stages, engineers must consider factors such as simplicity, appearance, ease and flexibility of use, maintenance requirements, production costs, lead time to product, and positioning strategy in the marketplace.

The design process of an instrument may follow well-ordered procedures from ideas to marketing of the final products. The process may be broken down into smaller tasks such as identifying specifications, developing possible solutions for these specifications, modeling, prototyping, installing and testing, making modifications, manufacturing, planning marketing and distribution, evaluating customer feedback, and making design and technological improvements. For example, many different specifications may be considered for particular product, which may include but are not limited to operational requirements, functionality, technological limitations, quality, installation, maintenance, documentation, servicing, and acceptance level usage by customers.

1.3.2 Testing and Use of Instruments

After the instrument is designed and prototyped, various evaluation tests may be conducted. These tests may be made under reference conditions or under simulated environmental conditions. Some examples of reference condition tests are accuracy, response time, drift, and warm-up time. Simulated environmental tests may be compulsory, being regulated by governments and other authorities. Some simulated environment tests include climatic test, drop test, dust test, insulation-resistance test, vibration test, electromagnetic compatibility tests, and safety and health hazard tests. Many of these are strictly regulated by national and international standards.

For maximum efficiency, an appropriate instrument for the measurement must be selected. Users should be fully aware of their application requirements, since instruments that do not fit their purposes will deliver false data resulting in wasted time and effort. For a particular application, users must

carefully study the documents about all the candidates and make comparisons among all options. While selecting the instrument, users must evaluate many factors such as accuracy, frequency response, electric and physical loading effects, sensitivity, response time, calibration intervals, power supply needs, spare parts, technology, and maintenance requirements. They must ensure compatibility with the existing equipment.

When selecting and implementing of instruments, quality becomes an important issue from both quantitative and qualitative perspectives. The quality of an instrument may be viewed differently depending on the people involved. For example, quality in the eyes of designer may be an instrument designed on sound physical principles, whereas from the users' point of view, it may be reliability, maintainability, cost, and availability.

1.3.3 Response and Drift

Instruments respond to physical phenomenon by sensing and generating signals. Depending on the type of instrument used and the physical phenomenon, the signals may be either slow or fast to change and may also contain transients. The response of the instruments to the signals can be analyzed in a number of ways by establishing static and dynamic performance characteristics. Although the static performances are relatively simple, the dynamic performances may be complex. More information on this can be found in Chapters 4, 11, and 12.

1.3.4 Accuracy and Errors

The performance of an instrument depends on its static and dynamic characteristics. The performance may be indicated by its *accuracy*, which may be described as the closeness of measured values to the real values of the variable. The total response is a combination of dynamic and static responses. If the signals generated by the physical variable are changing rapidly, then the dynamic properties of the instrument become important. For slow-varying systems, the dynamic errors may be neglected. Further information on accuracy can be found in Chapter 5.

The performance of an instrument may also be decided by other factors, such as the magnitudes of errors; the *repeatability*, which indicates the closeness of sets of measurements made in the short term; and the *reproducibility* of the instrument. The reproducibility is the closeness of sets of measurements when repeated in similar conditions over a long period of time.

The ideal or perfect instrument would have perfect sensitivity, reliability, and repeatability without any spread of values and would be within the applicable standards. However, in many cases, there will be imprecise and inaccurate results because of internal and external factors. The departure from the expected perfection is called the *error*. Often, sensitivity analyses are conducted to evaluate the effect of individual components that are causing these errors. Sensitivity to the affecting parameter can be obtained by varying that one parameter and keeping the others constant. This can be done practically by using the developed instruments or mathematically by means of appropriate models.

When determining the performance of an instrument, it is essential to appreciate how errors arise. There may be many sources of errors; therefore, it is important to identify these sources and draw up an error budget. In the error budget, there may be many factors, such as (1) imperfections in electric and mechanical components (e.g., high tolerances and noise or offset voltages), (2) changes in component performances (e.g., shift in gains, changes in chemistry, aging, and drifts in offsets), (3) external and ambient influences (e.g., temperature, pressure, and humidity), and (4) inherent physical fundamental laws (e.g., thermal and other electric noises, Brownian motion in materials, and radiation).

In instrumentation systems, errors can be broadly classified as systematic, random, or gross errors. For further information, readers can refer to Chapters 2 and 4.

1.3.5 Error Reduction

Controlling errors is an essential part of measurements and instrumentation. Various techniques are available to achieve this objective. The error control begins in the design stages by choosing the appropriate components, filtering, and bandwidth selection; by reducing the noise; and by eliminating the errors generated by the individual subunits of the complete system. In a good design, the errors of the previous group may be compensated adequately by the following groups.

The accuracy of instruments can be increased by postmeasurement corrections. Various calibration methods may be employed to alter parameters slightly to give correct results. In many cases, calibration graphs, mathematical equations, tables, the experiences of the operators, and the like are used to reduce measurement errors. In recent years, with the application of digital techniques and intelligent instruments, error corrections are made automatically by the computers or the devices themselves. More information is available in Chapters 8 and 9.

In many instrumentation systems, the application of compensation strategy is used to increase static and dynamic performances. In the case of static characteristics, compensations can be made by many methods, including introducing opposing nonlinear elements into the system, using isolation and zero environmental sensitivity, opposing compensating environmental inputs, using differential systems, and employing feedback systems. On the other hand, dynamic compensation can be achieved by applying these techniques as well as by reducing harmonics, using filters, adjusting bandwidth, using feedback compensation techniques, and the like. Further information on dynamic error measurements can be found in Chapter 12.

1.3.6 Calibration of Instruments

The calibration of all instruments is essential for checking their performances against known standards. This provides consistency in readings and reduces errors, thus validating the measurements universally. After an instrument is calibrated, future operation is deemed to be error bound for a given period of time for similar operational conditions. The calibration procedure involves comparison of the instrument against primary or secondary standards. In some cases, it may be sufficient to calibrate a device against another one with a known accuracy.

Many nations and organizations maintain laboratories with the primary functions of calibrating instruments and field measuring systems that are used in everyday operations. Examples of these laboratories are National Association of Testing Authorities (NATA) of Australia and the British Calibration Services (BCS). Detailed information on calibration is available in Chapter 8.

Calibrations may be made under static or dynamic conditions. A typical calibration procedure of a complex process involving many instruments is illustrated in Figure 1.2. In an ideal situation, for

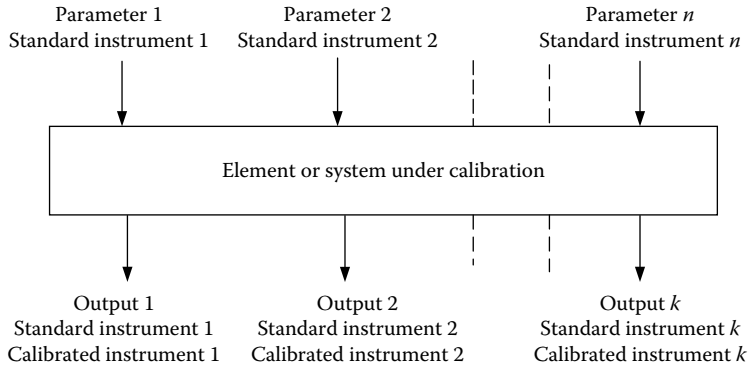


FIGURE 1.2 Instruments need to be frequently calibrated sequentially for all affecting inputs. Calibrations are made under static or dynamic conditions by varying a single input and observing the corresponding output while keeping all the other inputs constant until all inputs are covered.

an instrument that response to a multitude of physical variables, a commonly employed method is by keeping all the inputs constant except one. The input is varied in increments in increasing and decreasing directions over a specified range. The observed output then becomes a function of that single input. The calibration is continued in a similar manner until all other inputs are covered. For better results, this procedure may be repeated by varying the sequences of inputs, thus developing a family of relationships between the inputs and the outputs. As a result of these calibration readings, the input and output relation usually demonstrates statistical characteristics. From these characteristics, appropriate calibration curves can be obtained, and other statistical techniques can be applied.

1.4 Analog and Digital Instruments

Instruments can be analog or digital or a combination of the two. Nowadays, most instruments are produced to be digital because of the advantages they offer. However, the front end of majority of instruments is still analog; that is, most of the sensors and transducers generate analog signals. The signals initially are conditioned by analog circuits before converting into digital form for further signal processing. It is important to mention that nowadays, digital instruments operating purely on digital principles are being developed. For instance, today's smart sensors contain the complete signal condition circuits in a single chip integrated with the sensor itself. The output of smart sensors can be interfaced directly with other digital devices. More information on smart sensors can be found in Chapter 9.

1.4.1 Analog Instruments

Analog instruments are characterized by continuous signals. A purely analog system measures, transmits, displays, and stores data in analog form. The signal conditioning is usually made by integrating many functional blocks such as bridges, amplifiers, filters, oscillators, modulators, offsets and level converters, and buffers, as illustrated in Figure 1.3. Generally, in the initial stages, the signals produced by the sensors and transducers are conditioned mainly by analog electronics, even if they are configured as digital instruments later.

In analog instruments, the changes in amplitudes, phases, or frequencies or a combination of the three convey the useful information in response to physical variables. These signals can be deterministic or nondeterministic. As in the case with all signal-bearing systems, there are useful signals that respond to the physical phenomena and unwanted signal resulting from various forms of noise requiring extensive filtering and other signal processing as explained in the chapters in *Electrical, Optical, Chemical, and Biomedical Measurement*.

Analog signals can also be nondeterministic; that is, the future state of the signal cannot be determined. If the signal varies in a probabilistic manner, its future can be determined only by statistical methods. The mathematical and practical treatment of analog and digital signals, having foreseen, stochastic, and nondeterministic properties, is a very lengthy subject, and a vast body of information can be found in the literature; therefore, they will not be treated here.

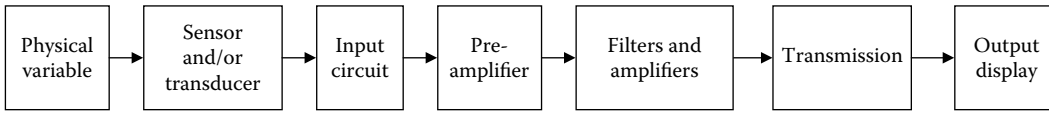


FIGURE 1.3 Analog instruments measure, transmit, display, and store data in analog form. The signal conditioning involves such components as bridges, amplifiers, filters, oscillators, modulators, offsets and level converters, buffers, and so on.

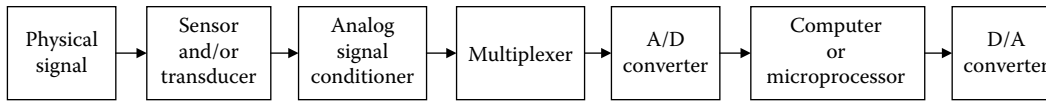


FIGURE 1.4 Digital instruments have more signal-processing components than analog counterparts. However, they have the advantage of data handling, storing, displaying, and transmitting.

1.4.2 Digital Instruments

In modern instruments, the original data acquired from the physical variables are usually in analog form. This analog signal is converted to digital before being passed onto the other parts of the system. For conversion purposes, analog-to-digital (A/D) converters are used together with appropriate sample-and-hold and multiplexing devices. The typical components of a digital instrument are illustrated in Figure 1.4. The digital systems are particularly useful in performing mathematical operations and storing and transmitting data.

A/D conversion involves three stages: sampling, quantization, and encoding. The Nyquist sampling theorem must be observed during sampling; that is, "the number of samples per second must be at least twice the highest frequency present in the continuous signal." As a rule of thumb, depending on the significance of the high frequencies, the sampling must be about 5 to 10 times the highest frequency of the signal. The next stage is the quantization, which determines the resolution of the sampled signals. The quantization error decreases as the number of bits increases. In the encoding stage, the quantized values are converted to binary numbers to be processed digitally. Once in digital form, the data can further be processed by employing various techniques such as FFT analysis, digital filtering, sequential or logical decision making, correlation methods, spectrum analysis, and more.

1.5 Sensors and Transducers

A sensor is a device that responds to a changing phenomenon. A transducer is a device that transfers energy from one form to another.

Sensors and transducers can be categorized in a number of ways depending on factors such as the energy input and output, input variables, sensing elements, and electric or physical principles. From the energy input and output point of view, there are three types: the modifiers, the self-generators, and the modulators.

In modifiers, a particular form of energy is modified rather than converted; therefore, the same form of energy exists in both the input and the output stages. In self-generators, electric signals are produced from nonelectric inputs without the application of external energy. Typical examples are piezoelectric transducers and photovoltaic cells. Modulators, on the other hand, produce electric outputs from nonelectric inputs, but they require an external source of energy. Strain gages are typical examples of such devices. Some examples of sensors and measurements are listed as follows:

- Spatial variables, discussed in Part II of this book
- Displacements, discussed in Part III of this book
- Mechanical variables, discussed in Part IV of this book
- Acoustics, discussed in Part V of this book
- Flow measurements and sensors, discussed in Part VI of this book
- Thermal and temperature, discussed in Part VII of this book
- Radiation, Part VIII of this book
- Electric variables, discussed in Part II of *Electrical, Optical, Chemical, and Biomedical Measurement*
- Electromagnetic variables discussed in Part III of *Electrical, Optical, Chemical, and Biomedical Measurement*

- Time and frequency, discussed in Part IV of *Electrical, Optical, Chemical, and Biomedical Measurement*
- Optical variables, Part V of *Electrical, Optical, Chemical, and Biomedical Measurement*
- Chemical variables, discussed in Part VI of *Electrical, Optical, Chemical, and Biomedical Measurement*
- Medical, biomedical, and health, Part VII of *Electrical, Optical, Chemical, and Biomedical Measurement*
- Environmental sensors and measurements, Part VIII of *Electrical, Optical, Chemical, and Biomedical Measurement*

The present trend in sensor technology has been shifted toward IC sensors in the form of microsystems, intelligent sensors, nanosensors, and others. The usefulness of semiconductor-based IC sensors is enhanced considerably by the integrations of microprocessors, microcontroller, converters, logic circuits, and other digital components in the same chip. Further, micromachining techniques combined with semiconductor processing technology provide a multiple range of sensors all integrated in the same chip for mechanical, optical, magnetic, chemical, biological, and other types of measurements.

IC devices refer to the dimensions of devices in micrometer (10^{-6} m) ranges, whereas nanotechnology refers to the dimensions of devices in nanometer (10^{-9} m) ranges. The microsystems technology (MST) is well established and simply known as the MST. A subset of MST is the *microelectromechanical systems* (MEMS). Another subset of MST is the *microelectro-optical systems* (MEOMs) and *systems-on-chip* (SOC) devices. Most of the sensors manufactured by MEMs and MEOMs are 3D devices with dimensions in the order of few micrometers.

For example, single-chip microsensors and microinstruments are being developed and used extensively for temperature, pressure, and radiation measurements, as well as mechanical, chemical, environmental, biomedical, biological variables, and implantable sensors, and many others. A typical example is illustrated in Figure 1.5, and detailed information on such devices can be found in Part I of *Electrical, Optical, Chemical, and Biomedical Measurement*. This particular single-chip implementation of microinstrumentation system is based on complementary metal–oxide–semiconductors (CMOS) and other technologies. It incorporates voltage, current, and capacitive-sensor interface; a temperature sensor; a 10-channel 12 bit A/D converter; and an 8 bit microcontroller with a 16 bit hardware multiplier and a 40 bit accumulator. This device operates on 3 V power supply drawing 16 mA when fully powered or 850 μA at standby.

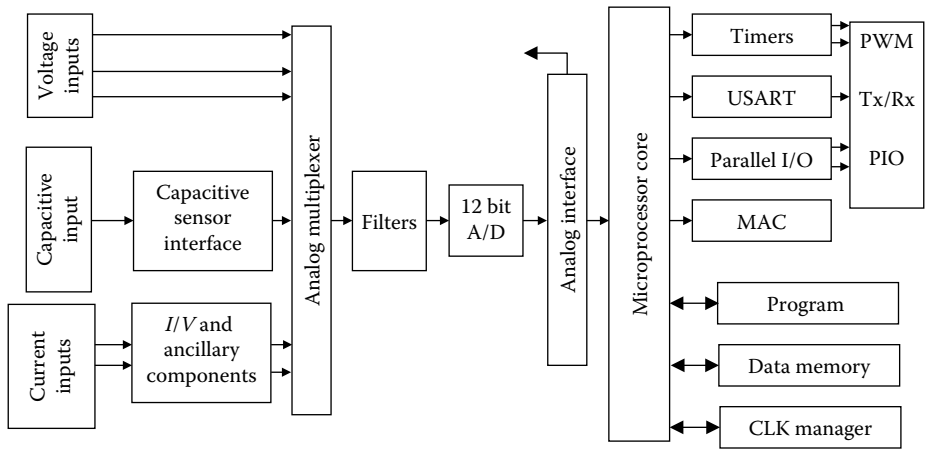


FIGURE 1.5 Block diagram of a typical micro-instrument.

Modern microsensors and microinstruments are fabricated by making full use of properties of semiconductors and their associated technologies. In the manufacturing process, the use of other materials and the deposition of thick and thin films are often required to give the sensing materials useful properties otherwise they would not have. For example, piezoelectric material films applied to silicon wafers provide piezoelectric properties. There are several methods of depositing thin and thick films on substrates or semiconductor wafers. Some of the methods are spin casting, vacuum deposition, sputtering, electroplating, screen printing, etc.

1.5.1 Smart Sensors

A conventional sensor measures physical, biological, or chemical parameters and converts these parameters into electric signals. They require extensive external circuits and components for signal processing and display. The term *smart sensor* was adopted in the mid-1980s to differentiate a new class of sensors from the conventional ones. Smart sensors have intelligence of some form and can convert a raw sensor signal into a level that makes them much more convenient to use. They provide value-added functions, thus increasing the quality of information rather than just passing the raw signal. They can perform functions such as self-identification, self-testing, lookup tables, calibration curves as well as ability to communicate with other devices. These functions are conducted by the integration of sensors with microcontrollers or microprocessor or logic circuits in the same chip. Understandably, the microprocessor contains RAM and ROM and can conveniently be programmed externally. Smart sensors also include signal amplification, conditioning, processing, and A/D conversions.

A variety of smart sensors are manufactured with the neural network and other intelligence techniques programmed and held onboard the chip. These sensors are capable of assimilating a large quantity of data; hence, they are capable of taking autonomous and appropriate actions to achieve goals in any dynamically changing environment. They are adaptable in anticipating events and complexities of the process; therefore, sensing, learning, and self-configurations are the key elements. Intelligent sensors appear in the marketplace as pressure sensors and accelerometers, biosensors, chemical sensors, optical sensors, magnetic sensors, and so on. Intelligent vision systems and parallel processors–based sensors are typical examples of such devices.

1.5.2 Wireless and Autonomous Sensors and Instruments

Sudden growth in the wireless communication technology has prompted the expansion of wireless industry by order of magnitudes. This is largely supported by improvement in digital and RF circuit fabrication methods, advances in signal-processing theory and applications, and emergence of new large-scale wireless-related ICs and other supporting devices. Particularly, the new IC technology makes radio equipment smaller, cheaper, and more reliable. In parallel to industrial expansion, consumer acceptance and need for products permit widespread deployment of the wireless communication systems. Trends will continue at an even greater pace in the coming years.

Wireless technology is used extensively in modern sensors and instrumentation. A typical wireless sensor contains sensing elements, signal-processing circuits, and wireless communication components in the same chip. Several miniaturization techniques are available for wireless sensors, including SOC, MEMS, and ASICs. In a wireless sensor, there are five main components that need to be integrated for a complete system; these components are sensor, signal-processing circuitry, radio, battery, and package. Radio, sensors, and signal-processing circuitry can be reduced in size through hybrid circuits, MEMS, or mixed-signal ASIC design. However, the power supply components must be dealt with separately as explained in Chapter 90 of this book and Chapter 13 of *Electrical, Optical, Chemical, and Biomedical Measurement*.

Wireless instruments consist of five main components—sensors and signal conditioners, programmable digital hardware, memory and storage, input/output and communication components, and others, for

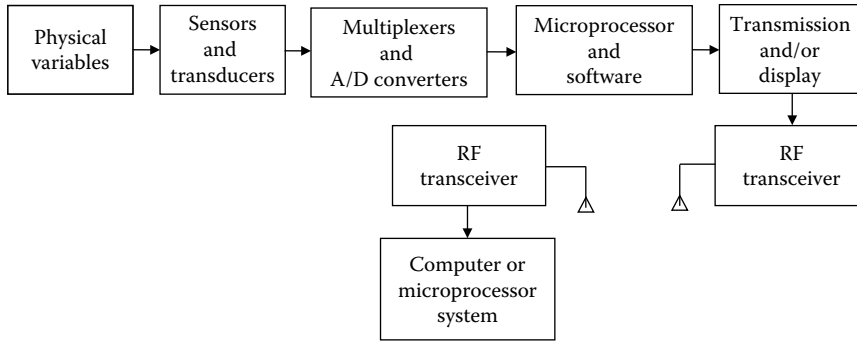


FIGURE 1.6 Components of a wireless instrument.

example, displays, keypads, and power supplies. A typical wireless instrument is illustrated in Figure 1.6. Instruments differ from each other by the way that they handle, transmit, and display information. Further information can be found in Part IX, which is dedicated to wireless technology for instruments and sensors.

Autonomous sensors are self-powered measurement devices that are capable of communicating wirelessly. They serve as the nodes in distributed data acquisition systems and wireless sensor networks (WSN) finding applications such as in health care, aerospace, and environmental monitoring. To increase their autonomy, autonomous sensors seek to reduce their average power consumption by working in low power modes whenever possible. They spend most of the time in *sleep* (standby) mode and only *wake up* to perform specific actions—namely, measurement, processing, and transmission/reception of data.

Autonomous sensors are composed of sensors, signal conditioners, processors, and transceivers. Sensors convert a signal from a physical or chemical quantity to a corresponding signal in the electric domain. Often, commercial transceivers are used for wireless communication. They transmit in the free-licensed ISM bands and can use a proprietary or standard (e.g., IEEE 802.15.4) protocol. Further information on autonomous sensors can be found in Chapter 90.

1.6 Instrumentation and Networks

Networking of hardware and software resources is essential to bring multiple sensors and instruments together for exchange of information, collaborative operations, and sharing of functions of equipment and devices. Networks are made by the collection of devices, the medium that links these devices, and the software that supports the networking of the entire system. A *system* is made from a group of interrelated parts with the focus of establishing an interrelationship between them to improve efficiency, to facilitate integration of the application, and to share the resources.

Due to recent progress in communications technology, sensors and instruments can easily be networked. Many processes require measurements of hundreds and perhaps thousands of parameters employing many instruments. The resulting arrangement for performing the overall measurement in a complex processes is called the *measurement system*. In measurement systems, instruments operate autonomously but in a coordinated manner. Information generated by each instrument may be communicated between the instruments themselves and the controllers or between instruments and other digital devices such as recorders, display units, printers, routers, base stations, or host computer.

Connecting devices together to form networks is not a new concept, and it has been operational for many years in a diverse range of applications. In the earlier networks, almost all the communicating devices were connected by wires; hence, they were largely fixed in space. The devices in modern networks, as discussed in this book, can be configured by using wireless communication technology and related software; hence, they can have mobility in space while still maintaining feasible networks. Therefore, modern networks can be viewed as (1) *wired networks* in which the communication devices are connected by wires, hence largely fixed in space, (2) *wireless networks* in which devices communicate wirelessly,

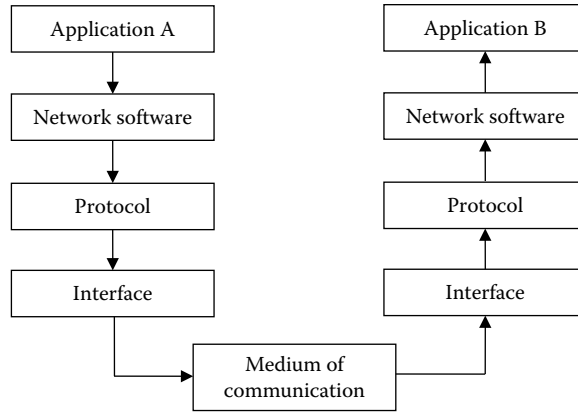


FIGURE 1.7 Process of communication of networked devices.

hence can move in space, and (3) *hybrid networks* in which both wired and wireless techniques are used in combination. At the moment, mobile networks based on wireless techniques provide primarily voice-based services, but they are increasingly handling data and other forms of information. Wireless networks can match similar functions as fixed networks plus they offer many advantages such as the reduction in cost for initial setup and maintenance.

Devices need network software to issue the requests and responses that allow them communicate with each other successfully. A communication process between two devices is illustrated in Figure 1.7. In this case, communication is taking place in simplex form; device A is sending information to device B.

In many networks, communicating devices invoke a layer of codes, which is called *network operating systems* (NOSs). NOSs control the access to network recourses. Examples of common NOSs used in computers are Windows.NET, WindowsXP, Novell's NetWare, etc.

Most network software packages come with modules for logging on and off the network. Logging on and logging off network modules may include considerations such as password security, validation of user access to specific files and software, automatic log-on feature for some devices, help menus, and error messages.

More information on instrument networks can be found in Part IX of this book and Part I of *Electrical, Optical, Chemical, and Biomedical Measurement*.

1.7 Software Support for Sensors and Instruments

The *software* is a term for describing the role of programs, procedures, and documentation in programmable digital devices and computer systems. Software support is an integral part of modern instrumentation system. It can be divided into three major groups: (1) system software, (2) programming software, and (3) application software. All are developed using programming languages, scripting languages, machine languages or assembly codes, or FPGA configurations. Some examples of programming languages are C or C++, Java, and Basic.

The software in instrumentation systems includes

- *Application software* such as Excel, word processors, video games, and Open Office that provides tasks for users.
- *Firmware* is software permanently stored in programmable memory devices on the main boards or other types of integrated hardware carriers.
- *Middleware* controls and coordinates distributed systems.
- *System software* such as operating systems that interface with hardware to provide a coordinated service for applications such as Linux and Microsoft Windows.
- *Web pages* developed by technologies such as HTML, Perl, JSP, XML, PHP.

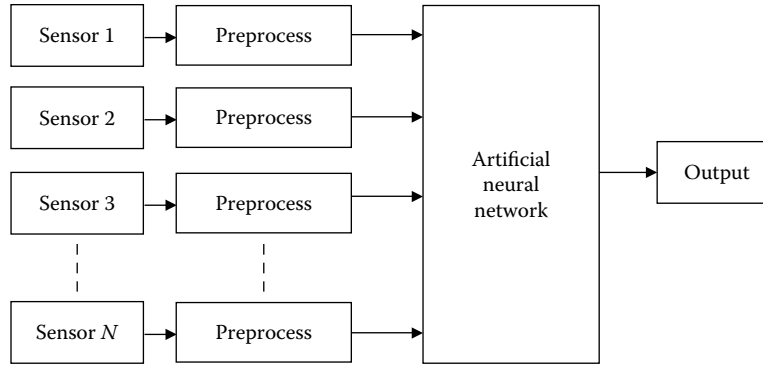


FIGURE 1.8 AI-controlled single-chip multisensor array.

At the lowest level, software consists of a *machine language* specific to an individual processor. A machine language consists of groups of binary values signifying processor instructions that change the state of the computer from its preceding state. Software is an ordered sequence of instructions for changing the state of hardware in a particular sequence. It is usually written in high-level programming languages that are easier and more efficient for humans to use than machine language. High-level languages are compiled or interpreted into machine language object code. Software may also be written in an assembly language that is mnemonic representation of a machine language using a natural language alphabet. Assembly language must be assembled into object code via an assembler.

For example, the microinstrument illustrated in Figure 1.5 is supported by C programming language. It includes hardware support for a single break point and trace event for code debugging. A development system interface provides instructions for halting the processor, single stepping through code, and reading and writing system registers. Single-cycle interrupt response and direct memory access (DMA) allow application of this device in time-critical applications. A clock manager is used to divide the system clock to reduce power, assign a slow clock, or halt the clock to the different sections of the chip until a programmed event occurs.

The integration of sensors with complex analog and digital signal-processing circuits and microprocessors in the same chip has enabled extensive development of the supporting software. The use of digital signal-processing circuits and the integration of intelligent techniques such as the artificial neural networks (ANNs) and other techniques serve as nonlinear signal-processing tools leading to convenient and easy to use devices. The onboard operating system and the additional decision-making software such as artificial intelligence (AI) and complex logic circuits result in much faster and efficient, fault-tolerant and reliable systems. A typical system configuration of an AI-controlled single-chip, multisensor array is shown in Figure 1.8. In *fully integrated decision and control systems*, an ANN performs both sensor enhancement and intelligent control. Fully integrated systems find extensive applications in aerospace, defense, consumer products, and industrial needs.

There are many different software tools to implement applications of sensors and instruments such as the LabWindows and LabVIEW from National Instruments and the VEE from Hewlet-Packard. Further information can be found in Chapter 10.

1.8 Application Examples

Many different instruments are offered to fulfill specific measurement requirements. Here, some examples will be discussed briefly. In this book, detailed information can be found on specific instruments, sensors, and measurements in the proceeding chapters.

In medical technology, there are three basic types of instruments—imaging, physiological measurements, and laboratory analysis. In imaging and physiological measurements, the instruments are closely

linked with patients. Some examples of these instruments are x-ray tomography, magnetic resonance imaging (MRI) and magnetic resonance tomography, ultrasound imaging, thermography, brain and nervous system sensors, and respiratory sensors. Many instruments are based on radiation and sound, force and tactile sensing, electromagnetic sensing, and chemical and bio-analytical sensors.

Power plants are instrumented for maximum availability, operational safety, and environmental planning. Therefore, their measurements must be as accurate as possible and reliable. Instruments are used for temperature, pressure, flow, level, vibration measurements, and water, steam, and gas analysis. For example, gas analysis requires instruments to measure carbon compounds, sulfur and nitrogen compounds, and dust and ash contents.

Environmental monitoring requires a diverse range of instruments for air, water, and biological monitoring. Instruments are used for measuring various forms of radiation, chemical hazards, air pollutants, and organic solvents. Many sophisticated instruments are also developed for remote monitoring via satellites, and they operate on optical, microwave, and RF electromagnetic radiation principles.

Industry uses instruments extensively for online monitoring purposes and off-line analysis. Specific instruments are used commonly for sensing variables such as temperature, pressure, volumetric and mass flow rate, density, weight, displacement, pH levels, color, absorbency, viscosity, material flow, dew point, organic and inorganic components, turbidity, solid and liquid level, humidity, and particle-size distribution.

In automobiles, instruments are used to assist drivers by sensing variables such as cooling, braking, fuel consumption, humidity control, speed, travel route monitoring, and position sensing. Instruments also find applications for safety and security purposes, such as passenger protection and locking and antitheft systems. Recently, with the advent of micromachined sensors, many diverse instruments such as engine control, fuel injection, air regulation, and torque sensing are developed.

The manufacturing industry, especially automated manufacturing, requires a diverse range of instruments. Machine diagnosis and process parameters are made by instruments that are based on force, torque, pressure, speed, temperature, and electric-parameter sensing. Optics, tactile arrays, and acoustic scanning instruments are used for pattern recognition. Distance and displacement measurements are made by many methods (e.g., inductive, capacitive, optical, and acoustic techniques).

Aerospace instrumentation requires an accurate indication of physical variables and the changes in these variables. Instruments are designed to suit specific conditions of operations. Some of the measurements are gas temperature and pressure, fluid flow, aircraft velocity, aircraft icing, thrust and acceleration, load, strain and force, position, altitude sensing, and direction finding.

Bibliography

Bakshi, U. A. and A. V. Bakshi, *Electronic Instrumentation*, Technical Publications, Pune, India, 2008.

Eren, H., *Electronic Portable Instruments—Design and Applications*, CRC Press, LLC, Boca Raton, FL, 2004.

Eren, H., *Wireless Sensors and Instruments—Networks, Design and Applications*, CRC Press, LLC, Boca Raton, FL, 2006.

Kalsi, H. S., *Electronic Instrumentation*, 2nd edn., Tata McGrawhill, New Delhi, India, 2006.

Lazzaroni, M., V. Piuri, and C. Mazerio, Computer security aspects in industrial instrumentation and measurements, *Proceedings of 2010 IEEE International Instrumentation and Measurement Technology Conference, I2MTC 2010*, Austin, TX: pp. 1216–1221, 2010.

Liptak, B. and H. Eren, *Instrument Engineers Handbook: Process Software and Digital Networks*, Vol. 3, 4th edn., CRC Press, LLC, Boca Raton, FL, 2011.

Morris, A. S., *Measurement and Instrumentation Principles*, 3rd edn., Butterworth-Heinemann, Oxford, U.K., 2001.

Northrop, R. B., *Introduction to Instrumentation and Measurement*, 2nd edn., CRC Press, Boca Raton, FL, 2005.

Tanabe, K., M. Hirose, A. Otani, K. Mochizuki, and M. Uchida, Recent topics in instrumentation and measurement, *IEEJ, Transactions on Fundamentals and Materials*, 129(2), 70–74, 2009.

Partial List of Vendors and Suppliers

AMETEK Power Instruments, 255 N. Union St., Rochester, NY 14605, Phone: 585-263-7700, 800-881-4156 (toll free), Fax: 585-262-4777, www.ametekpower.com

Analtech, Inc., 75 Blue Hen Dr., Newark, DE 19713, Phone: 302-737-6960, 800-441-7540 (toll free), Fax: 302-737-7115, www.ichromatography.com

Davis/Inotek, 1236 Mark St. Bensenville, IL 60106, Phone: 410-358-3900, 800-548-9409 (toll free), Fax: 410-358-0252, www.davis.com

Dwyer Instruments, Inc., Hwy. 212 at 12 P.O. Box 373 Michigan City, IN 46361, Phone: 219-879-8000, 800-872-9141 (toll free), Fax: 219-872-9057, www.dwyer-inst.com

Electro-Sensors, Inc., 6111 Blue Circle Dr. Minnetonka, MN 55343-9108, Phone: 952-930-0100, 800-328-6170 (toll free), Fax: 952-930-0130, www.electro-sensors.com

Honeywell Sensing and Control, 1985 Douglas Dr. North, MN10-192B Golden Valley, MN 55422, Phone: 763-954-4818, 800-446-6555 (toll free), sensing.honeywell.com/index.cfm/ci_id/15…

Instrumented Sensor Technology, Inc., 4704 Moore Street Okemos, MI 48864, Phone: 517-349-8487, Fax: 517-349-8469, www.isthq.com

Jewell Instruments, LLC, 850 Perimeter Rd. Manchester, NH 03103, Phone: 603-669-6400, 800-638-3771 (toll free), Fax: 603-669-5962, www.jewellinstruments.com

Macro Sensors, Division of Howard A. Schaevitz Technologies, 7300 US Rte. 130 N. Bldg. 22 Pennsauken, NJ 08110, Phone: 856-662-8000, Fax: 856-317-1005, www.macrosensors.com

Micro Sense, LLC, 70 Industrial Ave. E. Lowell, MA 01852, Phone: 978-843-7673, www.microsense.net

MKS Instruments, Inc., 2 Tech Dr., Suite 201 Andover, MA 01810, Phone: 978-645-5500, 800-227-8766 (toll free), Fax: 978-557-5100, www.mksinst.com

MTI Instruments, Inc., 325 Washington Ave., Extension Albany, NY 12205-5505, Phone: 518-218-2550, Fax: 518-218-2506, www.mtiinstruments.com

OTP Industrial Solutions, 1900 Jetway Blvd. Columbus, OH 43219, Phone: 614-342-6123, Fax: 614-342-6490, www.otpnet.com

PCB Piezotronics, Inc., 3425 Walden Ave. Depew, NY 14043-2417, Phone: 716-684-0001, 800-828-8840 (toll free), Fax: 716-684-0987, www.pcb.com

Piezo Technologies, 8431 Georgetown Rd., Suite 300 Indianapolis, IN 46268, Phone: 317-876-4670, Fax: 317-876-4681, www.piezotechnologies.com

Schneider Electric-Hyde Park, 1875 Founders Dr. Dayton, OH 45420, Phone: 937-514-7207, Fax: 937-258-5830, www.sesensors.com

Silicon Designs, Inc., 1445 NW Mall St. Issaquah, WA 98027, Phone: (425) 391-8329, Fax: (425) 391-0446, www.silicondesigns.com/contact.html

Turck, Inc., 3000 Campus Dr. Plymouth, MN 55441-2619, Phone: 763-553-7300, 800-554-4989 (toll free), Fax: 763-553-0708, www.turck.com

Warren-Knight Instrument Co., 2045 Bennett Dr. Philadelphia, PA 19116-3019, Phone: 215-464-9300, Fax: 215-464-9303, www.warrenind.com

Wilcoxon Research Inc., 20511 Seneca Meadows Pkwy. Germantown, MD 20876, Phone: 301-330-8811, 800-945-2696 (toll free), Fax: 301-330-8873, www.wilcoxon.com

2

Characteristics of Instrumentation

2.1 Simple Instrument Model **2**-1
Passive and Active Sensors • Calibration • Modifying and Interfering Inputs • Accuracy and Error • Systematic Error Sources (Bias) • Random-Error Sources (Noise) • Error Reduction • Sensor Fusion • Estimation • Testing of Instruments
Bibliography **2**-9

John R. Hansman, Jr.
Massachusetts Institute of Technology

In addressing measurement problems, it is often useful to have a conceptual model of the measurement process. This chapter presents some of the fundamental concepts of measurement in the context of a simple generalized instrument model.

In abstract terms, an *instrument* is a device that transforms a *physical variable* of interest (the *measurand*) into a form that is suitable for recording (the *measurement*). In order for the measurement to have broad and consistent meaning, it is common to employ a standard system of *units* by which the measurement from one instrument can be compared with the measurement of another.

An example of a basic instrument is a ruler. In this case, the measurand is the length of some object, and the measurement is the number of units (meters, inches, etc.) that represent the length.

2.1 Simple Instrument Model

Figure 2.1 presents a generalized model of a simple instrument. The physical process to be measured is in the left of the figure, and the measurand is represented by an observable physical variable $\boldsymbol{X}$. Note that the observable variable $\boldsymbol{X}$ need not necessarily be the measurand but simply related to the measurand in some known way. For example, the mass of an object is often measured by the process of *weighing*, where the measurand is the mass but the physical measurement variable is the downward force the mass exerts in the Earth's gravitational field. There are many possible physical measurement variables. A few are shown in Table 2.1.

The key functional element of the instrument model shown in Figure 2.1 is the *sensor*, which has the function of converting the *physical variable input* into a *signal variable output*. Signal variables have the property that they can be manipulated in a transmission system, such as an electrical or mechanical circuit. Because of this property, the signal variable can be transmitted to an output or recording device that can be remote from the sensor. In electric circuits, voltage is a common signal variable. In mechanical systems, displacement or force is commonly used as a signal variable. Other examples of signal variable are shown in Table 2.1. The signal output from the sensor can be displayed, recorded, or used as an input signal to some secondary device or system. In a basic instrument, the signal is transmitted to a *display* or recording device where the measurement can be read by a human observer. The observed

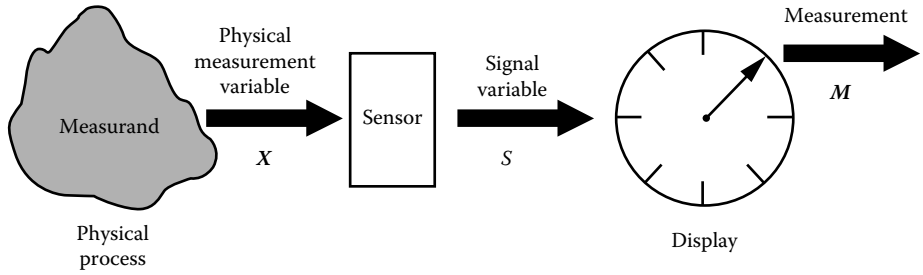


FIGURE 2.1 Simple instrument model.

TABLE 2.1 Examples of Physical Variables and Typical Signal Variables

Common Physical Variables	Typical Signal Variables
Force	Voltage
Length	Displacement
Temperature	Current
Acceleration	Force
Velocity	Pressure
Pressure	Light
Frequency	Frequency
Capacity	—
Resistance	—
Time	—
…	—

output is the measurement ***M***. There are many types of display devices, ranging from simple scales and dial gages to sophisticated computer display systems. The signal can also be used directly by some larger system of which the instrument is a part. For example, the output signal of the sensor may be used as the input signal of a closed-loop control system.

If the signal output from the sensor is small, it is sometimes necessary to amplify the output shown in Figure 2.2. The amplified output can then be transmitted to the display device or recorded, depending on the particular measurement application. In many cases, it is necessary for the instrument to provide a digital signal output so that it can interface with a computer-based data acquisition or communication system. If the sensor does not inherently provide a digital output, then the analog output of the sensor is converted by an analog-to-digital converter (ADC) as shown in Figure 2.2. The digital signal is typically sent to a computer processor that can display, store, or transmit the data as output to some other system, which will use the measurement.

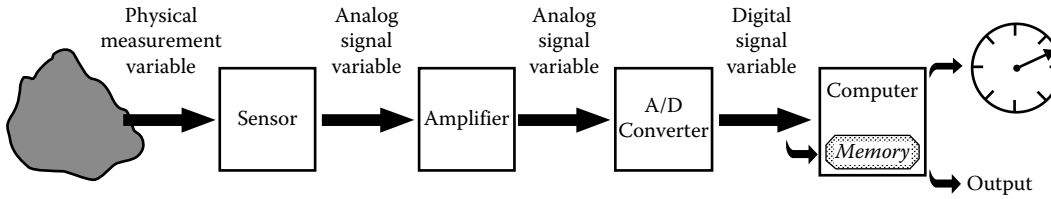


FIGURE 2.2 Instrument model with amplifier, ADC, and computer output.

2.1.1 Passive and Active Sensors

As discussed earlier, sensors convert physical variables to signal variables. Sensors are often transducers in that they are devices that convert input energy of one form into output energy of another form. Sensors can be categorized into two broad classes depending on how they interact with the environment they are measuring. *Passive sensors* do not add energy as part of the measurement process but may remove energy in their operation. One example of a passive sensor is a thermocouple, which converts a physical temperature into a voltage signal. In this case, the temperature gradient in the environment generates a thermoelectric voltage that becomes the signal variable. Another passive transducer is a pressure gage where the pressure being measured exerts a force on a mechanical system (diaphragm, aneroid, or Borden pressure gage) that converts the pressure force into a displacement, which can be used as a signal variable. For example, the displacement of the diaphragm can be transmitted through a mechanical gearing system to the displacement of an indicating needle on the display of the gage.

Active sensors add energy to the measurement environment as part of the measurement process. An example of an active sensor is a radar or sonar system, where the distance to some object is measured by actively sending out a radio (radar) or acoustic (sonar) wave to reflect off of some object and measure its range from the sensor.

2.1.2 Calibration

The relationship between the physical measurement variable input and the signal variable (output) for a specific sensor is known as the *calibration* of the sensor. Typically, a sensor (or an entire instrument system) is calibrated by providing a known physical input to the system and recording the output. The data are plotted on a calibration curve such as the example shown in Figure 2.3. In this example, the sensor has a linear response for values of the physical input less than $\boldsymbol{X}_0$. The *sensitivity* of the device is determined by the slope of the calibration curve. In this example, for values of the physical input greater than $\boldsymbol{X}_0$, the calibration curve becomes less sensitive until it reaches a limiting value of the output signal. This behavior is referred to as *saturation*, and the sensor cannot be used for measurements greater than its saturation value. In some cases, the sensor will not respond to very small values of the physical input variable. The difference between the smallest and the largest physical inputs that can reliably be measured by an instrument determines the *dynamic range* of the device.

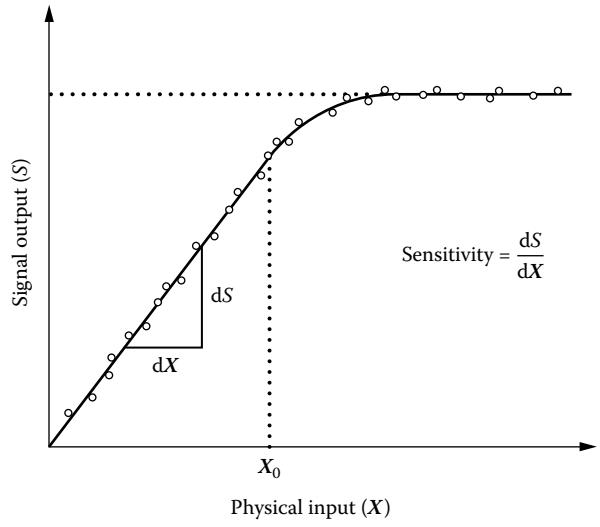


FIGURE 2.3 Calibration curve example.

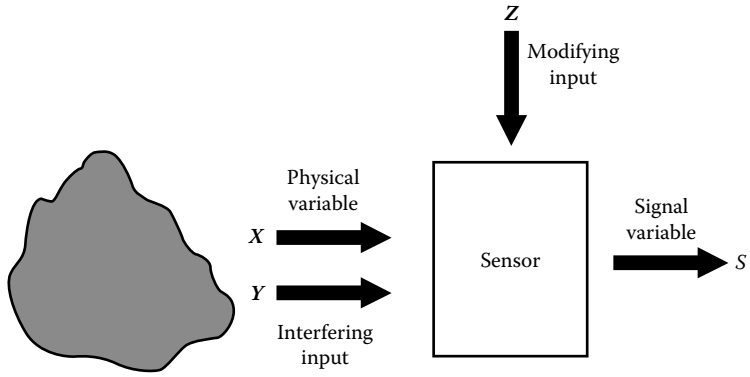


FIGURE 2.4 Interfering inputs.

The accuracy of instruments can be increased by calibration. In many cases, calibration graphs, mathematical equations, tables, and experience of the operators can be utilized. In recent years, with the application of digital techniques and intelligent instruments, error corrections are made automatically by computers or within the devices and sensors themselves.

2.1.3 Modifying and Interfering Inputs

In some cases, the sensor output will be influenced by physical variables other than the intended measurand. In Figure 2.4, ***X*** is the intended measurand, ***Y*** is an *interfering input*, and ***Z*** is a *modifying input*. The interfering input ***Y*** causes the sensor to respond in the same manner as the linear superposition of ***Y*** and the intended measurand ***X***. The measured signal output is therefore a combination of ***X*** and ***Y***, with ***Y*** interfering with the intended measurand ***X***. An example of an interfering input would be a structural vibration within a force measurement system.

Modifying inputs changes the behavior of the sensor or measurement system, thereby modifying the input/output relationship and calibration of the device. This is shown schematically in Figure 2.5. For various values of ***Z*** in Figure 2.5, the slope of the calibration curve changes. Consequently, changing ***Z*** will result in a change of the apparent measurement even if the physical input variable ***X*** remains constant. A common example of a modifying input is temperature; it is for this reason that many devices are calibrated at specified temperatures.

2.1.4 Accuracy and Error

The *accuracy* of an instrument is defined as the difference between the *true value* of the measurand and the *measured value* indicated by the instrument. Typically, the true value is defined in reference to some absolute or agreed upon standard. For any particular measurement, there will be some error due to *systematic* (*bias*) and *random* (*noise*) error sources. The combination of systematic and random error can be visualized by considering the analogy of the target shown in Figure 2.6. The total error in each shot results from both systematic and random errors. The systematic (bias) error results in the grouping of shots being offset from the bulls eye (presumably a misalignment of the gunsight or wind). The size of the grouping is determined by random-error sources and is a measure of the *precision* of the shooting.

2.1.5 Systematic Error Sources (Bias)

There are a variety of factors that can result in systematic measurement errors. One class of cause factors are those that change the input–output response of a sensor resulting in miscalibration. The modifying inputs and interfering inputs discussed earlier can result in sensor miscalibration. For example,

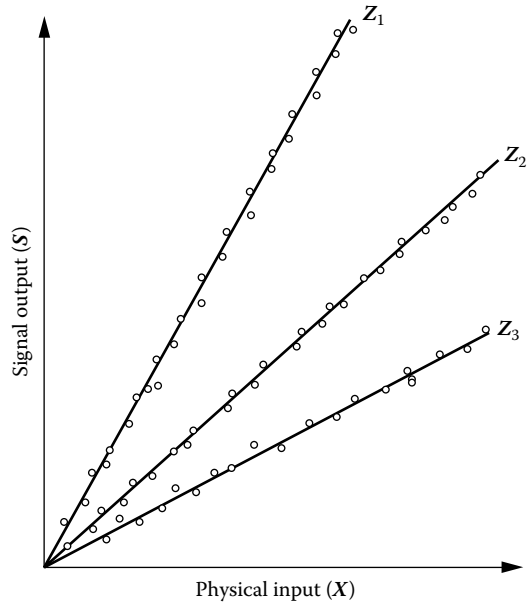


FIGURE 2.5 Illustration of the effect of a modifying input on a calibration curve.

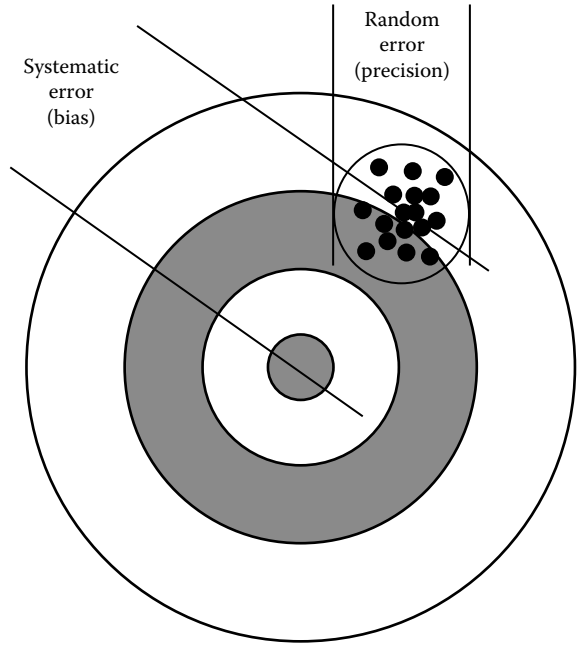


FIGURE 2.6 Target analogy of measurement accuracy.

if temperature is a modifying input, using the sensor at a temperature other than the calibrated temperature will result in a systematic error. In many cases, if the systematic error source is known, it can be corrected for by the use of *compensation methods*.

There are other factors that can also cause a change in sensor calibration resulting in systematic errors. In some sensors, aging of the components will change the sensor response and hence the calibration.

Damage or abuse of the sensor can also change the calibration. In order to prevent these systematic errors, sensors should be periodically recalibrated.

Systematic errors can also be introduced if the measurement process itself changes the intended measurand. This issue, defined as *invasiveness*, is a key concern in many measurement problems. Interaction between measurement and measurement device is always present; however, in many cases, it can be reduced to an insignificant level. For example, in electronic systems, the energy drain of a measuring device can be made negligible by making the input impedance very high. An extreme example of invasiveness would be to use a large warm thermometer to measure the temperature of a small volume of cold fluid. Heat would be transferred from the thermometer and would warm the fluid, resulting in an inaccurate measurement.

Systematic errors can also be introduced in the signal path of the measurement process shown in Figure 2.4. If the signal is modified in some way, the indicated measurement will be different from the sensed value. In physical signal paths such as mechanical systems that transmit force or displacement, friction can modify the propagation of the signal. In electric circuits, resistance or attenuation can also modify the signal, resulting in a systematic error.

Finally, systematic errors or bias can be introduced by human observers when reading the measurement. A common example of observer bias error is *parallax error*. This is the error that results when an observer reads a dial from a non-normal angle. Because the indicating needle is above the dial face, the apparent reading will be shifted from the correct value.

2.1.6 Random-Error Sources (Noise)

If systematic errors can be removed from a measurement, some error will remain due to the random-error sources that define the precision of the measurement. Random error is sometimes referred to as *noise*, which is defined as a signal that carries no useful information. If a measurement with true random error is repeated a large number of times, it will exhibit a *Gaussian distribution*, as demonstrated in the example in Figure 2.7 by plotting the number of times values within specific ranges are measured. The Gaussian distribution is centered on the true value (presuming no systematic errors), so the mean or average of all the measurements will yield a good estimate of the true value.

The precision of the measurement is normally quantified by the standard deviation (σ) that indicates the width of the Gaussian distribution. Given a large number of measurements, a total of 68% of the measurements will fall within $\pm 1\sigma$ of the mean, 95% will fall within $\pm 2\sigma$, and 99.7% will fall within $\pm 3\sigma$.

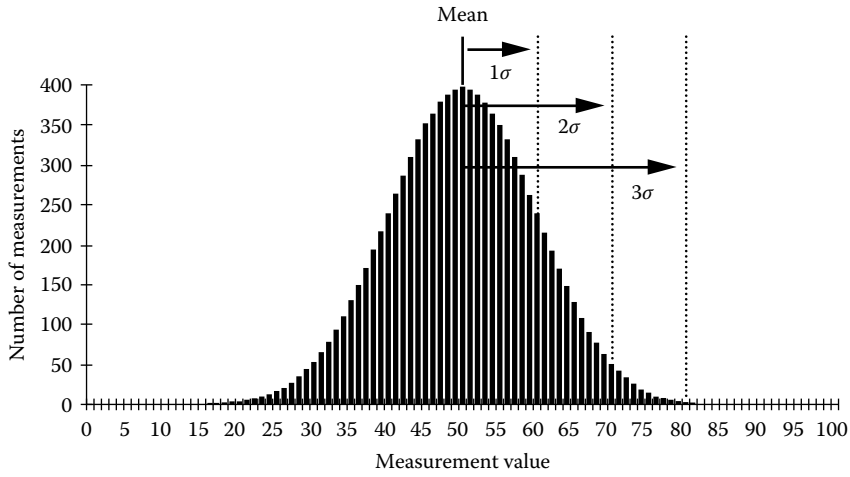


FIGURE 2.7 Example of a Gaussian distribution.

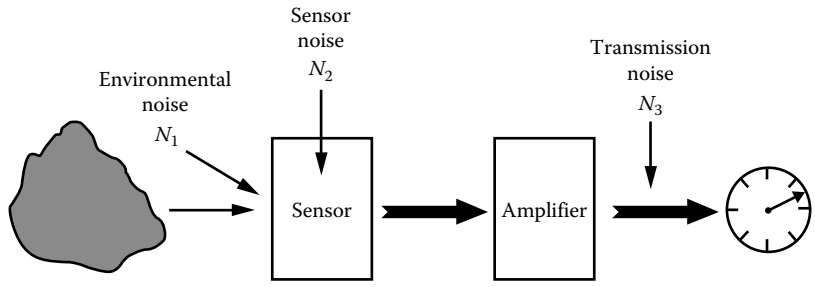


FIGURE 2.8 Instrument model with noise sources.

Smaller standard deviation leads to more precise the measurement. For many applications, it is common to refer to the 2σ value when reporting the precision of a measurement. However, for some applications such as navigation, it is common to report the 3σ value, which defines the limit of likely uncertainty in the measurement.

There are a variety of sources of randomness that can degrade the precision of the measurement—starting with the repeatability of the measurand itself. For example, if the height of a rough surface is to be measured, the measured value will depend on the exact location at which the measurement is taken. Repeated measurements will reflect the randomness of the surface roughness.

Random error generating noise can also be introduced at each stage in the measurement process, as shown schematically in Figure 2.8. Random interfering inputs will result in noise from the measurement environment N_1, which are introduced before the sensor, as shown in the figure. An example would be background noise received by a microphone. Sensor noise N_2 can also be introduced within the sensor. An example of this would be thermal noise within a sensitive transducer, such as an infrared sensor. Random motion of electrons, due to temperature, appears as voltage signals, which are apparently due to the high sensitivity of the device. For very sensitive measurements with transducers of this type (e.g., infrared detectors), it is common to cool the detector to minimize this noise source.

Noise N_3 can also be introduced in the transmission path between the transducer and the amplifier. A common example of transmission noise in the United States is 60 Hz interference from the electric power grid that is introduced if the transmission path is not well grounded or if an inadvertent electric grand loop causes the wiring to act as an antenna.

It is important to note that the noise will be amplified along with the signal as it passes through the amplifier shown in Figure 2.8. As a consequence, the figure of merit when analyzing noise is not the level of the combined noise sources but the level of the *signal-to-noise ratio* (*SNR*), defined as the ratio of the signal power to the power in the combined noise sources. It is common to report SNR in decibel units.

The SNR is ideally much greater than 1 (0 dB). However, it is sometimes possible to interpret a signal that is lower than the noise level if some identifying characteristics of that signal are known and sufficient signal processing power is available. The human ability to hear a voice in a loud noise environment is an example of this signal processing capability.

2.1.7 Error Reduction

Controlling errors is an essential part of instruments and instrumentation systems. Error control begins in the design stages by choosing the appropriate components, filtering, and bandwidth selection and by reducing noise. Minimizing the errors generated by the individual subunits of the complete system will result in lowering the overall error of the instrument.

In many instrumentation systems, the application of compensation strategy is used to increase static and dynamic performances. In the case of static characteristics, compensations can be made by different methods, including the introduction of opposing nonlinear elements into the system; by using isolation and zero environmental sensitivity conditions, opposing compensating environmental inputs; and by

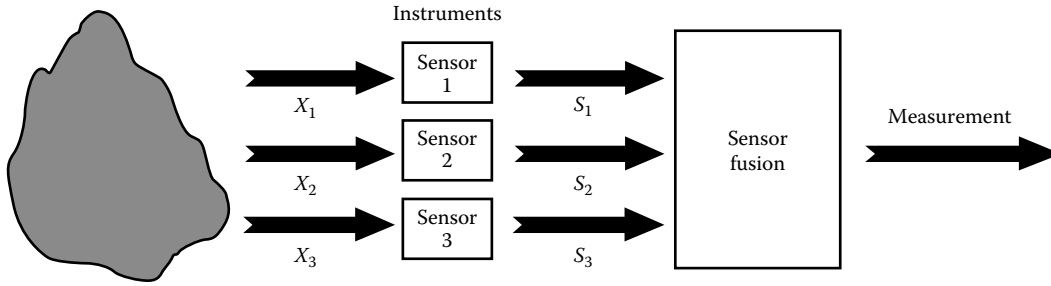


FIGURE 2.9 Example of sensor fusion.

using differential and feedback systems. At the same time, dynamic compensation can be achieved by applying these techniques as well as by taking measures to reduce harmonics, using filters, adjusting bandwidths, using feedback compensation techniques, and so on. Further information can be found in Chapter 1.

2.1.8 Sensor Fusion

The process of *sensor fusion* is modeled in Figure 2.9. In this case, two or more sensors are used to observe the environment, and their output signals are combined in some manner (typically in a processor) to provide a single enhanced measurement. This process frequently allows measurement of phenomena that would otherwise be unobservable. One simple example is thermal compensation of a transducer where a measurement of temperature is made and used to correct the transducer output for modifying effects of temperature on the transducer calibration. Other more sophisticated sensor-fusion applications range to image synthesis where radar, optical, and infrared images can be combined into a single enhanced image. Sensor fusion has become more prevalent with the advent of data bus architectures (such as the Controlled Area Network [CAN] bus) where sensor information can be easily shared between devices through the data bus.

2.1.9 Estimation

With the use of computational power, it is often possible to improve the accuracy of a poor-quality measurement through the use of *estimation techniques*. These methods range from simple-averaging or low-pass filtering to cancel out random fluctuating errors to more sophisticated techniques such as Wiener or Kalman filtering and model-based estimation techniques. The increasing capability and lowering cost of computation makes it increasingly attractive to use lower performance sensors with more sophisticated estimation techniques in many applications.

2.1.10 Testing of Instruments

After the instrument is designed and prototyped, various evaluation tests are conducted. These tests may be made under reference conditions or under simulated environmental conditions. Some examples of reference condition tests are accuracy, response time, drift, and warm-up time. Simulated environmental tests may be compulsory, thus being regulated by the governments and other authorities. Some simulated environment tests include climatic test, drop test, insulation-resistance test, vibration test, electromagnetic compatibility tests, safety and health hazard tests, and ingress progress (IP) tests against water and dust. The IP tests are particularly important for

electronic portable instruments since they are most likely to be used during their operations in unsteady environments containing rain, water, humidity, and dust. Many of these tests are strictly regulated by national and international standards.

Bibliography

Eren, H. *Electronic Portable Instruments—Design and Applications*, CRC Press, LLC, Boca Raton, FL, 2004.

Kalsi, H. S. *Electronic Instrumentation*, 6th edn., McGraw-Hill Companies, Delhi, India, 2006.

Klaassen, K. B. and S. Gee, *Electronic Measurement and Instrumentation*, Cambridge University Press, Cambridge, U.K., Reprint, 2002.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

3

Operational Modes of Instrumentation

3.1 Null Instrument **3**-1
3.2 Deflection Instrument **3**-3
3.3 Analog and Digital Sensors **3**-4
3.4 Analog and Digital Readout Instruments **3**-5
3.5 Input Impedance **3**-6
Defining Terms **3**-7
Further Information **3**-7

Richard S. Figliola
Clemson University

3.1 Null Instrument

The null method is one possible mode of operation for a measuring instrument. A null instrument uses the null method for measurement. In this method, the instrument exerts an influence on the measured system so as to oppose the effect of the measurand. The influence and the measurand are balanced until they are equal but opposite in value, yielding a null measurement. Typically, this is accomplished by some type of feedback operation that allows the comparison of the measurand against a known standard value. The key features of a null instrument include an iterative balancing operation using some type of comparator, either a manual or automatic feedback used to achieve balance, and a null deflection at parity.

A null instrument offers certain intrinsic advantages over other modes of operation (e.g., see deflection instruments). By balancing the unknown input against a known standard input, the null method minimizes interaction between the measuring system and the measurand. As each input comes from a separate source, the significance of any measuring influence on the measurand by the measurement process is reduced. In effect, the measured system sees a very high input impedance, thereby minimizing loading errors. This is particularly effective when the measurand is a very small value. Hence, the null operation can achieve a high accuracy for small input values and with a low loading error. In practice, the null instrument will not achieve perfect parity due to the usable resolution of the balance and detection methods, but this is limited only by the state of the art of the circuit or scheme being employed.

A disadvantage of null instruments is that an iterative balancing operation requires more time to execute than simply measuring sensor input. Thus, this method might not offer the fastest measurement possible when high-speed measurements are required. However, the user should weigh achievable accuracy against needed speed of measurement when considering operational modes. Further, the design of the comparator and balance loop can become involved such that highly accurate devices are generally not the lowest cost measuring alternative.

An equal-arm balance scale is a good mechanical example of a manual balance-feedback null instrument, as shown in Figure 3.1. This scale compares the unknown weight of an object on one side against a set of standard or known weights. The known values of weight are added iteratively to one side to

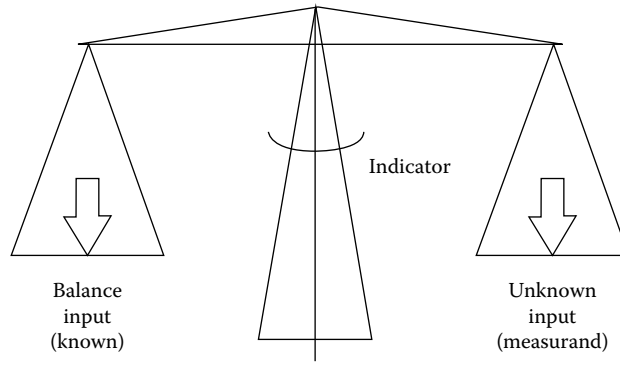


FIGURE 3.1 Measurand and the known quantities balance one another in a null instrument.

exert an influence to oppose the effect of the unknown weight on the opposite side. Until parity, a high or low value is noted by the indicator providing the feedback logic to the operator for adding or removing weights in a balancing iteration. At true parity, the scale indicator is null; that is, it indicates a zero deflection. Then, the unknown input or measurand is deduced to have a value equal to the balance input, the amount of known weights used to balance the scale. Factors influencing the overall measurement accuracy include the accuracy of the standard weights used and resolution of the output indicator and the friction at the fulcrum. Null instruments exist for measurement of most variables. Other common examples include bridge circuits, often employed for highly accurate resistance measurements and found in load cells, temperature-compensated transducers, and voltage-balancing potentiometers used for highly accurate low-voltage measurements.

Within the null instrument, the iteration and feedback mechanism is a loop that can be controlled either manually or automatically. Essential to the null instrument are two inputs: the measurand and the balance input. The null instrument includes a differential comparator, which compares and computes the difference between these two inputs. This is illustrated in Figure 3.2. A nonzero output from the comparator provides the error signal and drives the logic for the feedback correction. Repeated corrections provide for an iteration toward eventual parity between the inputs and the results in the null condition where the measurand is exactly opposed by the balance input. At parity, the error signal is driven to zero by the opposed influence of the balance input, and the indicated deflection is at null, thus lending the name to the method. It is the magnitude of the balance input that drives the output reading in terms of the measurand.

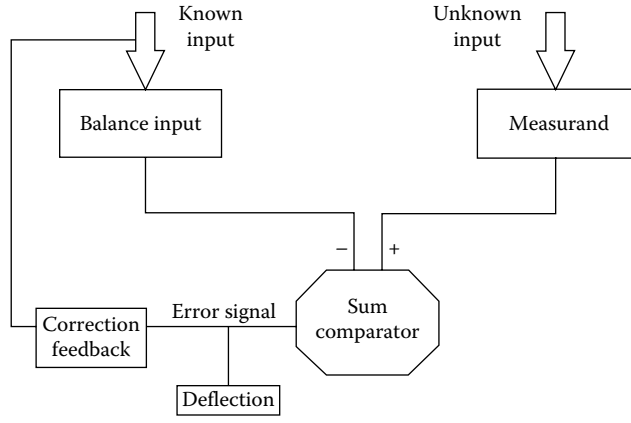


FIGURE 3.2 Null instrument requires input from two sources for comparison.

3.2 Deflection Instrument

The deflection method is one possible mode of operation for a measuring instrument. A deflection instrument uses the deflection method for measurement. A deflection instrument is influenced by the measurand so as to bring about a proportional response within the instrument. This response is an output reading that is a deflection or a deviation from the initial condition of the instrument. In a typical form, the measurand acts directly on a prime element or primary circuit so as to convert its information into a detectable form. The name is derived from a common form of instrument where there is a physical deflection of a prime element that is linked to an output scale, such as a pointer or other type of readout, which deflects to indicate the measured value. The magnitude of the deflection of the prime element brings about a deflection in the output scale that is designed to be proportional in magnitude to the value of the measurand.

Deflection instruments are the most common of measuring instruments. The relationship between the measurand and the prime element or measuring circuit can be a direct one, with no balancing mechanism or comparator circuits used. The proportional response can be manipulated through signal-conditioning methods between the prime element and the output scale so that the output reading is a direct indication of the measurand. Effective designs can achieve a high accuracy, yet sufficient accuracy for less demanding uses can be achieved at moderate costs.

An attractive feature of the deflection instrument is that it can be designed for either static or dynamic measurements or both. An advantage to deflection design for dynamic measurements is in the high dynamic response that can be achieved. A disadvantage of deflection instruments is that by deriving its energy from the measurand, the act of measurement will influence the measurand and change the value of the variable being measured. This change is called a loading error. Hence, the user must ensure that the resulting loading error is acceptable. This usually involves a careful look at the instrument input impedance for the intended measurement.

A spring scale is a good, simple example of a deflection instrument. As shown in Figure 3.3, the input weight or measurand acts on a plate spring. The plate spring serves as a prime element. The original position of the spring is influenced by the applied weight and responds with a translational displacement, a deflection x. The final value of this deflection is a position that is at equilibrium between the downward force of the weight, W, and the upward restoring force of the spring, kx. That is, the input force is balanced against the restoring force. A mechanical coupler is connected directly or by linkage to a pointer. The pointer position is mapped out on a corresponding scale that serves as the readout scale. For example, at equilibrium $W = kx$ or by measuring the deflection of the pointer, the weight is deduced by $x = W/k$.

The flow-diagram logic for a deflection instrument is rather linear, as shown in Figure 3.4. The input signal is sensed by the prime element or primary circuit and thereby deflected from its initial setting. The deflection signal is transmitted to signal conditioners that act to condition the signal into

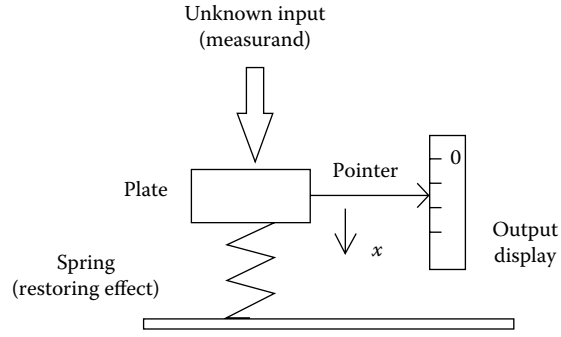


FIGURE 3.3 Deflection instrument requires input from only one source but may introduce a loading error.

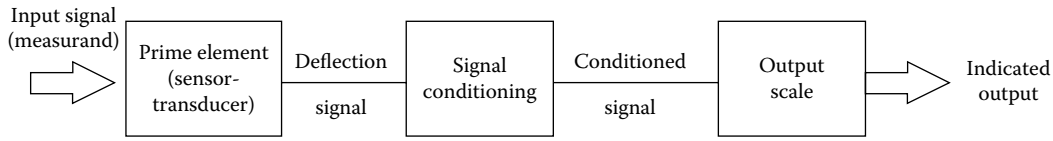


FIGURE 3.4 Logic flowchart for a deflection instrument is straightforward.

a desired form. The examples of signal conditioning are to multiply the deflection signal by some scaler magnitude, such as in amplification or filtering, or to transform the signal by some arithmetic function. The conditioned signal is then transferred to the output scale, which provides the indicated value corresponding to the measurand value.

3.3 Analog and Digital Sensors

Analog sensors provide a signal that is continuous in both its magnitude and its temporal (time) or spatial (space) content. The defining word for analog is "continuous." If a sensor provides a continuous output signal that is directly proportional to the input signal, then it is analog.

Most physical variables, such as current, temperature, displacement, acceleration, speed, pressure, light intensity, and strain, tend to be continuous in nature and are readily measured by an analog sensor and represented by an analog signal. For example, the temperature within a room can take on any value within its range, will vary in a continuous manner in between any two points in the room, and may vary continuously with time at any position within the room. An analog sensor, such as a bulb thermometer or a thermocouple, will continuously respond to such temperature changes. Such a continuous signal is shown in Figure 3.5, where the signal magnitude is analogous to the measured variable (temperature) and the signal is continuous in both magnitude and time.

Digital sensors provide a signal that is a direct digital representation of the measurand. Digital sensors are basically binary ("on" or "off") devices. Essentially, a digital signal exists at only discrete values of time (or space). And within that discrete period, the signal can represent only a discrete number of magnitude values. A common variation is the discrete sampled signal representation, which represents a sensor output in a form that is discrete both in time or space and in magnitude.

Digital sensors use some variation of a binary numbering system to represent and transmit the signal information in digital form. A binary numbering system is a number system using the base 2. The simplest binary signal is a single bit that has only one of two possible values, a 1 or a 0. Bits are like electrical "on–off" switches and are used to convey logical and numerical information. With appropriate input, the value of the bit transmitted is reset corresponding to the behavior of the measured variable. A digital sensor that transmits information one bit at a time uses serial transmission. By combining bits or transmitting bits in groups, it is also possible to define logical commands or integer numbers beyond a 0 or 1. A digital sensor that transmits bits in groups uses parallel transmission. With any digital device, an M-bit signal can express 2^M different numbers. This also provides the limit for the different values that a digital device can discern. For example, a 2 bit device can express 2^2 or 4

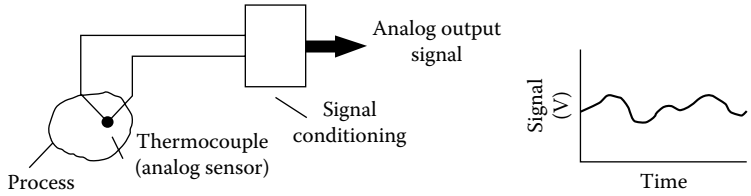


FIGURE 3.5 Thermocouple provides an analog signal for processing.

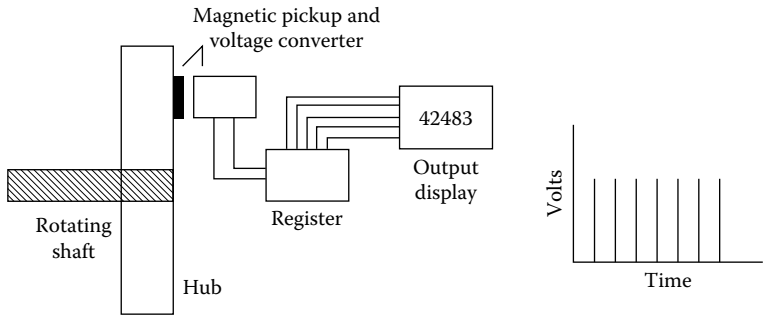


FIGURE 3.6 Rotating shaft with a revolution counter produces a digital signal.

different numbers, 00, 01, 10, and 11, corresponding to the values of 0, 1, 2, and 3, respectively. Thus, the resolution in a magnitude discerned by a digital sensor is inherently limited to 1 part in 2^M.

The concept of a digital sensor is illustrated by the revolution counter in Figure 3.6. Such devices are widely used to sense the revolutions per minute of a rotating shaft. In this example, the sensor is a magnetic pickup/voltage converter that outputs a pulse with each pass of a magnetic stud mounted to a hub on the rotating shaft. The output from the pickup normally is "off" but is momentarily turned "on" by the passing stud. This pulse is a voltage spike sent to a digital register whose value is increased by a single count with each spike. The register can send the information to an output device, such as the digital display shown. The output from the sensor can be viewed in terms of voltage spikes with time. The count rate is related to the rotational speed of the shaft. As seen, the signal is discrete in time. A single stud with pickup will increase the count by one for each full rotation of the shaft. The fractions of a rotation can be resolved by increasing the number of studs on the hub. In this example, the continuous rotation of the shaft is analog but the revolution count is digital. The amplitude of the voltage spike is set to activate the counter and is not related to the shaft rotational speed.

3.4 Analog and Digital Readout Instruments

An analog readout instrument provides an output indication that is continuous and directly analogous to the behavior of the measurand. Typically, this might be the deflection of a pointer or an ink trace on a graduated scale or the intensity of a light beam or a sound wave. This indicated deflection may be driven by changes in voltage or current; or by mechanical, magnetic, or optical means; or combinations of these. The resolution of an analog readout is defined by the smallest usable increment on its readout scale. The span of the readout is defined by the difference between the minimum and the maximum values that it can indicate. Its range specifies the minimum and maximum values that it can indicate.

A digital readout instrument provides an output indication that is a discrete value. The value is directly related to the value of the measurand. The digital readout is typically in the form of a numerical value that is either a fixed number or a number that is updated periodically. One means of displaying a digital number is the seven-segment digital display chip, shown in Figure 3.7, whose output can be updated by altering the grounding inputs A through G. The resolution of a digital readout is given by its least count, the equivalent amount of the smallest change resolved by the least significant digit in the readout. The span and the range are defined as for analog instruments.

Many digital devices combine features of an analog sensor with a digital readout or, in general, convert an analog signal to a discrete signal, which is indicated through a digital output. In such situations, an analog-to-digital converter (ADC) is required. This hybrid device has its analog side specified in terms of its full-scale analog range, E_{FSR}, which defines the analog voltage span over which the device will operate. The digital side is specified in terms of the bit size of its register. An M-bit device will output an M-bit binary number. The resolution of such a device is given by $E_{FSR}/2^M$.

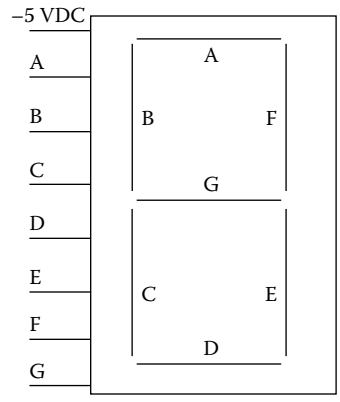


FIGURE 3.7 Seven-segment display chip can display any digit from 0 to 9.

3.5 Input Impedance

In the ideal sense, the very act of measurement should not alter the value of the measured signal. Any such alteration is a loading error. Loading errors can occur at any junction along the signal chain but can be minimized by impedance matching of the source with the measuring instrument. The measuring instrument input impedance controls the energy that is drawn from the source, or measured system, by a measuring instrument. The power loss through the measuring instrument is estimated by $P = E^2/Z_2$ where Z_2 is the input impedance of the measuring instrument and E is the source voltage potential being measured. Thus, to minimize the power loss, the input impedance should be large.

This same logic holds for the two instruments in a signal chain as the subsequent instrument draws energy from the previous instrument in the chain. As a general example, consider the situation in Figure 3.8 in which the output signal from one instrument provides the input signal to a subsequent device in a signal chain. The open circuit potential, E_1, is present at the output terminal of source device 1 having output impedance, Z_1. Device 2 has input impedance Z_2 at its input terminals. Connecting the output terminals of Device 1 to the input terminals of Device 2 creates the equivalent circuit also shown in Figure 3.7. The potential actually sensed by Device 2 will be

$$E_2 = E_1 \frac{1}{1 + Z_1/Z_2}$$

The difference between the actual potential E_1 at the output terminals of Device 1 and the measured potential E_2 is a loading error brought on by the input impedance of measuring Device 2. It is clear that a high input impedance Z_2 relative to Z_1 minimizes this error. A general rule is for the input impedance to be at least 100 times the source impedance to reduce the loading error to 1%.

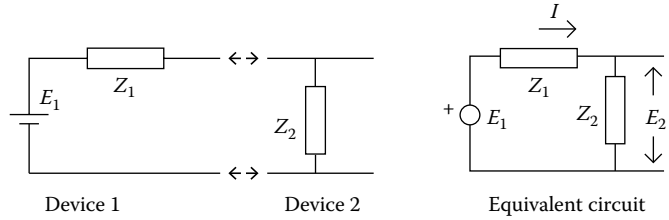


FIGURE 3.8 Equivalent circuit is formed by applying a measuring instrument to the output terminals of an instrument.

In general, null instruments and null methods will minimize loading errors. They provide the equivalent of a very high input impedance to the measurement, minimizing energy drain from the measured system. Deflection instruments and deflection measuring techniques will derive energy from the process being measured and therefore require attention to proper selection of input impedance.

Defining Terms

Analog sensor: Sensors that output a signal that is continuous in both magnitude and time (or space).

Deflection instrument: A measuring device whose output deflects proportional to the magnitude of the measurand.

Digital sensor: Sensors that output a signal that is discrete (noncontinuous) in time and/or magnitude.

Input impedance: The impedance measured across the input terminals of a device.

Loading error: That difference between the measurand and the measuring system output attributed to the act of measuring the measurand.

Measurand: A physical quantity, property, or condition being measured. Often, it is referred to as a measured value.

Null instrument: A measuring device that balances the measurand against a known value, thus achieving a null condition. A null instrument minimizes measurement loading errors.

Readout: This is the display of a measuring system.

Resolution: This is the least count or smallest detectable change in measurand capable.

Sensor: The portion of a measurement system that responds directly to the physical variable being measured.

Further Information

Doebelin, E. O., *Measurement Systems,* 4th edn., New York: McGraw-Hill, 1990.

Figliola, R. S. and D. E. Beasley, *Theory and Design for Mechanical Measurements,* 5th edn., New York: Wiley, 2011.

Northrop, R. B., *Introduction to Instrumentation and Measurement,* Boca Raton, FL: CRC Press, 2005.

Wobschall, D., *Circuit Design for Electronic Instrumentation: Analog and Digital Devices from Sensor to Display,* 2nd edn., New York: McGraw-Hill, 1987.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

4

Static and Dynamic Characteristics of Instrumentation

4.1 Static Characteristics of Instrument Systems **4**-4
Output/Input Relationship • Drift • Hysteresis and Backlash • Saturation • Bias • Error of Nonlinearity

4.2 Dynamic Characteristics of Instrument Systems **4**-11
Dealing with Dynamic States • Forcing Functions • Characteristic Equation Development • Response of the Different Linear Systems Types • Zero-Order Blocks • First-Order Blocks • Second-Order Blocks

4.3 Calibration of Measurements **4**-18
Self-Calibrations

References **4**-21

Peter H. Sydenham
University of South Australia

Before we can begin to develop an understanding of the static and time-changing characteristics of measurements, it is necessary to build a framework for understanding the process involved, setting down the main words used to describe concepts as we progress.

Measurement is the process by which relevant information about a system of interest is interpreted using the human thinking ability to define what is believed to be the new knowledge gained. This information may be obtained for purposes of controlling the behavior of the system (as in engineering applications) or for learning more about it (as in scientific investigations).

The basic entity needed to develop the knowledge is called *data*, and it is obtained with physical assemblies known as sensors that are used to observe or sense system variables. The terms *information* and *knowledge* tend to be used interchangeably to describe the entity resulting after data from one or more sensors have been processed to give more meaningful understanding. The individual variables being sensed are called *measurands*.

The most obvious way to make observations is to use the human senses of seeing, feeling, and hearing. This is often quite adequate or may be the only means possible. In many cases, however, sensors are used that have been devised by man to enhance or replace our natural sensors. The number and variety of sensors is very large indeed. Examples of man-made sensors are those used to measure temperature, pressure, or length. The process of sensing is often called *transduction*, being made with transducers. These man-made sensor assemblies, when coupled with the means to process the data into knowledge, are generally known as (measuring) instrumentation.

The degree of perfection of a measurement can only be determined if the goal of the measurement can be defined without error. Furthermore, instrumentation cannot be made to operate perfectly. Because of these two reasons alone, measuring instrumentation cannot give ideal sensing performance, and it must be selected to suit the allowable error in a given situation.

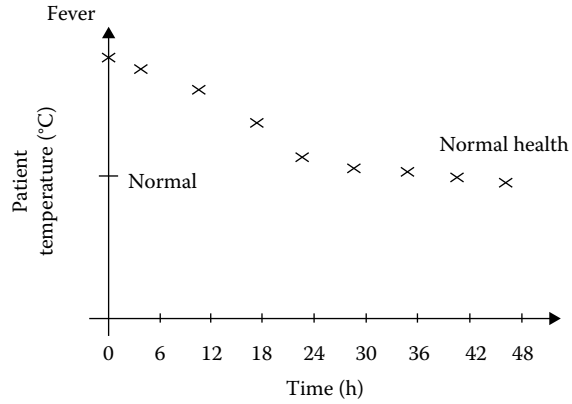


FIGURE 4.1 A patient's temperature chart shows changes taking place over time.

Measurement is a process of mapping actually occurring variables into equivalent values. Deviations from perfect measurement mappings are called *errors*: what we get as the result of measurement is not exactly what is being measured. A certain amount of error is allowable provided it is below the level of uncertainty we can accept in a given situation. As an example, consider two different needs to measure the measurand, time. The uncertainty to which we must measure it for daily purposes of attending a meeting is around a min in 24 h. In orbiting satellite control, the time uncertainty needed must be as small as milliseconds in years. Instrumentation used for the former case costs a few dollars and is the watch we wear; the latter instrumentation costs thousands of dollars and is the size of a suitcase.

We often record measurand values as though they are constant entities, but they usually change in value as time passes. These "dynamic" variations will occur either as changes in the measurand itself or where the measuring instrumentation takes time to follow the changes in the measurand—in which case it may introduce unacceptable error.

For example, when a fever thermometer is used to measure a person's body temperature, we are looking to see if the person is at the normally expected value and if it is not, to then look for changes over time as an indicator of his or her health. Figure 4.1 shows a chart of a patient's temperature. Obviously, if the thermometer gives errors in its use, wrong conclusions could be drawn. It could be in error due to incorrect calibration of the thermometer or because no allowance for the dynamic response of the thermometer itself was made.

Instrumentation, therefore, will only give adequately correct information if we understand the static and dynamic characteristics of both the measurand and the instrumentation. This, in turn, allows us to then decide if the error arising is small enough to accept.

As an example, consider the electronic signal amplifier in a sound system. It will be commonly quoted as having an amplification constant after feedback if applied to the basic amplifier of, say, 10. The actual amplification value is dependent on the frequency of the input signal, usually falling off as the frequency increases. The frequency response of the basic amplifier, before it is configured with feedback that markedly alters the response and lowers the amplification to get a stable operation, is shown as a graph of amplification gain versus input frequency. An example of the open-loop gain of the basic amplifier is given in Figure 4.2. This lack of uniform gain over the frequency range results in error—the sound output is not a true enough representation of the input.

Before we can delve more deeply into the static and dynamic characteristics of instrumentation, it is necessary to understand the difference in meaning between several basic terms used to describe the results of a measurement activity.

The correct terms to use are set down in documents called *standards*. Several standardized metrology terminologies exist, but they are not consistent. It will be found that books on instrumentation and statements of instrument performance often use terms in different ways. Users of measurement information need to be constantly diligent in making sure that the statements made are interpreted correctly.

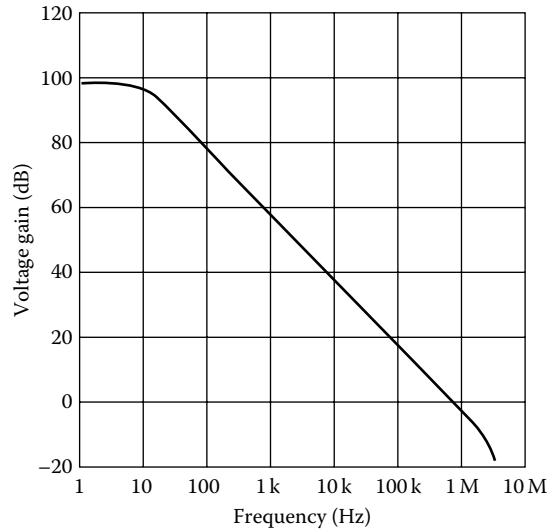


FIGURE 4.2 This graph shows how the amplification of an amplifier changes with input frequency.

The three companion concepts about a measurement that need to be well understood are its *discrimination*, its *precision*, and its *accuracy*. These are too often used interchangeably—which is quite wrong to do because they cover quite different concepts, as will now be explained.

When making a measurement, the smallest increment that can be discerned is called the *discrimination*. (Although now officially declared as wrong to use, the term *resolution* still finds its way into books and reports as meaning discrimination.) The discrimination of a measurement is important to know because it tells if the sensing process is able to sense fine enough changes of the measurand.

Even if the discrimination is satisfactory, the value obtained from a repeated measurement will rarely give exactly the same value each time the same measurement is made under conditions of constant value of measurand. This is because errors arise in real systems. The spread of values obtained indicates the precision of the set of the measurements. The word *precision* is not a word describing a quality of the measurement and is incorrectly used as such. Two terms that should be used here are *repeatability*, which describes the variation for a set of measurements made in a very short period, and the *reproducibility*, which is the same concept but now used for measurements made over a long period. As these terms describe the outcome of a set of values, there is a need to be able to quote a single value to describe the overall result of the set. This is done using statistical methods that provide for calculation of the "mean value" of the set and the associated spread of values, called its *variance*.

The *accuracy* of a measurement is covered in more depth elsewhere, so only an introduction to it is required here. Accuracy is the closeness of a measurement to the value defined to be the true value. This concept will become clearer when the following illustrative example is studied, for it brings together the three terms into a single perspective of a typical measurement.

Consider then the situation of scoring an archer shooting arrows into a target as shown in Figure 4.3a. The target has a central point—the bull's-eye. The objective for a perfect result is to get all arrows into the bull's-eye. The rings around the bull's-eye allow us to set up numeric measures of less-perfect shooting performance.

Discrimination is the distance at which we can just distinguish (i.e., discriminate) the placement of one arrow from another when they are very close. For an arrow, it is the thickness of the hole that decides the discrimination. Two close-by positions of the two arrows in Figure 4.3a cannot be separated easily. Use of thinner arrows would allow finer detail to be decided.

Repeatability is determined by measuring the spread of values of a set of arrows fired into the target over a short period. The smaller the spread, the more precise is the shooter. The shooter in Figure 4.3a is more precise than the shooter in Figure 4.3b.

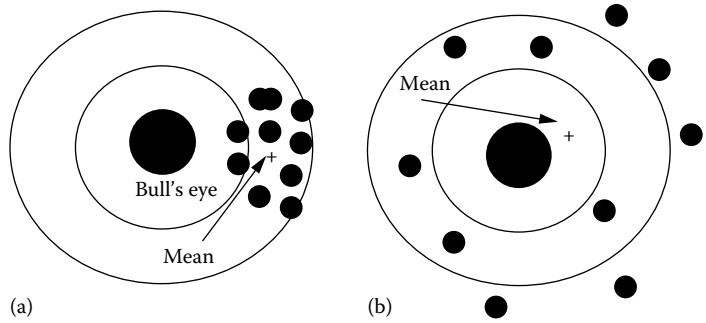


FIGURE 4.3 Two sets of arrow shots fired into a target allow understanding of the measurement concepts of discrimination, precision, and accuracy. (a) The target used for shooting arrows allows investigation of the terms used to describe the measurement result. (b) A different set of placements.

If the shooter returned to shoot each day over a long period, the results may not be the same each time for a shoot made over a short period. The mean and variance of the values are now called the *reproducibility* of the archer's performance.

Accuracy remains to be explained. This number describes how well the mean (the average) value of the shots sits with respect to the bull's-eye position. The set in Figure 4.3b is more accurate than the set in Figure 4.3a because the mean is nearer the bull's-eye (but less precise).

At first sight, it might seem that the three concepts of discrimination, precision, and accuracy have a strict relationship in that a better measurement is always that with all three aspects made as high as is affordable. This is not so. They need to be set up to suit the needs of the application.

We are now in a position to explore the commonly met terms used to describe aspects of the static and the dynamic performance of measuring instrumentation.

4.1 Static Characteristics of Instrument Systems

4.1.1 Output/Input Relationship

Instrument systems are usually built up from a serial linkage of distinguishable building blocks. The actual physical assembly may not appear to be so, but it can be broken down into a representative diagram of connected blocks. Figure 4.4 shows the block diagram representation of a humidity sensor. The sensor is activated by an input physical parameter and provides an output signal to the next block that processes the signal into a more appropriate state.

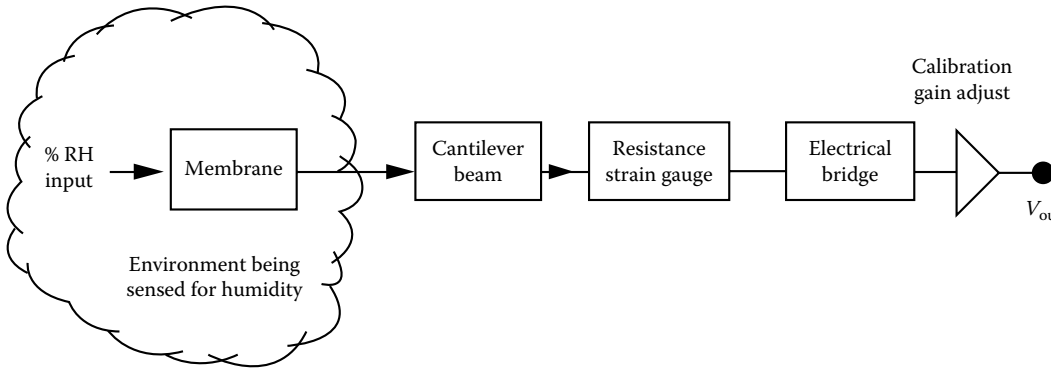


FIGURE 4.4 Instruments are formed from a connection of blocks. Each block can be represented by a conceptual and mathematical model. This example is of one type of humidity sensor.

A key generic entity is, therefore, the relationship between the input and output of the block. As was pointed out earlier, all signals have a time characteristic, so we must consider the behavior of a block in terms of both the static and dynamic states.

The behavior of the static regime alone and the combined static and dynamic regime can be found through use of an appropriate mathematical model of each block. The mathematical description of system responses is easy to set up and use if the elements all act as linear systems and where addition of signals can be carried out in a linear additive manner. If nonlinearity exists in elements, then it becomes considerably more difficult—perhaps even quite impractical—to provide an easy to follow mathematical explanation. Fortunately, general description of instrument systems responses can be usually adequately covered using the linear treatment.

The output/input ratio of the whole cascaded chain of blocks 1, 2, 3, etc. is given as

$$\left[\frac{\text{Output}}{\text{Input}}\right]_{\text{total}} = \left[\frac{\text{Output}}{\text{Input}}\right]_1 \times \left[\frac{\text{Output}}{\text{Input}}\right]_2 \times \left[\frac{\text{Output}}{\text{Input}}\right]_3 \ldots$$

The output/input ratio of a block that includes both the static and dynamic characteristics is called the *transfer function* and is given the symbol *G*.

The equation for *G* can be written as two parts multiplied together. One expresses the static behavior of the block, that is, the value it has after all transient (time-varying) effects have settled to their final state. The other part tells us how that value responds when the block is in its dynamic state. The static part is known as the *transfer characteristic* and is often all that is needed to be known for block description.

The static and dynamic response of the cascade of blocks is simply the multiplication of all individual blocks. As each block has its own part for the static and dynamic behavior, the cascade equations can be rearranged to separate the static from the dynamic parts, and then by multiplying the static set and the dynamic set, we get the overall response in the static and dynamic states. This is shown by the sequence of Equations 4.1 through 4.4:

$$G_{\text{total}} = G_1 \times G_2 \times G_3 \ldots \tag{4.1}$$

$$= [\text{Static} \times \text{Dynamic}]_1 \times [\text{Static} \times \text{Dynamic}]_2 \times [\text{Static} \times \text{Dynamic}]_3 \ldots \tag{4.2}$$

$$= [\text{Static}]_1 \times [\text{Static}]_2 \times [\text{Static}]_3 \ldots \times [\text{Dynamic}]_1 \times [\text{Dynamic}]_2 \times [\text{Dynamic}]_3 \ldots \tag{4.3}$$

$$= [\text{Static}]_{\text{total}} \times [\text{Dynamic}]_{\text{total}} \tag{4.4}$$

An example will clarify this. A mercury-in-glass fever thermometer is placed in a patient's mouth. The indication slowly rises along the glass tube to reach the final value, the body temperature of the person. The slow rise seen in the indication is due to the time it takes for the mercury to heat up and expand up the tube. The static *sensitivity* will be expressed as so many scale divisions per degree and is all that is of interest in this application. The dynamic characteristic will be a time-varying function that settles to unity after the transient effects have settled. This is merely an annoyance in this application but has to be allowed by waiting long enough before taking a reading. The wrong value will be viewed if taken before the transient has settled.

At this stage, we will now consider only the nature of the static characteristics of a chain; dynamic response is examined later.

If a sensor is the first stage of the chain, the static value of the gain for that stage is called the *sensitivity*. Where a sensor is not at the input, it is called the *amplification factor* or *gain*. It can take a value less than unity where it is then called the *attenuation*.

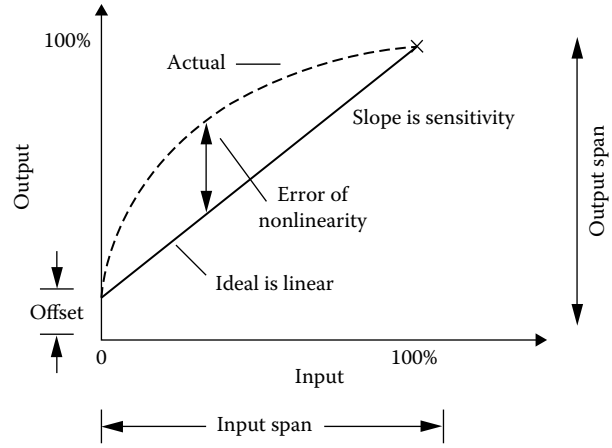


FIGURE 4.5 The graph relating input to output variables for an instrument block shows several distinctive static performance characteristics.

Sometimes, the instantaneous value of the signal is rapidly changing, yet the measurement aspect part is static. This arises when using AC signals in some forms of instrumentation where the amplitude of the waveform, not its frequency, is of interest. Here, the static value is referred to as its *steady-state* transfer characteristic.

Sensitivity may be found from a plot of the input and output signals, wherein it is the slope of the graph. Such a graph (see Figure 4.5) tells much about the static behavior of the block.

The intercept value on the *y-axis* is the *offset* value being the output when the input is set to zero. Offset is not usually a desired situation and is seen as an error quantity. Where it is deliberately set up, it is called the *bias*.

The range on the *x*-axis, from zero to a safe maximum for use, is called the *range* or *span* and is often expressed as the zone between the 0% and 100% points. The ratio of the span that the output will cover for the related input range is known as the *dynamic range*. This can be a confusing term because it does not describe dynamic time behavior. It is particularly useful when describing the capability of such instruments as flow rate sensors—a simple orifice plate type may only be able to handle dynamic ranges of 3–4, whereas the laser Doppler method covers as much as 10^7 variation.

4.1.2 Drift

It is now necessary to consider a major problem of instrument performance called *instrument drift*. This is caused by variations taking place in the parts of the instrumentation over time. Prime sources occur as chemical structural changes and changing mechanical stresses. Drift is a complex phenomenon for which the observed effects are that the sensitivity and offset values vary. It also can alter the accuracy of the instrument differently at the various amplitudes of the signal present.

Detailed description of drift is not at all easy, but it is possible to work satisfactorily with simplified values that give the average of a set of observations, this usually being quoted in a conservative manner. The first graph (*a*) in Figure 4.6 shows typical steady drift of a measuring spring component of a weighing balance. Figure 4.6b shows how an electronic amplifier might settle down after being turned on.

Drift is also caused by variations in environmental parameters such as temperature, pressure, and humidity that operate on the components. These are known as *influence parameters*. An example is the change of the resistance of an electrical resistor, this resistor forming the critical part of an electronic amplifier that sets its gain as its operating temperature changes.

Unfortunately, the observed effects of influence parameter–induced drift often are the same as for time-varying drift. Appropriate testing of blocks such as electronic amplifiers does allow the two to be

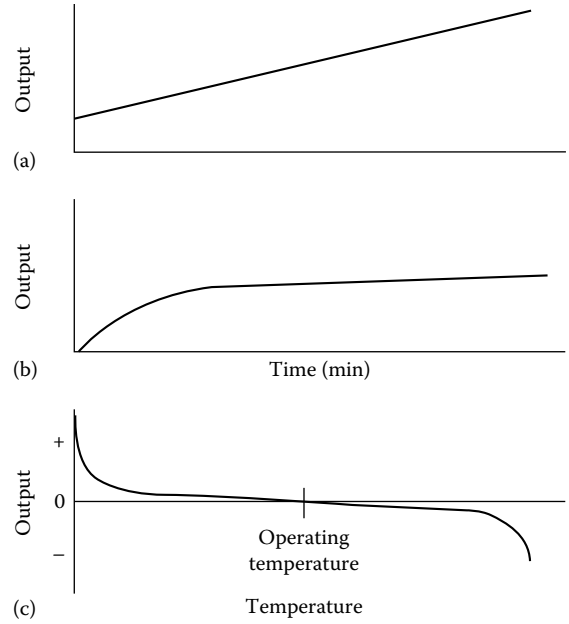


FIGURE 4.6 Drift in the performance of an instrument takes many forms: (a) drift over time for a spring balance, (b) how an electronic amplifier might settle over time to a final value after power is supplied, and (c) drift, due to temperature, of an electronic amplifier varies with the actual temperature of operation.

separated to some extent. For example, altering only the temperature of the amplifier over a short period will quickly show its temperature dependence.

Drift due to influence parameters is graphed in much the same way as for time drift. Figure 4.6c shows the drift of an amplifier as temperature varies. Note that it depends significantly on the temperature of operation, implying that the best designs are built to operate at temperatures where the effect is minimum.

Careful consideration of the time and influence parameter causes of drift shows they are interrelated and often impossible to separate. Instrument designers are usually able to allow for these effects, but the cost of doing this rises sharply as the error level that can be tolerated is reduced.

4.1.3 Hysteresis and Backlash

Careful observation of the output/input relationship of a block will sometimes reveal different results as the signals vary in the direction of the movement. Mechanical systems will often show a small difference in length as the direction of the applied force is reversed. The same effect arises as a magnetic field is reversed in a magnetic material. This characteristic is called *hysteresis*. Figure 4.7 is a generalized plot of the output/input relationship showing that a closed loop occurs. The effect usually gets smaller as the amplitude of successive excursions is reduced, this being one way to tolerate the effect. It is present in most materials. Special materials have been developed that exhibit low hysteresis for their application—transformer iron laminations and clock spring wire being examples.

Where this is caused by a mechanism that gives a sharp change, such as caused by the looseness of a joint in a mechanical joint, it is easy to detect and is known as *backlash*.

4.1.4 Saturation

So far, the discussion has been limited to signal levels that lie within acceptable ranges of amplitude. Real system blocks will sometimes have input signal levels that are larger than allowed. Here, the dominant errors that arise—*saturation* and *crossover distortion*—are investigated.

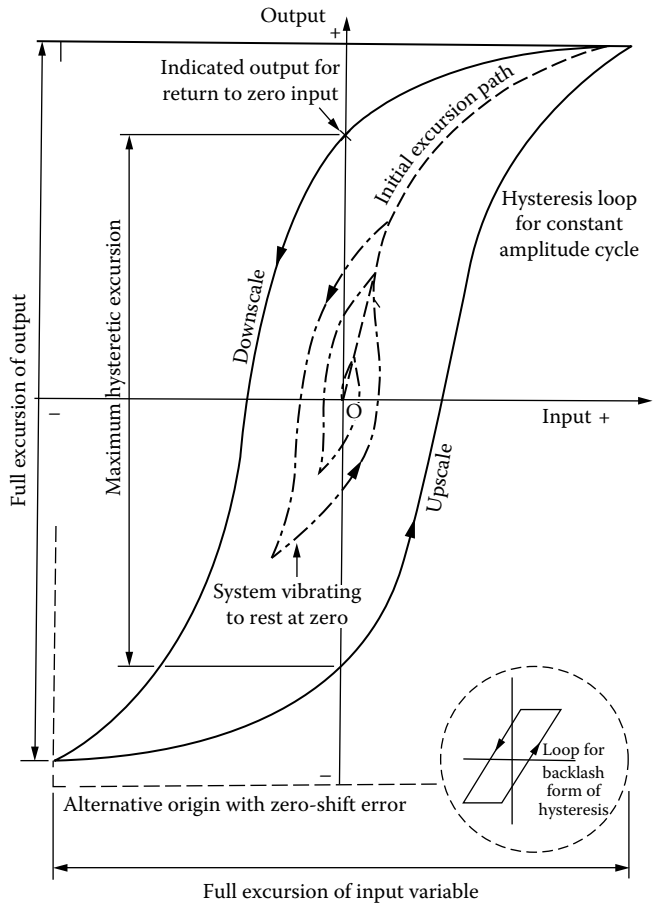


FIGURE 4.7 Generalized graph of output/input relationship where hysteresis is present. (Sydenham, P.H.: *Handbook of Measurement Science, Vol. 2*. 1983. Copyright Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced with permission.)

As mentioned earlier, the information-bearing property of the signal can be carried as the instantaneous value of the signal or be carried as some characteristic of a rapidly varying AC signal. If the signal form is not amplified faithfully, the output will not have the same linearity and characteristics.

The gain of a block will usually fall off with increasing size of signal amplitude. A varying amplitude input signal, such as the steadily rising linear signal shown in Figure 4.8, will be amplified differently according to the gain/amplitude curve of the block. In uncompensated electronic amplifiers, the larger amplitudes are usually less amplified than at the median points.

At very low levels of input signal, two unwanted effects may arise. The first is that small signals are often amplified more than at the median levels. The second error characteristic arises in electronic amplifiers because the semiconductor elements possess a dead zone in which no output occurs until a small threshold is exceeded. This effect causes crossover distortion in amplifiers.

If the signal is an AC waveform (see Figure 4.9) then the different levels of a cycle of the signal may not all be amplified equally. Figure 4.9a shows what occurs because the basic electronic amplifying elements are only able to amplify one polarity of signal. The signal is said to be *rectified*. Figure 4.9b shows the effect when the signal is too large and the top is not amplified. This is called *saturation* or *clipping*. (As with many physical effects, this effect is sometimes deliberately invoked in circuitry, an example being where it is used as a simple means to convert sine-waveform signals into a square waveform.) Crossover distortion is evident in Figure 4.9c as the signal passes from negative to positive polarity.

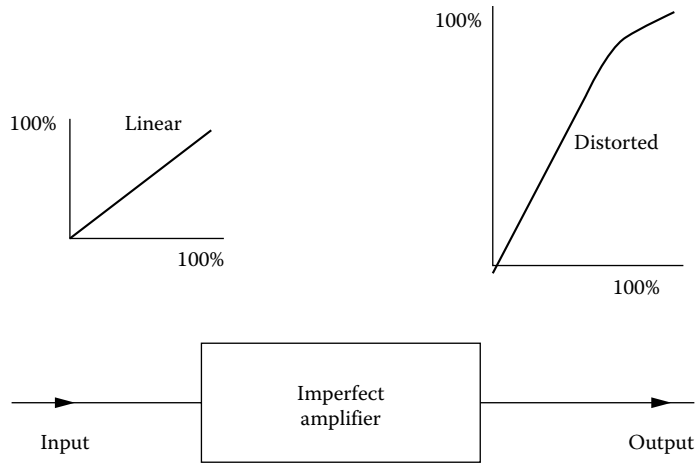


FIGURE 4.8 Nonlinear amplification can give rise to unwanted output distortion.

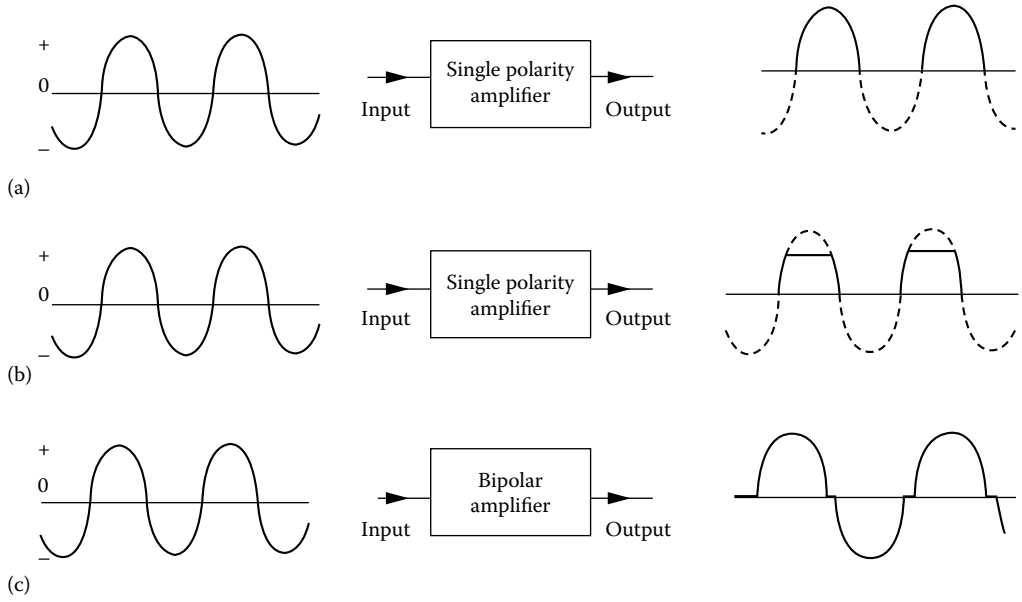


FIGURE 4.9 Blocks can incorrectly alter the shape of waveforms if saturation and crossover effects are not controlled: (a) rectification, (b) saturation, and (c) crossover distortion.

Where input signals are small, such as in sensitive sensor use, the form of analysis called *small signal* behavior is needed to reveal distortions. If the signals are comparatively large, as for digital signal considerations, a *large signal* analysis is used. Design difficulties arise when signals cover a wide dynamic range because it is not easy to allow for all of the various effects in a single design.

4.1.5 Bias

Sometimes, the electronic signal-processing situation calls for the input signal to be processed at a higher average voltage or current than arises normally. Here a DC value is added to the input signal to raise the level to a higher state as shown in Figure 4.10. A need for this is met where only one polarity of signal can be amplified by a single semiconductor element. Raising the level of all of the waveform equally takes all

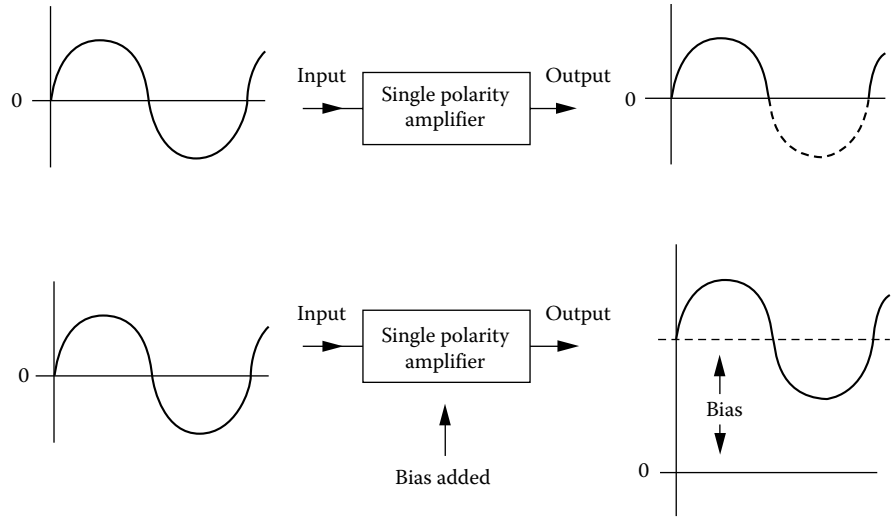


FIGURE 4.10 Bias is where a signal has all of its value raised by an equal amount. Shown here is an AC input waveform biased to be all of positive polarity.

parts into the reasonably linear zone of an amplifier, allowing more faithful replication. If bias were not used here, then the lower half cycle would not be amplified, resulting in only the top half appearing in the output.

4.1.6 Error of Nonlinearity

Ideally, it is often desired that a strictly linear relationship exists between input and output signals in amplifiers. Practical units, however, will always have some degree of nonconformity, which is called the *nonlinearity*. If an instrument block has constant gain for all input signal levels, then the relationship graphing the input against the output will be a straight line; the relationship is then said to be linear. *Linearity* is the general term used to describe how close the actual response is compared with that ideal line. The correct way to describe the error here is as the *error of nonlinearity*. Note, however, that not all responses are required to be linear; another common one follows a logarithmic relationship.

Detailed description of this error is not easy, for that would need a statement of the error values at all points of the plot. Practice has found that a shorthand statement can be made by quoting the maximum departure from the ideal as a ratio formed with the 100% value.

Difficulties arise in expressing error of nonlinearity, for there exist many ways to express this error. Figure 4.11 shows the four cases that usually arise. The difference arises in the way in which the ideal (called the "best fit") straight line can be set up. Figure 4.11a shows the line positioned by the usually calculated statistical averaging method of least squares fit; other forms of line-fitting calculation are also used. This will yield the smallest magnitude of error calculation for the various kinds of line fitting but may not be appropriate for how the stage under assessment is used. Other, possibly more reasonable, options exist. Figure 4.11b constrains the best-fit line to pass through the zero point. Figure 4.11c places the line between the expected 0% and the 100% points. There is still one more option, that where the theoretical line is not necessarily one of the above yet is the ideal placement (Figure 4.11d).

In practice then, instrument systems linearity can be expressed in several ways. Good certification practice requires that the method used to ascertain the error is stated along with the numerical result, but this is often not done. Note also that the error is the worst case and that part of the response may be much more linear.

The description of instrument performance is not a simple task. To accomplish this fully would require very detailed statements recording the performance at each and every point. That is often

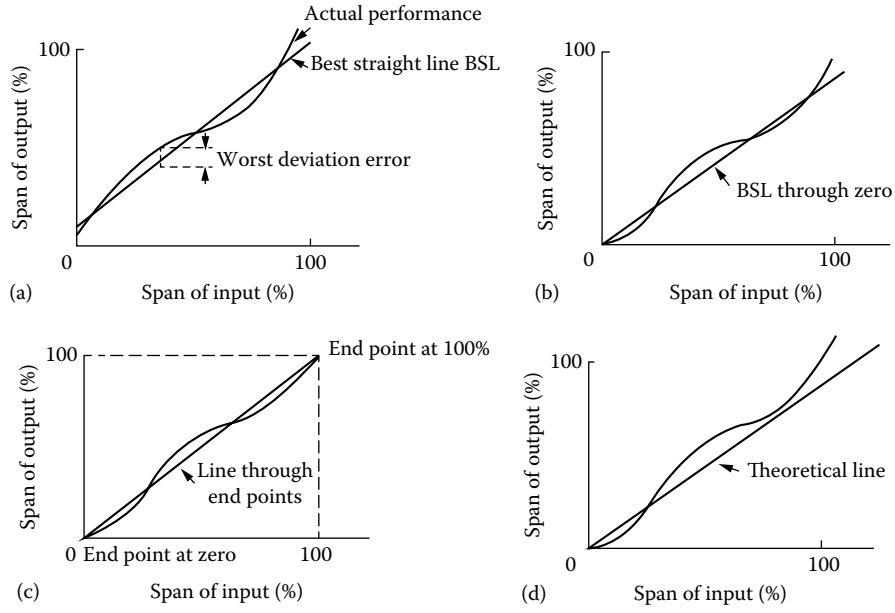


FIGURE 4.11 Error of nonlinearity can be expressed in four different ways: (a) best-fit line (based on selected method used to decide this); (b) best-fit line through zero; (c) line joining 0% and 100% points; and (d) theoretical line. (Sydenham, P.H.: *Handbook of Measurement Science, Vol. 2*. 1983. Copyright Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced with permission.)

too cumbersome, so the instrument industry has developed many short-form statements that provide an adequate guide to the performance. This guide will be seen to be generally a conservative statement.

Many other descriptors exist for the static regime of an instrument. The reader is referred to the many standards documents that exist on instrument terminology; for example, see Ref. [3].

4.2 Dynamic Characteristics of Instrument Systems

4.2.1 Dealing with Dynamic States

Measurement outcomes are rarely static over time. They will possess a dynamic component that must be understood for correct interpretation of the results. For example, a trace made on an ink pen chart recorder will be subject to the speed at which the pen can follow the input signal changes.

To properly appreciate instrumentation design and its use, it is now necessary to develop insight into the most commonly encountered types of dynamic response and to develop the mathematical modeling basis that allows us to make concise statements about responses.

If the transfer relationship for a block follows linear laws of performance, then a generic mathematical method of dynamic description can be used. Unfortunately, simple mathematical methods have not been found that can describe all types of instrument responses in a simplistic and uniform manner. If the behavior is nonlinear, then description with mathematical models becomes very difficult and might be impracticable. The behavior of nonlinear systems can, however, be studied as segments of linear behavior joined end to end. Here, digital computers are effectively used to model systems of any kind provided the user is prepared to spend time setting up an adequate model.

Now the mathematics used to describe linear dynamic systems can be introduced. This gives valuable insight into the expected behavior of instrumentation, and it is usually found that the response can be approximated as linear.

The modeled response at the output of a block G_{result} is obtained by multiplying the mathematical expression for the input signal G_{input} by the transfer function of the block under investigation $G_{response}$, as shown in Equation 4.5:

$$G_{result} = G_{input} \times G_{response} \tag{4.5}$$

To proceed, one needs to understand commonly encountered input functions and the various types of block characteristics. We begin with the former set: the so-called forcing functions.

4.2.2 Forcing Functions

Let us first develop an understanding of the various types of input signals used to perform tests. The most commonly used signals are shown in Figure 4.12. These each possess different valuable test features.

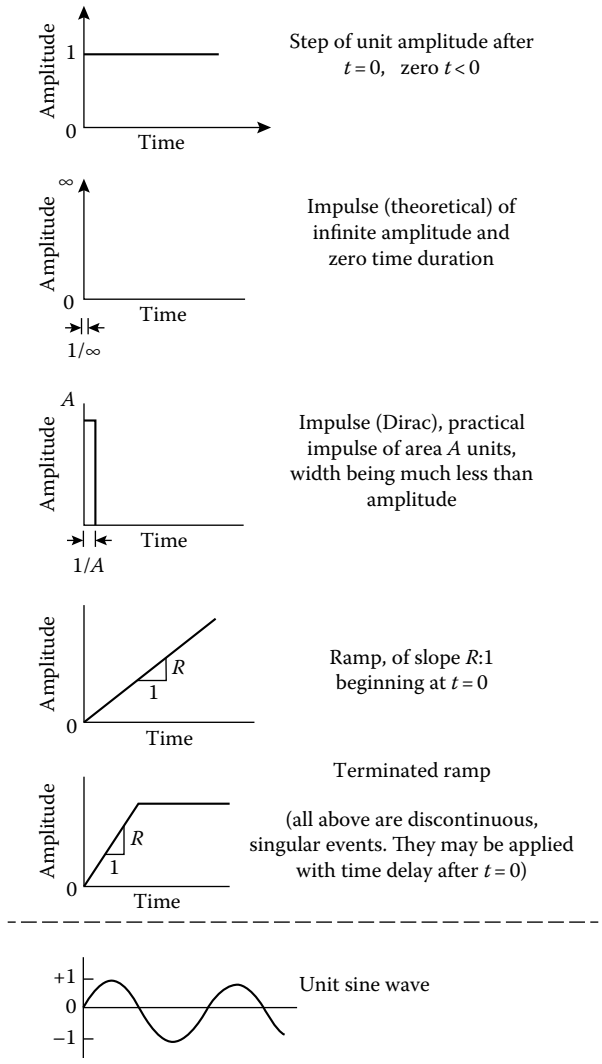


FIGURE 4.12 The dynamic response of a block can be investigated using a range of simple input forcing functions. (Sydenham, P.H.: *Handbook of Measurement Science, Vol. 2*. 1983. Copyright Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced with permission.)

For example, the sine wave is the basis of analysis of all complex wave shapes because they can be formed as a combination of various sine waves, each having individual responses that add to give all other wave shapes. The step function has intuitively obvious uses because input transients of this kind are commonly encountered. The ramp test function is used to present a more realistic input for those systems where it is not possible to obtain instantaneous step input changes, such as attempting to move a large mass by a limited size of force. Forcing functions are also chosen because they can be easily described by a simple mathematical expression, thus making mathematical analysis relatively straightforward.

4.2.3 Characteristic Equation Development

The behavior of a block that exhibits linear behavior is mathematically represented in the general form of expression given as Equation 4.6:

$$\cdots + \frac{a_2 d^2 y}{dt^2} + \frac{a_1 dy}{dt} + a_0 y = x(t) \tag{4.6}$$

Here, the coefficients a_2, a_1, and a_0 are constants dependent on the particular block of interest. The left-hand side of the equation is known as the *characteristic equation*. It is specific to the internal properties of the block and is not altered by the way the block is used.

The specific combination of forcing function input and block characteristic equation collectively decides the combined output response. Connections around the block, such as feedback from the output to the input, can alter the overall behavior significantly: such systems, however, are not dealt with in this section being in the domain of feedback control systems.

Solution of the combined behavior is obtained using Laplace transform methods to obtain the output responses in the time or the complex frequency domain. These mathematical methods might not be familiar to the reader, but this is not a serious difficulty, for the cases most encountered in practice are well documented in terms that are easily comprehended, the mathematical process having been performed to yield results that can be used without the same level of mathematical ability. More depth of explanation can be obtained from Ref. [1] or any one of the many texts on energy systems analysis. Space here only allows an introduction; this account is linked to Ref. [1] of Sydenham, Chapter 17, to allow the reader to access a fuller description where needed.

The next step in understanding block behavior is to investigate the nature of Equation 4.6 as the number of derivative terms in the expression increases (Equations 4.7 through 4.10):

$$\text{Zero order} \quad a_0 y = x(t) \tag{4.7}$$

$$\text{First order} \quad \frac{a_1 dy}{dt} + a_0 y = x(t) \tag{4.8}$$

$$\text{Second order} \quad \frac{a_2 d^2 y}{dt^2} + \frac{a_1 dy}{dt} + a_0 y = x(t) \tag{4.9}$$

$$n\text{th order} \quad \frac{a_n d^n y}{dt^n} + \frac{a_{n-1}\, d^{n-1} y}{dt^{n-1}} + \cdots + a_0 y = x(t) \tag{4.10}$$

Note that specific names have been given to each order. The zero-order situation is not usually dealt with in texts because it has no time-dependent term and is thus seen to be trivial. It is an amplifier (or attenuator) of the forcing function with gain of a_0. It has infinite bandwidth without change in the amplification constant.

The highest order usually necessary to consider in first-cut instrument analysis is the second-order class. Higher-order systems do occur in practice and need analysis that is not easily summarized here. They also need deep expertise in their study. Computer-aided tools for systems analysis can be used to study the responses of systems.

Another step is now to rewrite the equations after Laplace transformation into the frequency domain. We then get the set of output/input (Equations 4.11 through 4.14):

$$\text{Zero order} \quad \frac{Y(s)}{X(s)} = 1 \tag{4.11}$$

$$\text{First order} \quad \frac{Y(s)}{X(s)} = \frac{1}{(\tau s + 1)} \tag{4.12}$$

$$\text{Second order} \quad \frac{Y(s)}{X(s)} = \frac{1}{(\tau_1 s + 1)(\tau_2 s + 1)} \tag{4.13}$$

$$n\text{th order} \quad \frac{Y(s)}{X(s)} = \frac{1}{(\tau_1 s + 1)(\tau_2 s + 2)\ldots(\tau_n s + 1)} \tag{4.14}$$

The terms $\tau_1, \ldots, \tau_n$ are called *time constants*. They are key system performance parameters.

4.2.4 Response of the Different Linear Systems Types

Space restrictions do not allow a detailed study of all of the various options. A selection is presented to show how they are analyzed and reported, leading to how the instrumentation person can use certain standard charts in the study of the characteristics of blocks.

4.2.5 Zero-Order Blocks

To investigate the response of a block, multiply its frequency domain forms of equation for the characteristic equation with that of the chosen forcing function equation.

This is an interesting case because Equation 4.7 shows that the zero-order block has no frequency-dependent term (it has no time derivative term), so the output for all given inputs can only be of the same time form as the input. What can be changed is the amplitude given as the coefficient a_0. A shift in time (phase shift) of the output waveform with the input also does not occur as it can for the higher-order blocks.

This is the response often desired in instruments because it means that the block does not alter the time response. However, this is not always so because, in systems, design blocks are often chosen for their ability to change the time shape of signals in a known manner.

Although somewhat obvious (Figure 4.13), a resistive strain gage is given to illustrate zero-order behavior.

4.2.6 First-Order Blocks

Here, Equation 4.8 is the relevant characteristic equation. There is a time-dependent term, so analysis is needed to see how this type of block behaves under dynamic conditions. The output response is different for each type of forcing function applied. Space limitations only allow the most commonly encountered cases—the step and the sine-wave input—to be introduced here. It is also only possible here to outline the method of analysis and to give the standardized charts that plot generalized behavior.

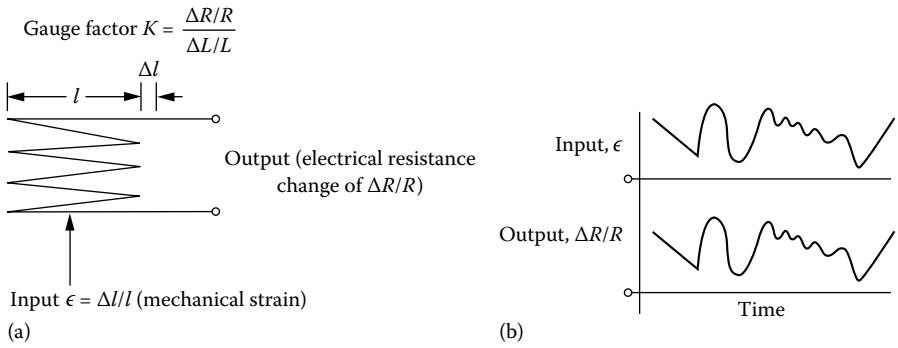


FIGURE 4.13 Input and output responses for a zero-order block: (a) strain gage physical and mathematical model and (b) responses. (Sydenham, P.H.: *Handbook of Measurement Science, Vol. 2*. 1983. Copyright Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced with permission.)

The step response of the first-order system is obtained by multiplying Equation 4.12 by the frequency domain equation for a step of amplitude A. The result is then transformed back into the time domain using Laplace transforms to yield the expression for the output, $y(t)$:

$$y(t) = AK(1 - e^{t/\tau}) \tag{4.15}$$

where

A is the amplitude of the step
K is the static gain of the first-order block
t is the time in consistent units
τ is the time constant associated with the block itself

This is a tidy outcome because Equation 4.15 covers the step response for all first-order blocks, thus allowing it to be graphed in normalized manner, as given in Figure 4.14. The shape of the response is

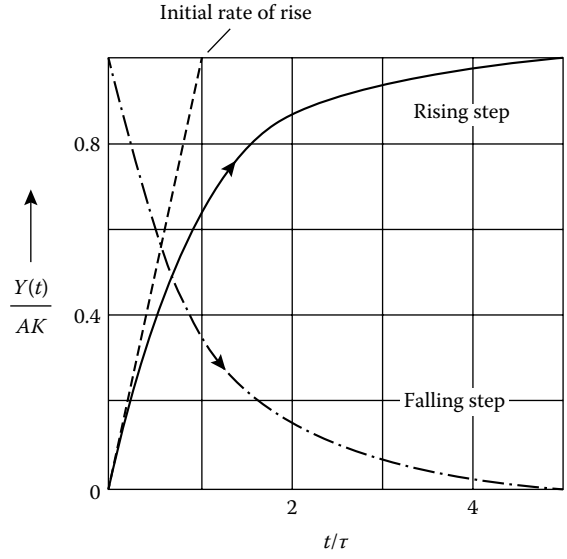


FIGURE 4.14 The step response for all first-order systems is covered by these two normalized graphs. (Sydenham, P.H.: *Handbook of Measurement Science, Vol. 2*. 1983. Copyright Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced with permission.)

always of the same form. This means that the step response of a first-order system can be described as having "a step of AK amplitude with the time constant τ."

If the input is a sine wave, the output response is quite different, but again, it will be found that there is a general solution for all situations of this kind. As before, the input forcing equation is multiplied by the characteristic equation for the first-order block and Laplace transformation is used to get back to the time-domain response. After rearrangement into two parts, this yields

$$y(t) = \left[\frac{AK\tau\omega e^{-t/\tau}}{(\tau^2\omega^2 + 1)}\right] + \left[\frac{AK}{(\tau^2\omega^2 + 1)^{1/2} \cdot \sin(\omega t + \phi)}\right] \tag{4.16}$$

where

ω is the signal frequency in angular radians, $\phi = \tan^{-1}(-\omega t)$
A is the amplitude of the sine-wave input
K is the gain of the first-order block
t is the time in consistent units
τ is the time constant associated with the block

The left side of the right-hand bracketed part is a short-lived, normally ignored, time transient that rapidly decays to zero, leaving a steady-state output that is the parameter of usual interest. Study of the steady-state part is best done by plotting it in a normalized way, as has been done in Figure 4.15.

These plots show that the amplitude of the output is always reduced as the frequency of the input signal rises and that there is always a phase lag action between the input and the output that can range from 0° to 90° but never can be more than 90°. The extent of these effects depends on the particular coefficients of the block and input signal. These effects must be well understood when interpreting measurement results because substantial errors can arise with using first-order systems in an instrument chain.

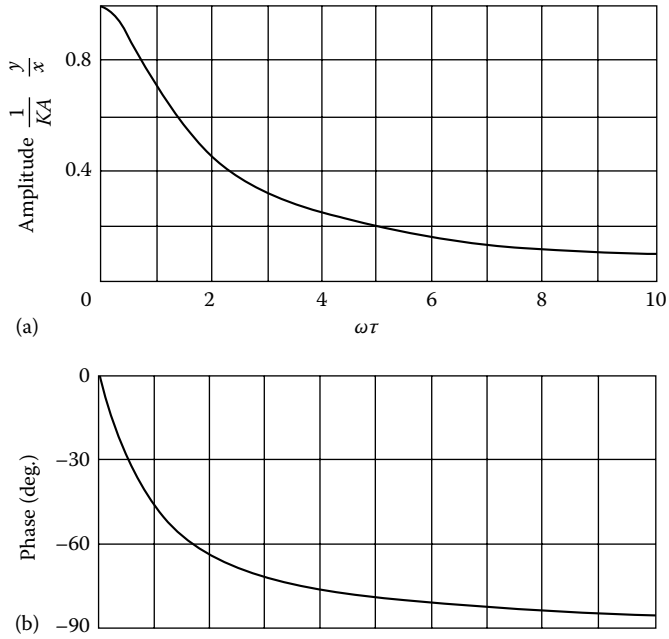


FIGURE 4.15 The amplitude and phase shift of the output of all first-order systems to a sine-wave input is shown by these two normalized curves: (a) amplitude and (b) phase. (Sydenham, P.H.: *Handbook of Measurement Science, Vol. 2*. 1983. Copyright Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced with permission.)

4.2.7 Second-Order Blocks

If the second-order differential term is present, the response of a block is quite different, again responding in quite a spectacular manner with features that can either be wanted or unwanted.

As before, to obtain the output response, the block's characteristic function is multiplied by the chosen forcing function. However, to make the results more meaningful, we first carry out some simple substitution transformations.

The steps begin by transforming the second-order differential Equation 4.6 into its Laplace form to obtain

$$X(s) = a_2 s^2 Y(s) + a_1 s Y(s) + a_0 Y(s) \tag{4.17}$$

This is then rearranged to yield

$$G(s) = \frac{Y(s)}{X(s)} = \frac{1}{a_0} \cdot \frac{1}{\left\{(a_2/a_0)s^2 + (a_1/a_0)s + 1\right\}} \tag{4.18}$$

The coefficients can then be expressed in system performance terms as follows:

$$\text{Angular natural frequency} \quad \omega_n = \left(\frac{a_0}{a_2}\right)^{1/2} \tag{4.19}$$

$$\text{Damping ratio} \quad \zeta = \frac{a_1}{2(a_0 \cdot a_2)^{1/2}} \tag{4.20}$$

$$\text{Static gain} \quad K = \frac{1}{a_0} \tag{4.21}$$

These three variables have practical relevance, as will be seen when the various responses are plotted.

Using these transformed variables, the characteristic equation can be rewritten in two forms ready for investigation of output behavior to step and sine-wave inputs, as

$$G(s) = \frac{K}{\left\{(1/\omega_n^2)s^2 + (2\zeta/\omega_n)s + 1\right\}} \tag{4.22}$$

and then as

$$G(s) = \frac{K}{(\tau^2 s^2 + 2\zeta\tau s + 1)} \tag{4.23}$$

We are now ready to consider the behavior of the second-order system to the various forcing inputs.

First consider the step input. After forming the product of the forcing and characteristic functions, the time-domain form can be plotted as shown in Figure 4.16.

This clearly shows that the response is strongly dependent on the damping ratio ζ, value. If it is less than unity, it exhibits an oscillatory movement settling down to the final value. If the damping value is greater than unity, the response moves to the final value without oscillation. The often preferred state is to use a damping factor of unity, *critical damping*. The choice of response depends strongly on the applications, for all levels of damping ratio have use in practice, ranging from needing an oscillation that never ceases (zero damping) to the other extreme where a very gradual rate of change is desired.

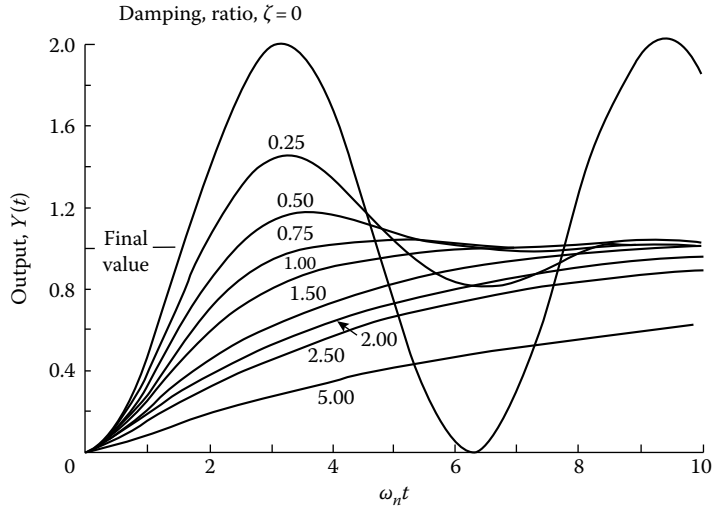


FIGURE 4.16 The response of second-order systems to a step input is seen from this normalized plot. (Sydenham, P.H.: *Handbook of Measurement Science, Vol. 2*. 1983. Copyright Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced with permission.)

A similar analysis is used to see how the second-order system responds to the sine-wave input. The two response plots obtained are shown in Figure 4.17: one for the amplitude response and the other showing how the phase shifts as the frequency changes.

The most unexpected result is seen at the point where the gain rises to infinity for the zero damping state. This is called *resonance* and it occurs at the block's *natural frequency* for the zero damping state. Resonance can be a desirable feature, as in detecting a particular frequency in a radio-frequency detection circuit, or it may be most undesirable, as when a mechanical system resonates, possibly to destruction. It can be seen that it is mostly controlled by the damping ratio. Note also that the phase shift for the second-order system ranges from 0° to 180°. This has important implications if the block is part of a feedback loop because as the frequency rises, the phase shift from the block will pass from stable negative feedback (less than 90°) to positive feedback (greater than 90°), causing unwanted oscillation.

More detail of the various other situations, including how to deal with higher orders, cascaded blocks of similar kind, and ramp inputs are covered elsewhere [1].

4.3 Calibration of Measurements

We have already introduced the concept of accuracy in making a measurement and how the uncertainty inherent in all measurements must be kept sufficiently small. The process and apparatus used to find out if a measurement is accurate enough is called *calibration*. It is achieved by comparing the result of a measurement with a method possessing a measurement performance that is generally agreed to have less uncertainty than that in the result obtained. The error arising within the calibration apparatus and process of comparison must necessarily be less than that required. This means that calibration is often an expensive process. Conducting a good calibration requires specialist expertise.

The method and apparatus for performing measurement instrumentation calibrations vary widely. An illustrative example of the comparison concept underlying them all is given in the calibration of flowmeters, shown diagrammatically in Figure 4.18.

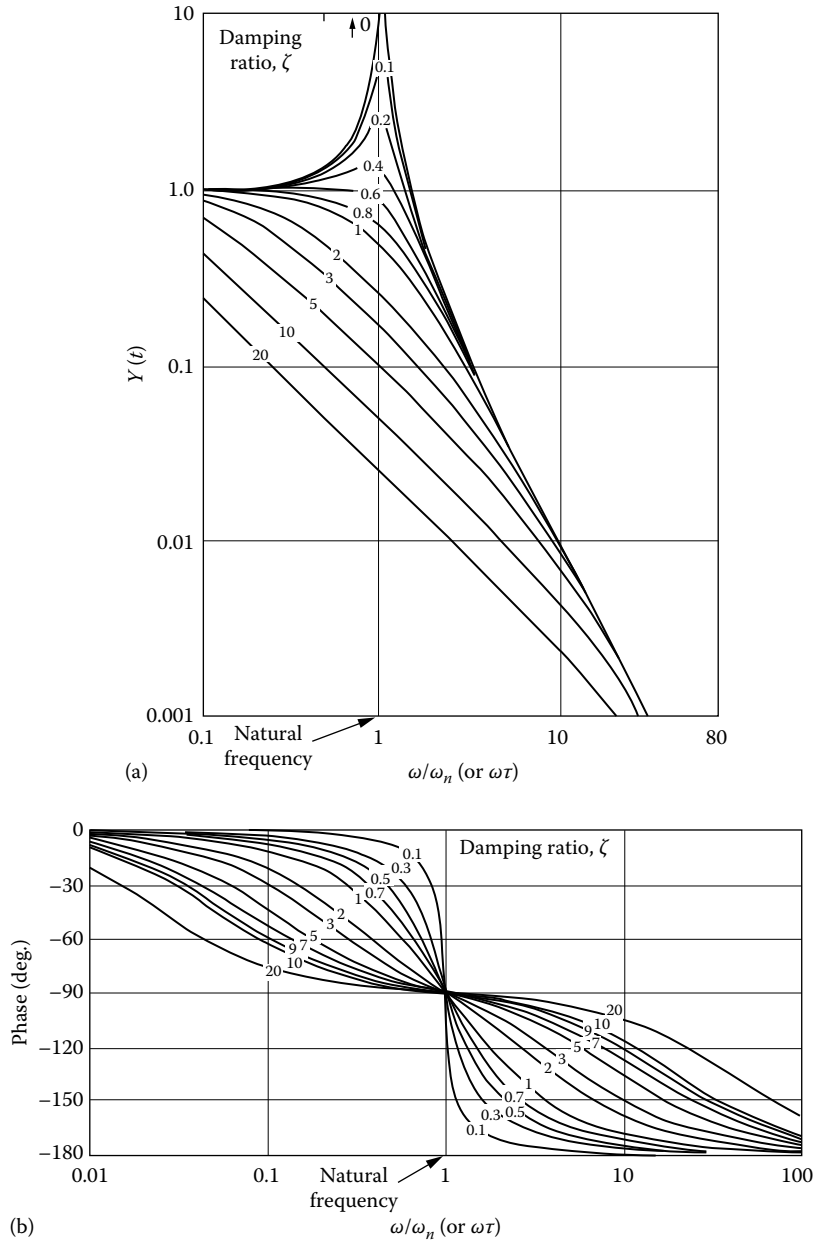


FIGURE 4.17 These two plots allow the behavior of second-order blocks with sine-wave inputs to be ascertained: (a) amplitude and (b) phase. (Sydenham, P.H.: *Handbook of Measurement Science, Vol. 2*. 1983. Copyright Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced with permission.)

By the use of an overflowing vessel, the top tank provides a flow of water that remains constant because it comes from a constant height. The meter to be calibrated is placed in the downstream pipe.

The downstream is either deflected into the weigh tank or back to the supply. To make a measurement, the water is first set to flow to the supply. At the start of a test period, the water is rapidly and precisely deflected into the tank. After a given period, the water is again sent back to the supply.

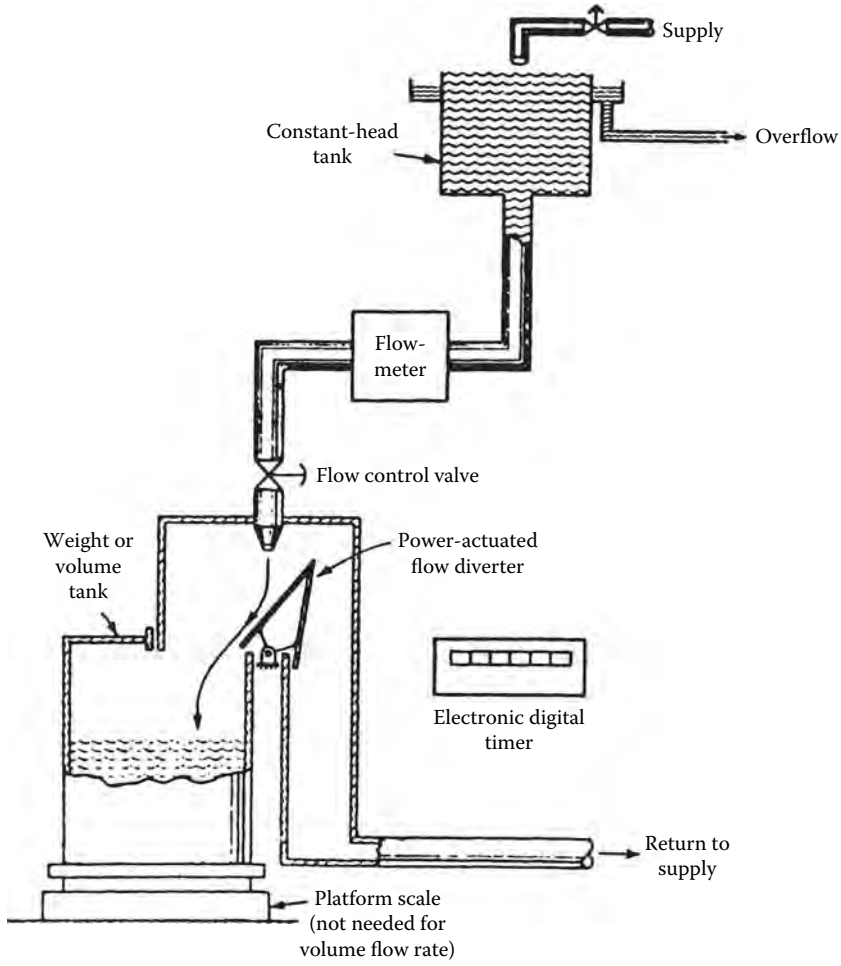


FIGURE 4.18 This practical example illustrates how flowmeters are calibrated by passing a known quantity of fluid through the meter over a given time. (From Sydenham, P. H., *Transducers in Measurement and Control*, Adam Hilger, Bristol, U.K., IOP Publishing, 1980.)

This then has filled the tank with a given amount of water for a given time period of flow. Calculations are then undertaken to work out the quantity of water flowing per unit time period, which is the *flow rate*. The meter was already registering a flow rate as a constant value. This is then compared with the weighed method to yield the error. Some thought will soon reveal many sources of error in the test apparatus, such as that the temperature of the water decides the volume that flows through, and thus this must be allowed for in the calculations.

It will also be clear that this calibration may not be carried out under the same conditions as the measurements are normally used. The art and science and difficulties inherent in carrying out quality calibration for temperature sensors are well exposed [2].

Calibration of instrumentation is a must, for, without it, measurement results may be misleading and lead to costly aftermath situations. Conducting good calibration adds overhead cost to measurement, but it is akin to taking out insurance. If that investment is made properly, it will assist in mitigating later penalties. For example, an incorrectly calibrated automatic cement batcher was used in making concrete for the structural frame of a multistory building. It took several days before concrete strength tests revealed the batcher had been out of calibration for a day with the result that the concrete already

poured for three floors was not of adequate strength. By then, more stories had been poured on top. The defective floors had to be fully replaced at great cost. More resource put into the calibration process would have ensured that the batcher was working properly.

4.3.1 Self-Calibrations

Due to advances in digital systems, most modern instruments have the self-calibrating capabilities. In self-calibrations, postmeasurement corrections are made, and the magnitudes of various errors are stored in the memory to be recalled and used in laboratory and in field applications. Further information on calibrations of instruments can be found in Chapter 8 and Refs. [4–6].

References

1. P. H. Sydenham, *Handbook of Measurement Science*, Vol. 2, Chichester, U.K.: John Wiley & Sons, 1983.
2. J. V. Nicholas and D. R. White, *Traceable Temperatures*, Chichester, U.K.: John Wiley & Sons, 1994.
3. British Standard Institution, *PD 6461: Vocabulary of Metrology*, London, U.K.: BSI, 1995.
4. S. V. Gupta, *Measurement Uncertainties: Physical Parameters and Calibration Instruments*, Berlin, Germany: Springer-Verlag, 2012.
5. Fluke, Calibration and Repair Services, http://www.fluke.com/fluke/usen/service/calibration/default.htm (accessed on March 13, 2013).
6. NIST, Calibrations, http://www.nist.gov/calibrations/ (accessed on March 13, 2013).
7. P. H. Sydenham, *Transducers in Measurement and Control*, Adam Hilger, Bristol, U.K., IOP Publishing, 1980.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

5

Measurement Accuracy

5.1 Error: The Normal Distribution and the Uniform Distribution ... **5**-1
Uncertainty (Accuracy)
5.2 Measurement Uncertainty Model ... **5**-4
Purpose • Classifying Error and Uncertainty Sources • ISO Classifications • Engineering Classifications • Random • Systematic • Symmetrical Systematic Standard Uncertainties
5.3 Calculation of Expanded Uncertainty ... **5**-10
ISO Total and Expanded Uncertainty • Engineering System Expanded Uncertainty • High Degrees of Freedom Approximation • Calculation Example
5.4 Summary ... **5**-13
Defining Terms ... **5**-13
References ... **5**-13
Further Information ... **5**-14

Ronald H. Dieck
Ron Dieck Associates, Inc.
Pratt & Whitney

All test measurements are taken so that data that are useful in decision making may be acquired. No tests are run and no measurements made when the "answer" is already known. For data to be useful, it is necessary that their measurement errors be small in comparison to the changes or effect under evaluation. Measurement error is unknown and unknowable. This chapter addresses the techniques used to estimate, with some confidence, the expected limits of the measurement errors.

5.1 Error: The Normal Distribution and the Uniform Distribution

Error is defined as the difference between the measured value and the true value of the measurand [1]. That is,

$$E = (\text{measured}) - (\text{true}) \tag{5.1}$$

where

E is the measurement error
(measured) is the value obtained by a measurement
(true) is the true value of the measurand

It is only possible to estimate, with some confidence, the expected limits of error. The most common method for estimating those limits is to use the *normal distribution* [2]. It is

$$Y = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2(X-\mu)^2/\sigma^2} \tag{5.2}$$

where

X is the input variable, here the value obtained by a measurement
μ is the average of the population of the X variable
σ is the standard deviation of the population

expressed as

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{N}} \tag{5.3}$$

where

X_i is the ith X measurement
N is the number of data points measured from the population

Typically, N, μ, and σ are not known.

Figure 5.1 illustrates this distribution. Here, for an infinite population ($N = \infty$), the standard deviation, σ, would be used to estimate the expected limits of a particular error with some confidence. That is, the average, plus or minus 2σ divided by the square root of the number of data points, would contain the true average, μ, 95% of the time.

However, in test measurements, one typically cannot sample the entire population and must make do with a sample of N data points. The sample standard deviation, S_X, is then used to estimate σ_X, the expected limits of a particular error. (That sample standard deviation divided by the square root of the number of data points is the starting point for the confidence interval estimate on μ.) For a large dataset (defined as having 30 or more degrees of freedom), plus or minus $2S_X$ divided by the square root of the number of data points in the reported average, M, would contain the true average, μ, 95% of the time. That S_X divided by the square root of the number of data points in the reported average is called the standard deviation of the average and is written as

$$S_{\bar{X}} = \frac{\sqrt{\sum_{i=1}^{N} (X_i - \bar{X})^2 / N - 1}}{\sqrt{M}} = S_X / \sqrt{M} \tag{5.4}$$

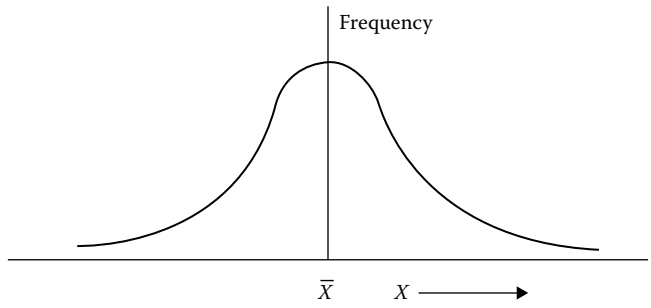


FIGURE 5.1 Normal distribution curve.

where

$S_{\bar{X}}$ is the standard deviation of the average, the sample standard deviation of the data divided by the square root of M

S_X is the sample standard deviation

$\bar{X}$ is the sample average, that is,

$$\bar{X} = \sum_{i=1}^{N}\left(\frac{X_i}{N}\right) \tag{5.5}$$

where

X_i is the ith data point used to calculate the sample standard deviation and the average, $\bar{X}$, from the data

N is the number of data points used to calculate the standard deviation

$(N - 1)$ is the degrees of freedom of S_X and $S_{\bar{X}}$

M is the number of data points in the reported average test result

Note in Equation 5.4 that N does not necessarily equal M. It is possible to obtain S_X from historical data with many degrees of freedom ([$N - 1$] greater than 30) and to run the test only M times. The test result, or average, would therefore be based on M measurements, and the standard deviation of the average would still be calculated with Equation 5.4.

Note that the sample standard deviation, S_X, is simply

$$S_X = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N-1}} \tag{5.6}$$

In some cases, a particular error distribution may be assumed or known to be *a uniform or rectangular distribution* (Figure 5.2) instead of a normal distribution. For those cases, the sample standard deviation of the data is calculated as

$$S_X = \frac{L}{\sqrt{3}} \tag{5.7}$$

where L is the plus/minus limits of the uniform distribution for a particular error [3].

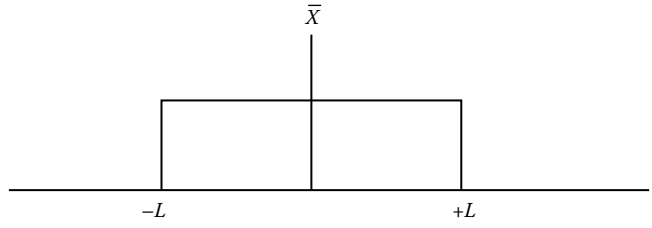


FIGURE 5.2 Uniform distribution curve.

For those cases, the standard deviation of the average is written as

$$S_{\bar{X}} = \frac{L/\sqrt{3}}{\sqrt{M}} \tag{5.8}$$

Although the calculation of the sample standard deviation (or its estimation by some other process) is required for measurement uncertainty analysis, all the analytical work computing the measurement uncertainty uses only the standard deviation of the average for each error source.

5.1.1 Uncertainty (Accuracy)

Since the error for any particular error source is unknown and unknowable, its limits, at a given confidence, must be estimated. This estimate is called the *uncertainty*. Sometimes, the term accuracy is used to describe the quality of test data. This is the positive statement of the expected limits of the data's errors. Uncertainty is the negative statement. Uncertainty is, however, unambiguous. Accuracy is sometimes ambiguous. (For example, what is twice the accuracy of ±2%? ±1% or ±4%?) For this reason, this chapter will use the term *uncertainty* throughout to describe the quality of test data.

5.2 Measurement Uncertainty Model

5.2.1 Purpose

One needs an estimate of the uncertainty of test results to make informed decisions. Ideally, the uncertainty of a well-run experiment will be much less than the change or test result expected. In this way, it will be known, with high confidence, that the change or result observed is real or acceptable and not a result of the errors of the test or measurement process. The limits of those errors are estimated with uncertainty, and those error sources and their limit estimators, the uncertainties, may be grouped into classifications to ease their understanding.

5.2.2 Classifying Error and Uncertainty Sources

There are two classification systems in use. The final expanded uncertainty calculated at a confidence is identical no matter what classification system is used. The two classifications utilized are the *ISO classifications* (International Standards Organization) and the *engineering classifications*. The former groups errors and their uncertainties by type, depending evidently on whether or not there are data available to calculate the sample standard deviation for a particular error and its uncertainty. The latter classification groups errors and their uncertainties by their effect on the experiment or test. That is, the engineering classification groups errors and uncertainties by random and systematic types, with subscripts used to denote whether there are data to calculate a standard deviation or not for a particular error or uncertainty source. For this reason, engineering classification groups usually are comprehensive and therefore more useful and recommended.

5.2.3 ISO Classifications

This error and uncertainty classification system is not recommended in this chapter, but will yield an expanded uncertainty in complete agreement with the recommended classification system—the engineering classification system. In this ISO system, errors and uncertainties are classified as type *A* if there are data available to calculate a sample standard deviation and type *B* if there are not [4]. In the latter case, the sample standard deviation might be obtained from engineering estimates, experience, or manufacturer's specifications, to name three possibilities.

The impact of multiple sources of error is estimated by root-sum-squaring their corresponding elemental uncertainties. The operating equations are as follows.

For Type *A*, data are available for the calculation of the standard deviation:

$$U_A = \left[\sum_{i=1}^{N_A} \left(\theta_i u_{A_i} \right)^2 \right]^{1/2} \tag{5.9}$$

where

u_{A_i} is the standard deviation (based on data) of the average for uncertainty source *i* of type *A* each with its own degrees of freedom

U_A is in units of the test or measurement result. It is an $S_{\bar{X}}$

N_A is the number of parameters with a type *A* uncertainty

θ_i is the sensitivity of the test or measurement result, *R*, to the *i*th type *A* uncertainty. θ_i is the partial derivative of the result with respect to each ith independent measurement

The uncertainty of each error source in units of that source, when multiplied by the sensitivity for that source, converts that uncertainty to result units. Then the effect of several error sources may be estimated by root-sum-squaring their uncertainties as they are now all in the same units. The sensitivities, θ_i, are obtained for a measurement result, *R*, which is a function of several parameters, P_i. The basic equations are

$$R = f(P_1, P_2, P_3, \ldots P_N) \tag{5.10}$$

where

R is the measurement result

P_i is the measurement parameter *i* used to calculate the result, *R*

The sensitivity, θ_i, also called error propagation or uncertainty propagation is found by differentiation as $\partial R/\partial P_i$.

Type *B* (no data for standard deviation) calculation

$$U_B = \left[\sum_{i=1}^{N_B} \left(\theta_i u_{B_i} \right)^2 \right]^{1/2} \tag{5.11}$$

where

u_{B_i} is the standard deviation of the average (based on an estimate, not data) for uncertainty source *i* of type *B*

U_B is in units of the test or measurement result, *R*. It is an $S_{\bar{X}}$

N_B is the number of parameters with a type *B* uncertainty

θ_i is the sensitivity of the test or measurement result, *R*, to the ith Type *B* uncertainty

θ_i is the $\partial R/\partial P_i$

For these uncertainties, it is assumed that the u_{B_i} represent one standard deviation of the average for one uncertainty source with an assumed normal distribution. (They also represent one standard deviation as the square root of the "*M*" by which they are divided is one, that is, there is only one Type *B* error sampled from each of these distributions.) The degrees of freedom assumed to apply to this standard deviation (also standard deviation of the average) are assumed to be infinity [5] in the absence of other information.

Note that θ_i, the sensitivity of the test or measurement result to the ith Type B uncertainty, is actually the change in the result, R, that would result from a change, of the size of the Type B uncertainty, in the ith input parameter used to calculate that result.

The degrees of freedom of the u_{A_i} and the u_{B_i} are needed to compute the degrees of freedom of the combined total uncertainty. It is calculated with the Welch–Satterthwaite approximation. The general formula for degrees of freedom [6] is

$$df_R = \nu_R = \frac{\left[\sum_{i=1}^{N}\left(S_{\bar{X}_i}\right)^2\right]^2}{\left[\sum_{i=1}^{N}\left(S_{\bar{X}_i}\right)^4 \Big/ (\nu_i)\right]} \tag{5.12a}$$

where

$df_R = \nu_R$ are the degrees of freedom for the result

ν_i is the degrees of freedom of the ith standard deviation of the average

For the ISO model, Equation 5.11 becomes

$$df_{R,\mathrm{ISO}} = \nu_{R,\mathrm{ISO}} = \frac{\left[\sum_{i=1}^{N_A}(\theta_i u_{A,i})^2 + \sum_{i=1}^{N_B}\left(\theta_i u_{B,i}\right)^2\right]^2}{\left[\sum_{i=1}^{N}\left(\theta_i u_{A,i}\right)^4 \Big/ \nu_i + \sum_{i=1}^{M}\left(\theta_i u_{B,i}\right)^4 \Big/ \nu_i\right]} \tag{5.12b}$$

The degrees of freedom calculated with Equation 5.12 are often a fraction. This may be truncated to the next lower whole number to be conservative [5].

Note that in Equations 5.9, 5.10, and 5.12, N_A and N_B need not be equal. They are only the total number of parameters with uncertainty sources of Types A and B, respectively.

In computing a total uncertainty, the uncertainties noted by Equations 5.10 and 5.11 are combined. For the ISO model [3], this is calculated as

$$U_{\mathrm{total,ISO}} = \pm\left[(U_A)^2 + (U_B)^2\right]^{1/2} \tag{5.13a}$$

Then the expanded 95% confidence uncertainty is obtained with Equation 5.13b:

$$U_{R,\mathrm{ISO}} = \pm t_{95}\left[(U_A)^2 + (U_B)^2\right]^{1/2} \tag{5.13b}$$

where

t_{95} is the Student's t for ν_r, 150 degrees of freedom

Student's t is obtained from Table 5.1

Note that alternative confidences are permissible. Ninety-five percent is recommended by the ASME [1], but 99% or 99.7% or any other confidence is obtained by choosing the appropriate Student's t. Ninety-five percent confidence is, however, recommended for uncertainty analysis.

In all the aforementioned, the errors were assumed to be independent. Independent sources of error are those that have no relationship to each other. That is, an error in a measurement from one source

TABLE 5.1 Student's *t* Statistic for 95% Confidence, t_{95}, Degrees of Freedom, ν

ν	t_{95}	ν	t_{95}	ν	t_{95}
1	12.706	11	2.201	21	2.080
2	4.303	12	2.179	22	2.074
3	3.182	13	2.160	23	2.069
4	2.776	14	2.145	24	2.064
5	2.571	15	2.131	25	2.060
6	2.447	16	2.120	26	2.056
7	2.365	17	2.110	27	2.052
8	2.306	18	2.101	28	2.048
9	2.262	19	2.093	29	2.045
10	2.228	20	2.086	≥30	2.000

Note: This is frequently written as $t_{95,\nu}$.

cannot be used to predict the magnitude or direction of an error from the other, independent, error source. Nonindependent error sources are related. That is, if it were possible to know the error in a measurement from one source, one could calculate or predict an error magnitude and direction from the other, nonindependent error source. These are sometimes called *dependent error sources*. Their degree of dependence may be estimated with the linear correlation coefficient. If they are nonindependent, whether Type *A* or Type *B*, Equation 5.13 becomes [7]

$$U_{R,\mathrm{ISO}} = \pm\, t_{95}\left\{\sum_{T=A}^{B}\sum_{i=1}^{N_{i,T}}\left[(\theta_i u_{i,T})^2 + \sum_{j=1}^{N_{j,T}}\theta_i\theta_j u_{(i,T),(j,T)}(1-\delta_{i,j})\right]\right\}^{1/2} \tag{5.14}$$

where

$u_{i,T}$ is the ith elemental uncertainty of Type T (can be Type A or B)
$U_{R,\mathrm{ISO}}$ is the total uncertainty of the measurement or test result
θ_i is the sensitivity of the test or measurement result to the ith Type T uncertainty
θ_j is the sensitivity of the test or measurement result to the jth Type T uncertainty
$u_{(i,T),(j,T)}$ is the covariance of $u_{i,T}$ on $u_{j,T}$
$\delta_{i,j}$ is the Kronecker delta. $\delta_{i,j} = 1$ if $i = j$, and $\delta_{i,j} = 0$ if not [7]

$$u_{(i,T)(j,T)} = \sum_{l=1}^{K} u_{i,T}(l)u_{j,T}(l) \tag{5.15}$$

where

l is the index or counter for common uncertainty sources
K is the number of common source pairs of uncertainties
T is the index or counter for the ISO uncertainty Type, A or B
$N_{i,T}$ is the number of error sources for Types A and B

Equation 5.15 equals the sum of the products of the elemental systematic standard uncertainties that arise from a common source (l).

This ISO classification equation will yield the same expanded uncertainty as the engineering classification, but the with ISO classification it is more difficult to gain insight into how to improve an experiment's or test's uncertainty. That is, whether to possibly take more data because the random standard uncertainties are too high or calibrate better because the systematic standard uncertainties are too large. The engineering classification now presented is therefore the preferred approach.

5.2.4 Engineering Classifications

The engineering classification recognizes that experiments and tests have two major types of errors that affect results and whose limits are estimated with uncertainties at some chosen confidence. These error types may be grouped as random and systematic. Their corresponding limit estimators are the random standard uncertainty and systematic standard uncertainties, respectively.

5.2.5 Random

The general expression for random standard uncertainty is the ($1S_{\bar{X}}$) standard deviation of the average [6]:

$$S_{\bar{X},R} = \left[\sum_{T=A}^{B}\sum_{i=1}^{N_{i,T}}\left(\theta_i S_{\bar{X}_{i,T}}\right)^2\right]^{1/2} = \left[\sum_{T=A}^{B}\sum_{i=1}^{N_{i,T}}\left(\frac{\theta_i S_{X_{i,T}}}{\sqrt{M_{i,T}}}\right)^2\right]^{1/2} \tag{5.16}$$

where

$S_{X_{i,T}}$ is the sample standard deviation of the ith random error source of Type T
$S_{\bar{X}_{i,T}}$ is the random standard uncertainty (standard deviation of the average) of the ith parameter random error source of Type T
$S_{\bar{X},R}$ is the random standard uncertainty of the measurement or test result
$N_{i,T}$ is the total number of random standard uncertainties, Types A and B, combined
$M_{i,T}$ is the number of data points averaged for the ith error source, Type A or B
θ_i is the sensitivity of the test or measurement result to the ith random standard uncertainty

Note that $S_{\bar{X},R}$ is in units of the test or measurement result because of the use of the sensitivities, θ_i. Here, the elemental random standard uncertainties have been root-sum-squared with due consideration for their sensitivities or influence coefficients. Since these are all random standard uncertainties, there is, by definition, no correlation in their corresponding error data, so these can always be treated as independent uncertainty sources.

(Note that the term "standard" is inserted to provide harmony with ISO terminology and to indicate that the uncertainties are standard deviations of the average.)

5.2.6 Systematic

The systematic standard uncertainty of the result, b_R, is the root sum square (RSS) of the elemental systematic standard uncertainties with due consideration for those that are correlated [5]. The general equation is

$$b_R = \left\{\sum_{T=A}^{B}\sum_{i=1}^{N_T}\left[(\theta_i b_{i,T})^2 + \sum_{j=1}^{N_T}\theta_i\theta_j b_{(i,T),(j,T)}(1-\delta_{i,j})\right]\right\}^{1/2} \tag{5.17}$$

where

$b_{i,T}$ is the ith parameter elemental systematic standard uncertainty of Type T
b_R is the systematic standard uncertainty of the measurement or test result
N_T is the total number of systematic standard uncertainties
θ_i is the sensitivity of the test or measurement result to the ith systematic standard uncertainty
θ_j is the sensitivity of the test or measurement result to the jth systematic standard uncertainty
$b_{(i,T),(j,T)}$ is the covariance of $b_{i,T}$ on $b_{j,T}$
$\delta_{i,j}$ is the Kronecker delta. $\delta_{i,j} = 1$ if $i = j$, and $\delta_{i,j} = 0$ if not [5]

$$= \sum_{l=1}^{M} b_{i,T}(l) b_{j,T}(l) \tag{5.18}$$

where

l is the index or counter for common uncertainty sources
T is the index or counter for the ISO uncertainty type, A or B

Equation 5.18 equals the sum of the products of the elemental systematic standard uncertainties that arise from a common source (l).

Here, each $b_{i,T}$ and $b_{j,T}$ is estimated as $1S_{\bar{X}}$ for an assumed normal distribution of errors with infinite degrees of freedom [5].

The random standard uncertainty, Equation 5.16, and the systematic standard uncertainty, Equation 5.17, must be combined to obtain a total uncertainty, Equation 5.19a:

$$U_{\text{total,ENG}} = \pm\left[(b_R)^2 + (S_{\bar{X},R})^2\right]^{1/2} \tag{5.19a}$$

Then the expanded 95% confidence uncertainty may be calculated with Equation 5.19b:

$$U_{\text{R,ENG}} = \pm t_{95}\left[(b_R)^2 + (S_{\bar{X},R})^2\right]^{1/2} \tag{5.19b}$$

Note that b_R in units of the test or measurement result is an $S_{\bar{X},R}$.

The degrees of freedom will be needed for the engineering system total and expanded uncertainties. It is accomplished with the Welch–Satterthwaite approximation, the general form of which is Equation 5.10, and the specific formulation here is

$$df_R = \nu_R = \frac{\left\{\sum_{T-A}^{B}\left[\sum_{i=1}^{N}(\theta_i S_{\bar{X},T,i})^2 + \sum_{i=1}^{M}(\theta_i b_{T,i})^2\right]\right\}^2}{\left\{\sum_{T-A}^{B}\left[\sum_{i=1}^{N}(\theta_i S_{\bar{X},T,i})^4 / \nu_{i,T} + \sum_{i=1}^{M}(\theta_i b_{T,i})^4 / \nu_i\right]\right\}} \tag{5.20}$$

where

N is the number of random standard uncertainties of Type T
M is the number of systematic standard uncertainties of Type T
ν_i is the degree of freedom for the ith uncertainty of Type T
$\nu_{i,T}$ is the degree of freedom for the ith uncertainty of type T and is infinity for all systematic standard uncertainties

5.2.7 Symmetrical Systematic Standard Uncertainties

Most times, all elemental uncertainties will be symmetrical. That is, their ± limits about the measured average will be the same. That is, they will be ±3 °C or ±2.05 kPa and the like and not +2.0 °C and −1.0 °C nor +1.5 kPa and −0.55 kPa. The symmetrical measurement uncertainty may therefore be calculated as follows (For an elegant treatment of nonsymmetrical uncertainties, see that section in Ref. [8]).

Note that throughout these uncertainty calculations, all the uncertainties are expressed in engineering units. All the equations will work with relative units as well. However, it is often easier to express all the uncertainties and the uncertainty estimation calculations in engineering units and then, at the end, with the expanded uncertainty, convert the result into relative terms. That is what this section recommends.

5.3 Calculation of Expanded Uncertainty

5.3.1 ISO Total and Expanded Uncertainty

The ISO expanded uncertainty for independent uncertainty sources (the most common) is Equation 5.13b:

$$U_{R,\mathrm{ISO}} = \pm t_{95}\left[(U_A)^2 + (U_B)^2\right]^{1/2} \tag{5.21}$$

where

$U_{R,\mathrm{ISO}}$ is the measurement uncertainty of the result
U_A is the Type *A* uncertainty for the result
U_B is the Type *B* uncertainty for the result
t_{95} is Student's t_{95} is the recommended multiplier to assure 95% confidence

The ISO uncertainty with some nonindependent uncertainty sources is Equation 5.14:

$$U_{R,\mathrm{ISO}} = \pm t_{95}\left\{\sum_{T=A}^{B}\sum_{i=1}^{N_{i,T}}\left[(\theta_i u_{i,T})^2 + \sum_{j=1}^{N_{j,T}}\theta_i\theta_j u_{(i,T),(j,T)}(1-\delta_{i,j})\right]\right\}^{1/2} \tag{5.22}$$

5.3.2 Engineering System Expanded Uncertainty

The engineering system equation for total uncertainty for independent uncertainty sources (the most common) is Equation 5.19b:

$$U_{R,\mathrm{ENG}} = \pm t_{95}\left[(b_R)^2 + (S_{\bar{X},R})^2\right]^{1/2} \tag{5.23}$$

The engineering system equation for uncertainty for nonindependent uncertainty sources (those with correlated systematic standard uncertainties) is also Equation 5.23 but remember to use the full expression for B_R, Equation 5.17:

$$b_R = \left\{\sum_{T=A}^{B}\sum_{i=1}^{N_T}\left[(\theta_i b_{i,T})^2 + \sum_{j=1}^{N_T}\theta_i\theta_j b_{(i,T),(j,T)}(1-\delta_{i,j})\right]\right\}^{1/2} \tag{5.24}$$

The degrees of freedom for Equations 5.21 through 5.23 are calculated with the Welch–Satterthwaite approximation, Equation 5.12b for the ISO system, and Equation 5.20 for the engineering system.

5.3.3 High Degrees of Freedom Approximation

It is often the case that it is assumed that the degrees of freedom are 30 or higher. In these cases, the equations for uncertainty simplify further by setting t_{95} equal to 2.000. This approach is recommended for a first-time user of uncertainty analysis procedures as it is a fast way to get to an approximation of the measurement uncertainty.

5.3.4 Calculation Example

The following calculation example is taken where all the uncertainties are independent and are in the units of the test result—temperature. It is a simple example that illustrates the combination of measurement uncertainties in their most basic case. More detailed examples are given in many of the references cited. Their review may be needed to assure a more comprehensive understanding of uncertainty analysis.

It has been shown [9] that there is no difference in the uncertainties calculated with the different models. The data from Table 5.2 [10] will be used to calculate measurement uncertainty with the two models. These data are all in temperature units and thus the influence coefficients, or sensitivities, are all unity.

Note the use of subscripts "*A*" and "*B*" to denote where data exist to calculate a standard deviation. Note too that in this example, all errors (and therefore uncertainties) are independent and that all degrees of freedom for the systematic standard uncertainties are infinity except for the reference junction whose degrees of freedom are 12.

Each uncertainty model will now be used to derive a measurement uncertainty.

For the U_{ISO} model, one obtains, via Equations 5.13a and b, the expressions

$$U_A = \left[(0.095)^2 + (0.045)^2 + (0.173)^2 + (0.032)^2\right]^{1/2} = 0.21 \tag{5.25}$$

$$U_B = \left[(0.03)^2 + (0.05)^2\right]^{1/2} = 0.058 \tag{5.26}$$

Thus,

$$U_{R,\text{ISO}} = \pm t_{95}\left[(U_A)^2 + (U_B)^2\right]^{1/2} = \pm t_{95}\left[(0.21)^2 + (0.058)^2\right]^{1/2} \tag{5.27}$$

Here, remember that the 0.21 is the RSS of the $1S_{\bar{x}}$ Type *A* uncertainties in Table 5.2 and 0.058 that for the $1S_{\bar{x}}$ Type *B* uncertainties. Also note that in most cases, the Type *B* uncertainties have infinite degrees of freedom and represent an equivalent $1S_X$.

TABLE 5.2 Temperature Measurement Uncertainties, °F

Defined Measurement Process	Systematic Standard Uncertainty, B_i	*df* for b_i	Standard Deviation, $S_{\bar{X},i}$	Number of Data Points N_i	Random Standard Uncertainty, $S_{\bar{X},i}$	Degrees of Freedom, *df*, ν_i
Calibration of *t*c	0.030_B	∞	0.30_A	10	0.095_A	9
Reference junction	0.032_A	12	0.10_A	5	0.045_A	4
Data acquisition	0.050_B	∞	0.60_A	12	0.173_A	11
RSS						

To obtain Student's t_{95}, the degrees of freedom must first be calculated. All the systematic components of type "*B*" have infinite degrees of freedom except for the 0.032, which has 12 degrees of freedom. Also, all the b_i, in Table 5.1, represents an equivalent $1S_X$. All Type *A* uncertainties, whether systematic or random in Table 5.1, have degrees of freedom as noted in the table. The degrees of freedom for U_{ISO} are then

$$\frac{\left[(0.095)^2+(0.045)^2+(0.173)^2+(0.03)^2+(0.032)^2+(0.05)^2\right]^2}{\left[(0.095)^4/9+(0.045)^4/4+(0.173)^4/11+(0.03)^4/\infty+(0.032)^4/12+(0.05)^4/\infty\right]}=22.51\approx 22 \tag{5.28}$$

t_{95} is, therefore, 2.07. $U_{R,ISO}$, the expanded uncertainty, is determined with Equation 5.13b:

$$U_{R,ISO}=\pm 2.07\left[(0.21)^2+(0.058)^2\right]^{1/2}=0.45 \quad \text{for 95\% confidence} \tag{5.29}$$

For a detailed comparison to the engineering system, here denoted as the $U_{R,ENG}$ model, three significant figures are carried so as not to be affected by round-off errors. Then

$$U_{R,ISO}=\pm 2.074\left[(0.205)^2+(0.0583)^2\right]^{1/2}=0.442 \quad \text{for 95\% confidence} \tag{5.30}$$

For the engineering system, $U_{R,ENG}$, model, one obtains the random standard uncertainty with Equation 5.16:

$$S_{\bar{X},R}=\left[(0.095)^2+(0.045)^2+(0.173)^2\right]^{1/2}=0.20 \tag{5.31}$$

The systematic standard uncertainty is obtained with Equation 5.17 greatly simplified as there is no correlated error or uncertainties in this example. Equation 5.17 then becomes

$$b_R=\left[(0.03)^2+(0.032)^2+(0.05)^2\right]^{1/2}=0.065 \tag{5.32}$$

$U_{R,ENG}$, the expanded uncertainty, is then obtained with Equation 5.19b:

$$U_{R,ENG}=\pm t_{95}\left[(0.065)^2+(0.20)^2\right]^{1/2} \tag{5.33}$$

The degrees of freedom need to be calculated just as in Equation 5.28. Therefore, the degrees of freedom were 22 and t_{95} equals 2.07. $U_{R,ENG}$ is then

$$U_{R,ENG}=\pm 2.07\left[(0.065)^2+(0.20)^2\right]^{1/2}=0.44 \quad \text{for 95\% confidence} \tag{5.34}$$

Carrying three significant figures for a comparison to $U_{R,ISO}$ not affected by round-off errors, one obtains

$$U_{R,ENG}=\pm 2.07\left[(0.0665)^2+(0.202)^2\right]^{1/2}=0.442 \quad \text{for 95\% confidence} \tag{5.35}$$

This is identical to $U_{R,ISO}$, Equation 5.30, as predicted.

5.4 Summary

Although these formulae for uncertainty calculations will not handle every conceivable situation, they will provide, for most experimenters, a useful estimate of test or measurement uncertainty. For more detailed treatment or specific applications of these principles, consult the references and the recommended "Further Information" section at the end of this chapter.

Defining Terms

Accuracy: The antithesis of uncertainty. An expression of the maximum possible limit of error at a defined confidence.

Confidence: A statistical expression of percent likelihood.

Correlation: The relationship between two data sets. It is not necessarily evidence of cause and effect.

Degrees of freedom: The amount of room left for error. It may also be expressed as the number of independent opportunities for error contributions to the composite error.

Error: [error] = [measured] − [true]. It is the difference between the measured value and the true value.

Expanded uncertainty: The 95% confidence interval uncertainty. It is the product of the total uncertainty and the appropriate Student's *t*.

Influence coefficient: See Sensitivity.

Measurement uncertainty: The maximum possible error, at a specified confidence, that may reasonably occur. Errors larger than the measurement uncertainty should rarely occur.

Nonsymmetrical uncertainty: An uncertainty for which there is an uneven likelihood that the true value lies on one side of the average or the other.

Propagation of uncertainty: An analytical technique for evaluating the impact of an error source (and its uncertainty) on the test result. It employs the use of influence coefficients.

Random error: An error that causes scatter in the test result.

Random standard uncertainty: An estimate of the limits of random error, one standard deviation of the average.

Sensitivity: An expression of the influence an error source has on a test or measured result. It is the ratio of the change in the result to an incremental change in an input variable or parameter measured.

Standard deviation of the average or mean: The standard deviation of the data divided by the number of measurements in the average.

Systematic error: An error that is constant for the duration of a test or measurement.

Systematic standard uncertainty: An estimate of the limits of systematic error, usually taken as 68% confidence for an assumed normal error distribution and infinite degrees of freedom.

Total uncertainty: The RSS combination of either the Type *A* or *B* uncertainties for the ISO error classifications or the random and systematic standard uncertainties for the engineering error classifications.

True value: The desired result of an experimental measurement.

Welch–Satterthwaite: The approximation method for determining the number of degrees of freedom in a combined uncertainty.

References

1. American National Standards Institute/American Society of Mechanical Engineers (ANSI/ASME) PTC 19.1-2005, *Instruments and Apparatus, Part 1, Test Uncertainty*, 1985, p. 5.
2. E. O. Doebelin, *Measurement Systems, Application and Design*, 4th edn., New York: McGraw-Hill, 1990, 38 ff.

3. International Standards Organization, *Guide to the Expression of Uncertainty Measurement*, Geneva, Switzerland, 1993, p. 16.
4. International Standards Organization, *Guide to the Expression of Uncertainty in Measurement*, Geneva, Switzerland, 1993, pp. 10, 11.
5. International Standards Organization, *Guide to the Expression of Uncertainty in Measurement*, Geneva, Switzerland, 1993, p. 64.
6. R. H. Dieck, *Measurement Uncertainty: Methods and Applications*, Research Triangle Park, NC: ISA, 1992, p. 45.
7. K. K. Brown, H. W. Coleman, W. G. Steele, and R. P. Taylor, Evaluation of correlated bias approximations in experimental uncertainty analysis, *Proceedings of 32nd Aerospace Sciences Meeting & Exhibit*, Reno, NV, AIAA paper no. 94-0772, January 10–13, 1996.
8. American National Standards Institute/American Society of Mechanical Engineers (ANSI/ASME) PTC 19.1-2005, *Instruments and Apparatus, Part 1, Test Uncertainty*, 1985, pp. 33–35.
9. W. T. Strike, III and R. H. Dieck, Rocket impulse uncertainty: An uncertainty model comparison, *Proceedings of 41st International Instrumentation Symposium*, Denver, CO, May 1995.
10. R. H. Dieck, Measurement uncertainty models, *Proceedings of 42nd International Instrumentation Symposium*, San Diego, CA, May 1996.

Further Information

Abernethy, R. B. and B. Ringhiser, The history and statistical development of the New ASME-SAE-AIAA-ISO measurement uncertainty methodology, *Proceedings of AIAA/SAE/ASME/ASME 21st Joint Propulsion Conference*, Monterey, CA, July 8–10, 1985.

Abernethy, R. B. et al., *Handbook-Gas Turbine Measurement Uncertainty*, AEDC, Tullahoma, TN, 1973.

Bevington, P. R. and D. K. Robinson, *Data Reduction and Error Analysis for the Physical Sciences*, New York: McGraw-Hill, 2003.

Dieck, R. H., *Measurement Uncertainty: Methods and Applications*, 4th edn., Research Triangle Park, NC: ISA, 2006.

Grabe, M., *Measurement Uncertainties in Science and Technology*, Berlin, Germany: Springer, 2005.

ICRPG Handbook for Estimating the Uncertainty in Measurements Made with Liquid Propellant Rocket Engine Systems, Chemical Propulsion Information Agency, No. 180, April 30, 1969.

Steele, W. G., R. A. Ferguson, and R. P. Taylor, Comparison of ANSI/ASME and ISO models for calculation of uncertainty, *Proceedings of 40th International Instrumentation Symposium*, Paper Number 94–1014, pp. 410–438, 1994.

6

Development of Standards

6.1 Introduction **6**-1
6.2 Standard Organizations **6**-2
Relations of Standards Organizations
6.3 Development of Standards **6**-4
The International Standards Organization • Structure of ISO • Obtaining ISO Standards
6.4 Obtaining and Using Standards **6**-7
6.5 Examples of Standards **6**-8
6.6 Software Standards **6**-8
Types of Software Standards • Instrumentation and Measurement Software Standards
Bibliography **6**-11

Halit Eren
Curtin University

6.1 Introduction

The word "standard" implies comparative evaluations of physical quantities, processes, and procedures.

Standards in instrumentation and measurement are implemented in a number of ways, as

- The fundamental physical standards
- Published standards (or the paper standards) that are published by national and international organizations and they are available for anyone to use
- De facto standards that are followed by convention or popular use
- De jure standards that are part of legally binding contracts, laws, or regulations

The standards and the implementation of standards are highly important in instrumentation and measurement systems. They are important documents that provide textual and illustrative information on what and how the processes and practices should be performed to achieve a particular task, for example, the measurement of temperature. The National Standards Policy Advisory Committee (http://www.nist.gov/) describes a standard as "a prescribed set of rules, conditions, or requirements concerning definitions of terms: classification of components; specification of materials, performance, or operations; delineation of procedures; or measurement of quantity and quality in describing materials, products, systems, services, or practices."

Published standards are necessary for worldwide trading, business practices, production of goods, manufacturing, scientific investigations, and research and development. They ensure consistency in implementation and interpenetration of different technologies in different organizations and sectors. Because of the national and international implications, there are many institutions that are responsible for developing, investigating, and maintaining the relevant standards to support worldwide scientific, commercial, and industrial activities. In this chapter, the published standards relevant to the instruments, instrumentation, and measurements will be explained in detail.

6.2 Standard Organizations

Standards are formulated by standards developing organizations (SDOs). These organizations can be government agencies, societies, international or regional organizations, private companies, professional institutions, and trade associations. World standards organizations can be categorized in four major groups as national, international, consortiums, and industrial and professional organizations, as illustrated in Figure 6.1.

A partial list of national and international standards organizations that deals with published standards is given in Table 6.1.

Each national or international organization has several specialist groups that encompass a broad range of scientific disciplines such as the measurements, measuring instruments, instrumentation and control, process control, safety, installations, electromagnetism, chemical engineering, mechanical engineering, radiation, and optical standards.

Common to all these organizations, the general terminology, vocabulary, and definitions concerning standardization and related activities are determined firstly. This acts as an agent to develop a mutual understanding among the members, development groups, and other concerned parties at the individual,

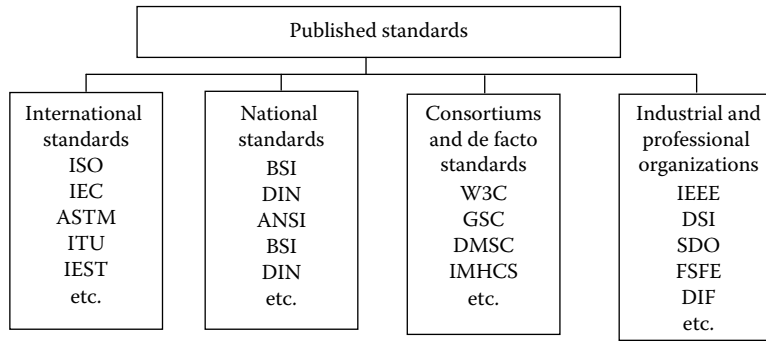


FIGURE 6.1 World standards organizations.

TABLE 6.1 Partial List of National and International Standards Organizations

Acronym	Organization Name and Web Address
ANSI	American National Standards Institute (http://www.ansi.org/)
AFNOR	Association Française de Normalization (http://www.afnor.fr/)
BSI	British Standards Institution (http://www.bsi-global.com/)
CEI	Comitato Elettrotecnico Italiano (http://www.ceiuni.it/)
CEN	Comité Européen de Normalisation (http://www.cenorm.be/)
CENELEC	Comité Européen de Normalisation Electrotechnique (http://www.cenelec.be/)
CSSN	China Standards Information Center (http://www.cssn.net.cn/)
DIN	Deutsches Institut für Normung (http://www2.din.de/)
IEC	International Electrotechnical Commission (http://www.iec.ch/)
ISO	International Organization for Standardization (http://www.iso.org/)
JISC	Japan Industrial Standards Committee (http://www.jisc.go.jp/)
NIST	American National Standards and Technology (http://www.nist.gov/)
SA	Standards Australia (http://www.standards.com.au/)
SCC	Standards Council of Canada (http://www.scc.ca/)
NZS	Standards New Zealand (http://www.standards.co.nz/)
TSE	Türk Standardlari Enstitüsü (http://www.tse.org.tr/)

governmental, and organizational levels. The aim is to prepare standards in such a manner that they can be interpreted, implemented, and referred to easily by all users. To stimulate this, standards organizations have strict guidelines to assist technical groups and committees in drafting and producing the standards. A typical preparation process has two main stages:

- *Preliminary work* that includes overall analysis, identification of the target groups, determination of subject delimitation, finding the relevant sources, determining the affecting concepts, choosing languages, and working out of the schedules.
- *Working procedure* that includes collecting and recording terminological data, establishing the term list, determining the concept fields and concept systems, formulating definitions, establishing and structuring the entries, streamlining the order of entries, indexing, graphic representation, and bibliography.

In many cases, the standards are produced by various bodies sometimes operating totally independent from each other. Hence, there may be different versions for the same subject. However, there are a number of international standards organizations such as the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) representing many member countries. These organizations have multiple internal departments, international committees, subcommittees, and working groups to support their activities.

6.2.1 Relations of Standards Organizations

Standards are important for scientific, research and development activities, and international trade; therefore, there is continuous relation and cooperation between national, regional, and international organizations. For example, in the case of ISO, there are many affiliated standards organizations, which represent member countries. The membership structure is as follows:

1. *Full members* are national bodies that are entitled to participate and exercise full voting rights on any technical committee and policy committee of the ISO.
2. *Correspondent members* that do not take an active part in the technical and policy development, but are entitled to be kept fully informed about their work of interest. These are countries that do not have their own standards organizations.
3. *Subscriber members* that represent the countries with small economies. The membership fees of the subscriber members are reduced but are allowed to be represented in some committees.

Apart from the international standards organizations, such as the ISO, that are structured with voting and nonvoting members, there are also numerous other mechanisms that maintain links between standards institutions and organizations. Examples are

- *The Asia-Pacific Legal Metrology Forum* (APLMF) is a grouping of legal metrology authorities in the Asia-Pacific Economic Cooperation (APEC). The objective of this forum is the development of legal metrology and the promotion of free and open trade in the region through the harmonization and removal of technical or administrative barriers.
- *Bureau International des Poids et Mesures* (BIPM) that ensures worldwide uniformity of measurements and their traceability to the International System of Units (SI).
- *World Trade Organization* (WTO) is an international agency whose purpose is to help trade flow as smoothly as possible in a system based on rules, to settle trade disputes between governments, and to organize related negotiations.
- *Joint technical committees* are formed to deal with the consequences of substantial overlap in areas of standardization and work related to information technology. For example, IEC has formed a joint technical committee known as the ISO/IEC JTC1, which was the first such joint committee and it is still fully functional.

Many standards organizations collaborate as consortiums to develop and implement particular standards, as in the case of fieldbus systems for industrial digital information transmissions and applications. The development of standards is a complex procedure involving many groups at the national, international, and industrial organizational levels. Some examples will be provided in the Section 6.3.

6.3 Development of Standards

SDOs undertake the task of producing relevant standards in their expertise areas. There are many different national and international SDOs; some of these organizations are as follows:

- National Electrical Manufactures Association (NEMA) (http://www.nema.org/)
- National Institute of Standards and Technology (NIST) (http://www.nist.gov/index.html)

At the individual organization level, each organization has its own different way of developing standards, which depends on its size, culture, aims, and structure. In this chapter, the ISO will be given as an example of a typical standardization body. More information can be found on ISO website (http://www.iso.org/) and indeed on the websites of many other organizations.

6.3.1 The International Standards Organization

The ISO plays an important role in development and proliferation of international standards. The organizational and governance structure is illustrated in Figure 6.2. It has worldwide representation having 110 full voting member countries, 43 corresponding members, and 10 subscriber members as it stands today in March 2012. It has 3335 technical bodies comprising 224 technical committees, 513 subcommittees, 2516 working groups, and 82 ad hoc study groups. Although being a nongovernment organization, it can set standards such that many of them become legislated laws either through international treaties or national governments. Therefore, here, further explanation will be provided about the ISO, but more information can be found on their website (http://www.iso.org/).

ISO promulgates worldwide proprietary industrial and commercial standards. It produces international standards as well as publishing technical reports, technical specifications, technical corrigenda, and guides as explained as follows.

One of the products of the ISO is the *technical reports*. Technical reports are produced by the technical committees or subcommittees after having collected substantial data on a particular subject.

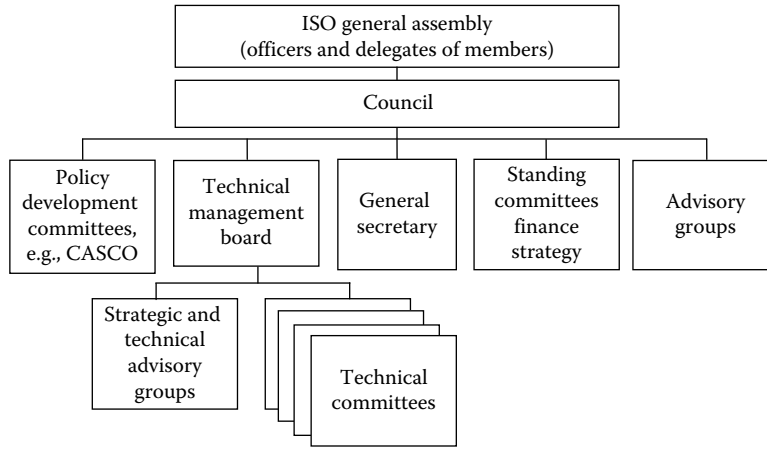


FIGURE 6.2 Branches of ISO.

These reports normally lead to the development of international standards and they are identified with *TR* on their titles. Some examples are

- ISO/TR 19033:2000. Technical product documentation—metadata for construction documentation
- SO/IEC TR 17799:2000. Code of practice for information security management

The ISO publishes *technical specifications* and *publicly available specifications* as intermediate information before the development of full international standard on a particular subject. Technical specifications are identified with *TS*, for example,

- ISO/TS 16952-1:2006. Technical product documentation—reference designation system—Part 1: General application rules
- ISO/PAS 11154:2006. Road vehicles—roof load carriers

ISO also issues *technical corrigendum*, which are amendments to existing standards because of minor technical flaws, improvements, or extend applicability in a limited way. Generally, technical corrigendum is issued with the expectation that the affected standard will be updated or withdrawn at its next review.

ISO *guides* are produced as meta-standards. They cover guides and rules for matters related to international standardization efforts, for example,

- ISO/IEC Guide 2:2004. Standardization and related activities—general vocabulary
- ISO/IEC Guide 65:1996. General requirements for bodies operating product certification

In order to understand how ISO develops standards, it is important to look at its organizational structure more closely.

6.3.2 Structure of ISO

A broad view of the structure of ISO is illustrated in Figure 6.2. Going to the details, it operates with the following committees and groups:

- *Policy development committees* that are established by the general assembly to draft policies. The committees are open to all full member and correspondent members, and they directly report to the council. Policy development committees have several branches such as the Committee on Conformity Assessment and Committee on Consumer Policy
- *Council* that operates the ISO itself. It consists of 18 elected representatives of full members and some permanent officers
- *Council standing committees* that advise the council on policy and strategic matters and raise any related issues
- *Technical management board* that examines proposals, monitors the work of technical committees, and establishes and dissolves technical advisory groups
- *Technical advisory groups* (TAGs) that advise the board on basic, sectoral, and cross-sectoral coordination; coherent planning; and the needs for new work. For example, TAG 4 is responsible for metrology matters
- *Technical committees* that are responsible with a particular area such as TC 12 for quantities, units, symbols, and conversion factors and TC 108 for mechanical vibration and shock
- *Ad hoc advisory groups* that recommend standardizations

The standards are developed according to the following principles:

- *Consensus*: The views of all interest manufacturers, vendors, users, consumer groups, testing laboratories, governments, engineering professions, and research organizations are taken into account.
- *Industry-wide*: Solutions intend to satisfy industries and customers worldwide.
- *Voluntary*: International standardization is market driven and therefore based on voluntary involvement of all interests in the marketplace.

There are three main phases in developing the standards, which can be explained as follows:

1. The need for a standard is usually expressed by an industry sector, which communicates this need to a national member body. The latter proposes the new work item to ISO as a whole. Once the need for an international standard has been recognized and formally agreed, the first phase involves definition of the technical scope of the future standard. This phase is usually carried out in working groups that comprise technical experts from countries interested in the subject matter.
2. Once agreement has been reached on which technical aspects are to be covered in the standard, a second phase is entered during which countries negotiate the detailed specifications within the standard. This is the consensus-building phase.
3. The final phase comprises the formal approval of the resulting draft international standard (DIS) (the acceptance criteria stipulate approval by two-thirds of the ISO members that have participated actively in the standards development process and approval by 75% of all members that have voted), following which the agreed text is published as an ISO international standard.

A normal procedure is that the standards to be internationalized can be brought to the attention of the ISO from a country or industry or business through one or more of the member organizations. The members propose the new work item to the ISO as a whole. If accepted, the work item is assigned to an existing technical committee. The technical committees are composed of experts appointed from industrial, technical, and business sectors. These experts may be joined by others with relevant knowledge, such as representatives of government agencies, testing laboratories, consumer associations, and environmentalists. Generally, the experts are selected from the national delegations nominated by the member institutes of the countries concerned. The ISO rules state that the member institutions are expected to take account of the views of a broad range of interested parties and other stakeholders so that acceptance of the proposal becomes consolidated in a consensus manner.

The national delegations of experts of a technical committee meet to discuss and debate until a consensus is reached on a draft agreement. The draft agreement is then circulated as a DIS to members for comments and balloting. Members are responsible for providing public review procedures for making draft standards known and available to interested parties as well as to the general public. Once the feedback is received from members, the draft standard is formulated in its final form to be voted by the members. If the voting is in favor, the document, with eventual modifications, is circulated to the ISO members as a final draft international standard (FDIS). If that vote is positive again, the document is then published as an international standard.

Most standards require periodic revision as several factors combine to render a standard out of date: technological evolution, new methods and materials, and new quality and safety requirements. To take account of these factors, ISO has established the general rule that all ISO standards should be reviewed at intervals of not more than 5 years. On occasion, it is necessary to revise a standard earlier.

Similar processes take place in other organizations depending on their internal cultures and structures. For example, in the case of IEC, there are seven stages to develop a full standard document (http://www.iec.org/):

1. Proposal stage—new work item proposal (NP).
2. Preparatory stage—working draft(s) (WD).
3. Committee stage—committee drafts (CD).
4. Inquiry stage—inquiry draft (DIS, CDV).
5. Approval stage (FDIS).
6. Publication stage.
7. If FDIS fails, the Technical Report 2 (TR2) is published and revised after 2 years.

TR2 status means publication of work is still under technical development or where for any other reason there is a future but not immediate possibility of agreement.

6.3.3 Obtaining ISO Standards

ISO documents are copyrighted and ISO charges for copies of most. ISO does not charge for most draft copies of documents in electronic format. Although useful, care must be taken using these drafts since there is the possibility of substantial change before it becomes finalized as an international standard. Some standards by ISO and its official U.S. representative (and the IEC via the U.S. National Committee) are made freely available.

Many of the ISO standards are ubiquitous, which has led to common use of "ISO" to describe the actual product that conforms to a standard. Some examples of this are as follows:

- ISO 9660. Standard file system has been adopted by many organizations; hence, the CD images are commonly referred to as "ISOs." Almost all computers with CD-ROM drives can read CDs that use this standard. Some DVD-ROMs also use ISO 9660 file systems.
- ISO 5800:1987. Photographic film's sensitivity to light and its film speed is often referred to as its "ISO number."

6.4 Obtaining and Using Standards

There are several hundred thousand standards and technical regulations in the world containing special requirements for a particular industry, country, or region. Finding relevant information on the specific standards can be a difficult task. Nevertheless, once the standard is identified for a particular application, it can be obtained in paper or electronic forms from various sources. For example, standards can be purchased online from standards institution or other providers using their catalogs. Some of the websites that provide full-text standards are given as follows, but the list is not exhaustive:

- ETSI, European Telecommunications Standards Institute, offers direct access to publicly available ETSI telecommunications standards (http://www.etsi.org/).
- IAEA, International Atomic Energy Agency, offers most recent IAEA publications online (http://www.iaea.org/).
- IEEE Xplore, Institute of Electrical and Electronics Engineers, offers standards for members (http://ieeexplore.ieee.org/).
- NISO, National Information Standards Organization, offers free full-text access to NISO information standards (http://www.niso.org/).
- ANSI, American National Standards Institute, offers a range of standards for purchasing (http://www.ansi.org/).
- Global Engineering Documents offers national and international standards (http://global.ihs.com/).
- Information Handling Services, Englewood, CO (IHS) offers most engineering standards (http://www.ihs.com/).
- World Standards Services Networks (WSSN) offers national and international standards (http://www.wssn.net/).
- National Standards Services Networks (NSSN) offers national and international standards (http://www.nssn.org/).

Standards used in measurement and metrology largely define guidelines for testing methodology, accuracy, data collection, and data dissemination, use and reporting criteria for consistency and accuracy. From guidelines, the necessary instrumentation, procedures, and control processes can be determined and used appropriately for the design, construction, operation, maintenance,

installations, safety, and other related areas. The use of standards in the design and the operations offer many advantages, such as

- Improves the communication between the designers and the operators
- Provides practical application of expert knowledge
- Represents accumulated experience, thus avoiding ground-up starts
- Enhances operability and interoperability thus ensures compatibility
- Ensures consistency, accuracy, and data interpretation
- Sets a benchmark for global applications
- Sets the recognized level of quality
- Increases market access and user acceptance
- Improves communication between peers in industry and other concerned parties
- Provides immediate access to technical resources
- Provides an understanding on the future trends
- Provides consistency in products and traceability
- Facilitates interchangeability of products and quality assurance
- Ensures safety
- Facilitates migration of products for integrated systems

6.5 Examples of Standards

As explained previously, due to the range of standards organizations, interest groups, and diversity on the subject matters, there are many standards available worldwide. This sentiment is summarized on the ISO website with the following information: ISO has developed 19,023 international standards to date on a variety of subjects, and some 1,208 new ISO standards have been published in the year 2011. The full range of technical fields can be seen from the listing international standards. Users can browse that listing to find bibliographic information on each standard and, in many cases, a brief abstract. The online ISO standards listing integrate both the ISO catalog of published standards and the ISO technical program of standards under development (http://www.iso.org/iso/iso_catalogue.htm#).

Some examples of standards relevant to instrumentation and measurement are given in Table 6.2.

6.6 Software Standards

Software is a set of instructions for digital systems, programmable devices, and computers. It describes computer programs, procedures, and documentation. Software is developed by programming languages like assembly language and C++. There are two types of software, being the system software and the application software. The system software enables the computer to execute its own internal operations as well as the application programs such as the word processors. System software includes an operating system to interface with hardware to provide the necessary services for applications.

Software is often regarded as anything except the hardware. Software encompasses a wide range of products and technologies, such as programming languages, FPGA configuration, web pages, HTML, and Microsoft Word, running under operating systems such as Linux and Microsoft Windows.

Software has its own niche industry made up of different entities and people. There are many software companies and programmers in the world specialized in certain areas such as mathematic applications, interfacing, measurements, and data processing. Also, there are many software standard organizations like the W3C and IETF (Internet Engineering Task Force). These organizations are primarily made up of people from large companies like Microsoft and IBM.

TABLE 6.2 Partial List of Standards on Instrumentation, Measurement, and Related Fields

ASTM B896-99	Standard test methods for evaluating connectability characteristics of electrical conductor materials
ASTM D4356	Standard practice for establishing consistent test method tolerances
ASTM E119	Standard test methods for fire tests of building construction and materials
ASTM G138	Standard test method for calibration of a spectroradiometer using a standard source of irradiance
ASTM 167	Standard test method for calibration of a pyranometer using pyrheliometer
ASTM G169	Standard guide for application of basic statistical methods to weathering tests
ASTM G187	Standard test method for measurement of soil resistivity using the two-electrode soil box
DIN1311-4	Vibration; vibrating continua, waves
DIN 12975-4	Thermal solar systems and components—solar collectors—Part 2: Test methods
DIN 19237	Measurement and control; control engineering vocabulary
DIN 43780	Electrical measuring instruments; direct acting indicating electrical measuring instruments and their accessories
DIN 45507	Measuring apparatus for frequency variations in sound recording equipment
DIN 54836	Testing of combustible materials; determination of ignition temperature
EN 10002-1	Method of test at ambient temperature
EN 13146	Railway applications—track—test methods for fastening systems
EN 55011:2000	Industrial, scientific, and medical RF equipment
IEC 60068	Environmental testing
IEC 60331	Test for electric cables under fire conditions
IEC 60601	Medical electric equipment
IEC 60793	Optical fibers
IEC 60801	EMI and RFI immunity
IEC 6095	Maritime navigation and radio communication equipment
IEC 61000	Electromagnetic compatibility (EMC)
IEC 61000-4 P1-30	Testing and measurement techniques
IEC 61043	Sound intensity meters with pairs of microphones
IEC 61326	Electrical equipment for measurement, control, and laboratory use
IEC 61724	Photovoltaic system performance monitoring
IEC 62366	Medical devices—application of usability engineering to medical devices
IEEE 82	Standard test procedure for impulse voltage test on insulated conductors
IEEE 299:1997	Measuring the effectiveness of the shielding enclosure
IEEE 1460	Guide for the measurement of quasi-static magnetic and electric fields
IEEE 1521	Standard for measurement of video jitter and wander
IEEE 1650	Test methods for measurement of electrical properties of carbon nanotubes
IEEE/IEC C57.12.20	Draft standard for nuclear power plants—instrumentation and control important to safety—electric equipment condition monitoring methods—Part 3
IEEE 61588	Standards for precision clock synchronization protocol for networked measurement and control systems
ISO 11:1987	Aircraft—ground pressure testing connections for pressure cabins
ISO ICS 01	Generalities, terminology, standardization, documentation—graphical symbols, technical drawings, info, science
ISO ICS 17	Metrology and measurement, physical phenomena
ISO 10993-4	Selection of tests for interactions with blood
ISO 14644	Clean rooms and associated controlled environment
ISO 15189	Medical laboratories—particular requirements for quality and competence
ISO 15970	Natural gas—measurement properties
ISO 16750	Road vehicles—environmental conditions and testing for electric and electronic equipment
ISO/IEC 19794	Information technology—biometric data interchange formats
ISO 22241-2	Diesel engines—NOx reduction agent AUS 32—test methods
ISO 7700-2:1987	Check of calibration of moisture meters

6.6.1 Types of Software Standards

Software standards are certain terms, concepts, and techniques agreed to by software creators. For instance, HTML, TCP/IP, SMTP, POP, and FTP are standards that all software designers must adhere to, if their software aims to interface within the standards. Complexity of a standard can vary depending on type of the problem and the people involved. Standard is often made to be as simple as possible, maintainable, understandable, and easy to follow.

There are two types of basic software standards:

1. *Closed standards*: In which the specifications are kept secret to prevent interference from third parties. It inhibits third-party software from being installed; it keeps third-party hardware from interoperating with it, and it prevents third-party enhancements from improving the product.
2. *Open standards*: Specifications for hardware or software that are developed by a standards organization or a consortium involved in supporting a standard. Available to the public for developing compliant products, open standards imply "open systems" that an existing component in a system can be replaced with that of another vendor. Although many vendors may have contributed to an open standard and one may have contributed more than others and even have great influence, an open standard is not completely controlled by a single vendor.

Some examples of software standards are given in Table 6.3.

6.6.2 Instrumentation and Measurement Software Standards

The use of software is an integral part of instrumentation and measurements for data collection, data processing, interpretation, and decision making. Therefore, the functional and structural quality of the supporting software bares ultimate importance. The ISO 25000:2005 quality model describes the structure, classification, and terminology of the attributes of the measurement and metric-related software.

TABLE 6.3 Examples of Software Standards

ISO/CDTS8000-1	Data quality—Part 1: Overview, principles, and general requirements
ISO/TS8000-100:2009	Data quality—Part 100: Master data: Overview
ISO8000-102:2009	Data quality—Part 102: Master data: Exchange of characteristic data: Vocabulary
ISO8000-110:2009	Data quality—Part 110: Master data: Exchange of characteristic data: Syntax, semantic encoding, and conformance to data specification
ISO/TS8000-120:2009	Data quality—Part 120: Master data: Exchange of characteristic data: Provenance
ISO/TS8000-130:2009	Data quality—Part 130: Master data: Exchange of characteristic data: Accuracy
ISO/TS8000-140:2009	Data quality—Part 140: Master data—Exchange of characteristic data: Completeness
ISO9506-1:2003	Industrial automation systems—manufacturing message specification—Part 1: Service definition
ISO9506-2:2003	Industrial automation systems—manufacturing message specification—Part 2: Protocol specification
ISO10303-1:1994	Industrial automation systems and integration—product data representation and exchange—Part 1: Overview and fundamental principles
ISO/PRFTS29002-20	Industrial automation systems and integration—exchange of characteristic data—Part 20: Concept dictionary resolution services
ISO/TS29002-31:2009	Industrial automation systems and integration—exchange of characteristic data—Part 31: Query for characteristic data
IEC62264-1:2003	Enterprise-control system integration—Part 1: Models and terminology
IEC62264-2:2004	Enterprise-control system integration—Part 2: Model object attributes
IEC62264-3:2007	Enterprise-control system integration—Part 3: Activity models of manufacturing operations management

The general requirements of the relevant software are reliability, efficiency, security, and maintainability. A number of examples of the existing standards addressing the instrumentation and measurements are given here, but the examples are not exhaustive:

ASTM E2439–09 (Standard Guide for Instrumentation, Sensors and Operating Software Used in Forensic Psychophysiological Detection of Deception [Polygraph] Examinations). This standard covers the minimum requirements for instrumentation (both analog and computerized systems), sensors, and operating software used in the forensic psychophysiological detection of deception (PDD).

IEEE–STD 1057 provides algorithms for fitting parameters of a sine wave to discrete time observations. It covers electronic digitizing waveform recorders, waveform analyzers, and digitizing oscilloscopes with digital outputs. It gives accurate estimation to Gaussian and quantization noise that commonly exists in measurement systems.

ISO 18629 standard is a process specification language (PSL) that can be used to represent manufacturing engineering and business processes including production planning, simulation, process modeling, and project management.

IEC 62304 is a standard that specifies the life cycle for the development of medical software within medical devices. It targets to harmonize the medical equipment produced by European Union and United States so that a benchmark is achieved to comply with regulatory requirements.

Bibliography

Alves, J. G. Developments in standards and other guidance for individual monitoring, *Radiation Measurements*, 43(2–6), 558–564, 2008.

American National Standards Institute (ANSI), http://www.ansi.org/ (accessed on February 5, 2012).

British Standards Institution (BSI), http://www.bsi-global.com/ (accessed on December 10, 2011).

Busch, L., *Recipes for Reality (Infrastructure)*, The MIT Press, Cambridge, MA, 2011.

Eren, H. Standards in process control and automation, *Process Software Digital Networks*, Eds., H. Eren. and B. Liptak, Vol. 3, 4th edn., CRC Press, Boca Raton, FL, pp. 150–157, 2011.

European Committee for Standardization (CEN), http://www.cen.eu/cen/pages/default.aspx (accessed on February 20, 2012).

German Institute for Standardization (DIN), http://www.din.de/cmd?level=tpl-home&languageid=en (accessed on February 5, 2012).

International Bureau of weights and measures, http://en.wikipedia.org/wiki/International_Bureau_of_Weights_and_Measures (accessed on February 10, 2012).

International Electrotechnical Commission (IEC), http://www.iec.org/ (accessed on July 10, 2011).

International Organization for Standardization (ISO), http://www.iso.org/ (accessed on January 26, 2012).

National Standards Services Networks (NSSN), http://www.nssn.org/ (accessed on March 7, 2012).

SIA, Global info store, ISO publications, http://infostore.saiglobal.com/store/Portal.aspx?publisher=ISO&gclid=CMX1sN_P4K4CFeRKpgodXFMKYw (accessed on February 12, 2012).

Techstreet Thomson, standards catalogs, http://www.techstreet.com/ (accessed on February 11, 2012).

Velychko, O. and Gordiyenko, T. The use of metrological terms and SI units in environmental guides and international standards, *Measurement*, 40(2), 202–212, 2007.

World Standards Services Networks (WSSN), http://www.wssn.net/ (accessed on November 18, 2011).



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

7

Measurement Standards

7.1 Historical Perspective 7-1
7.2 What Are Standards? 7-2
Standards of Practice (Protocol Standards) • Legal Metrology • Forensic Metrology • Standard Reference Materials
7.3 Conceptual Basis of Measurements 7-4
7.4 Need for Standards 7-5
7.5 Types of Standards 7-5
Basic or Fundamental Standards • Derived Standards • Measurement Assurance System
7.6 Numbers, Dimensions, and Units 7-6
7.7 Multiplication Factors 7-6
7.8 Conversions of Units 7-7
7.9 Examples of Defining Terms 7-7
Bibliography 7-12

DeWayne B. Sharp
Shape of Things

Measurement standards are those devices, artifacts, procedures, instruments, systems, protocols, or processes that are used to define (or to realize) measurement units and on which all lower echelon (less accurate) measurements depend. A measurement standard may also be said to store, embody, or otherwise provide a physical quantity that serves as the basis for the measurement of the quantity. Another definition of a standard is the physical embodiment of a measurement unit, by which its assigned value is defined and to which it can be compared for calibration purposes. In general, it is not independent of physical environmental conditions, and it is a true embodiment of the unit only under specified conditions. Another definition of a standard is a unit of known quantity or dimension to which other measurement units can be compared.

7.1 Historical Perspective

Many early standards were based on the human body: the length of man's hand, the width of his thumb, the distance between outstretched fingertips, the length of one's foot, a certain number of paces, etc. In the beginning, while groups were small, such standards were convenient and uniform enough to serve as the basis for measurements.

The logical person to impose a single standard was the ruler of the country—hence, our own 12 in. or other short measuring stick is still called a *ruler*. The establishment of measurement standards thus became the prerogative of the king or emperor, and this right has since been assumed by all governments.

History is replete with examples that show the importance of measurements and standards. In a report to the U.S. Congress in 1821, John Quincy Adams said, "Weights and measures may be ranked among the necessaries to every individual of human society." Our founding fathers thought them so important that the U.S. Constitution expressly gives the Congress the power to fix uniform standards of weights and measures. The need for weights and measures (standards) dates back to earliest recorded history and is even mentioned in the Old Testament of the Bible. Originally, they were locally decreed to

serve the parochial needs of commerce, trade, land division, and taxation. Because the standards were defined by local or regional authorities, differences arose that often caused problems in commerce and early scientific investigation. The rapid growth of science in the late seventeenth century highlighted a number of serious deficiencies in the system of units then in use and, in 1790, led the French National Assembly to direct the French Academy of Sciences to "deduce an invariable standard for all measures and all the weights." The academy proposed a system of units, the metric system, to define the unit of length in terms of the earth's circumference, with the units of volume and mass being derived from the unit of length. Additionally, they proposed that all multiples of each unit be a multiple of 10.

In 1875, the United States and 16 other countries signed the "Treaty of the Meter," establishing a common set of units of measure. It also established an International Bureau of Weights and Measures (called the BIPM). That bureau is located in the Parisian suburb of Sèvres. It serves as the worldwide repository of all the units that maintain our complex international system of weights and measures. It is through this system that compatibility between measurements made thousands of miles apart is currently maintained.

The system of units set up by the BIPM is based on the meter and kilogram instead of the yard and the pound. It is called the Système International d'Unités (SI) or the International System of Units. It is used in almost all scientific work in the United States and is the only system of measurement units in most countries of the world today.

Even a common system of units does not guarantee measurement agreement, however. Therein lies the crux of the problem. We must make measurements, and we must know how accurately (or, to be more correct, with what uncertainty) we made those measurements. In order to know that point, there must be standards. Even more important, everyone must agree on the values of those standards and use the same standards.

As the level of scientific sophistication improved, the basis for the measurement system changed dramatically. The earliest standards were based on the human body, and then attempts were made to base them on "natural" phenomena. At one time, the basis for length was supposed to be a fraction of the circumference of the earth, but it was "maintained" by the use of a platinum/iridium bar. Time was maintained by a pendulum clock but was defined as a fraction of the day and so on. Today, the meter is no longer defined by an artifact. Now, the meter is the distance that light travels in an exactly defined fraction of a second. Since the speed of light in a vacuum is now defined as a constant of nature with a specified numerical value (299,792,458 m/s), the definition of the unit of length is no longer independent of the definition of the unit of time.

Prior to 1960, the second was defined as 1/86,400th of a mean solar day. Between 1960 and 1967, the second was defined in terms of the unit of time implicit in the calculation of the ephemerides: "The second is the fraction 1/31, 556, 925.9747 of the tropical year for January 0 at 12 hours of ephemeris time." With the advent of crystal oscillators and, later, atomic clocks, better ways were found of defining the second. This, in turn, allowed a better understanding of things about natural phenomena that would not have been possible before. For example, it is now known that the earth does not rotate on its axis in a uniform manner. In fact, it is erratically slowing down. Since the second is maintained by atomic clocks, it is necessary to add "leap seconds" periodically so that the solar day does not gradually change with respect to the time used every day. It was decided that a constant frequency standard was preferred over a constant length of the day.

7.2 What Are Standards?

One problem with standards is that there are several kinds. In addition to "measurement standards," there are "standards of practice or protocol standards" that are produced by the various standard bodies such as the International Organization for Standardization (ISO), the International Electrotechnical Commission (IEC), the American National Standards Institute (ANSI), and the Standards Council of Canada (SCC) (see Figure 7.1).

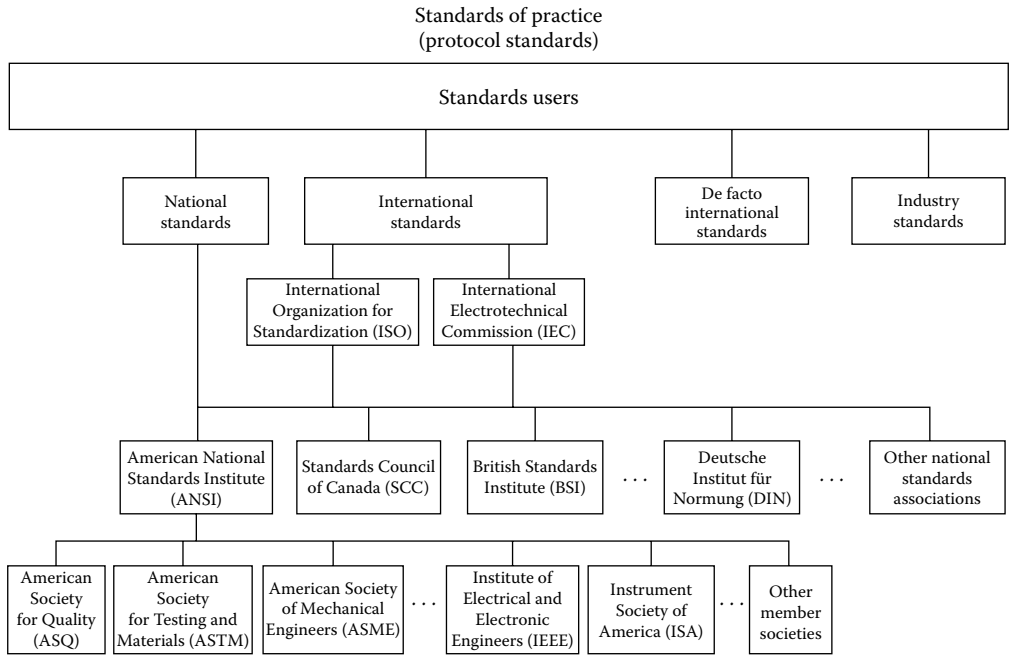


FIGURE 7.1 Standardization organizations and standards structure.

7.2.1 Standards of Practice (Protocol Standards)

These standards define everything from the dimensions and electrical characteristics of a flashlight battery to the shape of the threads on a machine screw and from the size and shape of an IBM punched card to the quality assurance requirements for measuring equipment. Such standards can be defined as documents describing the operations and processes that must be performed in order for a particular end to be achieved. They are called a "protocol" by Europeans to avoid confusion with a physical standard.

7.2.2 Legal Metrology

The application of measurement standards to the control of the daily transactions of trade and commerce is known as legal metrology; within the United States, it is more commonly known as weights and measures. Internationally, coordination among nations on legal metrology matters is, by international agreement, handled by a quasi-official body—the International Organization for Legal Metrology (OIML).

Within the United States, domestic uniformity in legal metrology matters is the responsibility of National Institute of Standards and Technology (NIST) acting through its Office of Weights and Measures. Actual enforcement is the responsibility of each of the 50 states and the various territories. These, in turn, generally delegate the enforcement powers downward to their counties and, in some cases, to large cities.

7.2.3 Forensic Metrology

Forensic metrology is the application of measurements and hence measurement standards to the solution and prevention of crime. It is practiced within the laboratories of law enforcement agencies throughout the world. Worldwide activities in forensic metrology are coordinated by Interpol (*International Police*; the international agency that coordinates the police activities of the member nations). Within the United States, the Federal Bureau of Investigation (FBI), an agency of the Department of Justice, is the focal point for most U.S. forensic metrology activities.

7.2.4 Standard Reference Materials

Another type of standard that should be mentioned here is the Standard Reference Materials (SRMs). SRMs are discrete quantities of substances or minor artifacts that have been certified as to their composition, purity, concentration, or some other characteristics useful in the calibration of the measurement devices and the measurement processes normally used in the process control of those substances. SRMs are the essential calibration standards in stoichiometry (the metrology of chemistry).

In the United States, the NIST, through its SRM program, offers for sale over 1300 SRMs. These range from ores to pure metals and alloys. They also include many types of gases and gas mixtures and many biochemical substances and organic compounds. Among the artifact devices available are optical filters with precise characteristics and standard lamps with known emission characteristics.

7.3 Conceptual Basis of Measurements

Lord Kelvin's oft-quoted statement may bear repeating here:

> I often say that when you can measure what you are speaking about, and can express it in numbers, you know something about it; but when you cannot measure it, cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind; it may be the beginnings of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science, whatever the matter may be. So therefore, if science is measurement, then without metrology there can be no science.
>
> **William Thomson (Lord Kelvin), May 6, 1886**

Lord Kelvin's statement has been quoted so many times that it has almost become trite, but looking at Figure 7.2 will show an interesting hierarchy. In order to achieve quality or "to do things right," it is

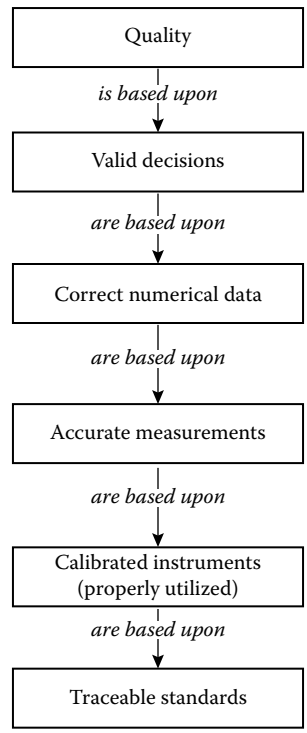


FIGURE 7.2 Decision-making process in applications of standards.

necessary to make some decisions. The correct decisions cannot be made unless there are good numerical data on which to base those decisions. Those numerical data, in turn, must come from measurements, and if "correct" decisions are really needed, they must be based on the "right" numbers. The only way to get "good" numerical data is to make accurate measurements using calibrated instruments that have been properly utilized. Finally, if it is important to compare those measurements to other measurements made at other places and other times, the instruments must be calibrated using traceable standards.

7.4 Need for Standards

Standards define the units and scales in use and allow comparison of measurements made in different times and places. For example, buyers of fuel oil are charged by a unit of liquid volume. In the United States, this would be the gallon, but in most other parts of the world, it would be the liter. It is important for the buyer that the quantity ordered is actually received and the refiner expects to be paid for the quantity shipped. Both parties are interested in accurate measurements of the volume and, therefore, need to agree on the units, conditions, and method(s) of measurement to be used.

Persons needing to measure a mass cannot borrow the primary standard maintained in France or even the national standard from the NIST in the United States. They must use lower-level standards that can be checked against those national or international standards. Everyday measuring devices, such as scales and balances, can be checked (calibrated) against working-level mass standards from time to time to verify their accuracy. These working-level standards are, in turn, calibrated against higher-level mass standards. This chain of calibrations or checking is called "traceability." A proper chain of traceability must include a statement of uncertainty at every step.

7.5 Types of Standards

7.5.1 Basic or Fundamental Standards

In the SI system, there are seven basic measurement units from which all other units are derived. All of the units except one are defined in terms of their unitary value. The one exception is the unit of mass. It is defined as 1000 g or 1 kg. It is also unique in that it is the only unit currently based on an artifact. The U.S. kilogram and hence all other standards of mass are based on one particular platinum/iridium cylinder kept at the BIPM in France. If that International Prototype Kilogram were to change, all other mass standards throughout the world would be wrong.

The seven basic units are listed in Table A.1. Their definitions are listed in Table A.2.

7.5.2 Derived Standards

All of the other units are derived from the seven basic units described in Table A.1. Measurement standards are devices that represent the SI standard unit in a measurement. (For example, one might use a Zener diode together with a reference amplifier and a power source to supply a known voltage to calibrate a digital voltmeter. This could serve as a measurement standard for voltage and be used as a reference in a measurement.)

Table A.3 lists the most common derived SI units, together with the base units that are used to define the derived unit. For example, the unit of frequency is the hertz; it is defined as the reciprocal of time. That is, 1 hertz (1 Hz) is one cycle per second.

7.5.3 Measurement Assurance System

Figure 7.3 illustrates the interrelationship of the various categories of standards throughout the world. While it gives more detail to U.S. structure, similar structures exist in other nations. Indeed, a variety of regional organizations exist that help relate measurements made in different parts of the world to each other.

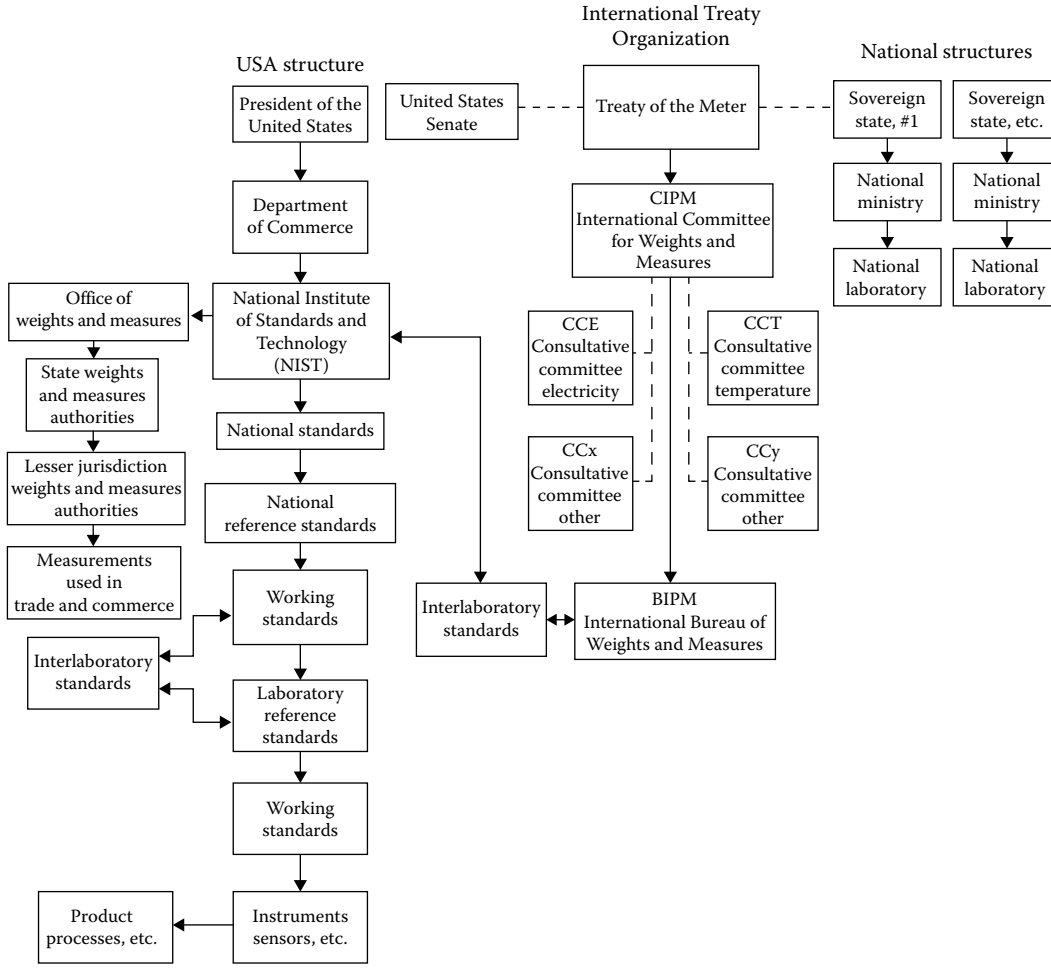


FIGURE 7.3 Measurement assurance systems.

7.6 Numbers, Dimensions, and Units

A measurement is always expressed as a multiple (or submultiple) of some unit quantity. That is, both a numeric value and a unit are required. If electric current were the measured quantity, it might be expressed as some number of milliamperes or even microamperes. It is easy to take for granted the existence of the units used, because their names form an indispensable part of the vocabulary.

7.7 Multiplication Factors

Since it is inconvenient to use whole units in many cases, a set of multiplication factors has been defined that can be used in conjunction with the units to bring a value being measured to a more reasonable size. It would be difficult to have to refer to large distances in terms of the meter; thus, one defines longer distances in terms of kilometers. Short distances are stated in terms of millimeters, micrometers, nanometers, etc. (see Table A.4).

7.8 Conversions of Units

Historically, many different units are involved in different places of the world or even within the same country; the relationships between different units are defined in fixed terms. For example, 1 lb = 453.59237 g. A partial list of conversions can be found in Table A.5. A more comprehensive list can be found in Liptak and Eren (2011).

Further information on standardization and processes can be found in Chapter 6.

7.9 Examples of Defining Terms

Most of the definitions in this listing were taken from the *International Vocabulary of Basic and General Terms in Metrology*, published by the ISO, 1993 (VIM) (ISO, 1993). They are indicated by the inclusion (in brackets) of their number designation in the VIM. The remainder of the definitions is not intended to represent any official agency but is ones widely accepted and is included to help in the understanding of this material. More detailed and rigorous definitions can be found in other works available from ANSI, IEC, ISO, and NIST. Words enclosed in parentheses "(…)" may be omitted from the term if it is unlikely that such omission will cause confusion.

Accuracy of measurement [3.5]: The closeness of the agreement between the result of a measurement and a true value of the measurand.

Notes:

1. *Accuracy* is a qualitative concept.
2. The term *precision* should not be used for *accuracy*. (Precision only implies repeatability.)

Note, that to say an instrument is accurate to 5% (a common way of stating it) is wrong. One would not find such an instrument very useful if it, in fact, were only accurate 5% of the time. What is meant when such a statement is made is that the instrument's inaccuracy is less than 5% and it is accurate to better than 95%. Unfortunately, this statement is almost as imprecise as "accurate to 5%." An instrument would not be useful if it were accurate only 95% of the time; but this is not what is implied by "5% accuracy." What is meant is that, (almost) all of the time, its indication is within 5% of the "true" value.

Calibration [6.11]: A set of operations that establish, under specified conditions, the relationship between values of quantities indicated by a measuring instrument or measuring system, or values represented by a material measure or a reference material, and the corresponding values realized by standards.

Notes:

1. The result of a calibration permits either the assignment of values of measurands to the indicators or the determination of corrections with respect to indications.
2. A calibration can also determine other metrological properties, such as the effect of influence quantities.
3. The result of a calibration can be recorded in a document, sometimes called a *calibration certificate* or a *calibration report*.

Calibration laboratory: A work space, provided with test equipment, controlled environment, and trained personnel, established for the purpose of maintaining proper operation and accuracy of measuring and test equipment. *Calibration laboratories* typically perform many routine calibrations, often on a production-line basis.

Certified reference material (CRM) [6.14]: A *reference material*, accompanied by a certificate, one or more of whose property values are certified by a procedure that established traceability to an accurate realization of the unit in which the property values are expressed and for which each certified value is accompanied by an uncertainty at a stated level of confidence

1. The definition of a *reference material* is given elsewhere in this vocabulary.
2. CRMs are generally prepared in batches for which the property values are determined within stated uncertainty limits by measurements on samples representative of the entire batch.
3. The certified properties of *CRM*s are sometimes conveniently and reliably realized when the material is incorporated into a specifically fabricated device, for example, a substance of known triple-point into a triple-point cell, a glass of known optical density into a transmission filter, and spheres of uniform particle size mounted on a microscope slide. Such devices can also be considered CRMs.
4. All CRMs lie within the definition of "measurement standards" given in the VIM.
5. Some RMs and CRMs have properties that, because they cannot be correlated with an established chemical structure or for other reasons, cannot be determined by exactly defined physical and chemical measurement methods. Such materials include certain biological materials such as vaccines to which an international unit has been assigned by the World Health Organization.

This definition, including the Notes, is taken from ISO Guide 30:1992.

Coherent (derived) unit (of measurement) [1.10]: A derived unit of measurement that may be expressed as a product of powers of base units with the proportionality factor one (1).

Note: Coherency can be determined only with respect to the base units of a particular system. A unit can be coherent with respect to one system but not to another.

Coherent system of units (of measurement) [1.11]: A system of units of measurement in which all of the derived units are coherent.

Conservation of a (measurement) standard [6.12]: A set of operations necessary to preserve the metrological characteristics of a measurement standard within appropriate limits.

Note: The operations commonly include periodic calibration, storage under suitable conditions, and care in use.

Interlaboratory standard: A device that travels between laboratories for the sole purpose of relating the magnitude of the physical unit represented by the standards maintained in the respective laboratories.

International (measurement) standard [6.2]: A standard recognized by an international agreement to serve internationally as the basis for assigning values to other standards of the quantity concerned.

International System of Units (SI) [1.12]: The coherent system of units adopted and recommended by the General Conference on Weights and Measures (CGPM).

Note: The SI is based at present on the following seven base units: meter, kilogram, second, ampere, Kelvin, mole, and candela.

Measurand [2.6]: A particular quantity subject to measurement.

Example: Vapor pressure of a given sample of water at 20 °C.

Note: The specification of a measurand may require statements about quantities such as time, temperature, and pressure.

Measurement [2.1]: A set of operations having the object of determining a value of a quantity.

Note: The operations may be performed automatically.

Method of measurement [2.4]: A logical sequence of operations, described generically, used in the performance of measurements.

Note: Methods of measurement may be qualified in various ways, such as

- Substitution method
- Differential method
- Null method

Metrology [2.2]: The science of measurement.

Note: Metrology includes all aspects, both theoretical and practical, with reference to measurements, whatever their uncertainty and in whatever fields of science or technology they occur.

National (measurement) standard [6.3]: A standard recognized by a national decision to serve, in a country, as the basis for assigning values to other standards of the quantity concerned.

National reference standard: A standard maintained by national laboratories such as the NIST in Gaithersburg, MD; the National Research Council (NRC) located in Ottawa, Canada; the National Physical Laboratory (NPL) in Teddington, U.K.; and the Physikalisch-Technische Bundesanstalt (PTB) at Braunschweig, Germany, which are the legal standards of their respective countries.

NIST: The U.S. national standards laboratory, responsible for maintaining the physical standards upon which measurements in the United States are based.

Primary standard [6.4]: A standard that is designated or widely acknowledged as having the highest metrological qualities and whose value is accepted without reference to other standards of the same quantity.

Note: The concept of primary standard is equally valid for base quantities and derived quantities.

Principle of measurement [2.3]: The scientific base of a measurement.

Examples

- The thermoelectric effect applied to the measurement of temperature
- The Josephson effect applied to the measurement of electric potential difference
- The Doppler effect applied to the measurement of velocity
- The Raman effect applied to the measurement of the wave number of molecular vibrations

Reference standard [6.6]: A standard, generally having the highest metrological quality available at a given location or in a given organization, from which measurements made there are derived.

Reference material [6.13]: A material or substance, one or more of whose property values are sufficiently homogeneous and well established to be used for the calibration of an apparatus, the assessment of a measurement method, or for assigning values to materials.

Note: A reference material can be in the form of a pure or mixed gas, liquid, or solid. Examples are water for the calibration of viscometers, sapphire as a heat-capacity calibrant in calorimetry, and solutions used for calibration in chemical analysis.

This definition, including the Note, is taken from ISO Guide 30:1992.

Repeatability (of results of measurements) [3.6]: The closeness of the agreement between the results of successive measurements of the same measurand carried out under the same conditions of measurement.

Notes:

1. These conditions are called *repeatability conditions*.
2. Repeatability conditions include
   a. The same measurement process
   b. The same observer
   c. The same measuring instrument, used under the same conditions
   d. The same location
   e. Repetition over a short period of time
3. Repeatability can be expressed quantitatively in terms of the dispersion of characteristics of the results.

Reproducibility (of results of measurements) [3.7]: The closeness of the agreement between the results of measurements of the same measurand carried out under changed conditions of measurement.

Notes:

1. A valid statement of reproducibility requires specification of the conditions changed.
2. The changed conditions include
   a. Principle of measurement
   b. Method of measurement
   c. Observer
   d. Measuring instrument
   e. Reference standard
   f. Location
   g. Condition of use
   h. Time
3. Reproducibility can be expressed quantitatively in terms of the dispersion characteristics of the results.
4. Results here are usually understood to be corrected results.

Secondary standard [6.5]: A standard whose value is assigned by comparison with a primary standard of the same quantity.

Standards laboratory: A work space, provided with equipment and standards, a properly controlled environment, and trained personnel, established for the purpose of maintaining traceability of standards and measuring equipment used by the organization it supports. Standard laboratories typically perform fewer, more specialized, and higher accuracy measurements than calibration laboratories.

Tolerance: In metrology, the limits of the range of values (the uncertainty) that apply to a properly functioning measuring instrument.

Traceability [6.10]: The property of the result of a measurement or the value of a standard whereby it can be related to stated references, usually national or international standards, through an unbroken chain of comparisons all having stated uncertainties.

Notes:

1. The concept is often expressed by the adjective *traceable*.
2. The unbroken chain of comparisons is called a *traceability chain*.

Even though the ISO has published (and accepted) the definition listed earlier, many practitioners endeavor to make this term more meaningful. They feel that the definition should introduce the aspect of evidence being presented on a continuing basis, to overcome the idea that if valid traceability is achieved, it could last forever. A definition similar to the following one would meet that requirement.

Traceability is a characteristic of a calibration or a measurement. A traceable measurement or calibration is achieved only when each instrument and standard, in a hierarchy stretching back to the national

(or international) standard, was itself properly calibrated and the results properly documented including statements of uncertainty on a continuing basis. The documentation must provide the information needed to show that all the calibrations in the chain of calibrations were appropriately performed.

Transfer standard [6.8]: A standard used as an intermediary to compare standards.

Notes: The term *transfer device* should be used when the intermediary is not a standard.

Traveling standard [6.9]: A standard, sometimes of special construction, intended for transport between locations.

Example: A portable battery-operated cesium frequency standard.

Uncertainty of measurement [3.9]: A parameter, associated with the result of a measurement that characterizes the dispersion of the values that could reasonably be attributed to the measurand.

Notes:

1. The parameter can be, for example, a standard deviation (or a given multiple of it) or the half-width of an interval having a stated level of confidence.
2. Uncertainty of measurement comprises, in general, many components. Some of these components can be evaluated from the statistical distribution of the results of series of measurements and can be characterized by experimental standard deviations. The other components, which can also be characterized by standard deviations, are evaluated from assumed probability distributions based on experience or other information.
3. It is understood that the result of the measurement is the best estimate of the value of the measurand and that all components of uncertainty (including those arising from systematic effects), such as components associated with corrections and reference standards, contribute to the dispersion.

This definition is that of the *Guide to the Expression of Uncertainty in Measurement*, in which its rationale is detailed (see, in particular, 2.2.4 and annex D) [4].

Value (of a quantity) [1.18]: The magnitude of a particular quantity generally expressed as a unit of measurement multiplied by a number.

Examples

- Length of a rod: 5.34 m or 534 cm
- Mass of a body: 0.152 kg or 152 g
- Amount of substance of a sample of water (H_2O): 0.012 mol or 12 mmol

Notes:

1. The value of a quantity can be positive, negative, or zero.
2. The value of a quantity can be expressed in more than one way.
3. The values of quantities of dimension one are generally expressed as pure numbers.
4. A quantity that cannot be expressed as a unit of measurement multiplied by a number can be expressed by reference to a conventional reference scale or to a measurement procedure or both.

Working standard [6.7]: A standard that is used routinely to calibrated or check material measures, measuring instruments, or reference materials.

Notes:

1. A working standard is usually calibrated against a *reference standard*.
2. A working standard used routinely to ensure that a measurement is being carried out correctly is called a *check standard*.

Bibliography

Alves, J. G., Developments in standards and other guidance for individual monitoring, *Radiation Measurements*, 43(2–6), 558–564, 2008.

Cochrane, R. C., Measures for progress—History of the National Bureau of Standards, United States Department of Commerce, Library of Congress Catalog Card Number: 65-62472, 1966.

JCGM/WG 2 Document N318, *International Vocabulary of Basic and General Concepts and Associated Terms, NIST*, http://www.nist.gov/pml/div688/grp40/upload/International-Vocabulary-of-Metrology.pdf (accessed on March 13, 2013).

Klein, H. A., *The Science of Measurement: A Historical Survey*, New York: Dover Publications, Inc., 1974.

Liptak, B. and H. Eren, *Instrument Engineers Handbook: Process Software and Digital Networks*, Vol. 3, 4th edn., LLC, Boca Raton, FL: CRC Press, pp. 953–961, 2011.

NIST Calibration Program, Calibration services users guide, SP 250 Appendix fee schedule 2011, NIST, http://www.nist.gov/calibrations/upload/feesch-11-2-2.pdf (accessed on March 13, 2013).

Publications, BIPM, http://www.bipm.org/en/publications/ (accessed on March 13, 2013).

Standards, Legistlation, Technical Information, Browse by the subject, SAI Global, http://infostore.saiglobal.com/store/ (accessed on March 13, 2013).

Taylor, B. N., NIST Special Publication 811, *Guide for the Use of the International System of Units* (*SI*), U.S. Department of Commerce, National Institute of Standards and Technology, Gaithersburg, MD, 1995.

Taylor, B. N. and C. E. Kuyatt, Guidelines for evaluating and expressing the uncertainty of NIST measurement results, NIST Technical Note 1297, 1994.

Velychko, O. and T. Gordiyenko, The use of metrological terms and SI units in environmental guides and international standards, *Measurement*, 40(2), 202–212, 2007.

8

Calibrations in Instrumentation and Measurements

8.1 Introduction **8**-1
8.2 Errors and Uncertainties in Calibrations **8**-2
8.3 Benefits of Calibrations **8**-3
8.4 Calibration Procedure and Personnel **8**-3
8.5 Calibration Methods **8**-4
Static Calibration • Dynamic Calibration
8.6 Laboratories and Institutions **8**-4
Calibration Records
8.7 Calibration Software Support **8**-5
Spreadsheets and Calibration Management Software
8.8 Cost of Calibrations **8**-6
Obtaining Information
8.9 Trends in Calibrations **8**-7
Electronic Calibration • Self-Calibration • Soft Calibration • Remote and e-Calibration, iCal
8.10 Calibration Examples **8**-9
Flow Calibration • Sensor Calibration • Calibration of Chemical Mixtures • Static Pressure Calibrations • Dynamic Pressure Calibrations
Bibliography **8**-11
Partial List of Calibration Service and Software Providers **8**-12

Halit Eren
Curtin University

8.1 Introduction

The Oxford English Dictionary explains the meaning and scope of calibration by defining it as follows: "(1) mark (a gauge or instrument) with a standard scale of readings, (2) compare the readings of (an instrument) with those of a standard, and (3) adjust (experimental results) to take external factors into account or to allow comparison with other data." Basically, calibration is the comparison between measurements and it is an essential component of measurements.

The ability of an instrument or a device to measure accurately depends on a number of factors, such as its duration in service, temperature, vibrations, humidity, environmental exposure, corrosion, electronic malfunctioning, drift, and changes in application conditions. Calibration quantifies the change in the measurements, and hence, periodical adjustments on the device become necessary to decrease or eliminate possible errors.

Calibration of instruments assures that the processes are well controlled and the products meet expected specifications since the measurements can drift away from their correct and calibrated values resulting in unreliable readings. For example, if the representation of process variables drifts away from their true range, it can result in costly production downtimes, create adverse safety issues, and lead to inferior quality goods being produced. Today, in our daily lives, we are very much dependent on the correct operations of instruments, for example, when driving vehicles, but we take them for granted with the expectation that they are all functioning well giving us correct readings, which may not be the case.

An important element in calibrations is the relationship between a single measurement and the reference base. The base unit of measurement is Le Système International d'Unités (SI) units maintained in the Bureau International des Poids et Mesures, Paris. These are *kilogram* for mass, *meter* for length, *second* for time, *candela* for luminous intensity, *Kelvin* for thermodynamics, *ampere* for current, and *mole* for amount of substance. Other reference bases such as *Newton* for force and *hertz* for frequency are derived from the base units and maintained by national standards institutions.

8.2 Errors and Uncertainties in Calibrations

Most instruments and sensors are designed to meet certain accuracy specifications; the process of adjusting an instrument to meet those specifications is known as the calibration. The device used to calibrate other instruments is known as a calibrator. Before the calibration is attempted, the calibrator itself must be calibrated to ensure that they themselves are performing error-free. Although calibration and accuracy are performed by the instrument manufacturers, it may also be necessary for the operators to perform a "user calibration." The user calibration includes the equipment settings, test setup, and verification of correct application to validate that the desired levels of precision and accuracy are achieved. The term *calibration standard* is used when an absolute measured value and the reference standards are traceable back to national and international primary standards.

Despite strict rules and procedures, most calibrations are likely to have errors and uncertainties in their final values. These calibration errors and uncertainties can be evaluated by implementing Type A and/or Type B techniques:

1. *Type A*: Evaluation applies to errors, uncertainties, and bias by using statistical techniques. The ISO guidelines, such as ISO 11095, give guidance on how to assess, correct, and calculate errors, uncertainties, and biases.
2. *Type B*: Evaluations can apply to errors, uncertainties, and bias too. The calculation of the uncertainty component is not based on a statistical analysis but on factors such as experience, scientific judgment, scant data, and the use of different laboratory assessments.

However, in many situations, it may be impossible to achieve a perfect calibration because of instrument and measurement biases and uncontrollable random errors. These can mathematically be expressed as

$$\text{Ideal value} = \text{Measured value} + \text{Bias} + \text{Error}$$

Similarly, reference value may be subject to bias and error:

$$\text{Ideal reference value} = \text{Reference value} + \text{Bias} + \text{Error}$$

This leads to a deficiency in the calibration as

$$\text{Deficiency} = \text{Ideal measured value} - \text{Ideal reference value}$$

Due to the randomness of errors, this deficiency may not be zero; therefore, in some cases, the calibration measurements may have to be repeated many times. Then the statistical techniques can be applied to work out the calibration curves, average readings, and standard deviations. The process of collecting data for creating the calibration curve is critical to the success of the calibration program.

8.3 Benefits of Calibrations

Usually, the calibration procedure involves comparison of the instrument against primary or secondary standards. In some cases, it may be sufficient to calibrate a device against another one with a known accuracy. After the calibration of a device or a process, future operation is considered to be error-bound for a period of time under similar operational conditions thus yielding to the following benefits:

- Confidence on the future measurements
- Consistency and compatibility
- Repeatability and reproducibility
- Products meeting their specifications, thus reducing legal liability
- Proper documentation to meet quality standards such as those set by the ISO
- Frequent calibrations can provide graphical view of the equipment uncertainty over time and lead to reliability in performance
- Measurements made within international standards promote global acceptance thus increasing competitiveness
- As the technology changes, the regulations and legislation of test and measuring instruments change continually and calibration helps compliance validity of measurements and processes under changing conditions

8.4 Calibration Procedure and Personnel

A successful calibration process requires hardware and software, special equipment, and skilled personnel. It is a process that assigns values to the response of an instrument relative to reference standards or to a designated process. The aim is to eliminate or reduce the errors, uncertainties, and biases in the measurements relative to the reference standard. In order to achieve this aim, a step-by-step calibration procedure and following this procedure with the detailed instructions are essential. In doing so, the following factors must be considered:

1. The type of calibration process to be employed
2. Calibration equipment and environment setup
3. Calibration cycles
4. Types of records, reports, and report keeping
5. Possible factors that may affect the calibration
6. Mathematical analysis and tools
7. In-house versus outsourced calibration
8. Reference standards to be used
9. Traceability issues
10. ISO 9000 compliance
11. Published standards to be used
12. Handling measurement uncertainties and errors
13. Identification of random and systematic errors

Once these factors have been noted, an appropriate calibration procedure must be followed. Calibrations must be repeatable and mathematically expressible.

Skilled personnel are essential to obtain successful calibrations. The personnel may have appropriate clothing (static-free clothing, gloves, face masks, etc.) for safety as well as to avoid the possibility of contamination. Although fully trained and experienced, it is possible for different operators to produce different results. To overcome this problem, measurements by different operators can be plotted and compared. Another solution may be to employ separate calibration curves by the same operator. This may not be a problem in automated calibrations.

8.5 Calibration Methods

Two basic calibration methods are the static calibration and the dynamic calibration.

8.5.1 Static Calibration

Static calibration aims the tuning of static characteristics of instruments under off-line conditions.

8.5.2 Dynamic Calibration

Dynamic calibration aims tuning of the dynamic characteristics, while the process is taking place. It makes use of multisampling points to dynamically decide stable and accurate operations of the process under different conditions.

In many dynamic applications, calibrated variables have multiple inputs and multiple outputs. In these cases, an input is varied in increments in increasing and decreasing directions over a specified range. The observed outputs then become a function of that single input. This procedure may be repeated by varying other inputs, thus developing a family of relationships between inputs and outputs. In these multivariable situations, the input/output relationship usually demonstrates statistical characteristics. From these characteristics, appropriate calibration curves can be obtained, and statistical techniques can be applied by using various mathematical tools, such as averaging, weighed average, and multi-fixed-range averaging or standard-deviation-range averaging.

8.6 Laboratories and Institutions

In many organizations, calibrations are conducted either on-site or in laboratories or remote locations. Large organizations may have several calibration laboratories dedicated for different instruments and processes. Laboratories are accredited by authorities in accordance with the guidelines such as the ISO Guide 58. Accreditation is a formal recognition that a particular laboratory is competent to conduct specific tests and/or calibrations.

During the calibration process, the readings of the test item are compared with the reference standards such as resistors, length standards, and voltage standards.

A successful laboratory calibration procedure requires the following basic steps:

- Selection of an appropriate reference standard with known values covering the range of interest
- Conducting calibration curves (i.e., least-squares fit) to establish the relationship between the measured and known values of the reference standard
- Correction of measurements by using calibration curves
- Preparation of the appropriate documentation of the calibration procedure, results, analysis, and interpretation of results for the client

Once the procedure is adapted, the calibration relies on the instrument continuing to respond consistently and in the same way during the calibrations. If the system drifts or takes unpredictable deviations, the calibrated values may not be easily corrected for errors and bias, thus degrading the accuracy of the measurements.

Many institutions provide comprehensive calibration services. A typical institution can provide calibration of multitude of variables, such as force, pressure, weight and mass, metrological values, dimensions, temperature, flow, and electronics test and measurement equipment.

Customers of an institution can be from a wide range of industries, such as automotive, aviation, contractor engineering, food and beverage, manufacturing, marine, metal and mining, nuclear, oil and gas, petrochemical and chemical, pharmaceutical, power and energy, and pulp and paper industries.

Some institutions develop, manufacture, and market calibration equipment, software, systems, and services. Their product range includes portable calibrators, workstations, calibration software, accessories, industry-specific solutions, and professional services. They comply with customer requirements for accuracy, versatility, efficiency, ease of use, compatibility, and reliability.

8.6.1 Calibration Records

Calibration history is an important part of knowing data quality. The records can be kept in software environment or in printed formats. Many companies provide software in data file format that enables the sharing of structured data across different information systems. Data fields such as position ID, device ID, location, serial number, and work order number can be transferred from one form to another suitable for engineers to evaluate or the managers to view. In addition, the data can include maximum errors, pass/fail notifications, calibration date and time, calibration frequency, and who carried out the calibration task.

Once calibration is completed, appropriate labeling is used for that instrument to supply information on the applicability of instruments. The labels indicate parameters of instrument, tolerances, and special conditions of use. Calibration labels need to conform to the requirements of ISO/IEC 17025 General Requirements for the Competence of Testing and Calibration Laboratories. ISO 17025 requires that the following conditions be met:

- All measurement equipment shall be securely and durably labeled
- The labels should clearly indicate the name of the calibration laboratory, date of calibration, due date, usage equivalent, and the authorized officer
- Information on label must be legible and durable under reasonable use and storage conditions
- When it is impractical to affix a label directly on an item, the label may be affixed to the instrument container
- Temper resistance seals may be used when necessary
- Functional labels should contain reference standards

8.7 Calibration Software Support

Calibrations are mostly conducted by using computers to capture and analyze data. Once the results are obtained, software packages assist in analyzing the information. Most packages use method of least squares for estimating the coefficients. Some of the packages are capable of performing weighted fit if the errors of the measurements are not constant over the calibration interval. The spreadsheets and software packages provide information such as the coefficients of the calibration curve, standard deviations, residual standard deviation of the fit, and goodness of the fit.

8.7.1 Spreadsheets and Calibration Management Software

A set of spreadsheets are used to perform random-number-driven simulations, including single standard, bracket, calibration curve, and standard addition methods. These simulations include additive and multiplicative interferences (systematic errors) and random errors in both signal and measurements. It is possible to observe how nonlinearity, interferences, and random errors combine and attempt to

optimize precision and accuracy of the measurement. Some of these spreadsheet models are linear fit, nonlinear fit, single standard addition, multiple standard addition, and others.

In addition to spreadsheets, there are many calibration management software (CMS) to schedule calibrations, track usage and locations, maintain histories, generate work orders, issue certificates, and print calibration labels. Most CMS will support web-based access. They support information on short- and long-term effects on the devices and their linearity and stability.

CMS may have automated features. An automated system promotes efficient management of plant assets and field instruments in one database. They generate future calibration schedules and alerts on the past failed calibrations. Most CMS complies with the ISO 17025 requirements. Some examples of the CMS are as follows:

1. *Ape Software* (calibration control): intuitive database program that organizes calibration data for gages and other test equipment
2. *AssetSmart*: web-based asset, calibration, material, and tool maintenance management software
3. *BDR Systems*: preventive calibration system that provides monitoring, tracking, and control of the calibration function
4. *Beamex*: CMS for documenting, planning, analyzing, and optimizing calibration work
5. *CompuCal Software*: online compliant calibration and maintenance management that ensure quality compliance with 21 CFR Part 11, GxP, ISO, and PAT initiatives
6. *CyberMetrics*: GAGEtrak—CMS
7. *Diversified Data Systems* (OpenMETRIC): Calibration and Metrology Data Management and Tracking Software for ISO 17025 and ISO 9000 with 21 CFR Part 11
8. *Honeywell International*: DocuMint automated calibration management system that promotes efficient management of plant assets and field instruments in one database
9. *IBM*: Maximo Calibration—provides all requirements for traceability and reverse traceability, all calibration history data, calibration data sheets, and reporting

8.8 Cost of Calibrations

The cost of calibration depends on what is calibrated and who is calibrating it. In simple cases where a single instrument is involved, the cost can be lower than one hundred dollars, but some complex cases can cost thousands of dollars. Calibration cost depends on where the calibration is carried out whether it takes place in special laboratories or on the plant. The cost of calibration may be dependent on the availability of expert in-house personnel or outsourcing it to third parties. Certification by ISO 10012-1, ISO 9001, MIL-STD 45662A, and MIL-HDBK-52B requires careful calibration of the measuring equipment used in the process. Also, in some cases such as the weighting systems, calibration is a statutory requirement.

One of the major factors for cost is the frequency of calibration. In most cases, a validity period is issued during which calibrated devices can be used without concern for major errors and uncertainties. Some organizations conduct calibrations in regular intervals, while others opt for conservative calibration intervals simply to meet the legal demands. Nevertheless, the use of uncalibrated instruments in an organization can be costly as it may affect the product quality and the quality of downstream operations.

Standards such as MIL-STD 45662A suggest regular and well-organized calibration intervals. As a rule of thumb, 85%–95% of all instruments returned for calibration must meet the calibration limits determined by the probability chart of age and failure data. Usually an instrument must be calibrated if failure rate increases or functionality deteriorates when compared to other standard devices. In this respect, a number of different mathematical techniques, such as the Weibull statistics and renewal equations, can be employed. A range of software, for example, the visualSMITH and Calibration Manager, is available for determining the calibration intervals and the cost analysis.

8.8.1 Obtaining Information

Information on calibration is available from various sources. These are

1. *Manufacturers* supply comprehensive information about calibration requirements of their products
2. *Regulating authorities and standards institutions* supply information about calibration requirements of instruments and devices. Calibration can be statutory particularly where health and safety are important
3. *Calibration services* provide information on calibration processes
4. *Organizations* provide rules and regulations for their equipments for assurance planning
5. *Books* provide information on calibration issues and processes

Many nations and organizations maintain laboratories with the primary functions of calibrating instruments and field measuring systems that are used in everyday operations. Examples of these laboratories are Standards Council of Canada (SCC), National Institute of Standards and Technology (NIST), National Association of Testing Authorities (NATA) of Australia and the British Calibration Services (BCS), and Australian Standards (AS). Some important information on calibration can be found in sources of ISO, IEC, IEEE, and national standards publications. Few of these publications will be briefly explained in the following.

ISO 17025 is an international standard for calibration and testing laboratories. It requires labs to demonstrate that they operate a quality system covering processes, documentation, and quality management. The laboratories need to generate technically valid results accounting of the equipment procedures and personnel. Information on ISO 1705 can be found in many books or ISO publications (http://www.fasor.com/iso25/).

The military standards MIL-STD-45662A is a standard that describes the requirements for creating and maintaining calibration systems for measurement and testing (available at http://store.milstandards.com/).

8.9 Trends in Calibrations

With the availability of advancing technology, the classical calibration process is changing at least in three fronts, these being

1. Electronic calibrations
2. Intelligent, soft calibrations, and self-calibrations
3. Remote and e-calibrations

These techniques will be explained next and some examples will be given.

8.9.1 Electronic Calibration

Many modern instruments offer features for closed-case calibrations so that electronic calibration can be employed. Electronic calibration is a single connection and one- or two-port calibration technique without tempering the components inside the case. Once the calibrating equipment, for example, computer, is linked with the device under calibration, appropriate software generates the necessary calibration information. Errors due to gains and offsets of the instrument are corrected mathematically within the instrument processor to obtain the correct measured values. Analog corrections can also be made via the adjustment of the parameters of some components such as the digital-to-analog converters. Corrected calibration constants are kept within the nonvolatile memory for permanent use.

As an example of this method, Agilent electronic calibration modules 8719, 8720, and 8722 microwave network analyzers provide a broad frequency range of calibrations from 10.0 MHz to 67.0 GHz

(http://we.home.agilent.com/). Similarly, Fluke offers electronic calibration facilities for multifunction process, pressure, and temperature instruments (http://www.fluke.com/).

8.9.2 Self-Calibration

With the wide applications of digital systems, many intelligent instruments are available and widely used in process industry. Many of these devices have self-calibrating features. In self-calibrations, post-measurement corrections are made, and the magnitudes of various errors are stored in the memory to be recalled and used in laboratory and in field applications.

Apart from the intelligent instruments, many intelligent (smart) sensors are being employed. Particularly, sensors complying with IEEE 1451.4 standards provide comprehensive Transducer Electronic Data Sheets (TEDS) that contain configuration, scaling, and calibration information necessary to make measurement through mixed-mode interface. Many smart sensors are capable of calibrating themselves, scaling the incoming data, computing statistics, and communicating with other digital systems on the network. Most commercially available software permits calibration of smart sensors and uploads the new parameters directly to the circuitry of the sensors. More information can be found on this topic in the chapter dedicated on intelligent instruments in this book.

In an application of smart sensors (e.g., the Atmos SSP14 Sensor Signal Processor family; http://www.sensorsmag.com/), the sensors have memory programmed at the factory with a set of default zero values and span curves defining the relationships with the physical phenomenon. These default curves represent average sensor output adjusted for the most accurate response at the room temperature. Each time calibration is done, an appropriate zero or span curve adjusted in the vicinity of the desired functionality points.

8.9.3 Soft Calibration

Soft calibration is a strategy for constructing a global multivariate calibration model that includes calibration samples measured over time on different days and is used in analysis.

8.9.4 Remote and e-Calibration, iCal

A new trend of calibration is the Internet calibration. Internet calibration technique is supported by web-accessible test procedures and appropriate hardware and software.

As examples of Internet calibration, Fluke, U.K., offer calibration systems for the Fluke 4950 multifunctional instruments (http://www.npl.co.uk/). On the other hand, Anritsu is active in investigating the possibilities of Internet-based calibration by using portable Optical Time Domain Reflectometer (OTDR) MW9076 (http://www.electronicstalk.com/news/anr/anr163.html). The OTDR is controlled via the Internet with the aid of appropriate software. The software can be controlled through a PC via modem, mobile telephone access, or PCI cards. Internet calibration is supported by appropriate software that includes the mathematical models of the interface electronics and optimizes calibration on that basis of this model.

Regular calibration is essential for quality and traceability to national and international standards. One way of achieving traceability is to send the instruments to be calibrated and acquiring certificate and correction values from a National Measurement Institute (NMI). In many cases, the value of settings can be affected by transport leading to uncertainty. This exercise can be costly and time consuming. Also, in many cases of the process industry, some of the equipment is bulky for transport and dynamic calibrations are necessary on-site. A solution to these problems is the e-calibrations through the Internet. This reduces the transportation needs, environmental effects, downtime, cost issues, and the need for in-house expertise.

Internet assists calibrations to take place in a number of ways:

1. Internet acts as a means of exchanging information between two remote sides with human operators at each end using interactive links
2. Remote monitoring of the sensors

Internet calibration consists of the operations to be performed for maintaining the system tuned for properly executing the process tasks in hand. Measurements rely on the available standards and the associated software suitable for the Internet calibrations. Typical requirements are

1. A computer connected to the Internet
2. Appropriate calibration software
3. A login name and password to use the service
4. Understanding of basic calibration principles
5. Understanding of the principles and functionality of the instrument under calibration
6. Understanding of scientific principles

Once a computer has been connected to the measurement system to be calibrated, the computer needs to log on to the calibration service provider. Services that offer online calibrations are known as *Internet Calibration Services* or iCals.

Once connected to the system, iCal instructs the operator, in the correct order, to perform the measurements. Effectively, it provides the operator with a soft procedure to perform the measurements. Once the measurement procedure is completed, the system generates the required final data with uncertainties ascertained from the measured and database information, providing the operator with a certificate if applicable.

During the measurement process, there is potential for the iCal system to provide measurement assistance screens with video or procedural details for specific parts of the process. In addition, it provides the administration with any measurement procedural changes through new international standards and ensures all groups are following common guidelines. The new data and historical records are automatically kept on the websites and they can be accessed and checked by the service users.

8.10 Calibration Examples

There are examples of many other Internet calibration and metrology in the process industry and medical and optical network markets. The optical communications network provides a method of a real standard transfer for wavelength measurements.

One of the first iCal services was developed by the United Kingdom's National Physical Laboratory (NPL). It was combined with the technology of remote monitoring, remote control, and NMI calibration techniques for application in a microwave frequency measurement system called *vector network analyzers* (VNAs).

In this application, the client enters the required measurement parameters and is offered options based on the knowledge that the NMI has about the client's equipment. From this point, the entire measurement process is controlled by NMI. VNAs provide a swept frequency measurement of the transmission and reflection coefficients for an electrical network. Calibration is performed using instrument firmware and a set of standard devices, all of which are assumed to be ideal and are available as standard items from the VNA manufacturer. The correction of the measurement data to that of the NMI comes via precision verification artifacts, air-spaced transmission lines, attenuators, and terminations, whose properties change little over time. Once the calibration is completed, all correction factors are stored in an online database. The standard is returned, all a client needs to do to measure a device with traceability.

A few other calibration examples are given in the following sections.

8.10.1 Flow Calibration

There are many devices available for the measurement of liquid, air, or solid flow. Once the method of measurement is determined by the use of appropriate flowmeter, static or dynamic calibrations can be carried. For example, in the case of static–gravimetric liquid flow, a calibration facility may include a reservoir, pumping system, piping system, flowmeter under test located on the pipeline, collection system, computers and interface, and supporting software. The calibration of flow of fluid through the meter can be determined by collecting prescribed mass of steady fluid flowing over a measured time interval.

The calibrations of flowmeters fall into three distinct categories:

- Factory calibration
- Initial site calibration
- Routine site recalibration

All size flowmeters, large or small, can be calibrated by mobile or static test rigs. Flowmeter calibration is very important in international trade of oil and other bulk fluid movements. Therefore, flowmeters require to be calibrated against accurate reference standards traceable to internationally accepted primary standards.

8.10.2 Sensor Calibration

Sensors may have to be calibrated after having integrated with a signal conditioning system. This process requires an injection of known input signal into the sensor. By observing the output, a correct output scale can be configured for that particular application. If the sensor is used for time-varying inputs, the dynamic calibration becomes necessary. In most cases, transient behavior of sensor top step response may be sufficient to assess the dynamic response of the sensor.

8.10.3 Calibration of Chemical Mixtures

As food contains many chemical substances, the calibration of devices becomes complex. For example, in the case of honey in food processing industry, the following parameters need to be identified by calibrating instruments: fructose, glucose, maltose, moisture level, acidity, etc.

8.10.4 Static Pressure Calibrations

For example, for aircraft pressure transducers, pressure up to 500 psi can be calibrated on computer-controlled automatic test facilities, using gas as the pressure medium. Ranges above 1000 psi can be calibrated manually using hydraulic media. Most calibration standards can have accuracies better than ±0.01%.

In an application, the computer-controlled calibrations are performed at zero and 20% increments of full scale to 100%. Manual calibrations are performed at zero and 25% increments of full scale. For all transducers, the test data are fed into a computer, which calculates the test parameters; plots nonlinearity, thermal zero shift, and thermal sensitivity shift; and prints out the test report. The computer then compares the test data with stored specification limits and accepts or rejects each transducer.

Measurements are first made at zero pressure, then 2 times full scale, then zero, to establish zero shifts after 2 × FSO. Then two complete cycles from zero to full scale and return are performed, measuring output at 20% (or 25%) increments. These data points are used for calculation of nonlinearity, hysteresis, and nonrepeatability.

8.10.5 Dynamic Pressure Calibrations

In some cases, dynamic characteristics can be established from periodic static calibrations. However, in many cases, the accuracy available in dynamic measurements cannot be extrapolated from only static calibrations. Several of the more commonly used methods of dynamic calibration are as follows:

- Comparison pressure calibrations can be performed dynamically with a range of sinusoidal pressure generators.
- Hydraulic pressure generators can be used. They incorporate a compression spring, piston and seismic mass assembly, hydraulic oil-filled chamber, and mounting cavities for the reference standard and test transducers. Sinusoidal vibratory motion applied to the generator housing imparts sinusoidal pressure to oil, which is exposed simultaneously to the sensing surfaces of both transducers providing a direct comparison calibration capability.
- Sinusoidal vibration of a vertically mounted liquid column provides a dynamic pressure that can be applied to a test transducer mounted at the bottom of the tube. By attachment to an electrodynamic vibrator, short liquid columns can be used to provide about 5 psi from about 50 to 2,000 Hz. Low frequency is generally limited by the vibrator and the high frequency is limited by the resonance frequency and damping of the liquid column.
- The inlet modulated pressure generator (IMPG) consists of a wheel with holes drilled through its periphery, which is rotated at high speed. Air is blown through the holes from one side of the wheel and there is a cavity on the opposite side of the wheel in which are located the transducer under test and a reference transducer. The frequency of the signal generated is directly proportional to the speed of rotation. Frequencies of up to 12 kHz can be generated with pressure amplitudes of 1.0 Bar at 1.0 kHz falling to 0.1 Bar at 12.0 kHz. Static mean pressures can be generated.
- The Galton whistle consists of a tube, which is sharp edged at one end and is closed with a moveable piston at the other. Air, which is blown over the edge of the tube, excites the first organ pipe resonance. The resonant frequency is adjusted by the position of the piston within the tube. Mounted in the piston are the reference pressure transducer and the transducer under test.
- Small shock tubes are often used to provide rise time and frequency response characteristics for transducers. Because of difficulties in determining the pressure level in the step, shock tubes are not usually used for pressure sensitivity calibration. Pressure rise times of about 1 μs are practical, which permit transducer characterization to frequencies beyond 100 kHz.

Bibliography

Alessio, C. et al., Quality management issues for PC-controlled calibration systems, *IEEE Instrumentation and Measurement Technology Conference*, *I2MTC*, Singapore, pp. 1610–1615, 2009.

Bucher, J. L., *Quality Calibration Handbook—Developing and Managing a Calibration Program*, ISA Publication, ASQ Quality Press, Milwaukee, WI, 2007.

Busoni, S. et al., Performance evaluation of a full line of medical diagnostic displays and test of a web-based service for remote calibration and quality assurance, *Proceedings of the SPIE—The International Society for Optical Engineering*, 7263, 72630A, 2009.

Commercial Services—National Physical Laboratory, Measurement Services, http://www.npl.co.uk (accessed on February 6, 2012).

Dixson, R. et al., Reference metrology in a research Fab: The NIST clean calibrations thrust, *Proceedings of the SPIE—The International Society for Optical Engineering*, 7272, 727209, 2009.

Eren, H., Calibration in process control, in: *Instruments Engineers' Handbook, The IEH*, Eds., B. Liptak and H. Eren, Vol. 3, 4th edn., Chapter 7, CRC Press, Boca Raton, FL, pp. 141–149, 2011.

Fluke testing and calibration, http://www.fluke.com (accessed on July 8, 2011).

Gupta, S. V., *Measurement Uncertainties: Physical Parameters and Calibration Instruments*, Springer-Verlag, Berlin, Germany, 2012.

Internet Enabled Calibration—Electronics Talk, http://www.electronicstalk.com/news/anr/anr163.html (accessed on December 19, 2011).

ISO/IEC/EN 17025, http://www.fasor.com/iso25 (accessed on February 18, 2012).

Menzies, T. et al., Accurate estimates without calibration? *Making Globally Distributed Software Development a Success Story—International Conference on Software Process, ICSP 2008*, Leipzig, Germany, Vol. 5007, LNCS, pp. 210–221, 2008.

MIL-Standards, Handbooks and specifications, http://store.mil-standards.com (accessed on January 16, 2012).

Technical Articles—Sensors, http://www.sensorsmag.com (accessed on January 24, 2012).

Test management products—Agilent Technologies, http://we.home.agilent.com (accessed on August 20, 2010).

Xuzhang, W. and R. Feng, Research on control methods of calibration program in tele-calibration system, *Proceeding of the International Conference on Networked Computing, INC2010*, Xiang, China, pp. 303–306, 2010.

Zagozdzon, M. and M. Turkowski, Computerization and automation of the water facility for flowmeters calibration, *Przeglad Elektrotechniczny*, 84(5), 195–199, 2008.

Partial List of Calibration Service and Software Providers

American Calibration, Inc., 4410 Rte 176, Suite 14, Crystal Lake, IL 60014, Phone: 815-356-5839, Fax: 815-356-5851, www.americancalibration.com

AMETEK Calibration Instruments, 8600 Somerset Dr., Largo, FL 33773, Phone: 727-536-7831, 800-527-9999 (toll free), Fax: 727-539-6882, www.ametekcalibration.com

Atlantic Instrument & Controls Service, Inc., 168 Old Bulgarmarsh Rd., Tiverton, RI 02878, Phone: 401-625-5778, Fax: 401-625-5730, www.aics-ri.com

Automated Precision, Inc., 15000 John Hopkins Dr., Rockville, MD, Phone: 800-537-2720 (toll free), Fax: 301-990-864820850, www.apisensor.com

Calibrators, Inc., 38 Morning Glory Rd., Cumberland, RI 02864, Phone: 401-769-0333, Fax: 401-769-0335, www.calibratorsinc.com

Cal Lab Co., Inc., 17035 Westview Ave., South Holland, IL 60473-2743, Phone: 708-596-5800, 800-373-1759 (toll free), Fax: 708-596-5802, www.callabco.com

Cal Tec Labs, Inc., 501 Mansfield Ave., Pittsburgh, PA 15205, Phone: 412-919-1377, www.cal-tec.com

CATLab-Accredited Calibration Laboratories, P.O. Box 6598, Williamsburg, VA 23188, Phone: 757-565-4767, Fax: 757-565-4767, catlab.net

CMM Technology, Inc., 1230 Puerta Del Sol, San Clemente, CA 92673, Phone: 949-366-0707, Fax: 949-366-6503, www.cmmtechnology.com

Endress+Hauser, 2350 Endress Place, Greenwood, IN 46143, Phone: 888-363-7377, 888-ENDRESS (toll free), Fax: 317-535-8498, www.us.endress.com

Exact Calibration, Inc., 264 Rancho Santa Fe Rd., Encinitas, CA 92024, Phone: 800-599-1497 (toll free), Fax: 760-753-7959, www.exactcalibration.com

Hayes Instrument Service, Inc., 530 Boston Rd., Billerica, MA 01821, Phone: 978-663-4800, Fax: 978-663-3812, www.hayesinstruments.com

Honeywell Sensing and Control, 1985 Douglas Dr. North, MN10-192B, Golden Valley, MN 55422, Phone: 763-954-4818, 800-446-6555 (toll free), www.sensing.honeywell.com/index.cfm/ci_id/

Indiana Standards Laboratory, 2919 Shelby St., Indianapolis, IN 46203, Phone: 317-787-6578, Fax: 317-787-6580, www.indianastandards.com

Inspec, Inc., 7282 Haggerty Rd., Suite 170, Canton, MI 48187, Phone: 734-451-8740, 800-246-7094 (toll free), Fax: 734-451-8741, www.inspec-inc.com

International Crystal Laboratories, 11 Erie St., Garfield, NJ 07026, Phone: 973-478-8944, Fax: 973-478-4201, www.internationalcrystal.net

Master Metrology, 1041 Cromwell Bridge Rd., Towson, MD 21286, Phone: 410-337-0687, 800-532-8020 (toll free), Fax: 410-337-0787, www.mastermetrology.com

Micro Quality Calibration, Inc., 20743 Marilla St., Chatsworth, CA 91311-4408, Phone: 818-701-4969, Fax: 818-341-9109, www.microqualitycalibration.com

MCMS Metrology, PO Box 232, Lewiston, CA 96052, Phone: 650-521-7678, Fax: 530-623-2289, www.mcmsmetrology.com

Precision Calibration Systems, LLC, 4901 Enka Hwy., Morristown, TN 37813, Phone: 423-278-0946, Fax: 865-281-0604, www.pcsllctn.com

Precision Metrology, 7350 N. Teutonia Ave., Milwaukee, WI 53209, Phone: 414-351-7420, 888-330-3303 (toll free), Fax: 414-351-7429, www.precisionmetrology.com

Restor Metrology, 921 Venture Ave., Leesburg, FL 34748, Phone: 877-220-5554 (toll free), www.restormetrology.com

RS Calibration Services, Inc., 1047 Serpentine Lane, Pleasanton, CA, 94566, Phone: 925-462-4217, www.rscalibration.com

Sears Calibration, 3845 Grader St., Suite A, Garland, TX 75041, Phone: 214-553-6787, 800-736-8830 (toll free), www.searstoolcalibration.com/

Transducer Techniques, Inc., 42480 Rio Nedo, Temecula, CA 92590, Phone: 951-719-3965, 800-344-3965 (toll free), Fax: 951-719-3900, www.transducertechniques.com

Tru Cal International, 401 Country Club Dr., Bensenville, IL 60106, Phone: 630-238-8100, 800-681-5540 (toll free), Fax: 630-238-8101, www.rucal.com/index.lasso

Warren-Knight Instrument Co., 2045 Bennett Dr., Philadelphia, PA 19116-3019, Phone: 215-464-9300, Fax: 215-464-9303, www.warrenknight.com



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

9

Intelligent Sensors and Instruments

9.1 Introduction **9**-1
9.2 Intelligent Sensors **9**-2
9.3 The IEEE 1451 Standard **9**-2
Transducer Electronic Data Sheet • Network Capable Application Processor
9.4 Communications of Intelligent Sensors **9**-4
9.5 Plug-and-Play Capability **9**-4
9.6 Applications and Examples of Intelligent Sensors **9**-5
9.7 Intelligent Instruments **9**-6
9.8 Examples of Intelligent Instruments **9**-7
References **9**-8
Bibliography **9**-9
Partial List of Manufacturers/Suppliers **9**-9

Halit Eren
Curtin University

9.1 Introduction

In measurement and monitoring, there always has been a need to employ devices with the classical requirements, such as the communication capabilities, data processing, fault analysis, configuration guidance, performance tracking and enhancement, calibrations, drift monitoring, convenient access to devices, network management, and security. Most modern intelligent sensors and instruments address these requirements as will be explained in this chapter. Emphasis will be given to intelligent sensors since they represent the state of art of this technology.

Intelligent and smart sensors are devices in a single chip equipped with the necessary digital electronics capable of delivering communications, data processing, and self-health-monitoring features with minimal cost and power overhead on the overall system. Successful integration of these device and their protocols to measurement systems enables a smooth and seamless networking capabilities as well as many additional features. They are capable of interfacing with other devices using agreed protocols such as the IEEE 1451 standards.

Intelligent instruments are computerized devices distinguished from nonintelligent ones by the inclusion of microcontrollers and microprocessor to fulfill the signal processing, communications, data handling, self-calibrations, and decision-making functions. Some examples that intelligent instruments can perform are the correction of the instrument output for biases that are caused by environmental variations (e.g., temperature, humidity) and ability to produce linear outputs and highly nonlinear variables and sensors.

9.2 Intelligent Sensors

A sensor is considered to be intelligent if

- It can provide data, a measure of the quality of the data, and a measure of the health of the sensor possibly based on some intelligent algorithms.
- It embodies specification/identification information as in the embedded form on the physical sensor or virtually on a remote node.
- It can communicate through a network using TCP/IP or similar protocols to support configuration and operation activities.

In order for sensors to integrate into networks successfully, they must embody networking capabilities that provide information flow and control. Currently, there is no defined common digital interface standard on how they should provide information flow. However, there is a strong push in the industry to harmonize the standards that enable networking and data acquisition of sensors.

9.3 The IEEE 1451 Standard

The IEEE-1451 is a set of standards for Smart Sensor Networks to unify the data acquisition control systems across various sensors and transducers [1]. A simplified diagram of this standard is illustrated in Figure 9.1. IEEE 1451 aims for easy connections by using the existing networking technologies, standardized connections, and common software architectures. The standard allows application software, field network, and transducer decisions to be made independently. It also offers flexibility to choose the products and vendors that are most appropriate for a particular application.

At physical connection level, the IEEE-1451 provides standard ways of creating self-describing measurements and control devices. Sensors complying with these standards are expected to have onboard information on the serial numbers, calibration factors, accuracy specifications, and so on. During the installation, location information can also be loaded so that the system can have self-describing properties.

The IEEE P1451 standards come in seven sections:

1. *1451.0-2007* IEEE Standard for a Smart Transducer Interface for Sensors and Actuators—Common Functions, Communication Protocols, and Transducer Electronic Data Sheet (TEDS) Formats
2. *1451.1-1999* IEEE Standard for a Smart Transducer Interface for Sensors and Actuators—Network Capable Application Processor (NCAP) information model
3. *1451.2-1997* IEEE Standard for a Smart Transducer Interface for Sensors and Actuators—Transducer to Microprocessor Communication Protocols and TEDS Formats
4. *1451.3-2003* IEEE Standard for a Smart Transducer Interface for Sensors and Actuators—Digital Communication and TEDS Formats for Distributed Multidrop Systems
5. *1451.4-2004* IEEE Standard for a Smart Transducer Interface for Sensors and Actuators—Mixed-Mode Communication Protocols and TEDS Formats

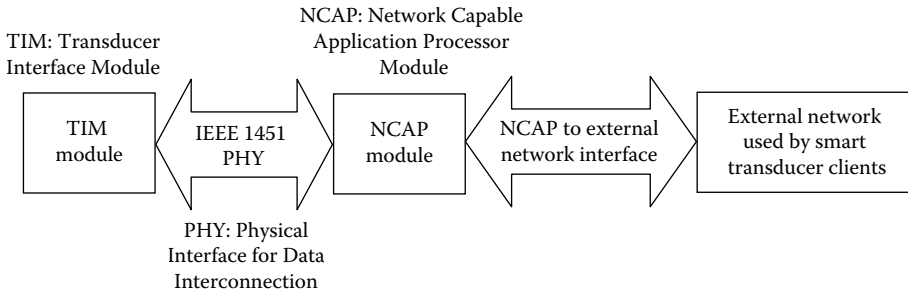


FIGURE 9.1 Simplified functionality of IEEE 1451 standards.

6. *1451.5-2007* IEEE Standard for a Smart Transducer Interface for Sensors and Actuators—Wireless Communication Protocols and TEDS Formats
7. *1451.7-2010* IEEE Standard for a Smart Transducer Interface for Sensors and Actuators—Transducers to Radio Frequency Identification (RFID) Systems Communication Protocols and TEDS Formats

IEEE 1451.1 is termed NCAP information model concerned with the software architecture that moves the intelligence to device level. IEEE 1451.1 uses the benefits of object-oriented technology to create flexible, natural software modules that allow designers to think at the level of real-world systems, not at the level of programming languages. In this way, flexibility can be achieved so that that system can be assembled, reassembled, or modified quickly.

IEEE 1451.2 concerns with the transducers and Microprocessor Communication Protocols and TEDS. It defines the transducer data and electronic interface of digital information direct from the sensors, thus creating a modular architecture to allow embedding of the module to any field networks automatically and transparently.

9.3.1 Transducer Electronic Data Sheet

The heart of IEEE 1451 standard is the definition of the TEDS, the information structure that contains the critical sensor information to enable plug-and-play (PnP) operation. The TEDS, which typically resides in an EEPROM embedded in the sensor, is accessed by the measurement system via a serial interface.

IEEE 1451.4 defines the TEDS structure to be compact yet flexible and extensible to be able to handle a wide range of sensor application requirements. The TEDS has two types of information: (1) the basic information that contains the sensor identification, details of the manufacturer, model number, and the serial number and (2) the standard information containing the specific "data sheet" (e.g., data needed to properly configure the electrical interface and other parameters) and conversion of the measurement data into engineering units. Typical parameters include measurement range, electrical output range, sensitivity, power requirements, and calibration data. IEEE provides standard templates as shown in Table 9.1.

TABLE 9.1 IEEE Standard Templates

Types	Template ID	Name of Template
Transducer templates	25	Accelerometer and force
	26	Charge amplifiers—accelerometer
	27	Microphone preamplifier—built-in
	28	Microphone preamplifier—attached
	29	Capacitive microphones
	30	High-level voltage output sensor
	31	Current loop output sensor
	32	Resistance sensor
	33	Bridge sensor
	34	AC linear/rotary differential transformer
	35	Strain gage
	36	Thermocouple
	37	Resistance temperature detector
	38	Thermistor
	39	Potentiometric voltage divider
	43	Charge amplifier—attached to force transducer
Calibration templates	40	Calibration table
	41	Polynomial calibration curve
	42	Frequency response table

Before installing intelligent sensors supported by Virtual TEDS technology, as an alternative one can configure traditional analog sensors by using binary files. These files can be used to store the information on identification, configuration, and calibration for an individual sensor.

9.3.2 Network Capable Application Processor

Defined in IEEE 1451, each node primarily consists of three fundamental components, which are the NCAP, the smart transducer interface module (STIM), and the physical transducer itself. The NCAP is responsible for providing intelligence and network communication with other nodes in the network. The STIM is responsible for controlling the transducers and their communication with the NCAP. The transducers are connected to the STIM. The NCAP and the STIM together form a networked smart transducer. The sensor nodes thus have the ability to take crucial decisions on their own and communicate more efficiently with other nodes in the network.

The NCAP hosts the processor that makes intelligent control decisions and is responsible for communication with the other NCAPs in the network. It consists of various blocks, namely, functional blocks, transducer blocks, and ports (e.g., application processor client port, publisher port, subscriber port). Each block is in charge of a unique group of functions and manages associated ports. NCAP block is mandatory but other blocks are optional. All of the objects residing in an NCAP process are state based. The blocks in the node would have different states and substates during the life cycle of the node, and only particular set of operations can be performed by a block in a particular state. IEEE 1451.1 defines classes for such blocks and ports. It also lists the operations performed in each class.

The STIM is responsible for housing the transducers and providing the information from the transducers to the NCAP. An STIM acts as an interface between an NCAP and a transducer. It performs the A/D conversion and houses the control unit as well. It usually also contains the TEDS, which provides calibration and other important information about the transducers and the STIM.

9.4 Communications of Intelligent Sensors

IEEE 1451.1 provides the object model of the networked transducer nodes and how they communicate among themselves. IEEE 1451 supports two paradigms of communication among the nodes, which are

1. Multicast publish–subscribe model
2. Point-to-point client–server model

In publish–subscribe communication model, a node publishes a message on a multicast address. This publication message is received by the nodes with subscription to the publication domain specified in the message. By definition for implementation, publish–subscribe messages use User Datagram Protocol (UDP) sockets in order to send and receive messages. This model is primarily used for node discovery and other operations.

Once the nodes are discovered and two nodes have adequate information (such as object tag and object dispatch address) about each other, the two nodes can start client–server communication between each other. Major operations performed by the NCAPs are client–server operations. The communication infrastructure uses TCP/IP for client–server communication.

9.5 Plug-and-Play Capability

A node is PnP if it becomes operational and networked immediately after it is turned on while physically connected to a network. PnP functionality includes self-announcement, self-configuration to a default setting, and acknowledgement to clients/servers of its availability to report sensor readings or to perform actuator functions. PnP capability is not complete unless an application layer communication

language is standardized, which is independent of the network and data link layer communication protocols (e.g., TCP/IP with ZigBee, TCP/IP with 802.11). PnP capability of intelligent sensors has found numerous applications [2,3].

A network becomes dynamic when all the elements in the network have PnP capability and a node is PnP node when it becomes operational and networked immediately after it is turned on and physically connected to a network. Thus, PnP capability plays a very important role in managing a dynamic network. For capability, a node has to

- Announce its presence to other nodes in the network
- Configure itself to the default settings during a start-up
- Acknowledge clients/servers in the network of its availability to report sensor readings or to perform actuator functions

The IEEE 1451.4 standard reduces the time and challenges associated with sensor configuration. The standard establishes a universally accepted method of giving sensors PnP capability, similar to the PnP capability of a USB mouse and personal computers. IEEE 1451.4 defines a mechanism for adding self-describing behavior to sensors with an analog signal interface.

In sensor networks, achievement of PnP provides the following advantages:

- Reduced configuration time by eliminating manual data entry
- Better sensor tracking by storing data sheets electronically
- Improved accuracy by providing detailed calibration information
- Simplified asset management by eliminating paper data sheets
- Reliable sensor location by identifying individual sensors electronically

9.6 Applications and Examples of Intelligent Sensors

Intelligent sensors find many industrial, consumers, and scientific applications, some of which are

- Manufacturing and assembly lines integrated with the machines
- Medicine for dedicated signal processing of the human sensory systems
- Transportation and automotive industries
- Consumer goods

Some examples are as follows.

In consumer industry, SONY has introduced an intelligent presence sensor in their BRAVIA brand televisions. This sensor is based on built-in camera, which detects the body movements and recognizes the facial patterns of up to five viewers (http://esupport.sony.com/docs/imanual/NA/EN/1a/ipsensor_uc_1a.html).

PCB PIEZOTRONICS produces a range of smart sensors. Some examples are 352A56 Series Accelerometers, 481 Series Multichannel Signal Conditioners, TEDS Sensor Interface Kits, and T422 Series In-Line Charge Converters. For instance, Model 352A56 accelerometer conforms to the IEEE 1451.4 standard to define a mixed-mode interface that retains the analog signal but adds a digital link to access TEDS embedded in the sensor for self-identification and self-description. It eliminates the need to manually input sensor parameters such as manufacturer, model number, serial number, calibration fate, and sensitivity when configuring the system (http://www.sensorsportal.com/HTML/SENSORS/TEDS_SensManuf.htm).

The DS2430A 256-bit 1-wire EEPROM produced by Maxim identifies and stores information about the product it is associated. It is capable of storing calibration constants, board identification, and product revision status. This chip finds TEDS compliant applications in smart devices such as piezoelectric force sensors, pressure sensors, and accelerometers. For instance, Model 2801, Smart Remote Charge Converter produced by ENDEVCO, is a TEDS-based low-noise, two-wire single-ended device to be used with nonsmart sensors such as some accelerometers to convert high impedance charge output to low impedance voltages.

9.7 Intelligent Instruments

Intelligent instruments have functionality of learning and adaptation capabilities. They have embedded processing to provide the computational features to perform complex sensing and actuating tasks. A typical structure of intelligent instrument is shown in Figure 9.2.

The functions of an intelligent instrument can be summarized as compensation, information processing, communications, integration, self-calibration, adaptation, and evaluation of the validity of collected data. Information processing encompasses the data-related processing that aims to enhance and interpret the collected data and maximize the efficiency of the system, through signal conditioning, data reduction, event detection, and decision making.

Communications component of intelligent instrument systems incorporates the standardized network protocol, which serves to link the distributed instruments in a coherent manner, enabling efficient communications and fault tolerance. Traditional task-specific instrument systems often contain a number of limitations in terms of complexity, cost, and flexibility. Intelligent instruments aim to overcome these limitations through the utilization of standardized transducer interfaces and communication protocols, resulting in autonomous, distributed, and reconfigurable instrument.

As far as the user is concerned, an intelligent instrument behaves like a black box, and no knowledge of its internal mode of operation is required for normal measurements. These types of instruments offer many advantages over their nonintelligent counterparts, principally because of the improvement in accuracy achieved by processing the output of transducers to correct for errors inherent in the measurement process.

Intelligent instruments usually provide many other functions, such as

1. Signal damping with selectable time constants
2. Switchable ranges (using several primary transducers within the instrument, which each measure over a different range)
3. Switchable output units (e.g., display in imperial or SI units)
4. Diagnostic facilities
5. Remote adjustment and control of instrument options
6. Self-calibrations
7. Various output options, 4–20 mA, Ethernet, Internet, wireless, etc.
8. Self-learning ability

In this chapter, we will not go into the details and intricacies of internal structures and principles of operations of intelligent instruments. Instead, some typical examples will be provided. The underlying technologies are also discussed at various other chapters of this book.

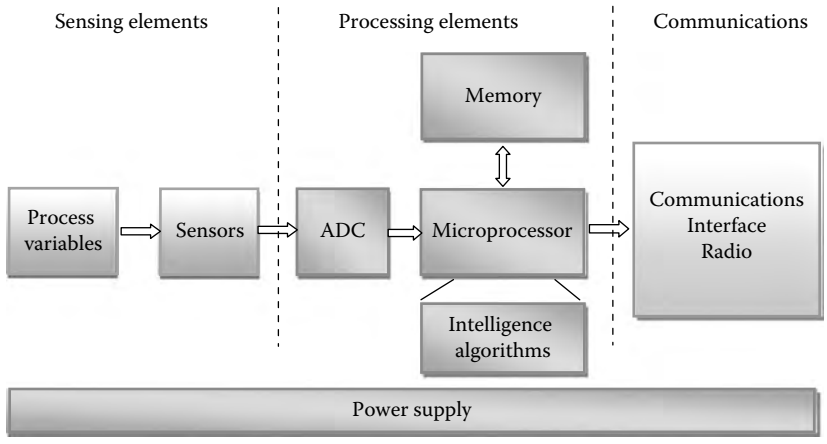


FIGURE 9.2 Architecture of an intelligent instrument.

9.8 Examples of Intelligent Instruments

There are many examples of intelligent instruments. Four typical examples are provided here.

Example 9.1

Intelligent microsurgical instruments [4,5]: They are designed as handheld instruments to be used particularly in microsurgery. Positioning error is inherent in a normal human such as tremor and jerk, which limit the accuracy of the hand motion of the surgeon. Intelligent behavior instruments sense their own motion and distinguish between desired and undesired motions using advanced filtering techniques. They can actively compensate for undesired motions by equal but opposite deflections of their own tips.

Example 9.2

Intelligent temperature transducers [6–8]: A typical intelligent temperature transducer is capable of measuring and controlling temperature. They can use fieldbus protocols to read measured values, set alarm limits, adjust the probes, and read firmware version. For this, appropriate ports are used to set alarm limits, time delays, e-mail addresses, probe description, refresh of www pages, select type of www pages, and set storing interval to history and automatic IP address assignment from a server. In case of alarm creation, warning e-mail message is sent to addresses defined by the user.

Another example of an intelligent temperature transducer is the Ethernet interface-capable temperature sensor. A typical device can sense display and communicate temperature, barometric pressure, relative humidity, dew point temperature, absolute humidity, specific humidity, mixing ratio, and specific enthalpy. Its communication modes are as follows:

- Modbus protocol enables to read measured values, set alarm limits, adjust the probe, and read firmware version.
- www pages: User-selectable design of www pages enabling to display curves of measurement history. User can design the look of www pages and select values to display.
- E-mail: In case of alarm creation, warning e-mail message is sent to addresses defined by the user.

Example 9.3

Intelligent flowmeters [9–13]: There are many intelligent flowmeters commercially available. They are based on different principles for different applications. Some of the intelligent flowmeters are coriolis mass flowmeters, electromagnetic flowmeters, ultrasonic flowmeters, variable area, vortex flowmeters, water flowmeters, oil flowmeters, metal rotameters, steam flowmeters, gas flowmeters, air flowmeters, mass flowmeters, and so on.

As an example, a typical intelligent vortex flowmeter contains four-piece piezoelectric sensors, two of which are used to detect the flow of fluid and the other two for detecting vibration signal of pipe. It can provide temperature and pressure compensation. The meter parameters and meter range can be changed by manual buttons. It provides 4–20 mA current signal output/pulse signal and communicates with Highway Addressable Remote Transducer communication protocol, PROFIBUS and Foundation Fieldbus. Liquid crystal displays can exhibit the instantaneous flow and accumulated flow of measured medium simultaneously.

Another example is the intelligent electromagnetic flowmeter used for high-viscosity uniform medium such as detecting paper pulp, mud, cement slurry, and mortar. It has compensation capability for fluid density, viscosity, temperature, and pressure change. It has double flow direction measurement and double direction quantum accumulating function.

Example 9.4

Intelligent pressure transmitters [14,15]: A typical intelligent pressure transmitter has onboard diagnostics, time-tracking functions, install date, calibration date, time in service, stress monitoring, point value, body temperature meter, as well as general reading functions. All diagnostics automatically run when the transmitter is powered and the information can be fed directly into the plant's digital control system (DCS) and asset management. This information remains in the transmitter for use in troubleshooting, in which it can measure absolute pressure, differential pressure, gage pressure, flange mount, and temperature.

Many other intelligent instruments have similar structure, software support, and functionalities. In some cases, intelligent instruments use artificial intelligence to learn from the past operations and external data.

References

1. IEEE Standard for a Smart Transducer Interface for Sensors and Actuators—Network Capable Application Processor (NCAP) Information Model, IEEE 1451.1 Standard, The Institute of Electrical and Electronics Engineers, Inc., New York.
2. Wall, R.W. and Huska, A., Design platform for plug-and-play IEEE 1451 traffic signal, *31st Annual Conference of the IEEE Industrial Electronics Society*, Montreal, Quebec, Canada, pp. 6–10, 2005.
3. Patra, J.C., Kot, A.C., and Panda, G., An intelligent pressure sensor using neural networks, *IEEE Transactions on Instrumentation and Measurement*, 49(4), 829–835, 2000.
4. Ang, W.T., Khosla, P.K., and Riviere, C.N., An intelligent hand-held microsurgical instrument for improved accuracy, *Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Istanbul, Turkey, Vol. 4, pp. 3450–3453, 2001.
5. Hunter, I.W., Lafontaine, S.R., Brenan, C.J.H., and Jones, L.A., Medical robots and micro machines, *Proceedings of the 6th International Symposium on Micro Machine and Human Science* (*MHS'95*), Nagoya, Japan, pp. 25–30, 1995.
6. Jingwei, Y. et al., The portable temperature and humidity monitor based on intelligent sensor, *1st International Conference on Information Science and Engineering* (*ICISE 2009*), Nanjing, China, pp. 5245–5247, 2009.
7. Xu, J., You, B., and Li, Q., Implementation of an IEEE 1451 smart quartz tuning fork temperature transducer for real-time distributed measurement and control system, *6th World Congress on Intelligent Control and Automation*, Dalian, China, pp. 6, 2006.
8. Zhaochun, L., Yuzhu, C., and Jin, H., Design of smart temperature sensor based on IEEE1451.2 standard, *International Forum on Information Technology and Applications (IFITA)*, Chengdu, China, pp. 312–314, 2009.
9. Yi, Y. and Huifeng, W., An improved intelligent calibration method for vortex flowmeter, *American Control Conference*, New York, pp. 2927–2931, 2007.
10. Yong-hui, H., Intelligent turbine flowmeter based on CAN bus, *Instrument Techniques and Sensor,* 12, 69–70, 2007.
11. Liao, J. and Liu, Y., Distributed flowmeter data acquisition system based on WirelessHART networks, *International Conference on Apperceiving Computing and Intelligence Analysis (ICACIA 2009)*, Chengdu, China, pp. 383–386, 2009.
12. Yuhui, W., A new-type high-speed data sampling circuit based on FPGA and its application in flowmeter, *2nd International Asia Conference on Informatics in Control, Automation and Robotics,* Wuhan, China, Vol. 2, pp. 454–457, 2010.
13. Marshall, R., Facts at your fingertips: Flowmeter selection, *Chemical Engineering,* 113(12), 23, 2006.

14. Hao, J., Li, Q., and Yang, L., Intelligent pressure transmitter based on HART protocol, *Instrument Techniques and Sensor,* 2, 16–22, 2007.
15. Li, Y.-L., Design of tyre pressure intelligence monitoring system based on wireless transmission technology, *Instrument Techniques and Sensor,* 11, 40–47, 2007.

Bibliography

Bhuyan, M., *Intelligent Instrumentation: Principles and Applications*, CRC Press, Boca Raton, FL, 2011.

Bruey, D., *SNMP: Simple Network Management Protocol*, Rane Corporation, RaneNote 161, written December 2005, http://www.rane.com/note161.html (accessed on January 28, 2012).

Casucci, A.T., *Intelligent Components and Instruments for Control Applications*, 2000, IFAC, Pergamon Press, Great Britain, U.K., 2001.

Case, J., Fedor, M., Schoffstall, M., and Davin, J., *RFC1157—Simple Network Management Protocol (SNMP)*, May 1990, http://www.faqs.org/rfcs/rfc1157.html (accessed on March 13, 2013).

Huddleston, C., *Intelligent Sensor Design Using the Microchip dsPIC*, Newness, Oxford, U.K., 2007.

Manufacturers of Plug-and-Play (TEDS) Smart Sensors (IEEE 1451), http://www.sensorsportal.com/HTML/SENSORS/TEDS_SensManuf.htm (accessed on March 24, 2012).

Meijer, G. C., *Smart Sensors Handbook,* Wiley, Chichester, U.K., 2008.

Randy, F., *Understanding Smart Sensors*, 2nd edn., Artech House, Boston, MA, 2000.

Sony Bravia i-Manual online, http://esupport.sony.com/docs/imanual/NA/EN/1a/ipsensor_uc_1a.html (accessed on March 20, 2012).

Partial List of Manufacturers/Suppliers

Cetacea Sound, Inc., 2950 Airways Blvd., Memphis, TN 38116, Phone: 901-363-3856, 800-489-1759 (toll free), Fax: 901-367-1420, www.cetacea.com

Endevco Corp., 30700 Rancho Viejo Rd., San Juan Capistrano, CA 92675-1748, Phone: 800-982-6732, Fax: 949-661-7231, www.endevco.com

Enga Tech, 607 Fairway Woods Dr., Chester, VA 23836, Phone: 888-492-5886 (toll free), Fax: 804-530-9223, www.engatechvision.com

Esensors, Inc., 4240 Ridge Lea Rd Ste 37, Amherst, NY 14226, Phone: 716-837-8719, Fax: 716-837-8740, www.eesensors.com

Fife Corp., P.O. Box 26508, Oklahoma City, OK 73126, Phone: 405-755-1600, 800-639-3433 (toll free), Fax: 405-755-8425, fife.maxcessintl.com

Honeywell Sensing and Control, 1985 Douglas Dr. North, MN10-192B, Golden Valley, MN 55422, Phone: 763-954-4818, 800-446-6555 (toll free), sensing.honeywell.com/index.cfm/ci_id/15…

Kistler North America, 75 John Glenn Dr., Amherst, NY 14228-2119, Phone: 716-691-5100, 888-KISTLER, 888-547-8537 (toll free), Fax: 716-691-5226, www.kistler.com

MAGPOWR, 222 W. Memorial Rd., Oklahoma City, OK 73114, Phone: 405-755-1600, 800-MAGPOWR (toll free), Fax: 405-755-8425, magpowr.maxcessintl.com

Microstrain, Inc., 459 Hurricane Lane, Suite 102, Williston, VT 05495, Phone: 802-862-6629, 800-449-3878 (toll free), Fax: 802-863-4093, www.microstrain.com

Monitor Technologies, LLC, 44W320 Keslinger Rd., Elburn, IL 60119, Phone: 630-365-9403, 800-601-5924 (toll free), Fax: 630-365-5646, www.monitortech.com

National Instruments Corp. 11500 N. Mopac Expwy., Austin, TX 78759-3504, Phone: 512-794-0100, 888-280-7645 (toll free), Fax: 512-683-8411, www.ni.com

PCB Piezotronics, Inc., 3425 Walden Ave., Depew, NY 14043-2417, Phone: 716-684-0001, 800-828-8840 (toll free), Fax: 716-684-0987, www.pcb.com

PICS, Inc., 81 Lancaster Ave., Suite 202, Malvern, PA 19355, Phone: 610-644-7701, Fax: 610-644-9297, www.picscontrols.com

Sensor Platforms, Inc., 1550 Airport Blvd Ste 220, Santa Rosa, CA 95403, Fax: 707-581-7361, www.sensorplatforms.com

Sensors Unlimited, Inc., 3490 Route 1 Bldg 12, Princeton, NJ 08540, Phone: 609-520-0610, Fax: 609-520-1663, www.sensorsinc.com

Sony Manufacturing Systems America, Inc., 20381 Hermana Cir., Lake Forest, CA 92630, Phone: 949-770-8400, Fax: 949-770-8408, www.sonypt.com

Steven Engineering, 230 Ryan Way, South San Francisco, CA 94080, Phone: 888-790-0022 (toll free), Fax: 800-314-0716, 800-314-0716 (toll free), stevenengineering.rtrk.com/?scid=1975164…

Technica-Tool, Inc., 553 Industrial Dr., Hartland, WI 53029, Phone: 262-367-8665, Fax: 262-367-0208, www.techna-tool.com

Turck, Inc., 3000 Campus Dr., Plymouth, MN 55441-2619, Phone: 763-553-7300, 800-554-4989 (toll free), Fax: 763-553-0708, www.turck.com

Unisys Corp., 20 Research Dr., Hampton, VA 23666, Phone: 757-865-0637, www.unisys.com/unisys/

Valin Corp., 555 E. California Ave., Sunnyvale, CA 94086, Phone: 408-370-7444, 800-922-6646 (toll free), Fax: 408-370-1356, www.automationcontrols.com

10

Virtual Instruments

10.1 Introduction **10**-1
10.2 Virtual Instrument Architecture **10**-1
10.3 Virtual Instrument Software **10**-2
Data Acquisition Software • Device Driver Software • Application Software
10.4 Supervisory Control and Data Acquisition **10**-8
Components of SCADA • SCADA Software
10.5 Conclusions **10**-12
Bibliography **10**-12
Partial List of Suppliers **10**-13

David Potter
National Instruments

Halit Eren
Curtin University

10.1 Introduction

The rapid adoption and technological evolution of the (personal computer) PC in the last 25 years or so catalyzed a revolution in instrumentation for test, measurement, and data acquisition (DAQ). There are at least two substantial developments in this area, being the concept of virtual instrumentation (VI) and the supervisory control and data acquisition (SCADA). In this chapter, primarily virtual instruments will be discussed and SCADA systems will be introduced briefly.

The virtual instruments consist of industry-standard computers or workstations equipped with application software, modular DAQ hardware such as plug-in boards, and driver software. Together, these components replicate the functions of traditional stand-alone instruments. Virtual instruments represent a fundamental shift from traditional hardware-centric instrumentation systems to software-centric systems that exploit the computing power, productivity, display, and connectivity capabilities of popular desktop computers and workstations.

10.2 Virtual Instrument Architecture

VI can be traced back to the birth of the International Business Machines Corporation (IBM) PC (in 1981) and the emergence of general purpose interface bus (GPIB) as a tool that enabled engineers to use a PC to control and communicate with instruments to automate test and measurement tasks. In 1986, National Instruments introduced the LabVIEW graphical development environment, which offered to acquire data, control instruments, and analyze and present data from a PC without sacrificing performance or functionality. Since then, the development and evolution of all aspects of VI—PC processors, memory, bus and networking technologies, software technologies, and DAQ hardware—had a phenomenal growth. As a result, VIs are widely used in everything from simple, low-cost applications, such as environmental temperature monitoring, to large-scale, real-time measurement and control applications, such as the real-time control of over a hundred collimators to help control the particle beam in the world's largest particle accelerator, the Large Hadron Collider (LHC) at CERN.

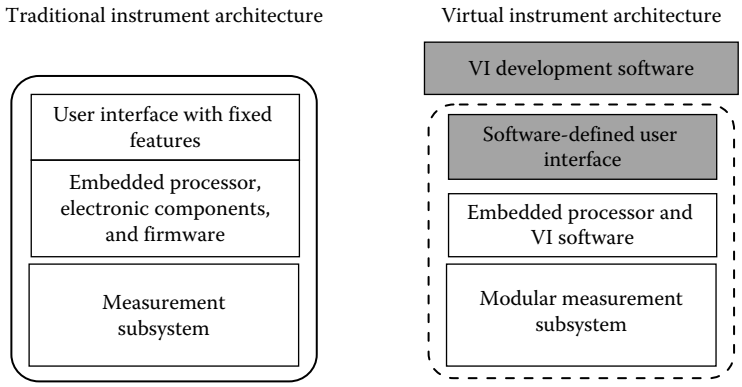


FIGURE 10.1 Traditional versus virtual instruments.

Traditionally, the stand-alone traditional instruments such as oscilloscopes, data recorders, waveform generators, and electronic measuring instruments are designed to perform one or more specific tasks defined by the vendors. While stand-alone traditional instruments can be very powerful, they also can be expensive and generally the users cannot extend or customize the instruments for their needs. The knobs and buttons on the instrument, the built-in circuitry, and the functions available to the user are all configured specifically for the instrument. In addition, special technology and components must be developed to build these instruments, making them expensive and slow to adapt to changing market conditions and the user needs. A comparison of the traditional and VI architecture is given in Figure 10.1.

10.3 Virtual Instrument Software

Virtual instruments are supported by a number of software used for different purposes such as DAQ, device driver, and application software. Some of this supporting software will be explained in Section 10.2.1.

10.3.1 Data Acquisition Software

In today's instruments and instrumentation systems, the software plays an important role as it determines the ultimate success or failure of the system. Nowadays, almost all aspects of modern instrumentation and DAQ system are software driven. This includes configuring and managing the hardware, controlling the flow and storage of the acquired data, processing and analyzing of the data, as well as visually displaying and communicating the data.

Essentially, there are two layers of the DAQ software as shown in Figure 10.2. Firstly, most vendors of DAQ hardware supply some type of device driver software for controlling and accessing the device. The device driver manages the low-level access and manipulation of the hardware, as well as management of the computer and operating system resources, hence simplifying the programming of the DAQ system.

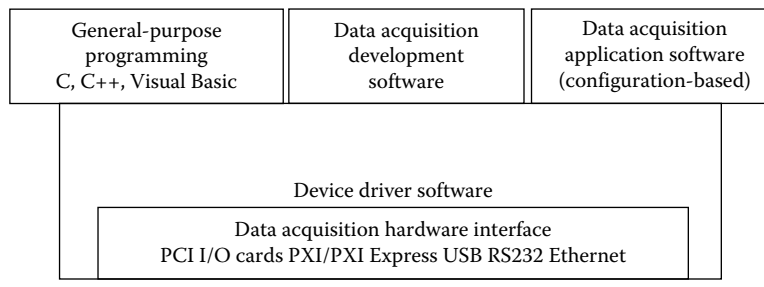


FIGURE 10.2 Levels of DAQ software application.

Secondly, the driver software is used in conjunction with a higher-level program, which may be developed using a general-purpose programming language or specialized DAQ software that includes tools for acquiring, analyzing, and managing the acquired data. DAQ software can further be divided into development software environments, which allow users to fully customize their systems and develop application-specific systems and configuration-based software. These are simpler, easier-to-use software packages designed to perform a fixed set of DAQ and analysis functions.

10.3.2 Device Driver Software

A DAQ device driver is software that allows higher-level computer programs to interact with the DAQ hardware. The driver communicates with the DAQ hardware through a bus or network, essentially coordinating the passing of commands and data to and from the hardware. The higher-level software invokes routines, or functions, in the driver to initiate interaction with the hardware. The collection of these functions and their formats is known as the application programming interface (API). In Microsoft Windows, drivers are usually distributed as dynamic link library (DLL) files that can be accessed by most Windows programming languages, such as Visual Basic, C++, LabVIEW, and C#.

Although driver software works "behind the scenes" and is not as visible as application software, DAQ device drivers can be very complex and critical to the effectiveness of the DAQ systems. Following is a brief description of the main functions and capabilities provided by driver software.

Configuration: Most DAQ and signal conditioning devices support multiple modes of operations and variable settings. Parameters such as A/D sampling rate, amplifier gain, filter settings, trigger mode, and sensor excitation are typically configurable via the driver software. In addition to these hardware settings, configuration of key acquisition and data scaling parameters must be set.

Driver software often includes a software utility, which you can use to easily and interactively set theses parameters, such as the Measurement and Automation Explorer utility shown in Figure 10.3.

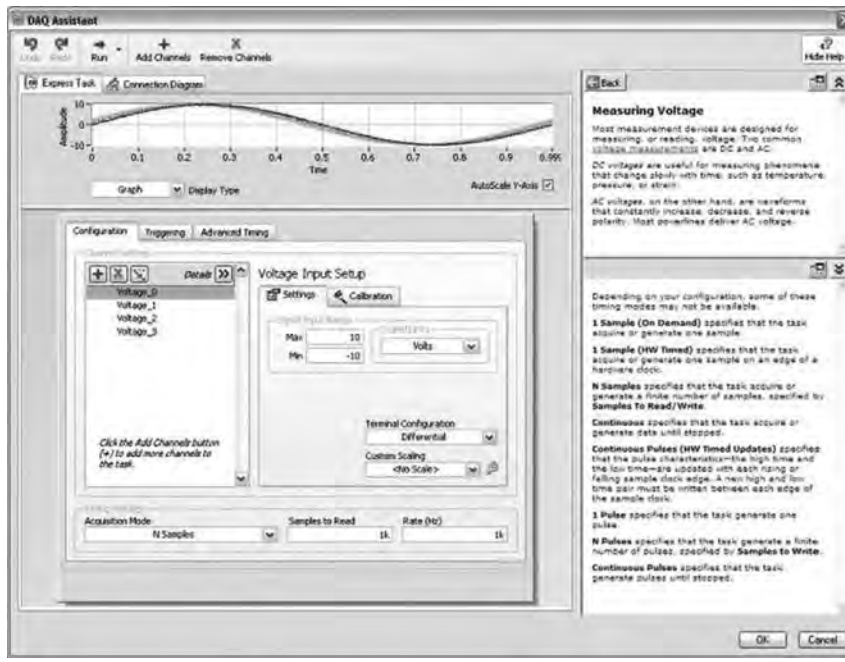


FIGURE 10.3 Example of a software configuration tool for setting up a DAQ system. (Courtesy of National Instruments, http://zone.ni.com/cms/images/devzone/tut/DAQ_Asst_20090321150309.png)

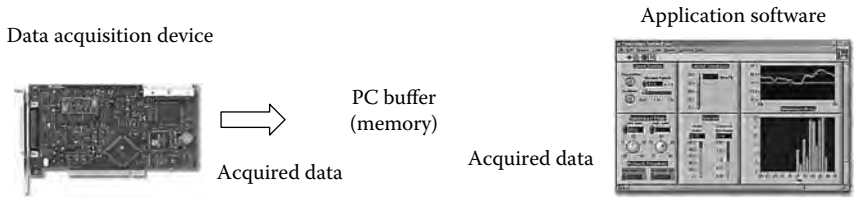


FIGURE 10.4 DAQ software must manage reliable buffering of acquired data.

The settings for the acquisition voltage measurements and the signal conditioning, and a data display or test panel to interactively check the settings. Additionally, many applications require the ability to programmatically set these parameters. For these applications, it is critical that the driver library exposes these configuration settings through its API.

Calibration: Calibration of a DAQ device consists of verifying its measurement accuracy by measuring a known value and adjusting for any measurement error. Modern DAQ devices often include the capability of self-calibration using an onboard reference voltage, which should be very stable and have high precision. A more complete calibration, often referred to as end-to-end calibration, encompasses the entire measurement path, from the sensor all the way through to the ADC. For example, an end-to-end calibration of a thermocouple would involve heating or cooling the thermocouple probe to a controlled, known temperature. By recording the difference in the measurement acquired by the DAQ device and the known temperature at multiple temperatures, this difference can be used in subsequent measurements to compensate for much of the gain and offset errors of the DAQ system.

Data buffering and management: Acquisition of dynamic waveforms can require streaming large amounts of data from the DAQ device to memory on the controller or PC. This PC memory acts as a data buffer to temporarily hold the data until the application software can access and process the data, store it to disk, etc., as shown in Figure 10.4. There are two basic modes of operation that DAQ driver software should handle effectively.

Finite buffered acquisition: The DAQ device is programmed to perform a finite number of A/D conversions and transfer the data into the PC buffer. When finished, the DAQ device is idle until programmed for another task.

Continuous buffered acquisition: With continuous acquisition, the DAQ device is programmed to perform acquisitions indefinitely as shown in flow diagram form in Figure 10.5. Data are streamed continuously to the PC memory buffer. However, because the PC buffer is of a finite size, data must be simultaneously pulled out and transferred to the application software for processing and more permanent storage. This requires that the driver manage the PC buffer as a circular buffer, meaning that when incoming data reach the end of the buffer, storage resumes at the beginning of the buffer. This process can continue indefinitely without losing acquired data as long as data are being simultaneously copied to the application software memory at a rate as fast as the data coming in from the DAQ device.

Advanced driver capabilities: In addition to the basic operations of sending configuration commands to the hardware and managing the transfer and buffering of acquired data, more advanced DAQ drivers also include advanced capabilities that can save significant development time and yield higher-performance systems.

Multithreading: DAQ multithreading is the concept of independently controlling more than one I/O function on the acquisition device. In many DAQ applications, analog output or digital I/O is needed for stimulus or control in addition to analog input measurements. For these applications, it is critical that the drivers support concurrent I/O operations and not limit the programmer to sequential tasks.

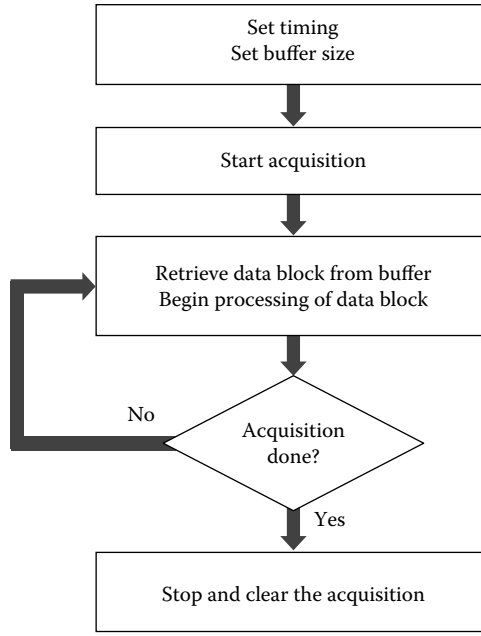


FIGURE 10.5 Flowchart for program for a continuous buffered acquisition.

Intelligent timing, triggering, and synchronization: Many DAQ devices include a number of timing and triggering mechanisms that can be configured in multiple ways. Setting up a synchronized I/O application on multiple devices can become very complex, involving complicated counter timer configurations and low-level signal routing of clock and trigger signals. More advanced drivers automatically handle this lower-level configuration and signal routing, thus simplifying the application program.

Hardware simulation: During development and system integration, the developer of the DAQ system may not always have access to operational DAQ hardware. Therefore, it can be very useful if the driver allows the use of a simulated device, allowing the developer to test and try out his code without having fully operational hardware setup.

Automatic data logging to disk: While most drivers manage the transfer of data to the PC or controller memory, and it is left to the application program to process and store the data to a nonvolatile location, such as the hard disk, more advanced driver software includes built-in functionality to simultaneously transfer data from the PC memory to a file on the hard disk.

10.3.3 Application Software

A real-world DAQ system typically includes much more than the acquisition of data. Systems very often include signal generation, digital I/O, and timing I/O, as well as processing and analysis of data, data visualization, user interfaces, data storage, and communications with distributed systems. Fortunately, developers of integrated DAQ systems do not have to be professional programmers who must code these systems from scratch using general-purpose programming tools such as C++ or Visual Basic. Specialized DAQ software development tools are available to engineers to greatly simplify the integration of DAQ systems.

While there is a wide range of application software packages, it is useful to categorize the options into two general types: (1) application development environments (ADEs) and (2) configuration-based programs.

10.3.3.1 Application Development Environments

ADEs offer the high level of customization and flexibility in developing DAQ systems. DAQ environments are specialized software development environments that integrate functions and tools for DAQ, analysis, visualization, and other functions common in applications. Users develop a DAQ program or script and are able to highly customize the program. Many of today's ADE software packages utilize some form of graphical development, which is generally considered more intuitive and easier to learn for nonprofessional programmers.

For example, the LabVIEW software is a popular development tool for DAQ system, which uses graphical programming language. Figure 10.6 shows a LabVIEW front panel. Execution of the program is determined by the structure of the graphical block diagram in which the programmer connects different icons, or function nodes, by drawing wires. The front panel serves as the user interface, allowing the user/operator to input data, interact with controls, or visualize outputs of the VI. The VIs can be embedded into higher-level programs or other VIs. In these cases, the front panel serves as the programmatic interface for I/O to the embedded VI.

The graphical programming language used in LabVIEW, also referred to as "G," is a dataflow programming language. These wires in the block diagram propagate variables, and any node can execute as soon as all its input data become available. Since this might be the case for multiple nodes simultaneously, G is inherently capable of parallel execution. Figure 10.7 shows a simple program with two parallel programming structures. Because no wires interconnect the two structures, the two loops will run in parallel and the program will automatically execute the two loops on multiple processors.

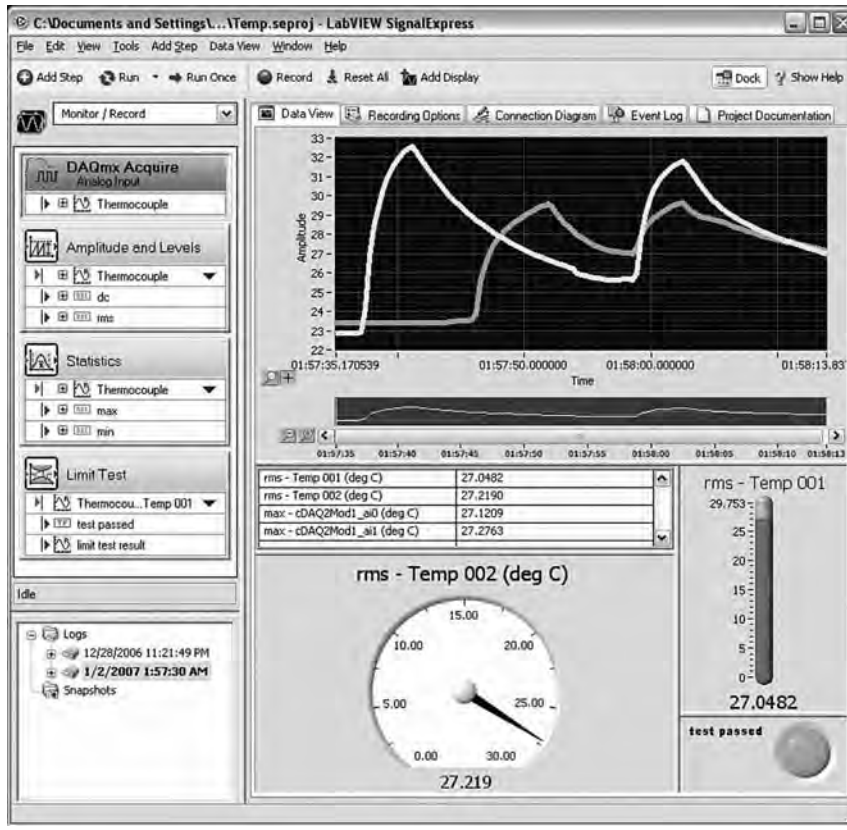


FIGURE 10.6 A typical VI front panel. (Courtesy of National Instruments, http://zone.ni.com/cms/images/devzone/tut/Figure%203.gif)

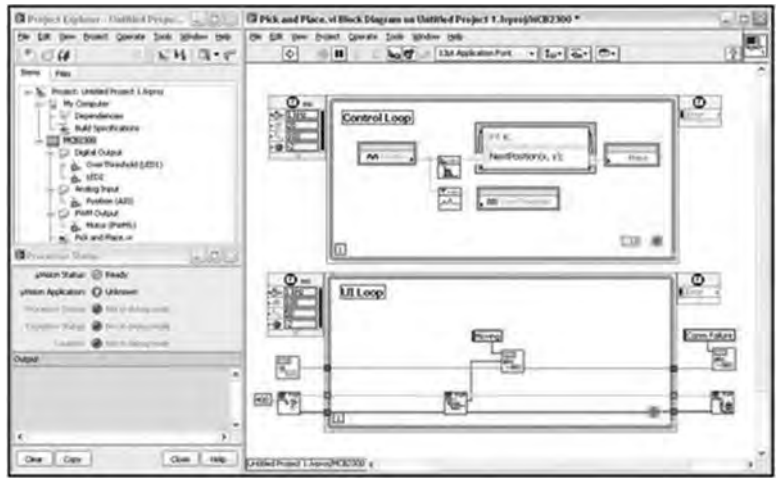


FIGURE 10.7 Graphical dataflow structure of LabVIEW programming. (Courtesy of National Instruments, http://zone.ni.com/cms/images/devzone/pub/arm1.jpg)

10.3.3.2 Configuration-Based Application Software

Graphical development applications can be made flexible. For simpler or smaller scope implementations, a number of application software packages are available for common DAQ tasks, such as acquiring and logging data to disk. These packages require no additional programming and are configured by the user, typically through intuitive menus and software wizards.

Figure 10.8 shows an example of a configure-based software package, LabVIEW SignalExpress. Acquisitions and some limited analysis routines are configured through the menu structure on the left side of the window. The acquired data can be displayed in configurable graphs and charts in real time, as well as stored to disk for later retrieval.

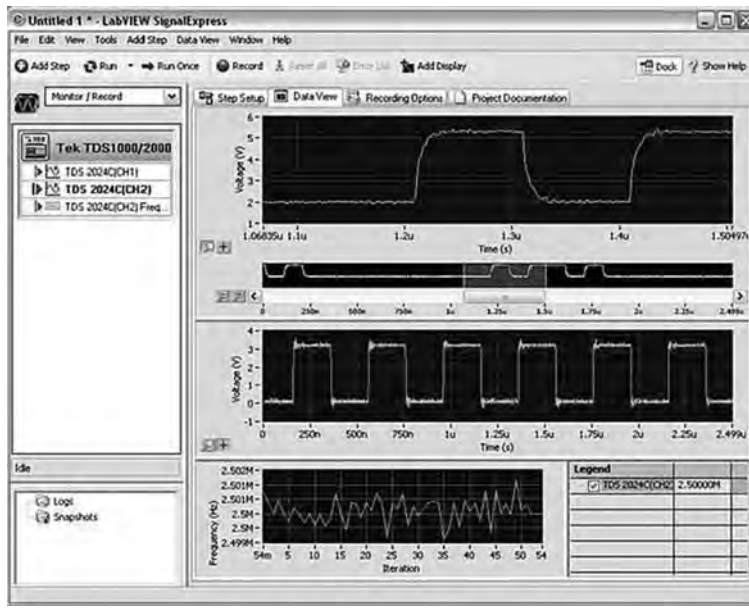


FIGURE 10.8 A configuration-based DAQ application where one can set up DAQ operations without any programming. (From http://www.tek.com/sites/tek.com/files/media/image/A002_8167-L.jpg)

TABLE 10.1 Types of Data Analysis and Processing Functions in Typical DAQ Applications

Measurement Analysis	Signal Processing Routines	Mathematical Processes
Amplitude, level, peak frequency, spectral, noise, distortion, pulse, transition, signal generation, time domain, tone measurements, etc. *Control*: PID, fuzzy control	Digital filters, convolution and correlation, frequency/spectral analysis, joint time-frequency analysis, sampling/resampling, decimation signal generation, transforms, time domain analysis, wavelet and filter bank design, windowing, etc.	Basic math, curve fitting, data modeling, differential equations, interpolation, extrapolation, linear algebra, nonlinear systems, optimization, root finding, statistics, and probability. *Geometry*: (2D, 3D), polynomial functions, etc.

10.3.3.3 Data Analysis and Visualization

The actual acquisition and storage of data are only a part of the DAQ system. In order to extract useful information from those data, make decisions on the process, and obtain results, the data need to be manipulated and analyzed. Therefore, data analysis and visualization are important components of data LabVIEW software systems, and many software packages include routines for data analysis and signal processing. Other packages require the data to be saved to disk or a database and retrieved later for processing using dedicated analysis software.

The type of analysis is, of course, very dependent on the goals of the application. Table 10.1 lists classes of analysis tools that might be integrated into a DAQ and analysis software program. This list is not all inclusive and only includes categories of analysis routines that are relevant for applications.

One key consideration for data analysis is whether the analysis is to be performed *inline* in real time or *offline* at a later time after the data have been stored. Inline analysis is used when decisions have to be made during run time and the results have direct consequences on the process. This is typically the case in control applications. When dealing with inline analysis, it is important to consider the amount of data acquired and the particular analysis routines that are performed on those data, as the processing could easily become computationally intensive and have an adverse effect on the performance of the application. The other key consideration for inline analysis is the use of point-by-point analysis algorithms. Most analysis software has analysis functions geared toward working on waveforms or arrays of data. Point-by-point analysis routines can simplify the program as they tend to more naturally match the flow of incoming data in an inline processing or control application.

Visualization is the ability of the software package to display the data and the results of the analysis in a meaningful way. Information and data may be plotted as graphs, charts, 2D images, and 3D graphs and even incorporate photos, videos, and 3D CAD models. Figures 10.9 is an example of a 3D analysis of a temperature distribution of a surface. Figure 10.10 is an example of LabVIEW application using sensor mapping function to map a 3D model of the analyzed data simultaneously displaying with synchronized sensor data.

10.4 Supervisory Control and Data Acquisition

SCADA is an industrial computer system monitoring and controlling large processes by using graphical user interface (GUI). The SCADA system consists of human machine interface (HMI), remote terminal units (RTUs), programmable logic controllers (PLCs), communications and network infrastructure, computers, embedded and other controllers, and appropriate software support at various levels. The term SCADA usually refers to centralized large systems that monitor and control entire sites or complex processes spread out over large areas. A typical SCADA network is illustrated in Figure 10.11. In this chapter, a brief description of SCADA is introduced.

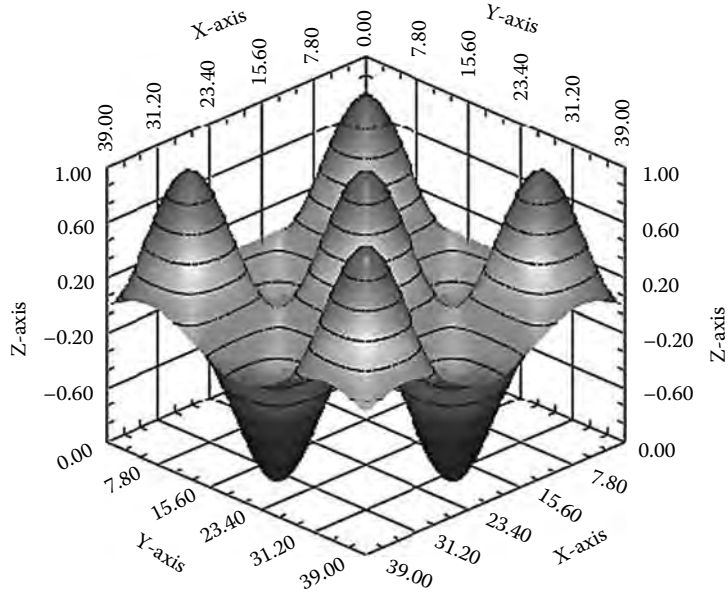


FIGURE 10.9 Example of 3D data graph of temperature distribution in LabVIEW. (Courtesy of National Instruments, http://www.ni.com/images/labview/2010/neutral/ui_3d_surface_graph.jpg)

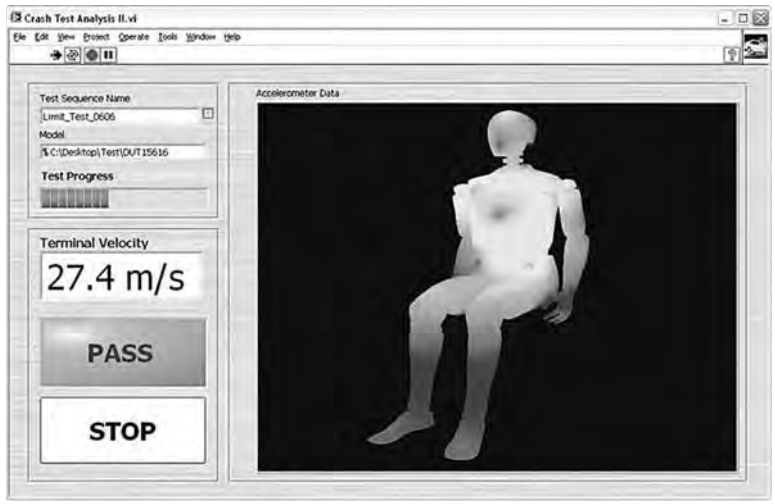


FIGURE 10.10 LabVIEW 3D sensor mapping forms the data analysis. (Courtesy of National Instruments, http://zone.ni.com/cms/images/devzone/tut/LV8.6_CrashTest_Analysis_fp_ss_p.jpg)

10.4.1 Components of SCADA

SCDA is a network of instrumentation systems. It has seven main components:

1. *Field I/O*: Digital and analog I/Os, rack mount with CPUs or networked.
2. *Logic solver*: Rack-mounted or RTU style wall mounted.
3. *Network and network components*: Ethernet and other networks.
4. *CPU, I/O, and network configuration and application development software*: Hardware platforms, such as Rockwell Automation and Allen-Bradley.

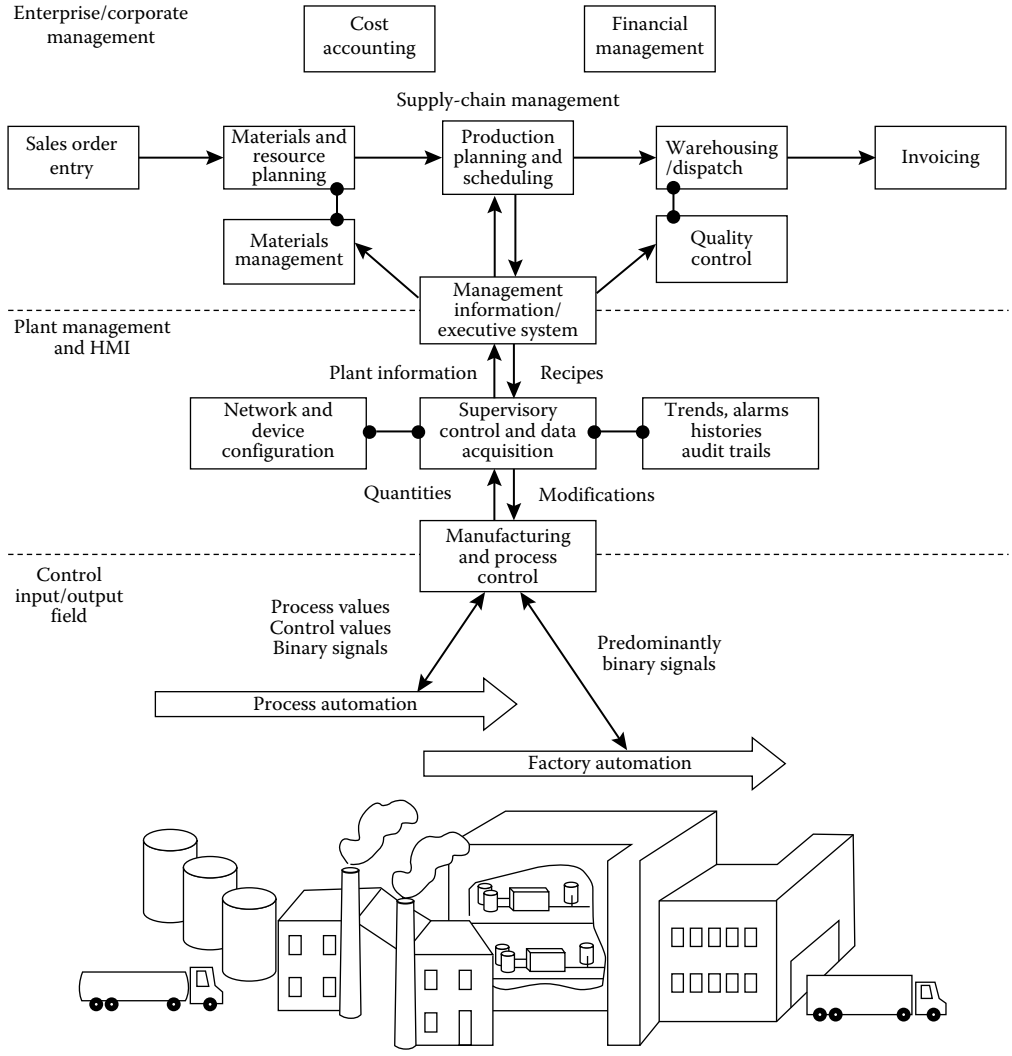


FIGURE 10.11 Typical SCADA network. (From Liptak, B. and Eren, H., *Instrument Engineers Handbook: Process Software and Digital Networks*, Vol. 3, 4th edn., CRC Press, LLC, Boca Raton, FL, 2011.)

5. *HMI hardware*: This can be industrial ruggedized PC hardware such as Versaview from GE Intelligent Platforms (GE-IP).
6. *HMI software*: Software, such as Wonderware, InTouch, or iFix for developing GUI.
7. *Engineering/maintenance workstation*: PC with all development software with the necessary licenses.

10.4.2 SCADA Software

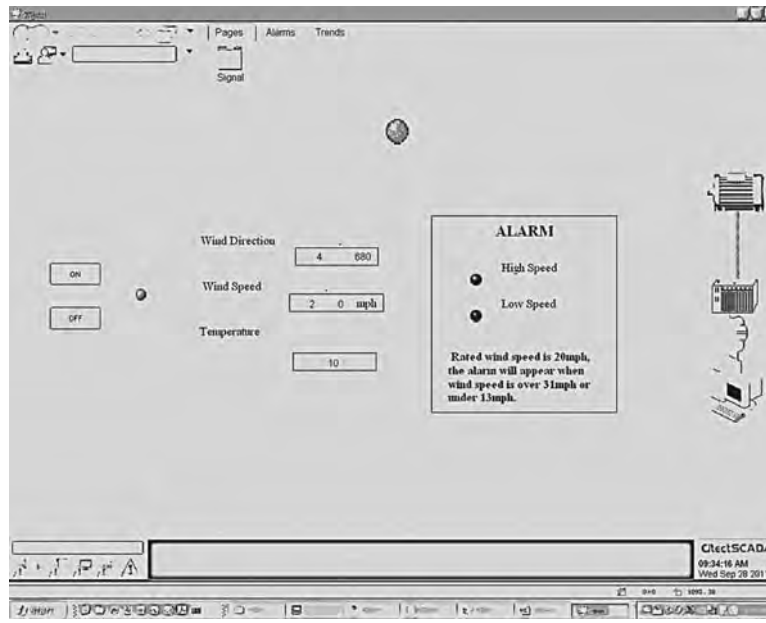
Software support of SCADA is configured in four layers:

- User machine interface
- Process and control
- Communication and memory
- Field drivers

User machine interface indicates the software for HMI and interactivity of operators with the process. The HMI programs can be very large and they are developed by following procedures such as

- HMI concept development—a mutually agreed requirement with the end user based on their existing standard or experience. It needs to define color scheme, navigation scheme, and individual equipment and device animation guidelines
- Preliminary system configuration—defines various nodes and configuration
- Middleware tag definitions and configuration—middleware is used such as object linking and embedding (OLE) for Process Control (OPC) or Modbus for live data
- Tag database loading—depends on completion of logic development because in the course of logic development, several memory tags are defined that may require animation
- Creation of building blocks of the process such as control valves, pumps, blowers, and tanks
- Creation of home screen and process screens by using process blocks and attaching tag names
- Creation of control panels or pop-ups that are detailed screens for operation of equipment
- Creation of setup parameter screens
- Creation of alarm management screen
- Creation of historical trends
- Creation of reports
- Basic navigation tests
- Connecting with controllers
- Testing set of screens with respective controller using forced tag values or jumpers

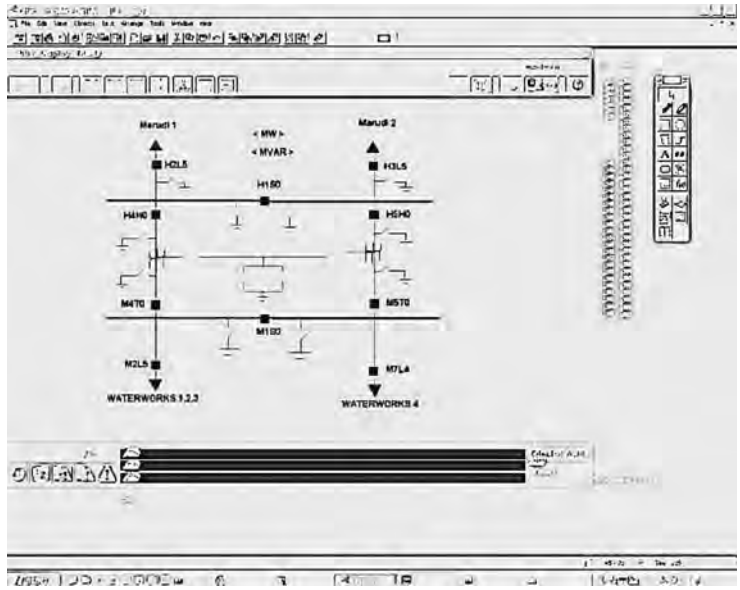
A simple example of a SCADA control system is illustrated in Figure 10.12a and b. In Figure 10.12a, the status and temperature and alarm condition of an electric motor are examined by calling a dedicated screen display. In Figure 10.12b, the operational characteristics of a substation in a large power distribution system are monitored.



(a)

FIGURE 10.12 SCADA displays for (a) to examine the status of a motor.

(*continued*)



(b)

FIGURE 10.12 (continued) SCADA displays for (b) monitoring a substation. (Developed by author at Curtin University, Bentley, Australia.)

10.5 Conclusions

Over the last few decades, DAQ systems have grown to fill an expanding role in process control, test, measurement, and automation. The wide adoption in particular of PC-based DAQ systems has been fueled by the technological evolution of PC technologies on all fronts. For example, a 65,000-point fast Fourier transform (FFT) would take over 15 min to execute on a commercial circa 1990 PC (Intel 386). On today's 3 GHz processors, equipped with multiple gigabytes of memory, the same FFT takes well under a second to execute. Those 1990 PCs utilized an 8 MHz parallel bus to move data. Today's PCs incorporate PCI Express data paths that can move data at rates up to 4 GB/s. DAQ software has grown to take advantage of the increased memory, speed, and multicore architectures of PCs, evolving into sophisticated graphical engineering tools such as the virtual instruments and SCADA for large systems involving hundreds and thousands of inputs and outputs.

Bibliography

Bentley, J. P., *Principles of Measurement Systems*, 4th edn., Prentice Hall, London, U.K., 2005.

Blume, P. A., *The Labview Style Book,* Prentice Hall, London, U.K., 2005.

Boyer, S. A., *SCADA: Supervisory Control and Data Acquisition*, 2nd edn., ISA, Research Triangle Park, NC, 1999.

Johnson, G. W. and Jennings, R., *Labview Graphical Programming: Practical Applications in Instrumentation and Control*, McGraw-Hill, New York, 2006.

Liptak, B. and Eren, H., *Instrument Engineers Handbook: Process Software and Digital Networks*, Vol. 3, 4th edn., CRC Press, LLC, Boca Raton, FL, 2011.

Mihura, B., *Labview Data Acquisition*, Prentice Hall, Upper Saddle River, London, U.K., 2001.

Potter, D. and Eren, H., Data acquisition fundamentals, *Instruments Engineers' Handbook*, The IEH, Eds., B. Liptak and H. Eren, CRC Press, Boca Raton, FL, Vol. 3, 4th edn., pp. 330–341, 2011.

Travis, J. and Kring, J., *Labview for Everyone—Graphical Programming Made Easy*, 3rd edn., Prentice Hall, London, U.K., 2007.

Partial List of Suppliers

ABB Automation, 2487 S. Commerce Dr., New Berlin, WI 53151, Phone: 262-785-3400, Fax: 262-785-0342, www.abb.us

Agilent Technologies, 140 Green Pond Rd., Rockaway, NJ 07866, Phone: 973-627-6400, Fax: 973-627-0570, www.home.agilent.com/

Alacron, Inc., 71 Spit Brook Rd Ste 200, Nashua, NH 03060, Phone: 603-891-2750, Fax: 603-891-2745, www.alacron.com

National Instruments Corp., 11500 N Mo Pac Expy, Austin, TX 75759, Phone (512) 338-9119, www.ni.com

Rockwell Automation, 1201 South Second Street, Milwaukee, WI USA 53204, Phone: 414-382-2000, Fax: 414-382-4444, http://www.rockwellautomation.com

Schneider Electric, Telemetry&Remote SCADA Solutions, 48 Steacie Dr., Kanata, ON K2K2A9, Phone: 613-591-1943, Fax: 613-591-1022, www.controlmicrosystems.com

Siemens Corp., 300 New Jersey Avenue, N.W., Suite 1000, Washington, DC 20004-2611, Phone: 202-434-4800, 800-SIEMENS (toll free), Fax: 202-347-4015, www.usa.siemens.com/entry/en/

Tektronix, Inc., 14200-T S.W. Karl Braun Dr., P.O. Box 500, Beaverton, OR 97077, Phone: 503-627-7111, 800-833-9200 (toll free), www.tek.com

Test Equipment Depot, 99 Washington St., Melrose, MA 02176-6024, Phone: 781-665-1400, 800-517-8431 (toll free), Fax: 781-665-0780, www.testequipmentdepot.com

Visidaq Solutions, Inc., 1970 Swarthmore Ave., Lakewood, NJ 08701, Phone: 732-901-6100, Fax: 732-901-2288, www.visidaq.com



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

11

Fail-Safe Instruments and Devices

Davide Quatrini
Italcertifer, Inc.
University of Rome "Tor Vergata"

Giuseppe Fazio
University of Rome "Tor Vergata" & Isatel Ltd.

Mauro Giaconi
University of Rome "Tor Vergata"

Adelio Salsano
University of Rome "Tor Vergata"

11.1 Introduction **11**-1
11.2 Fail-Safe Realization: Relevant Features **11**-2
Safety in Communications
11.3 Fail-Safe Realization: In-Depth Example **11**-3
General Architecture • Central Unit • I/O and Safety Board • I/O Ports • Fail-Safe Signal Generator • Auxiliary Clock and Timer Block • Fail-Safe Voltage Generator • Fail-Safe Procedures: Tests and Their Timing • Fail-Safe Relays • Reconfigurability
11.4 Case Study: Level 1 ERTMS **11**-9
11.5 Summary **11**-11
References **11**-11

11.1 Introduction

Some electronic applications are related to safety-of-life quality services, that is, services that must assure the safety of users or customers lives. Such an important task needs a life-cycle philosophy for the product strongly different from the one used in the development of other simpler commercial applications. For this reason all over the world, norms that describe the mandatory life-cycle management for fail-safe devices exist and must be applied by designers, verifiers, assessors, etc. The aim of this chapter is to provide the basic information needed for understanding how the world of fail-safe electronics is different from the one tied to standard applications in which human life is not involved.

A fail-safe device is essentially a device for which the final status in case of failure is absolutely known. This is a task that is hard to accomplish. Failures in fact can have several origins and several effects and can combine among themselves for producing unpredictable results. If we think, for example, of a railway traffic light, we must assure if we want it to be a real fail-safe device that all possible failures (including electrical and mechanical ones, both internal and external) lead to a situation in which the light is red or off, never green.

In this chapter, we intend to show the key concepts related to fail-safe instrument and device realization and how those concepts assemble to form a complete life cycle for such devices. The concepts are taken from the European norms and railway technical scenario, but they will be treated in a sufficiently generic way to be applied to the majority of devices that need strong safety performance. There are many fields in which the design philosophy here described should be applied (other than the already cited railway scenario): avionics, automotive, medical devices, nuclear plants, etc.

11.2 Fail-Safe Realization: Relevant Features

The first item on which safety is built is quality. The realization of a safe device must be done using the most modern methods and applying the most modern techniques of quality management, methods, and techniques that assure the minimization of human-related errors. Such a task is mainly achieved in a twofold way: a complete and correct documentation that follows the fail-safe device for the entirety of its life cycle and a deep and detailed technical analysis. The documentation, which should be prepared and maintained by a specialized team, should address at least the following issues from the very beginning of device life:

- The description of system architecture, with specific subsets related to the subsystem features
- A detailed description of system interfaces, both for the human user and other systems and also between the subsystems
- A complete tracking of system requirements, both functional and safety ones
- A detailed description of software and hardware functioning, with particular emphasis on their mutual relationships

Note that the software and hardware functioning description must detail also the RAMS parameters of the system, that is, its reliability, availability, maintainability, and safety, from both qualitative and (if possible) quantitative points of view.

The technical analysis, although based on the quality criteria detailed earlier, must be solidly focused on safety. For this reason, the analysis must start with a systematic evaluation of the faults, carried on using both top-down techniques like fault tree analysis (FTA) and bottom-up systems like failure mode effects analysis (FMEA). This approach should lead to a scenario in which the device remains safe if any single random fault occurs or, in other words, it is fail-safe. This goal could be achieved in several ways:

- *Composite fail safety*: The system is designed to be oversized/redundant. Every additional unit must be independent from other units, thus avoiding common-mode failures, and also for taking safety decisions based on a majority vote. An implementation in which the units are functionally equivalent but architecturally different is strongly recommended.
- *Reactive fail safety*: The system is formed by a single unit, but every possible fault of this unit is immediately detected and negated. This task is accomplished by an additional unit that is not functionally equivalent to the first (as in composite fail safety), but instead it is totally focused on the continual testing of the first.
- *Inherent fail safety*: The system is formed by a single unit, but its every possible fault has been analyzed and judged as not hazardous. An example could be a very simple analog hardware–based watchdog, in which all the possible faults, related to its single components, lead to a shutdown of the system (but only if the shutdown is the final, desired, safe status of the system).

It is evident that independently from the chosen approach (which can consist also in a combination of the aforementioned techniques), the practical realization of a fail-safe device requires a deep and complete knowledge of its single components, both software and hardware. Such a knowledge level is very hard to obtain using recent up-to-date VLSI products or powerful programming languages, so very often, the fail-safe designer must be a specialist of the old, well-tested, and completely debugged elements, for example, the microprocessors developed during the 1980s and the 1990s (or their DSP counterparts), the ANSI specification for the C programming language, or also even the assembly approach to software writing. Sometimes the necessity for a tailor-made ASIC could arise, so HDL skills are required. Use of commercial-off-the-shelf (COTS) products must be avoided, if possible, or minimized, and however, its employment is often regulated by strict norms. This choice is not based on technological reasons tied to specific performances (which can be obviously more easily obtained using the most recent multipurpose devices); the reason is always the necessity for the absolute knowledge of the system status in case of failure.

11.2.1 Safety in Communications

When dealing with secure communications, a general problem is detecting changes to raw data.

In this context, changes may be accidental or intentional. In case of accidental modification of data, cyclic redundancy check (CRC) is commonly used; the algorithms rely on cyclic codes and are simple to implement in hardware and very powerful in detecting errors due to noise in digital transmission, especially burst-type errors.

CRC algorithms are based on CRC polynomials of degree k of the variable x, specifically designed to protect messages of various lengths. The messages of n bits are considered polynomials of the variable x, with coefficients given by the values of bits, of degree $n + k - 1$, appending k zeroes at the end of messages. In order to compute the k-bit CRC, the polynomial message is divided by the CRC polynomial, appending at the end of polynomial message the remainder of the division. At the receiver, the division is performed again, verifying the equality of the received and newly calculated remainders.

In case of intentional changes, CRCs are totally useless and more powerful algorithms must be used, for instance, message integrity code (MIC) adopted by IEEE 802.11, based on secret temporal keys and Advanced Encryption Standard (AES).

11.3 Fail-Safe Realization: In-Depth Example

The cited safe systems can be realized in wired logic with electromechanical devices: the diagnostics are implemented using fail-safe relays and are based on the hypothesis that all possible faults could be individuated and the consequence would be to drive the system toward a "safer" state.

In the following, we will instead focus on the usage of programmable circuits in fail-safe systems. Since these systems can be considered as a subset of the self-checking ones, the typical diagnostic procedures of fault tolerance (Hayes and McCluskey 1980) can be used. Such procedures consist of the implementation of structures that must be self-checking as a whole: this goal can be obtained either with single electronic systems (Rennels 1978, Toy 1978) using a suitable redundancy or with two or more identical systems working synchronously whose outputs are compared in real time (Hopkins and Smith 1975). A solution of the second type is described in the following.

This solution consists of a system realized with two microcomputers and an arbitration unit; moreover, the correct flow of the software is controlled with a suitable hardware and software redundancy. A structure as the proposed one has the advantage of a high degree of software transparency with respect to other systems (Rennels 1980). In fact, as the only requirement on the software is some constraints on its flow, the whole system can be considered, from a functional point of view, as equivalent to a normal microcomputer.

11.3.1 General Architecture

The structure of the fail-safe computer is shown in Figure 11.1.

In this figure, the central unit, the input and output (I/O) blocks, and the circuits dedicated to the generation of fail-safe signals are recognizable. The central unit is a fault-tolerant system where the diagnostics are implemented through two identical microprocessor-based subsystems working synchronously. The following assumptions were assumed when developing the design:

- Both hardware and software are intrinsically "unsafe," and therefore, the final aim is to obtain a "fail-safe" system using unsafe instruments.
- The diagnostic signals that indicate a fault can arise anywhere in the system: from any origin, the design must warrant that the safe shutdown procedure would start, driving the system to the programmed safer state. This safer state corresponds to the absence of a secure supply voltage, which is normally generated through the relay R_{FS} of Figure 11.1.

Finally, the control signals and the procedures working in safety must be dynamic, to avoid faults of the "stuck-at" type.

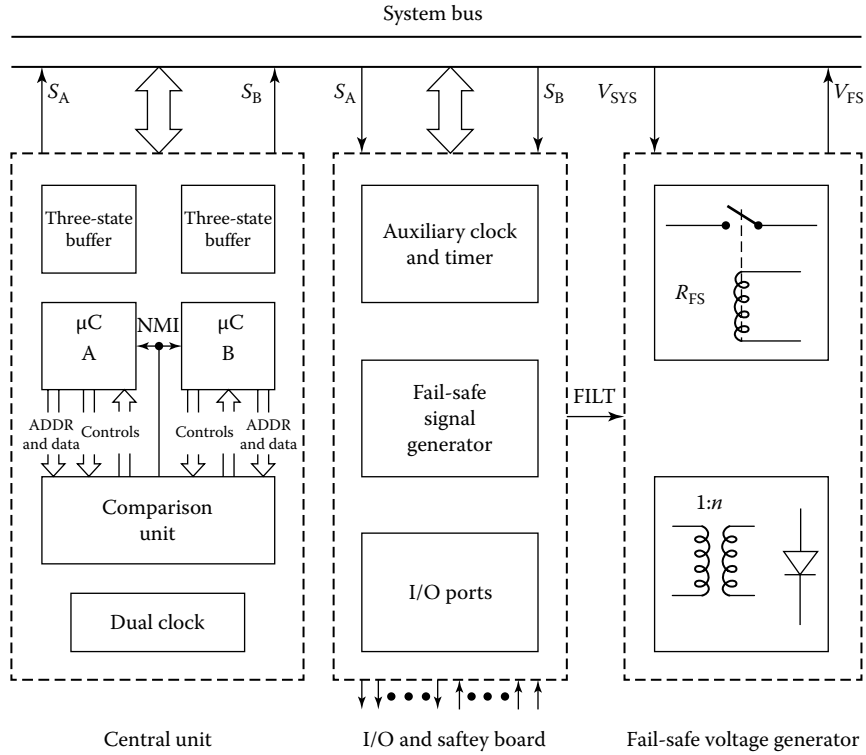


FIGURE 11.1 Structure of the fail-safe computer.

11.3.2 Central Unit

The central unit is realized with two Zilog Z80 CPUs, each with its own ROM and RAM, which run identical programs; output data and address are compared and possible disagreements are shown, causing the de-energization of the relay R_{FS}.

The correct flow of the program is verified by dividing the whole software, both user and system, into a known number of identical length blocks. This total number of blocks and their length are calculated by the "I/O and safety board" (described as follows) and compared with the theoretical ones.

The dual clock indicated in Figure 11.1 corresponds to the possible implementation of a real-time test procedure of the clock frequency, with the possible substitution of the faulty clock.

Finally, a periodic test procedure of the comparators is also implemented, to avoid stuck-at faults. Some different instructions are periodically inserted in the programs of both CPUs, and the outputs of the comparators are tested, looking now for the detection of the disagreement. The de-energization of the relay R_{FS} is correlated to the faulty operation of the comparators.

11.3.3 I/O and Safety Board

Figure 11.2 shows the block diagram of the board dedicated to the I/O and to the fail-safe controls.

This board, connected through the system bus to the central unit board,

- Manages the bidirectional communications with the external world
- Periodically checks the correct operation of the I/O devices
- Tests the duration of the software blocks and the frequency of the main clock
- Controls the energization of the fail-safe relay R_{FS}, dropping it out if some fault is detected, either in the same I/O or in the central unit board

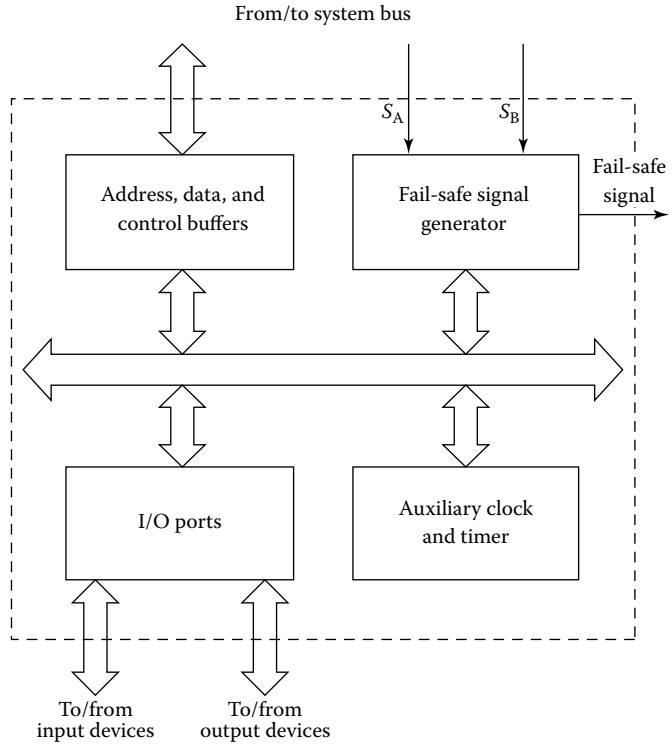


FIGURE 11.2 I/O and safety board.

11.3.4 I/O Ports

The system allows the safe transmission of remote controls and the safe acquisition of telesignals.

Two PIOs are used as shown in Figure 11.3, where Port A of the PIO 1 and the Port B of the PIO 2 are, respectively, dedicated to the output and to the test of the output, while the Port A of the PIO 2 and the Port B of the PIO 1 correspond to the input and to the control of the input. The output works as follows: The remote control signals, issued by the Port A of the PIO 1, drive the power transistors, whose outputs are read by the Port B of the PIO 2, and the system is considered faulty if no correspondence exists between issued and read signals. Moreover, the output is periodically checked to verify its capacity-changing state.

The input section is realized with optoisolators, which allow, varying one resistor, to work with digital signals of various values and protect the system from anomalous voltages. The test of this section is carried out by putting the output buffers in the low-impedance state and driving the input of the optoisolators with test signals issued by the Port B of the PIO 1. If appropriate ratios of R_{IN}/R are used, the signals read by the Port A of the PIO 2 must correspond to the issued ones: when such correspondence does not occur, the safe shutdown procedure will start. This test can be made either periodically or continuously, issuing after each input signal is received or the same signal is denied.

11.3.5 Fail-Safe Signal Generator

The safety characteristics of the whole system are correlated to the generation of a fixed frequency signal FILT, which is switched off when a fault is recognized. The frequency of FILT corresponds to the length of the blocks in which the whole program is divided and its generation is related to both hardware and software procedures.

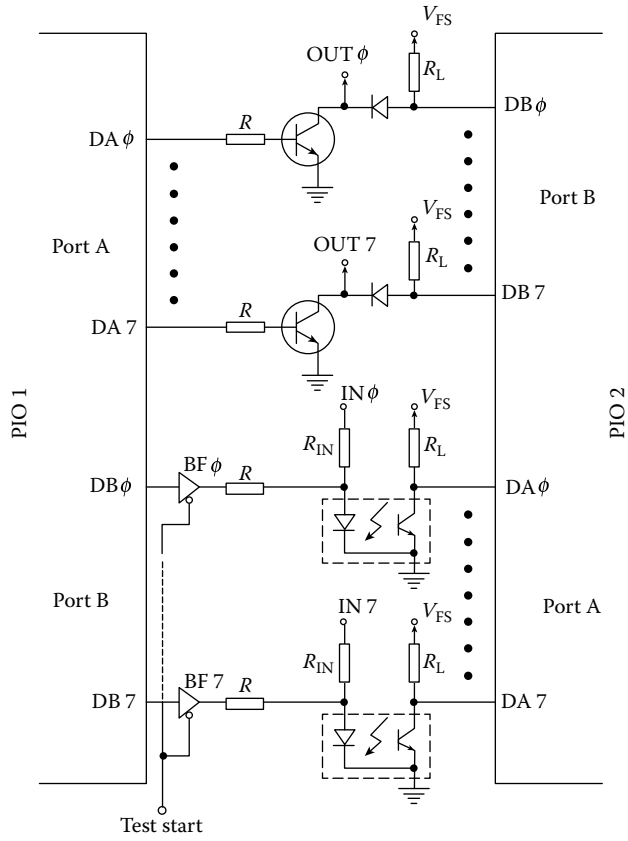


FIGURE 11.3 I/O configuration.

Each software block holds some instructions able to drive a hardware network that generates the basic signal *F*1. The block configuration of the circuit is shown in Figure 11.4. The *S* and *R* inputs of the flip-flop *F* are normally high: they go to zero at the end of each software block, when appropriate words on the address and data are generated by the program through arithmetic and logic operations. The output of the flip-flop *F* is the signal *F*1, while the final fail-safe signal FILT is obtained through SW_A and SW_B that halt the signal if some real-time test gives a wrong result. The control of the frequency of FILT is one of those real-time tests, made on the same I/O and safety board; therefore, the signal FILT outputted toward the fail-safe voltage generator can be the square wave, the correct frequency signal, or a zero-voltage signal, when a fault has occurred.

11.3.6 Auxiliary Clock and Timer Block

The auxiliary clock and timer block carries the appropriate time controls for the whole system. It is realized with a clock independent from the clock subsystem of the central unit and a CTC (or CTC-like circuits), and it calculates the number and the time length of the software blocks in which the program is divided and the frequency of the main clock.

Correct operation causes a signal to be forwarded to the fail-safe signal generator block, and consequently, the absence of this signal halts the signal FILT. This block can be easily duplicated in order to increase the reliability.

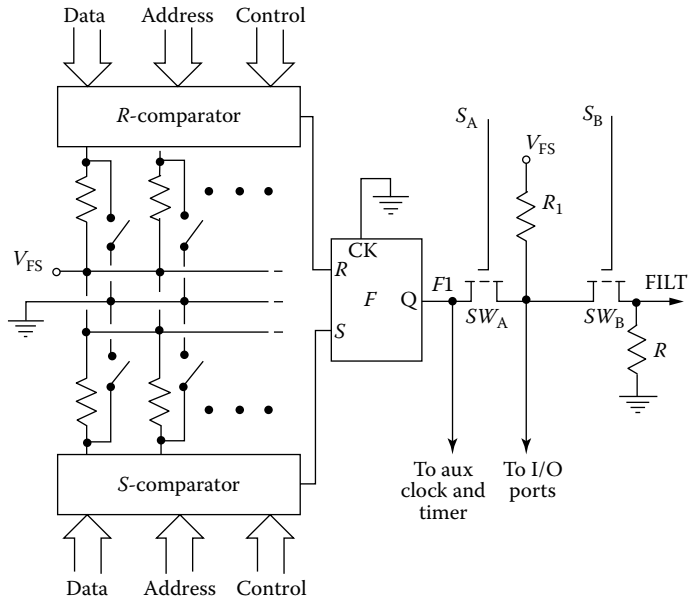


FIGURE 11.4 Fail-safe signal generator.

11.3.7 Fail-Safe Voltage Generator

The subsystem used to generate the fail-safe voltage is shown in Figure 11.5. This configuration prevents the input voltage from reaching the final fail-safe relay in the presence of any foreseeable fault of the transformer, diodes, etc.

The fail-safe voltage V_{FS} is used to supply through the system bus all the circuits, devices, and peripherals that must be halted when a possible fault is recognized, while the other circuits can be supplied directly by the voltage V_{SYS}.

Finally, the start of the system is assured using the fail-safe button B_{FS}, which bypasses the feedback control system of the supply voltage.

11.3.8 Fail-Safe Procedures: Tests and Their Timing

The starting point for the fail-safe operation of the whole system is the fail-safe relay R_{FS} that supplies voltage to all the parts of the system that must operate safely and defines the class of faults that must be

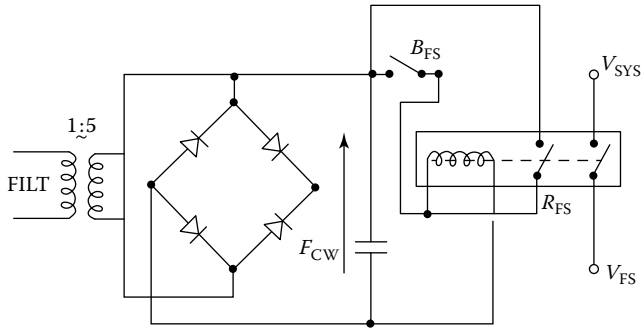


FIGURE 11.5 Subsystem used to generate the fail-safe voltage.

considered as foreseeable during the normal life of the system. To avoid the most common faults, the "stuck-at" faults, the fail-safe voltage generator with the relay R_{FS} is driven by a dynamic signal FILT; every foreseeable fault, both hardware and software, implies either a fixed value of $F1$ or a frequency different from the theoretical one and the consequent halt of the input signal FILT of the voltage generator (Figure 11.4).

The halt of FILT is tied to the correct operation of the switches SW_A and SW_B. The configuration shown in Figure 11.4 allows an easy test of the functionality of the MOS switches oriented to safety. Since the absence of FILT drives the system toward a safer state, the test must warrant against the permanent short circuit of the switches and it is accomplished, when FILT is low, driving both MOS switches to the OFF state. Every fault, either of SW_A and SW_B or of the signals S_A and S_B, will make the I/O Port read a LOW value instead of a HIGH one if an appropriate ratio R_1/R is used.

The tests implemented on the computer are of two types, continuous, or real-time, and periodic.

The continuous tests mainly consist of the comparison of address and output data from the CPUs A and B of Figure 11.1, made by the comparator C. When a disagreement is detected, an NMI signal is generated to both A and B CPUs. The CPUs forward on different lines the signals S_A and S_B that are independently able to drop out the switches SW_A and SW_B (Figure 11.4), and they also signal the I/O tristate buffers of the central unit board to the system bus in the high-impedance state, so halting the whole system.

The first periodic test considered is the test of comparators. Disagreement, which corresponds in this case to the right operation, always generates the NMI signal. But unlike the previous case, the system software prevents the tristate buffers from going into the high-impedance state, it doesn't allow the signals S_A and S_B to reach the MOS switches SW_A and SW_B (Figure 11.4), and it also actively contributes to the generation of the signal FILT. Another periodic test concerns the I/O ports: again when a fault is detected, the central unit board generates the signal S_A and S_B that stops the signal FILT.

Finally, a real-time test is implemented during output operations, in order to verify the agreement between the transmitted output and the signal received by the driven device: a disagreement causes the generation of the previously cited S_A and S_B that stop FILT.

The generation of FILT is obtained, using hardware and software procedures with the maximum possible degree of safety; moreover, each fault that causes, for instance, the absence of the alternating pulses at the inputs of the flip-flop F of Figure 11.5, determines a continuous value of FILT, not valid for the generation of V_{FS}. The frequency of FILT is directly related to the minimum reaction time requested for the safe operation of the whole system.

But all the previously cited test and procedures are useful only if they are "sure" that every fault will cause the de-energization of the relay R_{FS}. Such "sureness" has been reached, to a certain degree, making it possible that every class of faults will operate in different ways for preventing a square-wave signal FILT from reaching R_{FS}, not generating the signal $F1$ and doubling the lines S_A and S_B of Figure 11.1.

11.3.9 Fail-Safe Relays

Another assumption on which this work is based is that the fail-safe relays, if broken, go into a foreseeable state. In effect, if the relays are of the fail-safe type used, for instance, in the railway field, this assumption is commonly accepted and, moreover, contributes to the acceptability of the proposed system in fields connected with safety that are, rightly, very conservative.

However, alternative solutions based on the hypothesis, generally accepted in fault-tolerant systems, which two contemporary faults cannot occur, can be proposed.

An example consists of the use of MOS switches of the appropriate power mounted with a suitable redundancy that uses diagnostic methods for their periodic test. A possible configuration is shown in Figure 11.6, where the relay is replaced with eight MOS switches arranged in two parallel identical paths.

Considering the four driving MOS indicated with MD, the continuous fail-safe voltage is the supply, while the gates are driven by a fail-safe I/O configuration such as the one shown in Figure 11.3.

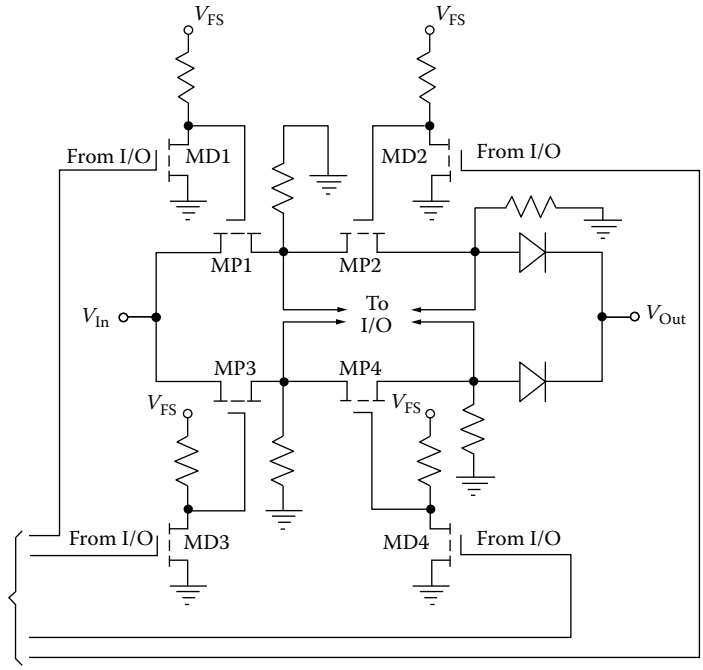


FIGURE 11.6 Solid-state fail-safe relay.

The test is made in sequence on the electronic switches MP, comparing I/O of each MOS, while the connection between the input and the output of the whole relay is assured by an independent test.

11.3.10 Reconfigurability

Although the described system is capable of a single kind of action, it is surely possible to modify it in the way in which its safe outcomes could be different on the basis of some configuration choices. For this reason, most recent fail-safe systems (e.g., European Railway Traffic Management System (ERTMS); see the following) include a configuration file in which parameters can be set at specific values, thus setting the performance of the system. Also for the production of the configuration files, fail-safe design techniques must be used, of course, from both design and communication points of view.

11.4 Case Study: Level 1 ERTMS

The ERTMS is a signaling system capable of maintaining the railway traffic safe in every condition, including the driver's incapacitation or death. Although it is mainly a European product, it is actually spreading all over the world, and its installations exist (or will exist very soon) in Asia, Africa, America, and Oceania. Its system requirement specifications are actually very solid: they are the product of a long effort, a work in which infrastructure managers, railway companies, technological companies, and various suppliers have been involved in for many years. From a very-high-level point of view, the system is formed by two segments: the trackside subsystem (composed of devices installed on the ground) and the onboard subsystem (i.e., installed on trains and called ETCS, i.e., European Train Control System). Both of the subsystems are modular and can be configured for using different technological solutions as substratum: low-frequency dedicated radio transmission, GSM for railway (GSM-R), and also satellite communication are all viable implementations (although the last is actually only at the prototype level). The combination of a specific subset of the system requirements specifications and one (or more) of the

technological solutions makes what is called a "level." ERTMS is, in fact, available in three levels: 1, 2, and 3 (plus Level 0 that is used for indicating ERTMS predisposition or transition from a legacy national train run protection system toward an ERTMS level). In the following paragraphs, a brief description of Level 1 in its simplest implementation is provided, with emphasis added on the practical implementation of the safety concepts shown before.

The trackside subsystem of Level 1 ERTMS is formed of passive devices installed on the railway lines, devices called "Eurobalises." Two types of Eurobalises exist: one with fixed informative content and one with variable informative content. The first is used, for example, for communicating to the onboard subsystem (and so to the driver) the exact position of the train, while the second is used for transmitting the information about the status of the line (e.g., free or occupied by another vehicle). Independent from the type, the transmission of the information is possible only through strict interaction between the trackside subsystem and the onboard subsystem. Eurobalises are passive devices and can transmit only when a certain quantity of energy is given to them by a specific antenna, part of ETCS, installed on trains. When the train passes on a Eurobalise (installed on the "sleepers" between the steel rails), its ETCS antenna, situated on the bottom of the locomotive, transmits electromagnetic energy to the balise itself that, being energized, "answers" sending a proper "telegram" to the same antenna, which is a transmitter/receiver. The "telegram," which is transmitted using a dedicated channel at MHz frequencies, contains the information about position or line status described before (the transmission uses the frequency-shift keying technique, although also amplitude-shift keying is available for retrocompatibility purposes). The telegram received by the ETCS antenna is then converted into a digital string by a specific ETCS module (the BTM, Balise Transmission Module) and then sent to the ETCS logic for elaboration. The elaboration leads to some actions by the ETCS logic, actions that often involve the driver's interaction with the train (usually requiring speed reduction or braking application) or automatic reaction in case of the driver's unavailability. As it is easy to see, the high-level architecture is simple (as always requested for fail-safe applications) but also sufficiently modular to allow a great range on configurability and usability. Let us now see how the fail-safe concepts are taken into account for the implementation of Level 1 ERTMS.

First of all, in Eurobalises, we can see a clear example of inherent fail safety in their design: they are passive (i.e., not connected to any power supply), so leading to a technical analysis that is exhaustive from the fault/failure point of views (for both internal and external causes). Besides, they are designed to transmit both default telegram or no telegram in case of vital/fatal error, leading to a safe reaction by the onboard subsystem (which is often the emergency braking, i.e., the one performed acting directly on the brake commanded by a pneumatic pipe and capable of blocking the train wheels). Also composite fail-safety techniques are used in their deployment: they are always installed in couples, in a full redundant configuration (a configuration called "balise group"). Finally, each Eurobalise, when transmitting its telegram, communicates also to the ETCS the position of the following balise group ("announcing" it): it is easy to imagine the ETCS reaction when an "announced" balise group is not encountered, both for its proper failure (e.g., on the antenna) or for balise fault. It is, of course, emergency braking.

ETCS: For a system that must be capable of "making decisions," that is, a system that has a complex hardware architecture with a software logic embedded and that must perform different actions on the basis of the information transmitted by the balises, inherent fail safety is not, obviously, a viable option, at least from a high-level point of view (it could be a possible design choice for single components, however, as seen in detail in Section 11.3). For this reason, the ETCS system is designed to be fully redundant, both at a system level (which is formed, substantially, by two twin systems installed on the same train-borne rack) and at a single-board level (each fully equipped with redundant logic units), characterizing itself as a composite fail-safe system. In addition, specific modules are built also with a reactive fail-safety approach. The antenna (used, as already said, for energizing the balises), for example, must include a subsystem capable of testing its transmitting/receiving performances at least at the system start-up (several antenna implementations are built having autotesting capabilities active also after the system start-up, however, and the tests are executed very frequently).

Note how the concepts related to fail safety are extensively used in ERTMS. Any single fatal fault or failure, either on the trackside segment or on the train-borne one, leads the system to a safe status (the application of the emergency brake). A task that is accomplished through the use of all the fail-safe design techniques described earlier, using composite systems that are also inherently safe on their low-level architecture or reactively safe because of the presence of external "verifiers."

11.5 Summary

A device is fail-safe only if its final status is always known, despite the kind of failure that may occur. This performance is necessary because human life depends on the safety of the device, and so exceptions are not acceptable. For this reason, the design approach for safety devices is radically different from the one used for not-safety-related applications. A very solid process, which must comply with the strictest quality standards, must be carried on by groups of experts rigidly separated by function (design, engineering, verification and validation, assessment). The components, both software and hardware, must be selected to be simple, fully known and debugged, and not extremely recent. They must be deeply analyzed from both top-down and bottom-up perspectives and combined together using techniques (composite, reactive, and inherent fail safety) capable of assuring the overall safety of the system, also formed by a large number of components or segments.

In this way, the possibility for building trains that guarantee the safety of the users' lives or medical devices that never cause harm to patients surely exists.

References

Hayes, J.P. and E.J. McCluskey, Testability considerations in microprocessor-based design, *Computer* 13: 17–25, 1980.

Hopkins, A.L. and T.B. Smith, The architectural elements of a symmetric fault-tolerant multiprocessor, *IEEE Transactions on Computer* C-24: 498–505, 1975.

Rennels, D.A., Architectures for fault tolerant spacecraft computers. *Proceedings of IEEE* 66: 1255–1268, 1978.

Rennels, D.A., Distributed fault-tolerant computer systems, *Computer* 13: 55–65, 1980.

Toy, W.N., Fault tolerant design of local ESS processors, *Proceedings of IEEE* 66: 1126–1145, 1978.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

12

Dynamic Error Measurements of Force Sensors

Akihiro Takita
Gunma University

Jin Tao
Gunma University

Yusaku Fujii
Gunma University

12.1 Introduction **12**-1
12.2 Problems and Solutions of Dynamic Calibrations **12**-1
Levitation Mass Method
12.3 Evaluation of Dynamic Characteristics of Force Sensors **12**-2
Structure of S-Shaped Force Sensor
12.4 Impulse Response **12**-7
12.5 Oscillation Response **12**-10
12.6 Conclusions **12**-14
References **12**-14

12.1 Introduction

Force sensors are extensively used in various research and industrial fields such as monitoring, testing, and manufacturing. Further, it is extremely important to calibrate the static and dynamic errors of sensors in order to improve their accuracy. However, only static calibration methods are available currently. It is easy to generate static forces precisely by using well-calibrated balance weights. However, this has a significant disadvantage, that is, the absence of a well-established method for generating and measuring dynamic force precisely. If we attempt to measure the dynamic forces acting on a force sensor for calibration, another force sensor with better dynamic properties and more accuracy than the tested sensor becomes necessary, thereby creating infinite loop and somewhat nonsensical situation. In this chapter, we introduce a method for calibrating force sensors without using other force sensors.

12.2 Problems and Solutions of Dynamic Calibrations

Although there are no well-established methods for the dynamic calibration of force sensors, a few researchers have previously attempted to develop them. Prof. Kumme has proposed and developed a method that uses the inertial force of the attached mass shaken by a shaker [1,2], and Prof. Park et al. have used this method to investigate the dynamic characteristics of multicomponent force–moment sensors [3,4]. This method enables the evaluation of the dynamic characteristics of force sensors under a condition in which the force acts as a continuous and single-frequency vibration. Therefore, this method was not suitable for evaluating the dynamic characteristics with impulse responses, which is important for analyzing crash tests.

A few researchers [5,6] have attempted to develop the dynamic calibration method for impulse responses as well; they concluded that the differences between the static and the dynamic characteristics

are explained by the inertial force of a part of the force sensor itself. However, their methods could only evaluate the response after an impact because the force could not be measured during the impact itself; moreover, the reference forces also cannot be measured.

Hence, we developed some methods for dynamic force calibration by using the levitation mass method (LMM). In the LMM, the force acting on the sensor can be measured precisely without using any other force sensors.

12.2.1 Levitation Mass Method

The LMM was proposed and developed by the authors. In this method, a rigid body is made to collide with the measurement object and a dynamic force is applied on it. The velocity of the rigid body during the collision is then measured, and as a result, the force acting on the object can be measured by using the equation of motion, $F = ma$, where F is the instantaneous force; m, the mass of the rigid body; and a, the instantaneous acceleration of the rigid body calculated by the differentiation of the instantaneous velocity of the rigid body. To move the rigid body linearly with little friction, an aerostatic linear bearing is used. Because the velocity and the mass can be measured precisely by using a laser Doppler interferometer and a well-calibrated balance, the force also can be measured precisely. Therefore, the LMM is suitable for calibrating the dynamic error of force sensors. The authors applied this method for dynamic force calibration under impact, oscillating, and step loads [7–11]. In the following section, the dynamic calibration methods using the LMM are reviewed.

The LMM can be applied not only to the dynamic calibration but also to the other force measurement applications. The authors developed evaluating method for material viscoelasticity, material friction, biomechanics, dynamic performances of linear motor, and human body weights under microgravity condition [12–23].

12.3 Evaluation of Dynamic Characteristics of Force Sensors

An example of the calibration and correction method of dynamic force measurements using the LMM is shown here. An S-shaped force sensor was used to measure the impact force with a pulse width of approximately 7 ms full width at half maximum (FWHM).

12.3.1 Structure of S-Shaped Force Sensor

Strain-gage force sensors are the most common force sensors because they are inexpensive but have high sensitivity. Figure 12.1a shows the schematic diagram of an S-shaped strain-gage force sensor. The sensor is formed by two metal blocks and a parallel leaf spring connecting the two blocks. Therefore, the sensor is modeled as a mass spring model, as shown in Figure 12.1b. When the force is acting on the sensing point, the spring is compressed in proportion to the force. The displacement of the sensing point is measured by strain gages and is converted into force by using Hooke's law, $F = -kx$, where F is the force; k, the spring constant; and x, the displacement. In the dynamic force measurements, the metal block, including the sensing point, is in motion, and its velocity is changed by the force acting on the sensing point. In such a case, the force that compresses the spring is not only the force acting on the sensing point but also the inertial force of the block. This inertial force is the primary cause of the error in the dynamic measurements with this type of force sensor.

Figure 12.2 shows a schematic diagram of the experimental setup for measuring the electric and mechanical responses of the force sensors against the impact force. A force sensor, whose rated force capacity is 200 N, was stiffly fixed to a base. The base was heavy and entirely fixed to the floor. An aerostatic linear bearing was used to enable horizontal linear motions with negligible friction. An impact force was generated and applied to the sensor by making the moving part of an aerostatic linear bearing to collide with the sensor. An initial velocity was manually imparted to the moving part. An extension

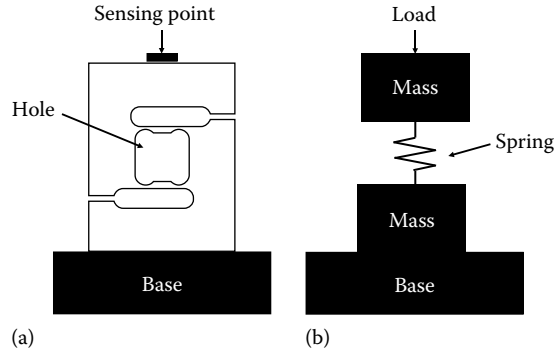


FIGURE 12.1 S-shaped strain-gage force sensor: (a) Schematic diagram, (b) spring mass model.

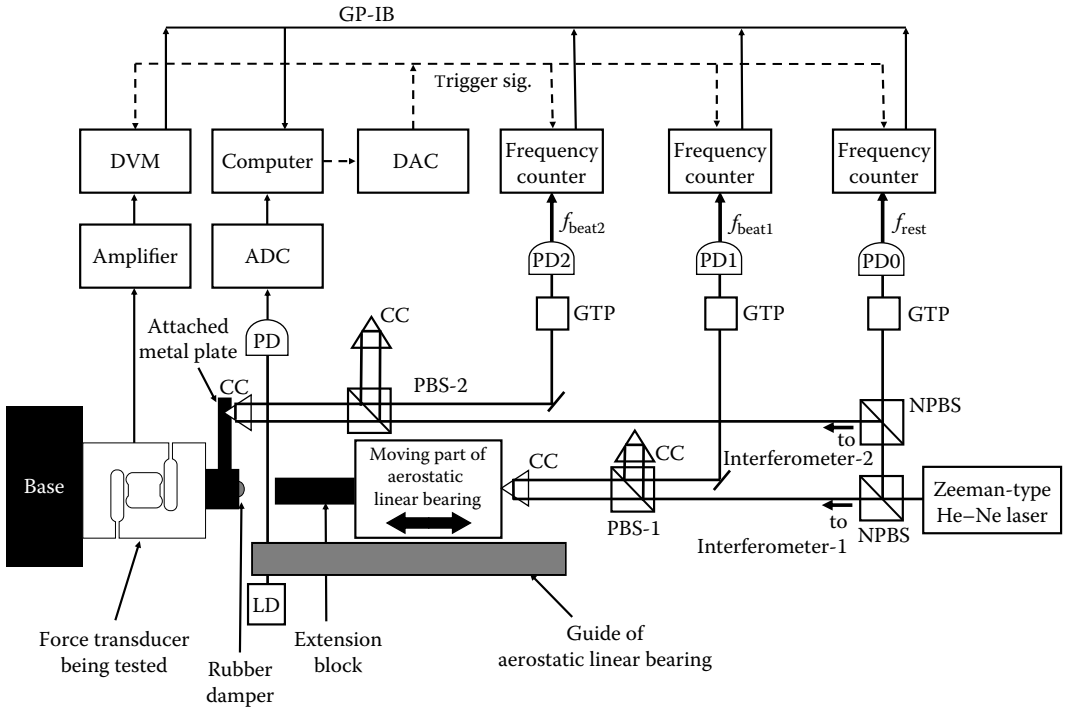


FIGURE 12.2 Experimental setup for calibrating the impact test of force sensors. CC, corner cube; PBS, polarized beam splitter; NPBS, nonpolarizing beam splitter; GTP, Glan–Thompson prism; ADC, analog-to-digital converter; DAC, digital-to-analog converter; DVM, digital voltmeter; PD, photodetector; LD, laser diode.

block for adjusting the collision position as well as a corner cube (CC) prism for turning the laser beam was attached to the moving part. The total mass of the moving part including the extension block and CC, M_1, was 2.653 kg.

The sensor being tested was statically calibrated with the standard uncertainty of approximately 0.2 N, that is, 0.1% against the rated force capacity. The output signal of the sensor was measured by using a digital voltmeter (DVM) (VP5481L, Panasonic Corp.). A CC prism was attached near the sensing point of the force sensor, as shown in Figure 12.2, in order to measure the displacement of the sensing point by an optical interferometer. The mass of the attached metal plate including the CC prism, M_2, was approximately 0.082 kg.

The force acting on the moving part of the linear bearing, F_{mass}, is calculated as a product of its mass and its acceleration, a_1, as follows:

$$F_{mass} = M_1 a_1 \tag{12.1}$$

Because the friction acting on the moving part of the bearing was negligible, the force acting on the moving part can be considered to be equal to the force acting on the force sensor. The acceleration was calculated from the velocity of the moving part. The velocity of the moving part was measured by the optical interferometer.

The instantaneous velocities of the moving part and the sensing point of the force sensor, v_1 and v_2, respectively, were measured by using two optical interferometers. A Zeeman-type two-frequency He–Ne laser was used as the light source. The frequency difference between the two frequencies was $f_{rest} = |f_1 - f_2| = 2.6$ MHz. The velocity was measured as the Doppler shift frequency, $f_{Doppler}$, of the signal beam. The velocity can be expressed as follows:

$$v_1 = \frac{\lambda_{air}(f_{Doppler1})}{2} \tag{12.2}$$

$$f_{Doppler1} = -(f_{beat1} - f_{rest}) \tag{12.3}$$

$$v_2 = \frac{\lambda_{air}(f_{Doppler2})}{2} \tag{12.4}$$

$$f_{Doppler2} = -(f_{beat2} - f_{rest}) \tag{12.5}$$

where

λ_{air} is the wavelength of the signal beam
f_{beat1} is the beat frequency detected by the photodetector (PD)-1
f_{beat2} is the beat frequency detected by the PD-2

Further, the directions of the coordinate system for the force, velocity, and acceleration for the moving part were toward the right (Figure 12.2). Three electric frequency counters (R5363, Advantest Corp.) measured and recorded the beat frequencies 14,000 times with the sampling interval of $T = 400/f$. The sampling rate of the counter was approximately 6.6 kHz for f of 2.6 MHz. The DVM and the three counters were triggered by a signal generated by a computer and initiated by an optical switch comprising an LD and a PD.

Results and Discussions: Figure 12.3 shows the force measured by the force sensor, F_{trans}; the force measured by the optical interferometer-1, F_{mass}; and their difference $F_{diff} = F_{trans} - F_{mass}$; further, the electric response of the force sensor under an impact load can be observed. The difference between F_{trans} and F_{mass} was caused mainly by the difference between the static and the dynamic characteristics of the force sensor, and the root mean square (rms) value of F_{diff} during the collision, 0 ms < t < 14 ms, was approximately equal to 3.97 N. The LMM can be used to obtain this result because its performance is the best among available methods.

Figure 12.4 shows the second-order derivative of the force measured by the force sensor, d^2F_{trans}/dt^2, and F_{diff}. The solid line shows a regression line expressed by $F_{reg} = C(d^2F_{trans}/dt^2) = (-3.33 \times 10^{-7})(d^2F_{trans}/dt^2)$. The coefficient C is obtained by the gradient of the regression line determined by the least-square method.

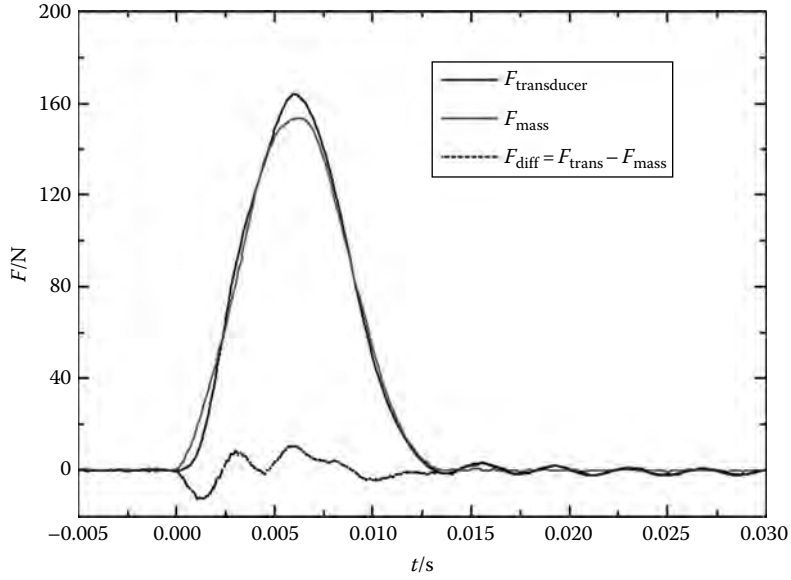


FIGURE 12.3 Force measured by the force sensor, F_{trans}; the force measured by the optical interferometer, F_{mass}; and their difference $F_{diff} = F_{trans} - F_{mass}$.

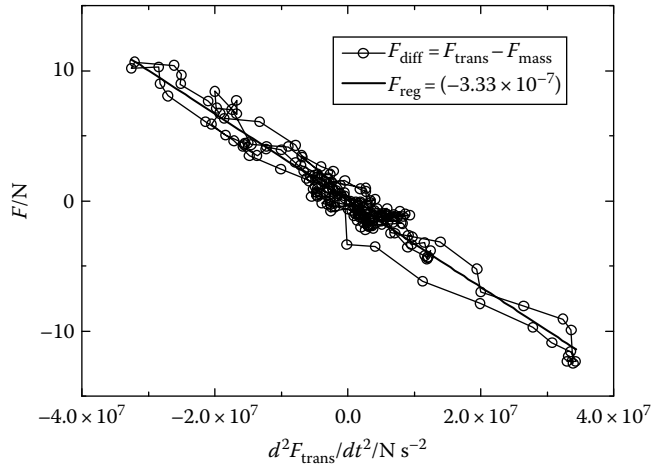


FIGURE 12.4 Relationship between the second-order derivative of the force measured by the force sensor, d^2F_{trans}/dt^2, and F_{diff}.

Figure 12.5 shows the changes in the difference F_{diff} between the forces measured by the force sensor and the optical interferometer and the force regressed by using F_{trans} and F_{reg}, against time. The measurement error F_{diff} is almost identical to the results of the regression analysis, F_{reg}. Therefore, F_{diff} can be represented by $F_{reg} = (-3.33 \times 10^{-7})$ (d^2F_{trans}/dt^2), and the force measured by the force sensor can be corrected using the output signal of the sensor itself by using the following equation:

$$F_{corrected} = F_{trans} - F_{reg} = F_{trans} - C\left(\frac{d^2F_{trans}}{dt^2}\right) \tag{12.6}$$

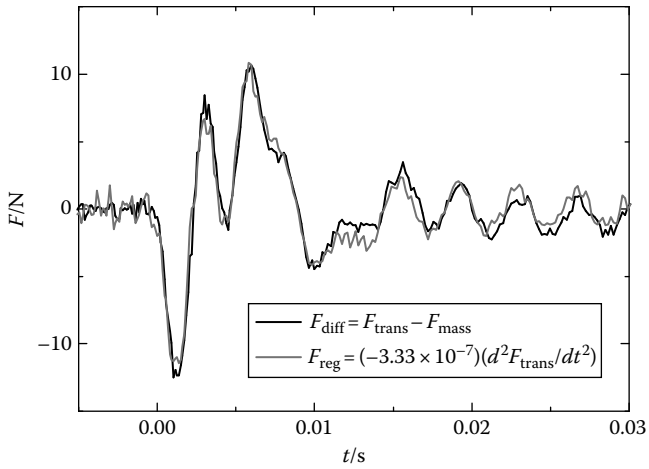


FIGURE 12.5 Changes in F_{diff} and F_{reg} against time.

This is the theory of the self-correction method. Once the coefficient C has been measured for a force sensor by using the LMM, the dynamic force measured by the force sensor can be corrected with output signal of the force sensor itself without the LMM setup.

Figure 12.6 shows the rms value of the errors of the measured force with and without self-correction; the results of the 10 impact tests are plotted. The regression coefficient of the single experiment (shown in Figure 12.4) is used for all the self-corrections. The rms value of the F_{diff} error drastically increases with the peak forces of F_{mass}, $F_{mass,\,max}$. The rms value of ($F_{corrected} - F_{mass}$) also increases with $F_{mass,\,max}$, but the increment is significantly smaller than that of the rms value of the error of F_{diff}. This result shows the efficiency of the self-correction method.

Figure 12.7 shows relationship between the second-order derivative of the force measured by the force sensor, d^2F_{trans}/dt^2, and the acceleration of the sensing point of the force sensor, a_2, measured by the optical interferometer-2. The plotted data are from the same experiment shown in Figure 12.3 through Figure 12.5. It is clear that a_2 is almost proportional to d^2F_{trans}/dt^2. The solid line shows a regression line expressed by $a_2 = (-1.03 \times 10^{-6})\ (d^2F_{trans}/dt^2)$. According to the operational principle of the force sensor,

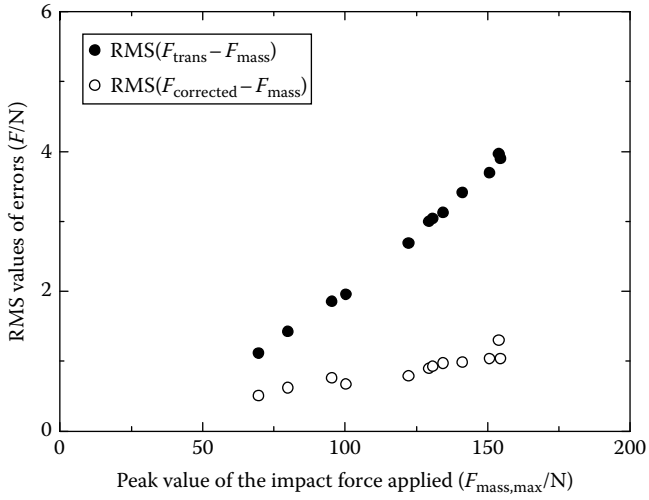


FIGURE 12.6 Rms values of the errors of measured force with and without self-correction.

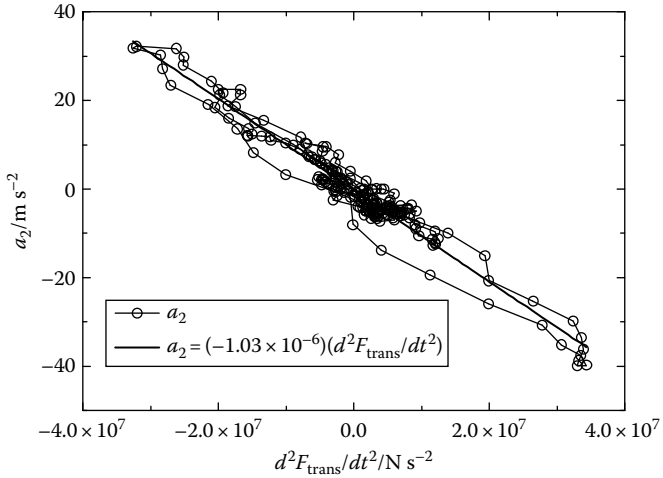


FIGURE 12.7 Relationship between d^2F_{trans}/dt^2 and the acceleration of the sensing point of the force sensor, a_2.

the force measured by the sensor is a product of the spring constant and displacement. Therefore, the acceleration of the sensing part of the force sensor can be obtained by the second time derivative of the F_{trans} as follows:

$$F_{trans} = -kx$$

$$-\frac{1}{k}\frac{d^2F_{trans}}{dt^2} = \frac{d^2x}{dt^2} = a_2$$

The slope of the regression line (-1.03×10^{-6}) is almost identical to $-1/k = -0.95 \times 10^{-6}$, where k is the nominal value of the spring constant of the force sensor. The relationship shown in Figure 12.7 is a validation of the aforementioned equation. Further, the aforementioned results prove that the tested force sensor can be modeled by a mass–spring model, and the primary reason for the dynamic error is the inertial force of the sensing part of the force sensor.

The other application of the LMM for calibrating force sensors is described in the sections as follows.

12.4 Impulse Response

Introduction: The impulse response refers to the response when the force is inputted as a Dirac delta function, and it is the most important dynamic characteristics for analyzing dynamic systems. Because the input of a Dirac delta function cannot be obtained experimentally, a very short pulse is substituted for the Dirac delta input. In our experiment, the FWHM of the pulse duration of approximately 1 ms and a maximum impact force of up to approximately 1 kN could be achieved by using our improved measurement method.

Experimental Setup: Figure 12.8 shows the schematic diagram of the experimental setup for evaluating the impulse responses of force sensors. This setup was roughly the same setup shown in Figure 12.2 except the lack of the second optical interferometer. In order not to prevent the sensor from fast motion, the external weight such as the CC and the metal plate were not attached to the sensor. Further, no damper was used to reduce the steepness of the impulse. The total mass of the moving part of the aerostatic linear bearing, M, is 4.5000 kg.

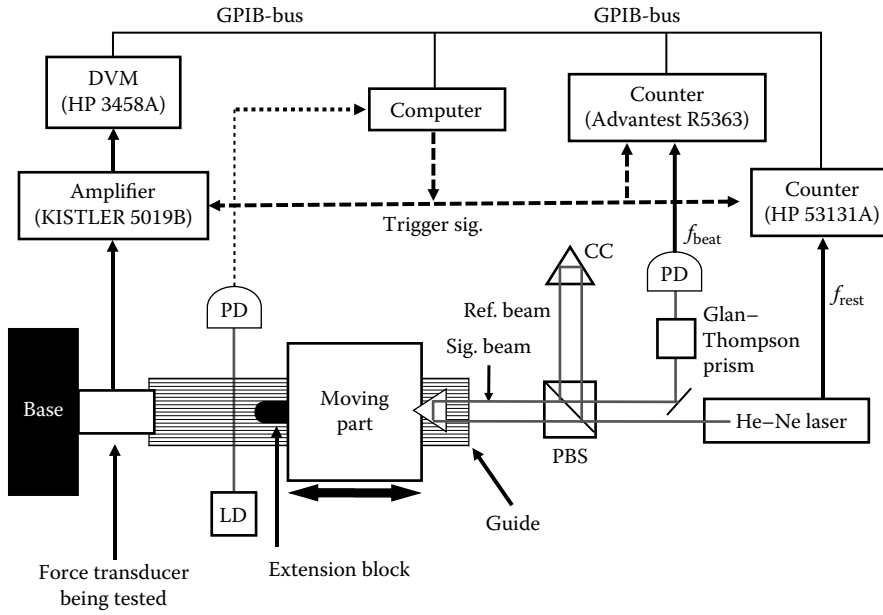


FIGURE 12.8 Experimental setup for evaluating impulse response. CC, corner cube; PBS, polarized beam splitter; PD, photodetector; LD, laser diode.

A piezoelectric force sensor (9213A, Kistler Instrument Corp.) was used for measuring the high-speed responses. The rated force capacity and the characteristic frequency of the sensor were 2.5 kN and approximately 200 kHz. The sensor was statically calibrated with the standard uncertainty of approximately 25 N, that is, 1% against the rated force capacity. The output signal of the sensor was measured by a digital multimeter (HP 3458A, Agilent Technologies Corp.) with the sampling rate of 50 kHz.

The instantaneous velocity of the moving part was measured by using an optical interferometer. A Zeeman-type two-frequency He–Ne laser was used as the light source, and the frequency difference between the two frequencies was $f_{rest} = |f_1 - f_2| = 2.6$ MHz. The velocity was measured using the Doppler shift frequency, $f_{Doppler}$, of the signal beam by using the following equation:

$$v = \lambda_{air} \frac{f_{Doppler}}{2} \tag{12.7}$$

$$f_{Doppler} = -(f_{beat} - f_{rest}) \tag{12.8}$$

where

λ_{air} is the wavelength of the signal beam

f_{beat} is the beat frequency detected by the PD

The directions of the coordinate system for the force, velocity, and acceleration on the moving part were toward the right (Figure 12.8). An electric frequency counter (R5363, Advantest Corp.) was used to measure and record the beat frequency. Another electric counter (HP53131A, Agilent Technologies Corp.) measured f_{rest} by using the trigger signal supplied by a photodiode embedded in the He–Ne laser.

In the experiment, the initial velocity of the moving part, v_1, changed to v_2 after the collision. The impulse responses of the force sensors were determined by comparing the impulse obtained by

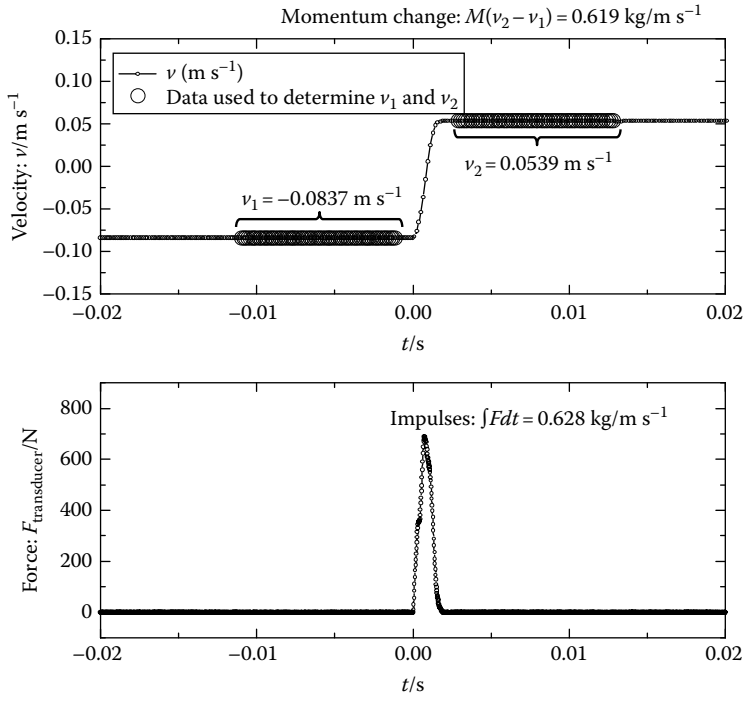


FIGURE 12.9 Changes in the velocity of the moving part and impact force detected by the force sensor.

integrating the output signal of the sensor, F_{sensor}, with the momentum changes of the moving part measured by the optical interferometer, $M(v_2 - v_1)$.

Results and Discussion: Figure 12.9 shows the velocity and force measured by the interferometer and force sensor. The velocities before and after the collision, v_1 and v_2, were measured as mean velocities during 10 ms before and after the collision. In this experiment, v_1 and v_2 were −0.0837 and 0.0539 m/s, respectively. The sample standard deviation of the velocities was $\mu_{v,1} = 4.0 \times 10^{-6}$ m/s and $\mu_{v,2} = 1.9 \times 10^{-5}$ m/s, respectively. The corresponding total velocity changes caused by the collision were 0.003% and 0.014%, respectively. These values are sufficiently smaller than the required uncertainty of 0.1%. The changes in momentum were approximately 0.619 kg m/s.

Figure 12.10 shows the instantaneous values of the impact force measured by the optical interferometer, $F_{inertial}$, and the force sensor, F_{sensor}. $F_{inertial}$ is calculated by using the following equation:

$$F_{inertial} = M\frac{d}{dt}(-f_{beat} + f_{rest}) \tag{12.9}$$

Although the time resolution of $F_{inertial}$ is not sufficient to determine the difference between $F_{inertial}$ and F_{sensor}, the response of F_{sensor} seems to overshoot near the peak of the pulse. The relative difference between the impulse measured by the sensor and the momentum change measured by the interferometer

$$\frac{\int F_{transducer} dt}{M(v_2 - v_1)} - 1 \tag{12.10}$$

is approximately 0.015. This error is mainly from the overshoot of F_{sensor}.

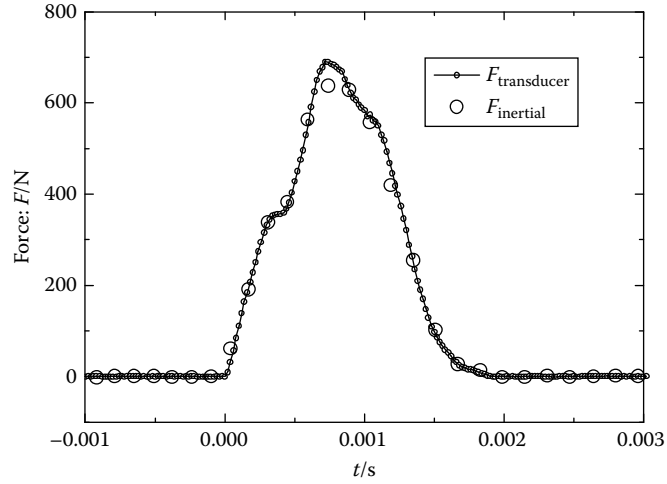


FIGURE 12.10 Instantaneous value of force measured by optical interferometer and force sensor.

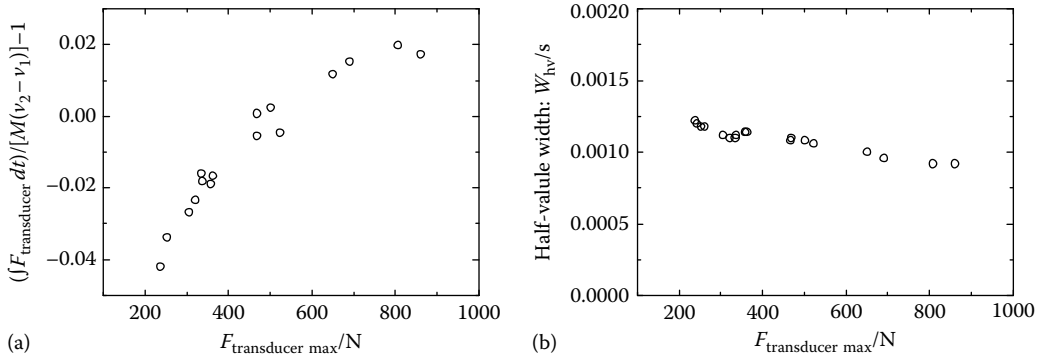


FIGURE 12.11 (a) Relative difference between impulse and change in momentum against the maximum force and (b) FWHM of pulse duration against the maximum force.

Figure 12.11a shows the relative difference between the impulse and the changes in momentum against the peak value of the force measured by the sensor, $F_{\text{sensor\_max}}$. This figure shows a clear relationship between the relative difference and $F_{\text{sensor\_max}}$.

Figure 12.11b shows the pulse duration in FWHM against $F_{\text{sensor\_max}}$. From the results, the pulse width can be estimated at approximately 1 ms in FWHM.

12.5 Oscillation Response

The continuous oscillation response of force sensor is also evaluated by using LMM.

Experimental Setup: Figure 12.12 shows a schematic diagram of the experimental setup for evaluating the oscillation responses of a force sensor. The components of the setup were almost identical to those for evaluating the impulse response. The major difference was that the rigid body—a moving part of an aerostatic linear bearing—was connected to the side of the tested force sensor with a steel spring. The other side of the sensor was fixed to the base. The moving part, the force sensor, and the spring comprised a mass–spring system. The inertial force of the moving part was used as the standard oscillation force and was compared to the output signal of the force sensor when the mass–spring system

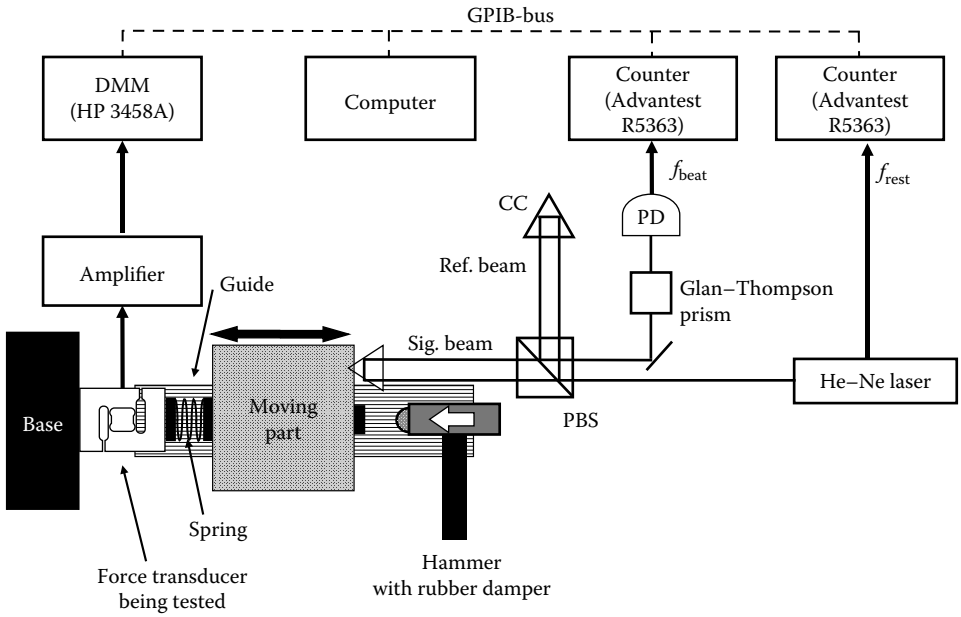


FIGURE 12.12 Experimental setup for evaluating oscillation response. CC, corner cube; PBS, polarized beam splitter; PD, photodetector.

was continuously oscillating. The oscillation was excited by hitting the moving part with a hammer. The inertial force acting on the moving part was measured accurately by measuring the velocity of the moving part with an optical interferometer.

The force acting on the moving part from the sensor, *F*, was equal to the inertial force of the moving part according to the principle of action and reaction if other forces, such as the inertial force of the spring and the friction of the aerostatic bearing, were ignored. In the case of the inertial force of the spring, half the mass of the spring was considered to be the moving part of the bearing, and the other half was considered to be the sensing part of the force sensor; the spring itself was treated as having uniform composition. The mass of the spring was approximately 0.0257 kg. The total mass of the moving part including the half part of the spring was approximately 4.6126 kg.

In this experiment, a conventional S-shaped sensor was used for the test. The rated force capacity was 200 N, and it was statically calibrated with the standard uncertainties of approximately 0.2 N, which corresponded to 0.1% of the rated force capacity. The output signal of the force sensor was measured by using a digital multimeter (HP 3458A, Agilent Technologies Corp.) with a sampling rate of 5 kHz.

The inertial force was measured by using an optical interferometer similar to the procedure described as follows. One of the two-frequency laser lights was modulated by the Doppler effects, and the Doppler frequency was calculated as $-(f_{beat} - f_{rest})$. The frequency difference between the signal beam and reference beam, f_{beat}, and the fundamental frequency difference, f_{rest}, were measured by employing counters (R5363, Advantest). The sample number and the sampling frequency of the counter were 14 k samples and *f*/400 Hz, where *f* is the instantaneous frequency for the measurement. HP3458A and R5363 were triggered by an external synchronizing signal. An aerostatic linear bearing was used so that the rigid body, which is the moving part, can move linearly with less friction.

Results: Figure 12.13 shows the changes in frequency, velocity, acceleration, and force for a set of measurements. At $t = 0.0$ s, an impulse was manually applied to the moving part with a hammer. The frequency of f_{beat} changed immediately. After the impact, f_{beat} began to oscillate at a frequency of approximately 15 Hz. By using the frequency measured from f_{beat}, the velocity, acceleration, and inertial force could be calculated.

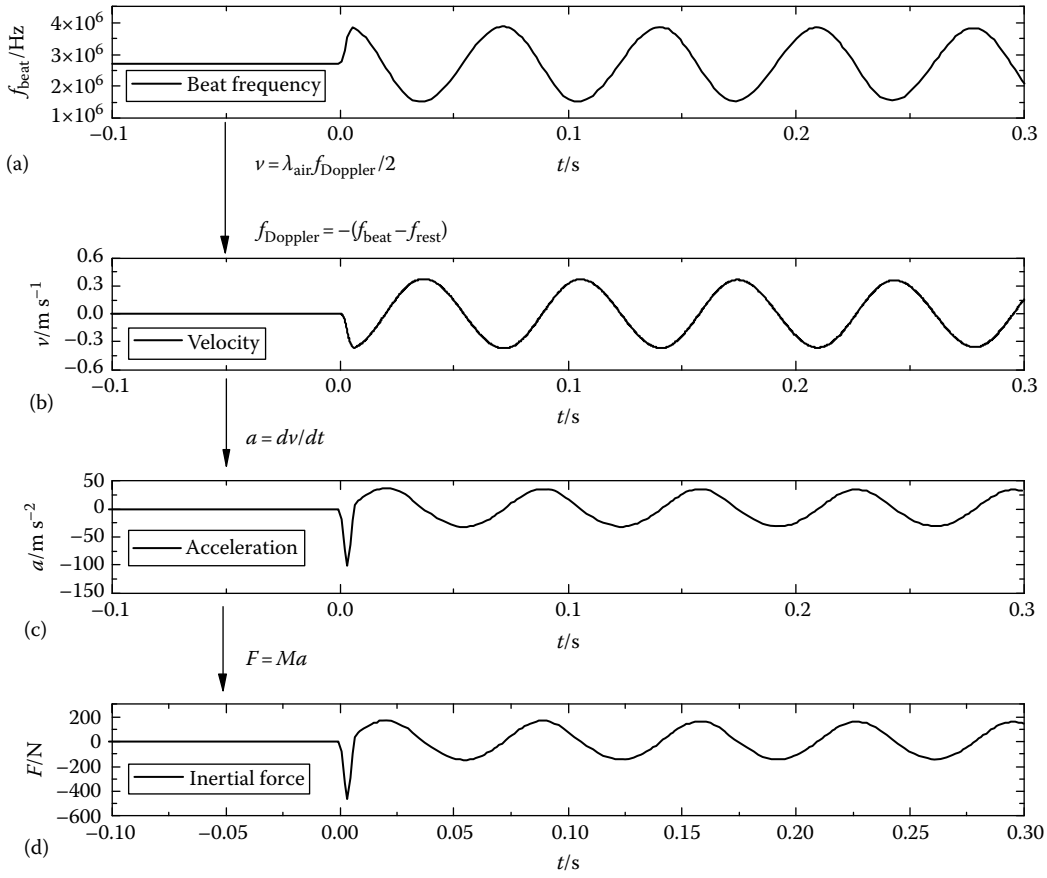


FIGURE 12.13 (a) Frequency, (b) velocity, (c) acceleration, and (d) force measured during oscillation.

Figure 12.14 shows the forces measured by the interferometer, $F_{inertial}$, as shown in Figure 12.13, and by the force sensor, F_{sensor}.

Figure 12.15 shows the details of the same data shown in Figure 12.14 in a different manner. In the first half wave, just after the impulse was manually applied by the hammer, a high-frequency oscillation was superimposed on an oscillation with a frequency of approximately 15 Hz, probably as a result of the vibration of the optical interferometer. The impact force applied by the hammer has a vertical component against the moving direction of the aerostatic linear bearing, and the force including the vertical component is propagated along the base plate on which the optical interferometer is constructed; this force causes the vibration of the optics.

To evaluate the responses of the sensor against the oscillation force, the impulse, that is, the time integration of the oscillation force during half a period, is introduced. "Half a period" is determined as the period from the time at which the oscillation force $f_{inertial}$ crosses the coordinate axis to the time at which $f_{inertial}$ again crosses the coordinate axis in the reverse direction. As schematically shown in Figure 12.15, consecutive numbers have been assigned to all the impulses except the first one because the first impulse includes a sharp impulse applied by the hammer.

Figure 12.16 shows the relative difference between the impulse measured by the sensor, $\int F_{transducer} dt$, and the impulse measured by the optical interferometer, $\int F_{inertial} dt$, against the peak value of $F_{inertial}$ in

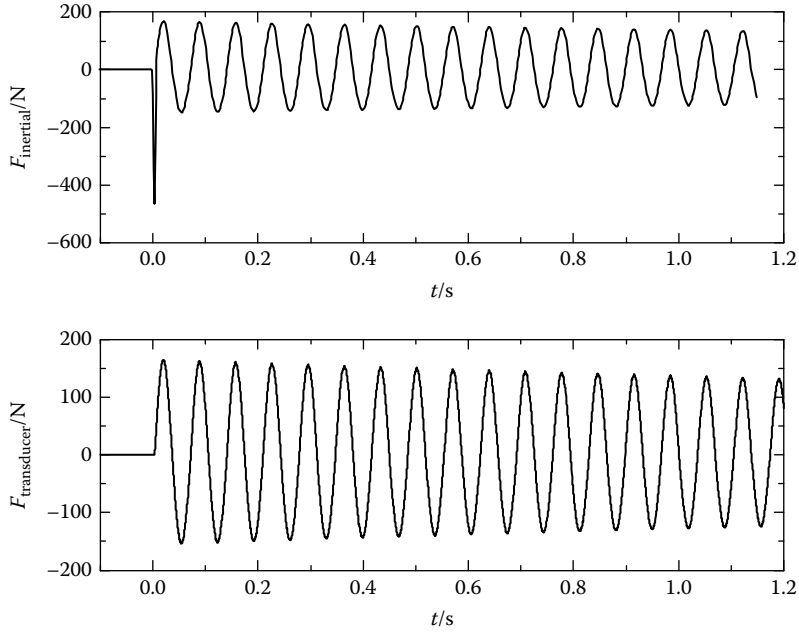


FIGURE 12.14 Oscillating forces measured by the interferometer, $F_{inertial}$, and by the force sensor, F_{sensor}.

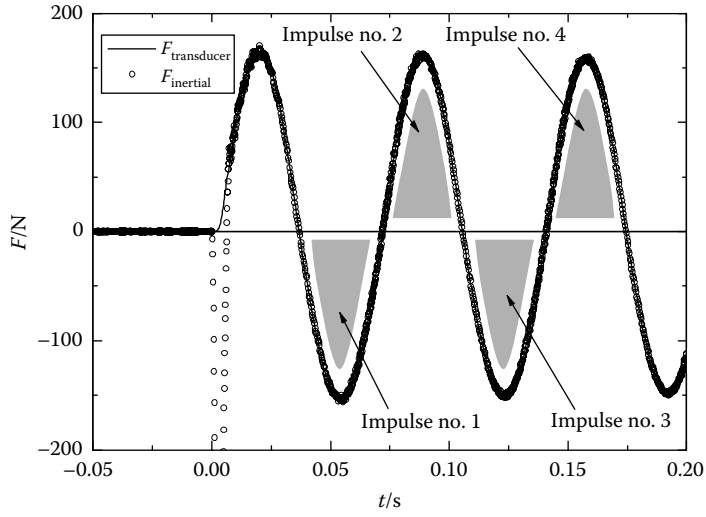


FIGURE 12.15 Changes in $F_{inertial}$ and F_{sensor}.

each sample. In the figure, the results of the 10 impact tests and 30 impulses for each test are plotted. The filled black circles show the data for the first three impulses of each impact test, and the circles show all the other data.

Discussion: As shown in Figures 12.13 through 12.15, because of the asymmetry of the spring constant of the spring, the shapes of the time series of the velocity, acceleration, and force are not perfectly symmetrical. Gradually damping is also observed. In this experiment, only the amplitude of the oscillation could be changed by changing the initial impulse applied by the hammer. For more accurate sensor

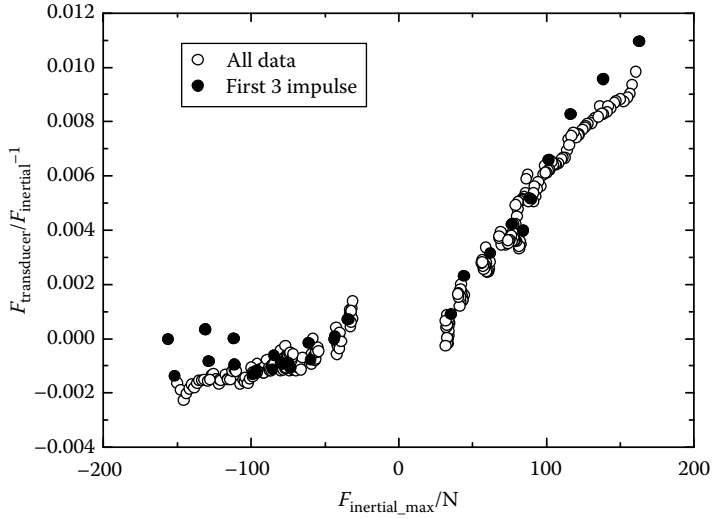


FIGURE 12.16 Relative differences between the impulse measured by the sensor and the impulse measured by the optical interferometer against the peak value of $F_{inertial}$ in each sample.

calibrations, more accurate amplitude controls are required, for example, using a linear actuator to impart a certain amount of the momentum on the hammer. Further, in order to change the oscillation frequency, an elastic body with an appropriate elastic modulus is required. A piezoelectric body, which can be controlled as if it is a perfect elastic body with a certain spring constant, could be suitable for realizing oscillations with a definite frequency and amplitude. However, in general, the control routine may become more sophisticated, and the number of uncertainty sources may increase.

In Figure 12.16, a clear relationship is observed in the case of the relative difference between the impulse measured by the sensor and those measured by the optical interferometer, and this indicates the validity and reproducibility of the method. Although the relative standard uncertainty of the statically calibrated sensor is approximately 0.1%, the relative difference changes by more than 1% along with the peak value of the force. In each test, the tendencies are different for the beginning part and the other parts, and these differences show that the first few impulses are in the transient state.

12.6 Conclusions

In this chapter, we have reviewed a calibration method for force sensors by using the LMM. The dynamic force acting on the force sensor was accurately measured and compared with the force measured by the sensor itself. This calibration method can be applied under other force conditions such as steep impulse and oscillating forces.

References

1. R. Kumme, Investigation of the comparison method for the dynamic calibration of force transducers, *Measurement* 23, 239–245, 1998.
2. R. Kumme and M. Dixon, The results of comparisons between two different dynamic force measurement system, *Measurement* 10, 140–144, 1992.
3. Y.-K. Park, R. Kumme, and D.-I. Kang, Dynamic investigation of a three-component force-moment sensor, *Measurement Science and Technology* 13, 654–659, 2002.
4. Y.-K. Park, R. Kumme, and D.-I. Kang, Dynamic investigation of a binocular six-component force-moment sensor, *Measurement Science and Technology* 13, 1311–1318, 2002.

5. Y. F. Li and X. B. Chen, On the dynamic behavior of a force/torque sensor for robots, *IEEE Transactions on Instrumentation and Measurement* 47(1), 304–308, 1998.
6. K. J. Xu and L. Jia, One-stage identification algorithm and two-step compensation method of Hammerstein model with application to wrist force sensor, *Review of Scientific Instruments* 73(4), 1949–1955, 2002.
7. Y. Fujii, Measurement of steep impulse response of a force transducer, *Measurement Science and Technology* 14, 65–69, 2003.
8. Y. Fujii, A method for calibrating force transducers against oscillation force, *Measurement Science and Technology* 14, 1259–1264, 2003.
9. Y. Fujii, Proposal for a step response evaluation method for force transducers, *Measurement Science and Technology* 14, 1741–1746, 2003.
10. Y. Fujii, Dynamic force calibration method for force transducers, *IEEE Transactions on Instrumentation and Measurement* 58, 2358–2364, 2009.
11. Y. Fujii, Toward dynamic force calibration, *Measurement* 42, 1039–1044, 2009.
12. Y. Fujii and T. Yamaguchi, Method for evaluating material viscoelasticity, *Review of Scientific Instruments* 75, 119–123, 2004.
13. Y. Fujii and T. Yamaguchi, Proposal for material viscoelasticity evaluation method under impact load, *Journal of Materials Science* 40, 4785–4790, 2005.
14. Y. Fujii and D. W. Shu, Impact force measurement of an actuator arm of a hard disk drive, *International Journal of Impact Engineering* 35, 98–108, 2008.
15. Y. Fujii and T. Yamaguchi, Optical method for evaluating material friction, *Measurement Science and Technology* 15, 1971–1976, 2004.
16. Y. Fujii, Method for measuring transient friction coefficients for rubber wiper blades on glass surface, *Tribology International* 41, 17–23, 2008.
17. Y. Fujii, T. Yamaguchi, and J. Valera, Impact response measurement of a human arm, *Experimental Techniques* 30, 64–68, 2006.
18. Y. Fujii and T. Yamaguchi, Method of evaluating the force controllability of human finger, *IEEE Transactions on Instrumentation and Measurement* 55, 1235–1241, 2006.
19. Y. Fujii, K. Maru, and T. Jin, Method for evaluating the electrical and mechanical characteristics of a voice coil actuator, *Precision Engineering* 34, 802–806, 2010.
20. Y. Fujii and K. Shimada, Instrument for measuring the mass of an astronaut, *Measurement Science and Technology* 17, 2705–2710, 2006.
21. Y. Fujii and K. Shimada, The space scale: An instrument for astronaut mass measurement, *Transactions of the Japan Society for Aeronautical and Space Sciences* 50, 251–257, 2008.
22. Y. Fujii, K. Shimada, M. Yokota, S. Hashimoto, Y. Sugita, and H. Ito, Mass measuring instrument for use under microgravity conditions, *Review of Scientific Instruments* 79, 056105, 2008.
23. Y. Fujii, K. Shimada, and K. Maru, Instrument for measuring the body mass of astronauts under microgravity conditions, *Microgravity Science and Technology* 22, 115–121, 2010.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

II

Spatial Variables

13 **Thickness Measurement** *John C. Brasunas, G. Mark Cushman, and Brook Lakew* **13**-1
Descriptions of the Relatively Mature Measuring Techniques • Future Directions in Thickness Measurement • References • Bibliography

14 **Distance Measurement** *W. John Ballantyne* **14**-1
Basic Distinctions between Range Measurement Techniques • Performance Limits of Ranging Systems • Selected Examples of Ranging, Range-Imaging, and Motion-Tracking Systems • A Sampling of Commercial Ranging, Range-Imaging, and Motion-Tracking Products • References

15 **Altitude Measurement** *Dimitris E. Manolakis* **15**-1
Pressure Altitude Measurement • Radar Altimeter • Height Measurement • Ground-Based Height Estimation • Onboard–Derived Height Estimation • Special Topics • References

16 **Attitude Measurement** *Mark A. Stedham, Partha P. Banerjee, Seiji Nishifuji, and Shogo Tanaka* **16**-1
Introduction to Attitude Sensors for Ships, Aircraft, and Crane Lifters • Attitude Sensors for Spacecraft Applications • Automatic On-Line Attitude Measurement for Ships and Crane Lifters • Attitude Measurement for Ships • Attitude Measurement for Crane Lifters • Aircraft Attitude Determination • Spacecraft Attitude Determination • PALADS • References

17 **Inertial Navigation** *Halit Eren* **17**-1
Principles of Inertial Navigation • Gyroscopes • Accelerometers • Errors and Stabilization in Inertial Navigation Systems • Vehicular Inertial Navigation • Applications of Inertial Navigation • Bibliography • Partial List of Manufacturers/Suppliers

18 **Level Measurement** *Detlef Brumbi* **18**-1
Measurements Using the Effects of Density • Time-of-Flight Measurements • Commonly Used Evaluation Methods • Level Measurements by Detecting Physical Properties • Level Measuring Instruments • References • Further Readings • Partial List of Manufacturers and Suppliers

19 **Area Measurement** *Charles B. Coulbourn and Wolfgang P. Buerner* **19**-1
Theory • Equipment and Experiment • Evaluation • Defining Terms • References • Further Information

20 **Volume Measurement** *René G. Aarnink and Hessel Wijkstra* **20**-1
Plethysmography Theory • Numerical Integration with Imaging • Indicator Dilution Methods • Water Displacement Volumetry • Equipment and Experiments • Evaluation • References • Further Information • Partial List of Manufacturers and Suppliers

21 Tilt Measurement *Adam Chrzanowski and James M. Secord*.....**21**-1
Tiltmeters or Inclinometers • Geodetic Leveling • Hydrostatic Leveling • Suspended and Inverted Plumb Lines • Integration of Observations • References • Further Readings • Partial List of Vendors and Suppliers

22 Proximity Sensing for Robotics *Ricardo E. Saad, Ben Benhabib, A. Bonen, and K.C. Smith* **22**-1
Proximity Definition • Typical Sensor Characteristics • Technologies for Proximity Sensing • Acknowledgments • References • Further Readings • Partial List of Vendors and Suppliers

13

Thickness Measurement

John C. Brasunas
Goddard Space Flight Center

G. Mark Cushman
Goddard Space Flight Center

Brook Lakew
Goddard Space Flight Center

13.1 Descriptions of the Relatively Mature Measuring Techniques **13**-2
Mechanical • Electronic Gages • Pneumatic Gaging • Optical: Focusing, Shadowing, and Comparing • Weighing • Capacitive Gaging • Inductive Gaging (Eddy-Current Sensing) • Magnetic Induction • Hall Effect Gage • Far Field/Time of Flight: Ultrasound, Radar, and Lidar • Far Field/Resonance: Ultrasound, Interferometry, and Ellipsometry • Far Field/Absorption, Scattering, and Emission • Destructive Techniques
13.2 Future Directions in Thickness Measurement **13**-8
Concerning Techniques Mentioned Earlier • THz Technology • Nanoscale Scanning Probe Microscopy
References **13**-10
Bibliography **13**-10

One can measure thickness on many scales. The galaxy is a spiral disk about 100 Em (10^{20} m) thick. The solar system is pancake-like, about 1 Tm (10^{12} m) thick. The rings of Saturn are about 10 km thick. Closer to home, the Earth's atmosphere is a spherical shell about 40 km thick; the weather occurs in the troposphere, about 12 km thick. The outermost shell of the solid Earth is the crust, about 35 km thick. The ocean has a mean depth of 3.9 km. In the Antarctic, the recently discovered objects believed to be microfossils indicative of ancient Martian life are less than 100 nm thick. In terms of the man-made environment, industry must contend with thickness varying from meters, for construction projects, to millimeters on assembly lines and to micrometers and nanometers for the solid-state, optical, and coating industries. Perhaps the most familiar way of measuring thickness is by mechanical means, such as by ruler or caliper. Other means are sometimes called for, either because both sides of an object are not accessible; the dimension is either too big or too small for calipers; the object is too fragile, too hot, or too cold for direct contact; or the object is in motion on an assembly line—it may not even be a solid. Thickness may also be a function of position, as either the object may have originally been made with nonuniform thickness, deliberately or not, or the thickness may have become nonuniform with time due to either corrosion, cracking, or some other deterioration. The thickness may also be changing with time due to deliberate growth or etching, as example for thin films. Thus, it follows that, in more general terms, measuring thickness might require measuring the topography or height profile of two surfaces and taking the difference. Alternatively, the measurement technique may produce a reading directly related to the difference. Table 13.1 lists some of the many techniques suited to determining thickness, together with the range of usefulness and some comments on accuracy and/or precision.

TABLE 13.1 Thickness Measuring Techniques

Technique	Range	Comments
Mechanical		
Caliper gage, micrometer	1 μm–100 mm	±3 μm accuracy
Electronic gages: LVDT	0–1 m	Precision depends on noise level
Pneumatic gaging	50 nm minimum	
Optical/focusing, shadowing, comparing		
Microscope	5 μm minimum	About 1% accuracy
Comparators/projectors	25–250 nm	
Laser caliper	100 μm–100 mm	Precision of 6 μm or better
Weighing	Range depends on area	
Capacitive gaging	From <1 μm to about 1 cm	
Inductive gaging (eddy-current sensing)	0–1.5 mm	Precision of 2.5 μm
Magnetic induction	0–4 mm	10% accuracy
Hall effect gage	0–10 mm	1%–3% accuracy
Far field/time of flight		
Sonar/ultrasound	0.5–250 mm	25 μm accuracy
Radar	0.1 to few hundred km	
Lidar	10 m–5 km	
THz technology		
Far field/resonance		
Resonant ultrasound		
Interferometry: spectral and spatial	1 nm–100 μm	Accuracy about $\lambda/50$
Ellipsometry	0.3 nm–10 μm	0.1 nm accuracy
Far field/absorption, scattering, emission		
Gamma-ray backscatter	Range to 25 mm	0.5% precision
Beta-transmission	2 μm–1 mm	0.2% precision
Beta-backscatter	100 nm–50 μm	3%–20% precision
X-ray fluorescence	0–30 μm	
Infrared absorption	Depends on material	
Scanning techniques: scanning probe microscopy		Precision better than 0.1 nm
Destructive techniques: electrolytic	15 nm–50 μm	

13.1 Descriptions of the Relatively Mature Measuring Techniques

The following descriptions will also refer to some of the relevant vendors, whose addresses are found in Table 13.2. Additional vendor information, with specific price or model number identification, is found in Table 13.3. The words "gage" and "gauge" are used interchangeably.

13.1.1 Mechanical

The fundamental tool for measuring thickness is the line-graduated instrument [1,2]. It is the only mechanical means to make direct measurements. Graduated spacing that represents known distances is used as direct comparisons to the unknown distance. Instruments include bars, rules, and tapes generically called rulers; caliper gages, which employ a positive contact device for improved alignment of the distance boundaries; and micrometers, which typically have greater precision due to a combination of

TABLE 13.2 Vendor Addresses

Vendor	Address
Bomem	Quebec, Canada
Brown & Sharpe	North Kingstown, RI
CMI International	Elk Grove Village, IL
Conductus	Sunnyvale, CA
deFelsko	Ogdensburg, NY
Digilab	Cambridge, MA
Digital Instruments	Santa Barbara, CA
Electromatic	Cedarhurst, NY
Fischer	Windsor, CT
Hewlett Packard	Englewood, CO
Kta-Tator	Pittsburgh, PA
Magnetic Analysis Corp.	Mount Vernon, NY
Mattson	Madison, WI
Measurex	Cupertino, CA
Micro Photonics	Allentown, PA
Midac	Costa Mesa, CA
Mitutoyo	Plymouth, MI
Moore Products Co.	Spring House, PA
NDC Systems	Irwindale, CA
Nicolet	Madison, WI
Ono Sokki	Addison, IL
Oxford Instruments	Concord, MA
Panametrics	Waltham, MA
Park Scientific Instruments	Sunnyvale, CA
Penny + Giles	Attleboro, MA
Perkin-Elmer	Norwalk, CT
Phase-Shift Technology	Tucson, AZ
Rudolf Instruments	Fairfield, NJ
Scantron	Dist. by Micro Photonics
Schaevitz	Pennsauken, NJ
Sentech	Dist. by Micro Photonics
SolveTech	Claymont, DE
Starrett	Athol, MA
Stresstel	Scotts Valley, CA
Transicoil Inc.	Valley Forge, PA
Trans-Tek	Ellington, CT
Willrich Precision Instrument Co.	Cresskill, NJ
J.A. Woollam Co., Inc.	Lincoln, NE
Wyko	Tucson, AZ
Zygo	Middlefield, CT

linear and circumferential scales. Caliper precision can be improved with Vernier scales or linear transducers. Fixed gages are often used to measure objects on a pass/fail basis. An object of fixed geometry (length, tapered bore, thread, etc.) is compared to a test piece typically for part inspection. Variations include the master gage, an object used to represent the nominal dimension of the part; the limit gage, an object used to represent the limit condition for tolerance dimensioning; and gage blocks or Johansson blocks, an object of fixed length used as a dimensional reference standard. Dial indicators are used to

TABLE 13.3 Instruments for Measuring Thickness

Manufacturer	Model Number	Price	Description
KTa-Tator	TI-12	$1,595	General-purpose ultrasonic gage, 0.75–75 mm range
NDC Systems	6100TC	$49,300	Backscatter gamma-gage for 60 in. web, 25 mm range
NDC Systems		$66,600	Transmission beta-gage, for continuous web products
Panametrics	25DL	$2,200–$3,800	Single-element ultrasonic gage, 50 mm range
Panametrics	26DL Plus	$1,400–$2,500	Dual-element ultrasonic gage, 250 mm range
Panametrics	8000	$6,500	Hall effect magnetic gage, for nonferrous materials, 6 mm range
DeFelsko	Positest 1000-N	$1,995	Eddy-current sensor, Apple Newton readout, measure out to 1.5 mm nonferrous, nonconducting coating on conducting substrate
Magnetic Analysis	Various	$1,500–$100,000	Ultrasonic, time-of-flight gages
Fischer	Deltascope MP2C	$1,200	Magnetic induction gage, measure nonmagnetic coating on ferromagnetic substrate
Fischer	IsoScope MP1C	$1,200	Eddy-current gage, measure nonconducting coating on nonferrous conducting substrate
Fischer	Fischerscope MMS	$6,500	Beta-backscatter system to measure coating thickness
Fischer	Fischerscope X-Ray	$34,000	X-ray fluorescence system to measure coating
	1020 video		Thickness
Fischer	Couloscope Sx	$2,500+ accessories	Electrolytic, destructive system to measure coating thickness
J.A. Woollam Co. Inc.	M-44	Application specific	Variable angle, multiwavelength spectroscopic ellipsometer
Rudolf Instruments	431A31WL633	$10,100	Manual, He–Ne wavelength ellipsometer
Rudolf Instruments	444A12	$34,000	Automatic, He–Ne wavelength ellipsometer
Hewlett Packard	HP8712C	$13,500	RF vector network analyzer, measure transmission/reflection frequency response to 1.3 GHz, optional to 3 GHz
Stresstel	T-Mike Programmable	$995	Dual-element ultrasonic system
Stresstel	TM1D	$1,795	Single-element ultrasonic system
Measurex	DMC480	Application specific	High-speed x-ray thickness gage
Bomem	MB series	$20,000 and up	1 cm^{-1} resolution FTR
Park Scientific Instrumens	Autoprobe CP	$65,000	Ambient SPM
Park Scientific Instruments	Autoprobe VP2	$130,000	UHV SPM
Digital Instruments	Nanoscope IIIa/D3000	$90,000	Small-sample SPM

sense displacement from a reference plane and display the deviation thereof. The display can be electronically coupled for amplification and/or display purposes. The range of a measuring instrument may be extended if multiple copies of the object to be measured are available. For example, the thickness of a sheet of paper may be measured by a simple ruler if 500 sheets of paper are stacked (vendors: Brown & Sharpe, Starrett, and Mitutoyo; also see [3]).

13.1.2 Electronic Gages

A linear variable differential transformer (LVDT) utilizes multiple toroidal transformers to sense axial displacement of an iron core that is attached to a measuring contact, either directly or by another joint (such as a lever). The displacement has a direct correlation to the distance that other electronics display. Thus, the LVDT serves as a replacement for a lined ruler or micrometer, incorporating an electrical readout (vendors: Penny + Giles, Schaevitz, Transicoil Inc., and Trans-Tek).

13.1.3 Pneumatic Gaging

Pneumatic gages have pressurized air exiting gage orifices. The air velocity differential or back pressure is a function of the separation of the gage and the part. In the direct or open jet method, the pressurized air experiences back pressure due to the impedances posed by the measured part. The typical scenario is that the gage head and the measured part have similar geometry (i.e., a cylindrical gage in a bored hole). By placing two gages on either side of a flat plate, the thickness may be inferred. In the indirect or contact method, the pressurized air pushes on a contact piece that directly contacts the part. Tolerances as small as 50 nm can be measured (vendors: Willrich Precision Instrument Co. and Moore Products Co.).

13.1.4 Optical: Focusing, Shadowing, and Comparing

This includes microscopes, which can determine thickness either by comparison with a known reference or by focusing on the front and rear surfaces of a sample, noting the difference in focus position. Comparators project onto a screen what might be noted through a microscope. Laser calipers retrieve dimensions by measuring the shadowing of a laser beam (vendors: NDC Systems for laser caliper and Scantron for laser profilometer).

13.1.5 Weighing

Given a plate of material with known density, first, measure the area with some type of calibrated video system. Then, a measurement of weight can be simply converted to an estimate of the thickness. As is common with this technique and most of the following techniques, estimating the thickness requires knowledge of some other property of the material to be measured—in this case, the density.

13.1.6 Capacitive Gaging

Capacitive gaging is realized by inserting a nonmetallic material into a known electric field. Knowing the gage sensor area and the material's dielectric constant, the thickness can be determined. Submicron thickness levels can be achieved (vendors: Ono Sokki and SolveTech).

13.1.7 Inductive Gaging (Eddy-Current Sensing)

The principle here is that AC currents in a coil induce eddy currents in a nearby conducting plate [4,5]. These eddy currents can be sensed by a pickup coil, which may be the exciting coil or a second coil. The presence of the eddy currents manifests itself as a modification of the apparent inductance and/or

the loss of the pickup coil. This technique is appropriate for nonferrous metals and is especially sensitive to thickness variations due to flaws such as cracks or corrosion. There is one particular instance in which it is common to measure thickness rather than variations. That would be the thickness of a nonconducting coating on a nonferrous conducting substrate. The coating thickness creates a gap (liftoff) between the exciting coil and the eddy currents, thereby affecting the eddy-current signal. The range of this technique would be about 1 mm. Fischer has an instrument designed for measuring the thickness of a newly laid road surface coating to a depth of 40 cm, by burying a conductive plate below the road (vendors: Fischer, deFelsko, CMI International).

13.1.8 Magnetic Induction

This technique is also used to measure coating thickness, in this case a nonmagnetic coating on a ferromagnetic substrate. The nonmagnetic coating creates a gap (liftoff) between the ferromagnetic substrate and a probe. One way to measure the gap and thereby the thickness is by measuring the force required to pull away a magnetic probe. Another technique would be to magnetically couple the ferromagnetic substrate to a transformer core, with a gap between the substrate and the core. This technique would have a range of about 4 mm. CMI International has an informative brochure describing the relative merits of measuring coating thickness via eddy current, magnetic induction, beta-backscatter, microresistance, and x-ray fluorescence; the choice of technique depends, among other things, on the material to be tested (vendors: Fischer, CMI International, Electromatic, and deFelsko).

13.1.9 Hall Effect Gage

This sensor measures the thickness of nonferrous materials with 1% accuracy by sandwiching the material being measured between a magnetic probe on one side and a small target steel ball on the other side [6]. It measures up to 10 mm. The Hall effect sensor is used to measure the magnetic field, as a DC measurement; the AC Hall effect measurements can be made more precisely because they eliminate bias and are done with less noise (vendor: Panametrics).

13.1.10 Far Field/Time of Flight: Ultrasound, Radar, and Lidar

Using the 1940s sonar principles and today's microprocessor technology, high-frequency (1–20 MHz) ultrasound waves can be used to measure thickness by sending pulsed sound waves through a material and measuring the transit time of the reflected signal [5,7]. Knowing the sound velocity of the material, the materials from 0.5 to 250 mm can be measured, often as fine as 25 μm. Media include metal, glass, ceramic, liquid, rubber, fiberglass, plastic, and concrete. Ultrasound can also be used to measure living tissues, as is often done in the agricultural and medical fields. The fat layers of cattle and pigs can influence marketability. Skin burn depths can direct treatment procedures. The depth of foreign objects in the body is useful for microsurgery. Ultrasonic thickness determination has expanded to include multidimensional echolocation applications, such as imagery and acoustic microscopes that can resolve in the submicron level. The principles behind ultrasound also apply to electromagnetic waves. In the gigahertz range, this is called radar. Radar can be used to estimate the thickness of atmospheric layers such as cloud layers. The light-wave version of radar, called lidar, can be used to measure the thickness of water vapor layers in the lower atmosphere (ultrasonic vendors: KTa-Tator, deFelsko, Stresstel, Magnetic Analysis Corp., Panametrics, and Electromatic).

13.1.11 Far Field/Resonance: Ultrasound, Interferometry, and Ellipsometry

The idea here is that when waves such as ultrasound impinge on a plane-parallel slab of material, there will be reflected power from both the front and the rear surfaces; depending on whether the

slab thickness is an odd or even number of quarter wavelengths, the reflected beams will be in constructive or destructive interference. If the frequency is swept, the distance in frequency between successive maxima and minima may be related to the slab thickness, if the index of refraction is known. Since the natural, or resonant, modes of an object depend on the properties and dimensions of an object, the knowledge of the properties enables estimation of dimensions from the resonant frequencies. Compared with time-of-flight ultrasound, resonant ultrasound is much less common. It has been used to characterize concrete and is quite sensitive to flaws, as anyone who has heard a cracked bell would know. Resonant techniques are much more common with visible [8], infrared, or microwave [4] radiation. Spectral interferometry would be appropriate to characterize the thickness of transparent substrates with reasonably flat surfaces, sufficiently parallel to one another. A common way to do this would be to measure a transmission spectrum with a spectrometer such as a Fourier transform spectrometer (FTS). The successive maxima and minima are here called the Fabry–Perot effect, and their appearance in a spectrum is called channeling. Thickness can also be measured with spatial interferometry, which is essentially a way of measuring surface topography. An example would be the phenomenon of Newton's rings, which occur when the surface to be tested is in contact with an optical flat. Using a transparent optical flat, transmit monochromatic light such as a mercury lamp through the flat and onto the interface between the flat and the test surfaces. If there are variations in the height of the test surface, then the two return beams from the optical flat and the test surface will alternate between constructive and destructive interferences, producing fringes or rings. The sensitivity is not limited to the scale of the wavelength λ: with sufficient stability and signal to noise, dimensions down to 1/1000 of a fringe can be measured. With sources of longer coherence length, such as lasers, the test surface and the optical flat need not be in direct contact.

Another optical way to measure thickness is with ellipsometry [9], typically used to measure properties of thin, transparent films from a few tenths of nanometers to several hundreds of nanometers thick. This includes metals, as long as the metal is sufficiently thin to be partially transparent. By measuring the change in polarization state for nonnormal incidence light, both the thickness and the refractive index of a thin layer may be inferred. Additional information (e.g., the properties of multiple layers) can be obtained by varying the angle of incidence and by observing at multiple wavelengths. The ability to estimate both thickness and refractive index is an important advantage of this technique, as often the refractive index of a material in thin film form is not the same as the bulk value and indeed may be a property of the deposition conditions. (FTS vendors: Bomem, Digilab, Mattson, Midac, Nicolet, and Perkin-Elmer. FTS system pricing may range from about $15,000 to over $100,000, depending on the application. Spatial interferometer vendors: Zygo, Wyko, and Phase-Shift Technology. Ellipsometer vendors: J.A. Woollam Co., Inc.; Rudolf Instruments; and Sentech. The cost of an ellipsometer may range from $10,000 for a manual, single-wavelength system to $200,000 for an automatic, multiwavelength system. Microwave resonance vendor: Hewlett-Packard.)

13.1.12 Far Field/Absorption, Scattering, and Emission

13.1.12.1 Beta, Gamma, X-Ray, and Infrared

These techniques depend on the extinction (scattering or absorption) or emission of photons or massive particles (electrons, protons, neutrons) when transiting the material to be measured. Typically, the extinction or emission shows an exponential dependence on thickness; the dependence becomes linear if the absorption is sufficiently low. These techniques, in particular gamma-ray backscatter and beta-ray transmission, are used to measure continuously moving web materials (paper, metals, fabrics) on assembly lines. Infrared absorption is also suitable if the moisture content is controlled. Beta-backscatter and x-ray fluorescence [10] are used for measuring coatings. In x-ray fluorescence, upon exposure to x-rays,

certain elements fluoresce (emit) x-rays at characteristic wavelengths. The strength of this emission is related to thickness. These absorption/emission techniques may sometimes be better suited than time-of-flight ultrasound to the dimensional measurement of objects with complex shapes. (Gamma-gage vendor: NDC Systems; x-ray absorption vendor: Measurex; x-ray fluorescence vendors: Fischer, NDC Systems, and CMI International; Beta-backscatter vendor: Fischer, Electromatic, CMI International, and Measurex; Infrared absorption vendor: NDC Systems. The prices for these systems will depend on the application; a typical system could cost $500,000.)

13.1.13 Destructive Techniques

Fischer markets a system that removes a coating into an electrolyte and then electrolytically deposits the removed coating. The electrical charge required for deposition is related to the coating thickness.

13.2 Future Directions in Thickness Measurement

13.2.1 Concerning Techniques Mentioned Earlier

Concerning capacitive sensors, the NASA Langley Research Center is developing sensors based on patterns of conductors sandwiched between insulating layers. The presence of ice over the conductors changes the capacitance, providing a way of sensing ice buildup on aircraft wings. With respect to eddy-current sensing, one limitation is that a nonsuperconducting sense coil responds best to high-frequency excitations and not at all to DC magnetic fields. This limits the technique to fairly high frequencies and thus low penetration depths, since the skin depth becomes shallower with increasing frequency. One possibility is to use a superconducting quantum interference detector (SQUID) as the sensor, since the SQUID is probably the most sensitive sensor of DC and low-frequency magnetic fields. One disadvantage of the SQUID has been the need for liquid helium for cooling for low-temperature superconductors; with the recent availability of high-temperature superconductors (HTS, above 90 K) and now HTS SQUIDS, cooling can be done with liquid nitrogen or single-stage mechanical coolers. In the area of spatial interferometry, the work at Lawrence Livermore National Laboratory replaces the reference surface with a single-mode fiber in a process called phase-shifting diffraction interferometry. A measurement accuracy of 1.44 nm rms is quoted, with a goal of 0.1 nm rms (HTS SQUID vendor: Conductus).

13.2.2 THz Technology

With the availability of femtosecond pulsed lasers, Bell Labs has been investigating a technique using 100 fs pulses to pulse an antenna in the range of 0.1–3.0 THz. The terahertz pulses are sent through the material to be tested and detected, and the received pulse shape is analyzed to extract constituent information. This technique may also provide information on thickness.

13.2.3 Nanoscale Scanning Probe Microscopy

Scanning probe microscopes (SPMs) are used in a wide variety of disciplines, including fundamental surface science, routine surface roughness analysis, and spectacular 3-D imaging—from atoms of silicon to micron-sized protrusions on the surface of a living cell [11]. The SPM is an imaging tool with a vast dynamic range, spanning the realms of optical and electron microscopes. It is also a profiler with unprecedented 3-D resolution. In some cases, SPMs can measure physical properties such as surface conductivity, static charge distribution, localized friction, magnetic fields, and elastic moduli. As a result, the applications of SPMs are very diverse. The scanning tunneling microscope (STM), the progenitor of SPMs, utilizes a sharp conductive tip with a bias voltage applied between the tip and the sample.

When the tip is within 1 nm of the sample, the electrons from the sample begin to tunnel through the 1 nm gap into the tip. If the bias voltage is reversed, the tunneling occurs into the sample. The tunneling current is a function of the separation. Both the tip and the sample must be conductors or semiconductors.

The atomic force microscope (AFM) utilizes a small tip at the end of a cantilever. Forces between the tip and the sample cause a deflection in the cantilever, which is translated into signals. The tip or sample can be scanned covering a large area, producing a topographical map. AFMs can be used on insulators or conductors. AFMs are used in two modes: contact and noncontact. In contact mode, the tip is brought within about 200 pm—about the length of a chemical bond. The electron clouds of the tip and sample atoms interact, netting a repulsive force. For this reason, the contact mode is also called repulsive. The vertical resolution of about 50 pm can be achieved. In noncontact mode, a vibrating cantilever is used in the attractive regime of the van der Waals interactions. The cantilever is typically 2–20 nm away from the sample surface and has low total force. Noncontact AFM is subsequently less sensitive; thus, sensitive AC detection systems must be employed. The low force does have the advantage of not contaminating the sample surface and is preferred for applications involving silicon wafers and soft or elastic tissues. In noncontact mode, the cantilever is resonated with a small amplitude. As the tip comes near the sample surface, the resultant force changes the spring constant, translating into a deviation of the resonance frequency. This change in resonance (or vibrational amplitude) reflects changes in the sample topology.

Intermittent contact mode is a combination of noncontact and contact modes and best suited for soft, adhesive, or fragile samples. Contact mode can damage the tip and the sample due to frictional or shear forces and/or create data artifacts from tip/surface adhesion. Noncontact mode produces lower amplitudes and hence lower resolution. Furthermore, surface monolayers of adsorbed gases such as water vapor can produce erroneous results. Intermittent contact mode avoids these pitfalls by placing the tip in contact with the surface, providing high resolution and then removing the tip to prevent dragging and/or lateral forces. The cantilever is resonated via a piezoelectric crystal (50–500 kHz in ambient, 5–40 kHz in fluids) overcoming the tip/sample adhesion forces.

In magnetic force microscopy (MFM), the noncontact mode is employed using a tip coated with a ferromagnetic film. Both magnetic and van der Walls interactions are present, but at larger tip/sample separations, the magnetic forces dominate. Multiple scans as a function of tip/sample distance allow differentiation of magnetic forces and topographic information. Magnetic domain structures are resolved to 50 nm via this technique. Current applications of MFM include data storage devices, imaging of micromagnetic structures, IC analysis, imaging of magnetotactic bacteria, and magnetic geophysics. Lateral force microscopy (LFM) is used to generate profiles of changes in surface friction and/or height variations. The probe tip is deflected laterally, indicating some sort of twist. Electronics measure the cantilever deflection. To differentiate between the two effects, LFM and AFM images should be obtained simultaneously. Phase detection microscopy or phase imaging is an extension of intermittent contact AFM. It utilizes the phase lag between the driving frequency (cantilever) and the output signal frequency, generating a map of specific mechanical properties such as adhesion, elasticity, and friction. The identification of contaminants, composite materials, and regions of hardness and low surface adhesion can be obtained at the nanometer scale. Additional techniques include force modulation microscopy, where a periodic signal is applied to the cantilever, generating a map of the sample's elastic modulus and/or contaminants; electrostatic force microscopy, where a charged tip is scanned over the sample, revealing the locally charged domains generating a map of the charge-carrier density; scanning capacitance microscopy, where a charged tip, kept at a constant tip/sample distance, generates a map of capacitance correlated information such as dielectric material thickness and subsurface charge-carrier distributions (i.e., dopant profiles of ion-implanted semiconductors); and thermal scanning microscopy, where the tip in noncontact mode and a bimetal cantilever are used to map the thermal conductivity of the sample (vendors: Park Scientific Instruments, Digital Instruments, and Oxford Instruments).

References

1. R. E. Green (ed.), *Machinery's Handbook,* 28th edn., New York: Industrial Press, 2008.
2. M. A. Curtis and F. T. Farago, *Handbook of Dimensional Measurement,* 4th edn., New York: Industrial Press, 2007.
3. T. Busch, *Fundamentals of Dimensional Metrology,* 2nd edn., Albany, NY: Delmar Publishers, 1989.
4. R. C. McMaster, P. McIntire, and M. L. Mester (eds.), *Nondestructive Testing Handbook,* Vol. 4, 2nd edn., Columbus, OH: American Society for Nondestructive Testing, 1986.
5. D. E. Bray and D. McBride, *Nondestructive Testing Techniques*, New York: John Wiley & Sons, 1992.
6. Application Notes, Wall Thickness Gaging in the Blow Molding in Industry, Olympus (http://www.olympus-ims.com/pl/applications/wall-thickness-gaging-blow-molding-industry/347-pos.56.html (accessed on March 14, 2013).
7. A. S. Birks, R. E. Green, and P. McIntire (eds.), *Nondestructive Testing Handbook,* Vol. 7, 2nd edn., Columbus, OH: American Society for Nondestructive Testing, 1991.
8. D. Malacara, *Optical Shop Testing*, New York: John Wiley & Sons, 1978.
9. J. A. Woollam, B. Johs, C. M. Herzinger, J. Hilfiker, R. Synowicki, and C. L. Bungay, "Overview of variable abgle spectroscopy ellipsometry (VASE): Basic theory and typical applications" Reviews of Optical Scinece and Technology (1999), Mendeley, http://www.mendeley.com/research/overview-variable-angle-spectroscopic-ellipsometer-vase-part-i-basic-theory-typical-applications/ (accessed on March 14, 2013).
10. H. H. Behncke, Coating thickness measurement by the X-ray fluorescence method, *Metal Finishing*, 82, 33–39, 1984.
11. R. Howland and L. Benatar, *A Practical Guide to Scanning Probe Microscopy*, Sunnyvale, CA: Park Scientific Instruments, 1993.

Bibliography

Dotson, C. L., *Fundamentals of Dimensional Metrology,* Vol. 5, Clifton Park, NY: Thomson Delmar Learning, 2006.

Moore, P. O., *Non-Destructive Testing Handbook: Infrared and Thermal Testing*, Vol. 3, 3rd edn., Columbus, OH: American Society of Nondestructive Testing, 2001.

de Silva, G. M. S., *Basic Metrology for ISO 9000 Certification*, Oxford, U.K.: Butterworth-Heinermann 2002.

14

Distance Measurement

14.1 Basic Distinctions between Range Measurement Techniques **14**-2
Contact or Noncontact • Active or Passive • Time of Flight, Triangulation, or Field Based • Time of Flight • Triangulation • Field-Based Approaches • Form of Energy • Sound • Stationary Magnetic Fields • Radio Frequencies • Light Frequencies • Coherent or Noncoherent Detection • Ranging, Range Imaging, or Position Tracking

14.2 Performance Limits of Ranging Systems **14**-8
Range Accuracy • Depth of Field • Maximum Range • Lateral Resolution • Rate of Acquisition

14.3 Selected Examples of Ranging, Range-Imaging, and Motion-Tracking Systems **14**-10
Laser-Based Active Triangulation Ranging and Range-Imaging Sensors • Laser-Based LIDAR Range-Imaging Sensors • Resonating LIDAR (Frequency-Based Detection) • Position Tracking with Active Targets • Active Target Triangulation • Magnetic Position Tracking

14.4 A Sampling of Commercial Ranging, Range-Imaging, and Motion-Tracking Products **14**-16

References **14**-16

W. John Ballantyne
Spar Aerospace Ltd.

The tools and techniques of distance measurement are possibly one of humankind's longest-running inventive pursuits. Figure 14.1 illustrates the enormous range of distances that science and engineering have an interest in measuring [1]. This chapter concerns itself with methods to measure a relatively small segment of this range—from centimeters to kilometers. Even within this limited segment, it would hardly be possible to list, much less describe, all of the distance measurement approaches that have been devised. Nevertheless, the small sampling of technologies that are covered here should be of help to a broad range of readers.

Distance measurement, at its most basic, is concerned with determining the length of a unidimensional line joining two points in 3-D space. Oftentimes, a collection of distance measurements is called for, so that the shape, the orientation, or the changes in position of an object can be resolved. Therefore, one must consider not only the measurement of distances but also their spatial and temporal distributions. The terminology "ranging" will be used in reference to systems that perform single sensor-to-target measurements, "range imaging" for systems that collect a dense map or grid of spatially distributed range measurements, and "position tracking" for systems that record the time history of distance measurement to one or several targets.

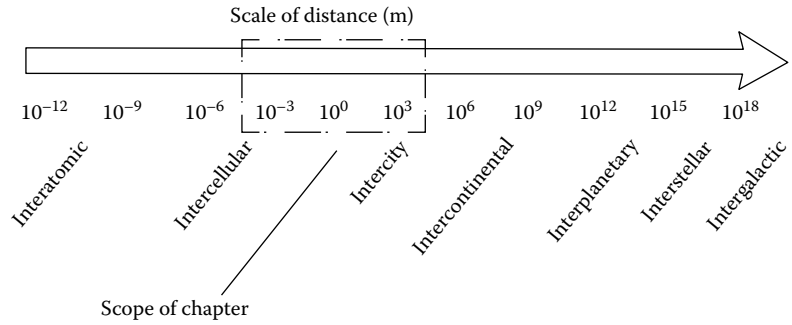


FIGURE 14.1 From the interatomic to the intergalactic, the range of measurable distances spans at least 30 orders of magnitude. The box outline indicates the relatively small segment that concerns this chapter.

14.1 Basic Distinctions between Range Measurement Techniques

Range measurement devices may be classified according to some basic distinctions. Generalization can be made based on these broad classes, thereby facilitating the process of comparison and selection. The following sections identify the fundamental bases for classification.

14.1.1 Contact or Noncontact

A common approach to measuring the distance to a point on an object is through a calibrated mechanical device that simultaneously connects the selected point to a reference position. Any tape measure, feeler gage, or dial gage may be considered an example of a simple contacting measurement device. Mechanical/electronic devices are available that allow a user to "digitize" discrete point positions on a 3-D surface. A gimbaled probe on the end of an *X-Y-Z* positioner or articulated arm is used to touch a specific point, and sensory information of the linear positions or joint articulations provides an accurate position estimate. Mechanical, contact-based methods are widely used in industry and can be extremely accurate. Some coordinate measuring machines (CMMs), for example, can achieve 1 μm repeatability.

The chief disadvantage of mechanical approaches is that they are usually restricted to distances and work volumes up to a few meters at maximum. This is due to fundamental scaling laws for mechanical structures. As the requirement to span larger distances increases, the mass and mechanical tolerancing requirements on the machine make designs impractical. Also, mechanical approaches are too slow to make multiple measurements in rapid succession, as is typically required in range imaging or position tracking, when the measurement involves large sets of spatially or temporally distributed data.

Noncontact techniques for performing ranging, range imaging, and position tracking are many and varied. Besl [2] reviews and compares several of these. In the centimeters to meters range, most do not approach the accuracy of CMMs, but at larger scales and for large quantities of data, they become a practical necessity. The rest of this chapter will review noncontact approaches only.

14.1.2 Active or Passive

Noncontact distance measurement may be divided into *active* or *passive* techniques. Active techniques involve some form of controlled energy (field or wave) linking a known reference location to the unknown target location. The source of energy is typically associated with the reference location, but in some cases the target, or both target and reference, may be active. Passive techniques rely on an externally occurring source of energy (e.g., sunlight or target/background temperature contrast) to make the target detectable.

An active approach can often simplify the distance measurement problem because it allows a greater degree of control over the many factors that can influence a measurement. For example, the choice of the form of energy and the power level of the active source can minimize the effect of uncontrolled variables like ambient illumination, weather, and atmospheric conditions. Furthermore, an active approach provides an opportunity to selectively localize the measurement spatially and temporally, eliminating possible ambiguity about which target point was measured at a given time. In contrast, passive systems (e.g., stereo ranging) sometimes suffer from the so-called correspondence problem, which is concerned with how to determine whether a given target point, detected from two or more viewpoints, or over two or more instants, is in fact the same physical point.

A common use of active approaches is to make range measurements "through" materials that are mechanically or optically impenetrable. Examples include medical imaging, where various forms of directed energy (ultrasound, x-rays) are used to build surface or volumetric maps of organs and bones; sonar, which penetrates water better than light does; and ground-penetrating radar, which can detect objects and their depth beneath ground surface.

Passive approaches, while not offering the same range of control and flexibility of active approaches, offer certain advantages. First, because they emit no energy, their existence cannot be detected by another remote detection system. This feature is very important in military applications. Second, passive systems can often collect multiple point range measurements more quickly because they are not limited by the rate at which they can direct an energy source toward a target point, as is the case with most active systems. For example, a stereo ranging system effectively collects all resolvable target points in its field of view simultaneously, while a scanning laser, radar, or sonar ranging system collects each measured point sequentially. Finally, the absence of a directed energy source is a simplification that can significantly reduce the size, cost, and hardware complexity of a device (although at the expense of increased signal-processing complexity).

14.1.3 Time of Flight, Triangulation, or Field Based

There are many different classes and instances of noncontact ranging devices, but with very few exceptions, they are based on one of the following three basic principles:

1. Energy propagates at a known, finite speed (e.g., the speed of light, the speed of sound in air).
2. Energy propagates in straight lines through a homogeneous medium.
3. Energy fields change in a continuous, monotonically decreasing, and predictable manner with distance from their source.

The techniques associated with these basic phenomena are referred to as time of flight (TOF), triangulation, and field based, respectively.

14.1.4 Time of Flight

TOF systems may be of the "round-trip" (i.e., echo, reflection) type or the "one-way" (i.e., cooperative target, active target) type. Round-trip systems effectively measure the time taken for an emitted energy pattern to travel from a reference source to a partially reflective target and back again. Depending on whether radio frequencies, light frequencies, or sound energy is used, these devices go by names such as RADAR, LIDAR, and SONAR. One-way systems transmit a signal at the reference end and receive it at the target end or vice versa. Some form of synchronizing reference must be available to both ends in order to establish the TOF.

A characteristic of many TOF systems is that their range resolution capability is based solely on the shortest time interval they can resolve and not the absolute range being measured. That is, whether an object is near or far, the error on the measurement is basically constant.

14.1.5 Triangulation

Triangulation techniques were known and practiced by the ancients. Triangulation is based on the idea that if one knows the length of one side of a triangle and two of its angles, the length of the other sides can be calculated. The known side is the "baseline." Lines of detection extend from either end of the baseline to the target point as shown in Figure 14.2. If the angles formed between these lines and the baseline can be determined, the distance is calculated as

$$R = \frac{b \sin \alpha_{\text{left}} \sin \alpha_{\text{right}}}{\sin(\alpha_{\text{right}} - \alpha_{\text{left}})} \tag{14.1}$$

Classical surveying is a passive range-finding technique based on the aforementioned formula. A surveyor uses a precision pointing instrument to sight a target from two positions separated by a known baseline. Reference [3] notes that the distance to a nearby star may be calculated by observing it through a pointing instrument at 6 month intervals and by using the diameter of Earth's solar orbit as the baseline. Stereo ranging, which compares the disparity (parallax) between common features within images from two cameras, is another form of passive triangulation. It is of interest to note that human vision estimates distance using a variety of cues, but two of the most important—stereopsis and motion parallax—are fundamentally triangulation based [4].

Active triangulation techniques use a projected light source, often laser, to create one side of the triangle, and the viewing axis of an optical detection means to create the second side. The separation between the projector and detector is the baseline.

A fundamental issue for all triangulation-based approaches is that their ability to estimate range diminishes with the square of the range being measured. This may be contrasted with TOF approaches, which have essentially constant error over their operating range. Figure 14.3 illustrates how, conceptually, there is a "crossover" distance where TOF techniques become preferable to triangulation techniques.

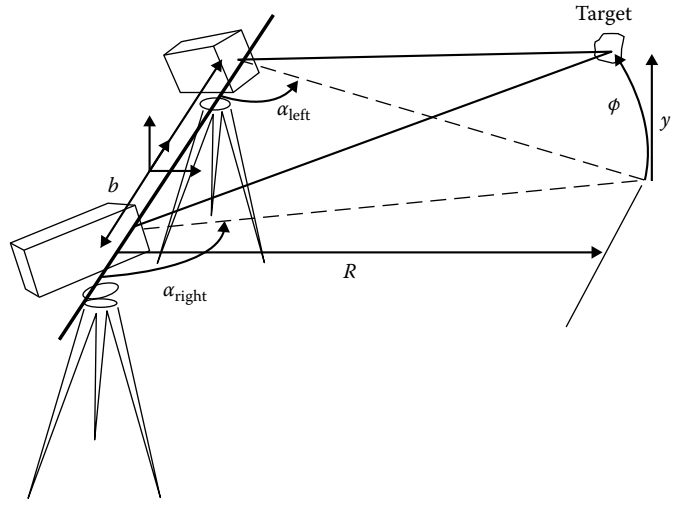


FIGURE 14.2 The basic triangulation geometry as used in classical surveying determines the distance to a remote point by sighting it from two locations separated by a known baseline. The pointing angles α_{left} and α_{right} are measured locally.

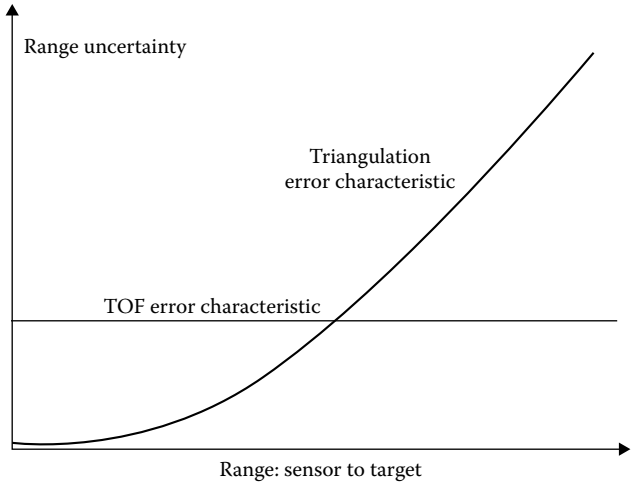


FIGURE 14.3 TOF and active triangulation techniques tend to exhibit error characteristics related to their fundamental principles of operation. The dominant error source in TOF systems is usually the shortest measurable time interval, but this is a detection issue and is essentially independent of distance. Active triangulation systems are typically more accurate at close distances, but geometry considerations dictate that the effects of their error sources will increase with the square of distance.

14.1.6 Field-Based Approaches

Whereas TOF and active triangulation techniques employ the wave propagation phenomena of a particular energy form, *field-based* approaches make use of the spatially distributed nature of an energy form. The intensity of any energy field changes as a function of distance from its source. Moreover, fields often exhibit vector characteristics (i.e., directionality). Therefore, if the location of a field generator is known and the spatial characteristics of the field that it produces are predictable, remote field measurements contain information that may be used to infer distance from the source.

An interesting distinction between field-based approaches and wave-based approaches is that the former, although they employ energy fields, do not rely on the propagation and conversion (and concomitant losses) of energy. That is, they may employ stationary fields, like those generated by a magnet or static charge. Such fields encode position information by their very shape. Sound and light, although having a wave nature, can be exploited in the same manner as stationary fields because of their distance-dependent intensity.

Field-based techniques must confront some basic issues that limit their range of application. First, the characteristics of most practically exploitable fields are typically influenced by objects or materials in the vicinity, and it is not always possible to ensure that these influences will remain constant. Second, the variation of fields through space is highly nonlinear (typically inverse square or inverse cube), implying that the sensitivity of a measurement is strongly affected by proximity to the source. Notwithstanding these concerns, devices have been developed and are available that perform very well in the situations for which they are intended [7].

14.1.7 Form of Energy

As discussed previously, all noncontact, active ranging devices employ some form of energy. This is true whether TOF, triangulation, or field-based principles apply. The following sections describe the various forms of energy employed and some generalizations about the effectiveness of each in various situations.

14.1.8 Sound

Ranging systems based on sound energy are usually of the pulsed-echo TOF type and employ carrier frequencies in the so-called "ultrasonic" (beyond audible) range of frequencies. Besides being inaudible (an obvious benefit), ultrasonic frequencies are more readily focused into directed beams and are practical to generate and detect using piezoelectric transducers. Ultrasonic signals propagate through air, but long-distance transmission is much more effective in liquids, like water, where higher density-to-viscosity ratios result in higher wave velocity and lower attenuation per unit distance. Ultrasonic ranging techniques (or SONAR, for sound navigation and ranging) were first developed for subsea applications, where sound is vastly superior to electromagnetic energy (including light) in terms of achievable underwater transmission distances [5]. Low-cost, portable sonar systems are widely used by sport fisherman as "fish finders" [6].

The frequencies typically used in sonic ranging applications are at a few tens of kilohertz to a few hundred kilohertz. A basic trade-off in the choice of ultrasonic frequency is that while high frequencies can be shaped into narrower beams and therefore higher lateral resolution is achieved, they tend to fade more quickly with distance. It may be noted that beam widths narrow enough for range-imaging applications (less than 10°) are effective in a fluid medium but attenuate too quickly to be practical in air. Interestingly, although sound energy attenuates more rapidly in air than in water, useful short-range signals can be generated in air with relatively low power levels because the much lower density of air requires smaller dynamic forces in the transducer for a given wave amplitude.

When comparing sound energy to electromagnetic energy for TOF-based techniques, one need to remember that sound, unlike light, propagates at not only much lower speeds but also with considerably more speed variation, depending on the type and state of the carrying media. Therefore, factors like air humidity and pressure will affect the accuracy of a TOF ranging device. For underwater applications, salinity and depth influence the measurement. The lower speed of sound has a detrimental impact on the rate at which range samples can be collected. For example, a target 10 m away takes at least 60 ms to measure through an air medium. This may not seem like a long time to wait for a single sample, but it becomes an issue if the application involves multiple sampling, as in motion tracking or collision avoidance sensing.

14.1.9 Stationary Magnetic Fields

Stationary or pseudostationary (i.e., low frequency) magnetic fields are only used in field-based approaches. An advantage of such fields is that they are easily and cheaply produced by either a permanent magnet or electrical coil. Since stationary fields do not transmit energy, the targets cannot be passive—they must actively sense the properties of the field at their particular location. A variety of sensing technologies may be used to make measurements of the direction and intensity of a magnetic field, including flux gate, Hall effect, and magnetostrictive type magnetometers. A comprehensive list of such technologies is given in Ref. [7].

14.1.10 Radio Frequencies

Echo-type TOF ranging systems based on the band of the electromagnetic spectrum between approximately 1 m and 1 mm wavelength are known as RADAR (radio detection and ranging). Radio waves can be used for long-distance detection in a variety of atmospheric conditions. As in the case of sound waves, there are few trade-offs to be addressed in the choice of frequency. Long waves tend to propagate better over long distances, but short waves can be focused into narrow beams capable of better lateral discrimination. An interesting application of short-range radar is ground-penetrating radar, which can be used to locate and image subsurface objects [8]. Here, the frequency versus range trade-off is particularly acute because of the need to balance reasonable imaging capability (narrow beam) with good depth penetration (long wave).

An example of a TOF one-way (active receiver) system that uses radio frequencies is the global positioning system (GPS). The distances between a receiver on land and satellites are determined by each of several orbiting satellites equipped with a transmitter and a very precise Cesium clock for synchronization. A good description of GPS and its use in vehicle navigation is available in Ref. [9].

14.1.11 Light Frequencies

Beyond the radio portion of the electromagnetic spectrum are the infrared, visible, and ultraviolet frequencies. These frequencies can be produced by lasers and detected by solid-state photosensitive devices and are useful for both TOF and active triangulation ranging. Echo-type TOF techniques are known as LIDAR (light detection and ranging), in keeping with the terminology introduced earlier.

While light frequencies attenuate more than radio frequencies through cloud and fog, they can have very narrow beam widths, allowing superior lateral resolution and target selectivity.

14.1.12 Coherent or Noncoherent Detection

Echo-type TOF devices, whether sonar, radar, or lidar, can be further classified according to whether the detection approach measures TOF directly (noncoherent) or exploits an inherent periodicity in the emitted energy to ascertain the flight distance (coherent).

Noncoherent techniques face the problem of timing short intervals. This is not a serious challenge in the case of sound waves, where a meter round trip corresponds to 6 ms, but is somewhat more problematic for light and radio waves, where that distance equates to only 6 ns. Accuracy of noncoherent detection typically relies on the averaging of repeated measurements.

Coherent detection is achieved by combining a portion of the emitted signal with the reflected signal to produce a third signal indicating the amount of phase delay. The signals are continuous wave (CW) as opposed to pulsed. Coherent detection techniques are classified as amplitude modulated (AMCW) or frequency modulated (FMCW).

A basic issue with coherent detection techniques is the inability to distinguish between integral multiples of the basic modulation wavelength. Any coherent detection system must employ techniques to resolve the so-called ambiguity interval. Noncoherent techniques do not face this problem.

14.1.13 Ranging, Range Imaging, or Position Tracking

Ranging devices are typically pointed toward a target to produce a single range reading. A common example of simple ranging is the feedback sensor used in autofocus cameras. There are many active ranging devices currently available based on TOF (i.e., RADAR, SONAR, LIDAR) and active triangulation principles.

Range-imaging devices use the same principles as ranging devices, except that they include some form of scanning that is employed to generate an array of spatially distributed range samples. Sometimes, the scanning action is accomplished by means of intrinsic to the sensor (e.g., spinning and nodding mirrors or phased-array antenna) so that the reference location remains fixed. In this case, the data are recorded in the polar form (range, elevation, azimuth) as shown in Figure 14.4. In other cases, the sensor might scan on only one axis internally, while the second scan dimension is realized by moving the sensor location through some set pattern. It is not uncommon to record the "intensity" or return energy associated with a range sample as well. The intensity map may be presented as a "gray scale" image and, like a black and white photograph, often contains additional information useful in interpreting a scene. Range images can be used to produce 3-D graphic representations of scenes and objects. A common use of range imaging is aerial terrain mapping.

Position-tracking devices are used to measure the change in an object's position and orientation over time. Basic issues in position tracking are the acquiring of, and locking on to, specific target points. These issues can be avoided by employing active targets, and most systems available today are of this type.

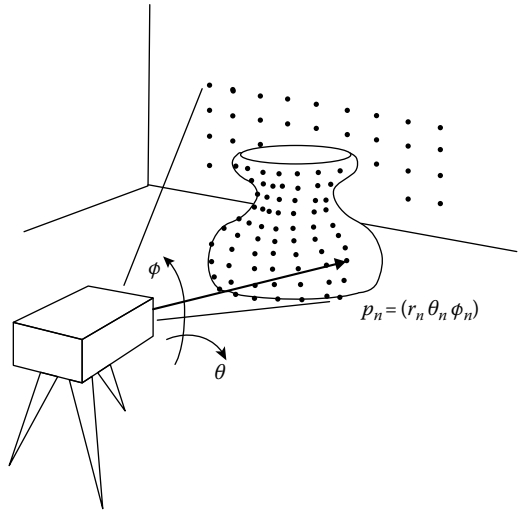


FIGURE 14.4 Range images are typically an array of individual range values sampled while changing the pointing direction (e.g., azimuth and elevation angles) of a ranging device. A digital range image of the polar form shown can be readily transformed into rectangular coordinates if required.

14.2 Performance Limits of Ranging Systems

The performance characteristics of available ranging systems vary widely, as do the requirements of the applications for which they are designed. The following sections review the most basic performance categories and the technical issues of performance limits.

14.2.1 Range Accuracy

As illustrated in Figure 14.3, TOF and active triangulation techniques differ fundamentally in their error versus distance characteristics. Currently available systems based on active triangulation achieve better repeatability and accuracy in the less than 1 m range than do TOF systems but are seldom used at distances of several meters. Hymarc Ltd. and Perceptron Inc. each offer laser triangulation systems with 3σ accuracy of 25 and 50 mm, respectively [10,11].

In principle, TOF systems could achieve accuracy rivaling active triangulation, but the most promising detection technique—a variation of laser interferometry, which solves the ambiguity interval problem [12]—has yet to make its commercial debut.

14.2.2 Depth of Field

Depth of field refers to the interval of distance through which a stationary reference ranging system can measure without resorting to a change in configuration. Large depth of field is often an important characteristic in practical applications. For example, if the distance to the target is poorly known a priori, then a large depth of field is desirable.

Passive optical triangulation approaches like stereography and photogrammetry tend to have restricted depth of field because they rely on camera-type imaging, which is inherently limited by depth of focus. Timed-interval TOF systems have excellent depth of field because they do not rely strongly on optical imaging except to concentrate the collected returned energy on the detector. Some active triangulation systems do rely on optical imaging of the projected laser spot, but the design employed by Hymarc Ltd. regains a large depth of field by tilting the detector array with respect to the lens plane [13].

14.2.3 Maximum Range

Any active ranging, range-imaging, or position-tracking system has a practical maximum distance that it can measure. This is because the controlled energy, whether propagated as a wave or established as a field, must spread before reaching the detector. The spreading inevitably increases with distance and all detectors, no matter what form of energy they measure, require a certain minimum amount to exceed their inherent "noise floor."

The "classical radar range equation" is introduced in many texts on radar (e.g., [14]). Jelalian [15] points out that the equation is equally applicable to lidar, which, after all, just employs a higher frequency version of electromagnetic wave. In fact, the same idea applies to sonar and to active triangulation systems as well. The equation computes the power of the received signal as

$$P_R = \frac{P_T G_T}{4\pi R^2} \times \frac{\rho A}{4\pi R^2} \times \frac{\pi D^2}{4} \times \eta_{atm}\eta_{sys} \tag{14.2}$$

where

P_R is the power at the receiver
P_T is the power transmitted
G_T is the transmitter gain
R is the range to target
ρ is the reflectivity of target
A is the effective area of target
D is the diameter of collecting aperture
η_{atm} is the atmospheric transmission coefficient
η_{sys} is the system transmission coefficient

Equation 14.2 applies when the target area is smaller than the footprint of the incident beam, which is often the case for radar and sonar ranging. However, in the case of laser-based systems, the relatively narrow beam usually means that the laser spot is small compared to the target. For a transmitted beam that spreads with a solid angle θ_T, the illuminated patch area is

$$\sigma_{spot} = \pi R^2 \theta_T^2 \tag{14.3}$$

The definition of transmitter gain is based on the notion of the solid angle beam width as compared to an omnidirectional transmitter:

$$G_T = \frac{4\pi}{\theta_T^2} \tag{14.4}$$

One can substitute for Equation 14.4 for G_T and Equation 14.3 for the variable s in Equation 14.2 to produce the range equation for a small spot size:

$$P_R = \frac{P_T}{R^2} \times \frac{\rho}{4} \times \frac{\pi D^2}{4} \times \eta_{atm}\eta_{sys} \tag{14.5}$$

The importance of this equation is primarily in the $1/R^2$ dependence. Any ranging system that works by bouncing energy off a diffuse reflective target encounters severe signal attenuation with

increasing distance. Given a detector with a fixed noise floor, the only way to improve maximum range is to increase the transmitted power or the collecting area. In practice, there are design constraints that limit both of these measures. For example, laser power must sometimes be limited for eye-safety considerations, and increased collecting area can imply a proportional increase in sensor packaging volume.

14.2.4 Lateral Resolution

In range-imaging applications, it is generally desirable to use the narrowest possible beam width to provide good lateral discrimination of target surface features. Lasers, because of their short wavelength, can be optically collimated to produce much narrower beam widths than are possible with radio sources. However, even lasers cannot produce arbitrarily narrow beams. The interested reader is referred to Ref. [13] for a discussion of Gaussian beam propagation and optimal focusing. There are basically two ways to project laser light. The beam can be "focused down" to produce the smallest possible spot at a particular point inside the measurement range, in which case the beam will diverge as the distance from that point increases, or the beam can be focused at infinity or some very distant point so as to minimize the divergence through the entire measurement range. The former approach provides higher lateral resolution at the focus distance, but by implication restricts the practical depth of field. The latter compromises spot sizes for the increased depth of the field.

14.2.5 Rate of Acquisition

The rate at which a ranging sensor can acquire range samples is important when the target object is changing shape or position or when the required sample density of a range image is very high. There are several potential factors that can limit sample acquisition rate: the amount of time required by the detector to integrate the weak return signal to a sufficient level (integration time), the time constant of any filtering or averaging that must be performed to realize an acceptably "clean" signal (smoothing time), the rate at which samples can be transferred through the signal-processing stages (transfer time), and the velocity limits of mechanical scanning apparatus (scanning bandwidth). Acquisition rates vary widely: from tens of hertz for acoustic ranging devices to tens of kilohertz for some laser-based systems. It is worth noting that, in general, there is a trade-off between rate of acquisition, accuracy, and maximum range. Some systems permit control over basic parameters so that this trade-off may be optimized for a particular application. The reader should be aware that data sheets may not be clear as to whether stated performance figures for these three specifications are valid in combination.

14.3 Selected Examples of Ranging, Range-Imaging, and Motion-Tracking Systems

The following sections review selected examples of some specific ranging, range-imaging, and position-tracking sensor systems. The list is by no means exhaustive but offers a reasonable sampling of available technologies.

14.3.1 Laser-Based Active Triangulation Ranging and Range-Imaging Sensors

14.3.1.1 Active Triangulation Basics

Figure 14.5 illustrates the basic active triangulation geometry. In this so-called "pinhole camera" model, practical aspects like lenses for projection and detection and mirrors for scanning are eliminated

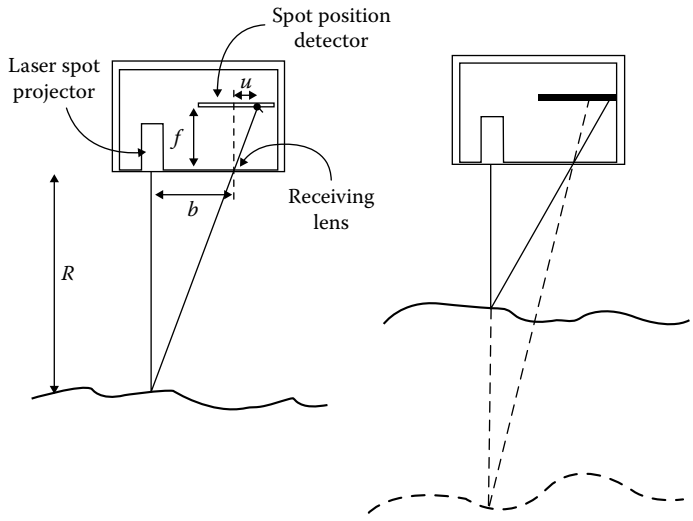


FIGURE 14.5 A simple pinhole camera model illustrates the basic active triangulation principle. As the distance R to the target surface changes, the spot position u on the detector changes, maintaining similarity between the large triangle outside the camera and the small triangle inside. There is an inverse relationship between R and u.

for clarity. It can be shown by means of similar triangles that the range is inversely proportional to the deflection of the imaged spot:

$$R = \frac{bf}{u} \tag{14.6}$$

where
R is the distance to object
b is the baseline distance
f is the lens-to-detector distance
u is the detected spot position in the image plane

The sensitivity of the range measurement, or the incremental change in u with R, is

$$\left|\frac{du}{dR}\right| = \frac{bf}{R^2} \tag{14.7}$$

The significance of Equation 14.7 is that range estimating performance is expected to fall as distance increases. Improvements in accuracy are realized by increasing the baseline or the lens-to-detector distance (i.e., the focal length).

14.3.1.2 Synchronized Scanning Principle

Lateral scanning of an active triangulation ranging sensor is accomplished by an elegant and effective technique developed at Canada's National Research Council and now marketed by Hymarc Ltd. under the name "HYSCAN" [10]. A two-sided oscillating mirror simultaneously steers the outgoing beam on one face and directs the collected light to the spot-imaging optics on the opposite face. By synchronously scanning both the beam and the axis of the detection system, rather than the beam only, as conventionally practiced, significant performance improvements are made. Figure 14.6 is a schematic illustration

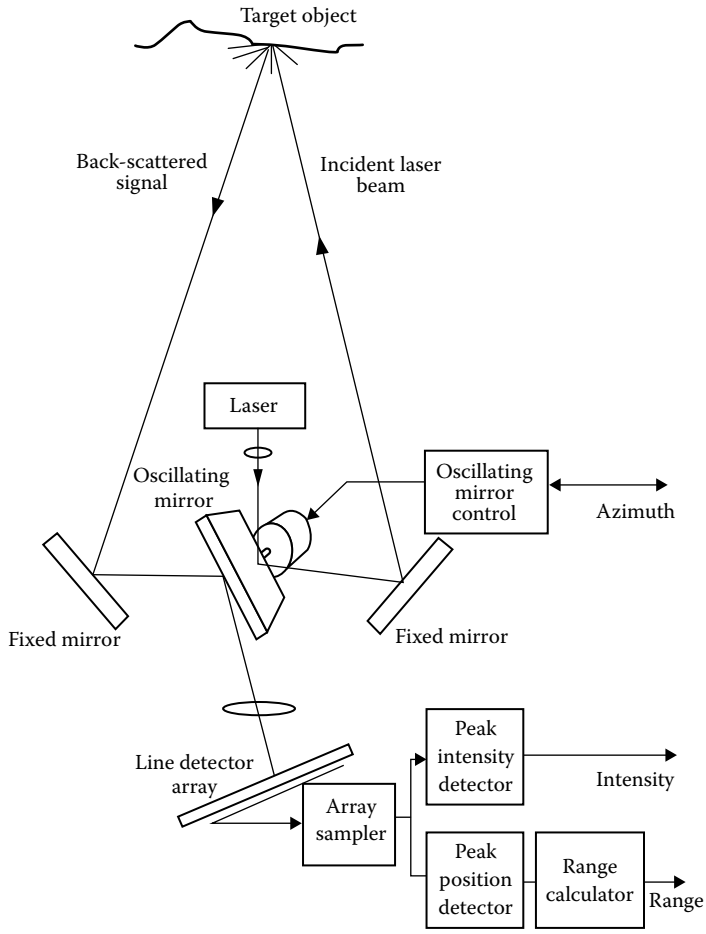


FIGURE 14.6 The Hymarc laser triangulation line scanner uses the synchronized scanning principle. Both sides of an oscillating mirror are used to sweep both the projected beam and the axis of detection over the target. The detector array is tilted to the lens plane to maximize the depth of focus.

of the approach. Note also that the detector plane is tilted with respect to the lens plane. This feature increases the depth of focus so that the ranging performance is maximized over the measuring volume. The HYSCAN sensor produces a single-axis sweep, or so-called line scan. Full-dimension range images are acquired by translating the sensor over a target surface with a controlled motion pattern.

14.3.1.3 Light Plane Principle

Perceptron Inc. offers a similar line-scan system under the name "TriCam" [11]. In this case, the laser is not swept. Instead, the beam is transformed to a focused plane by means of cylindrical lenses. A 2-D detector is used to generate range profiles through the analysis of a deformation of the laser line as the sensor is translated over the object surface.

14.3.2 Laser-Based LIDAR Range-Imaging Sensors

14.3.2.1 AM LIDAR (Phase-Based Detection)

Perceptron Inc. also offers a scanning LIDAR under the name "LASAR" that can produce high-resolution range images through a large measurement volume. The device uses a near-infrared laser that is projected

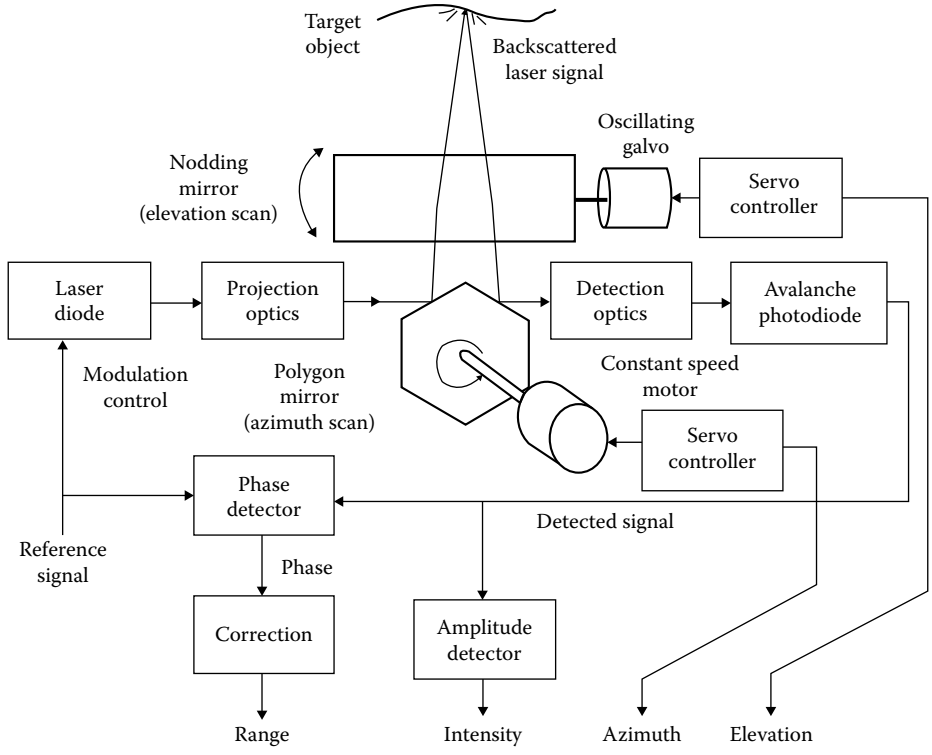


FIGURE 14.7 The Perceptron AM LIDAR system described in U.S. patent 5,006,721 uses a rotating polygon mirror for synchronized scanning. A "nodding mirror" is also added to sweep at a slower rate in the orthogonal direction, producing a raster scan pattern. Range measurement is determined by comparing the phase of the outgoing and returning AM laser signal.

through a collimating telescope to form a spot on the first surface encountered. The spot is swept over a programmable field of view in a raster pattern by means of a spinning polygon mirror and an oscillating "nodding mirror." Some of the backscattered light is collected and directed by means of an adjacent facet of the polygon mirror. The projected laser light is amplitude modulated at a reference frequency by controlling the power to the laser diode source. The return signal, although orders of magnitude weaker than the outgoing signal, is phase-compared to determine the range for a particular azimuth and elevation. The intensity of the return energy is also recorded. The Perceptron sales literature claims a maximum measurement volume of 60° × 72° × 40 m, a range image grid resolution of 1024 × 2048 pixels and a maximum acquisition rate of 360,000 pixels/s. A schematic diagram of the LASAR™ system is shown in Figure 14.7. Details of the Perceptron technical approach may be found in Ref. [16].

14.3.3 Resonating LIDAR (Frequency-Based Detection)

Acuity Research Inc. has developed a laser-based TOF ranging sensor based on a simple but effective idea. The detector controls the laser output such that the absence of a signal drives the laser on and the presence of a signal turns it off. The finite transit time of the light bounce turns this arrangement into a two-state resonator, with the period being proportional to the target distance. Rather than measuring the period, which is extremely short and difficult to time, the frequency is measured using conventional counting techniques for as many cycles as necessary to yield the required accuracy. The AccuRange 4000, as it is named, is also available in a 360° line-scanning arrangement suitable for robotic vehicle navigation applications [17]. Details of the technical approach may be found in Ref. [18].

14.3.4 Position Tracking with Active Targets

Active target approaches are not convenient in some applications, but they are an excellent way to track the changing positions of several target points simultaneously. Active targets are a way of getting around the "correspondence problem" mentioned earlier. The two systems introduced here are interesting to compare. One employs light energy and triangulation; the other uses a magnetic field-based approach. They are both used for real-time tracking and recording of human kinetics, robotics, and other moving objects.

14.3.5 Active Target Triangulation

The "OPTOTRAK" system offered by Northern Digital Ltd. [19] uses infrared light-emitting diodes (LEDs) as targets. The LEDs are multiplexed so that only one at a time can be seen by the camera system, avoiding the correspondence problem. The unique form of stereo ranging is based on three line detectors with lenses that transform the point source LED illumination into a focused line. The simplified triangulating geometry is shown in Figure 14.8. It may be shown from this geometry that the target position (x_p, y_p, z_p) can be determined from the detector outputs u_{left}, u_{right}, and v as follows:

$$x_p = \frac{b(u_{right} + u_{left})}{2(u_{right} - u_{left})} \tag{14.8}$$

$$y_p = \frac{bv}{(u_{right} - u_{left})} \tag{14.9}$$

$$z_p = \frac{fb}{(u_{right} - u_{left})} \tag{14.10}$$

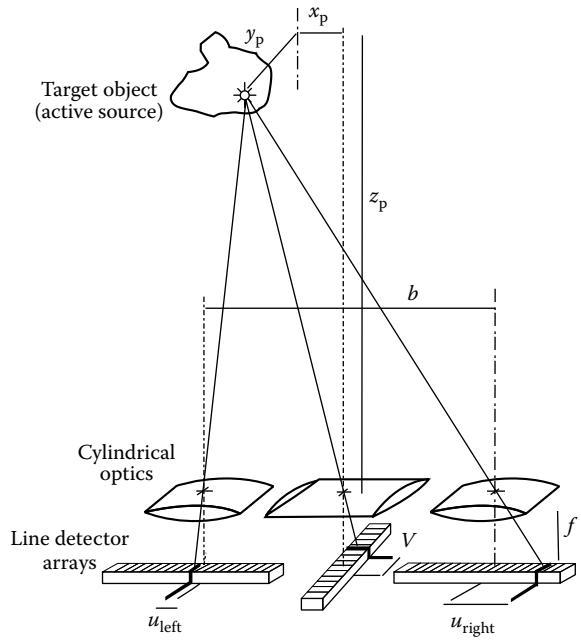


FIGURE 14.8 The OPTOTRAK position-tracking system employs a novel arrangement of cylindrical optics and 1-D detectors to triangulate the 3-D position of an infrared LED target. Up to 255 individual multiplexed targets can be tracked by the system.

where f and b are the lens-to-detector distance and the baseline separation, respectively. In practice, the image space to object space mapping is much more complicated than Equations 14.8 through 14.10 and involves a camera model with more than 60 parameters that are determined through a calibration process.

OPTOTRACK offers high sampling rate, large measurement volume, and high accuracy compared to many other position-tracking systems.

14.3.6 Magnetic Position Tracking

A position/orientation tracking sensor based on a three-axis magnetic dipole transmitter and a three-axis magnetic loop detector has been developed by Polhemus Inc. [20]. The transmitted fields are alternating current for ease of detection (i.e., transformer coupled) and time-multiplexed so that the field due to each axis can be distinguished from the others. Distance between transmitter and detector is determined by exploiting the $1/R^3$ relationship between field strength and distance from the source. Orientation of the detector is determined by exploiting the directionality of magnetic fields and the direction sensitivity of loop detectors.

An issue with respect to the use of ac fields is the distortions in field shape that occur if metal objects are present and the consequent effect on sensor accuracy. These distortions result from eddy currents in the conducting metal. Ascension Technology Corp. has developed a variation on the Polhemus sensor

TABLE 14.1 Ranging, Range-Imaging, and Position-Tracking Products and Vendors

Class	Trade Name	Principle	Features	Contact
Ranging (contact)	MicroScribe-3DX	Instrumented arm	50 in. spherical work volume, 0.3 mm accuracy	Immersion Corp., (408)467–1900, info@immerse.com
Ranging (noncontact)	LASERVISION	TOF, laser	50 m range, 4.9 mm accuracy at 15 m, integrated electronic level	ZIRCON Corp., (408) 866–8600
Range imaging (line scan)	HYSCAN	Active triangulation laser	40 mm depth of field, 70 mm swath, 0.025 mm accuracy, 10,000 points/s	Hymarc Ltd., (613)727–1584, info@hymarc.com
Range imaging (line scan)	TriCam	Active triangulation laser	120 mm depth of field, 60 mm swath, 0.05 mm accuracy	Perceptron Inc., (810) 478–7710, inquiry@perceptron.com
Range imaging (line scan)	ALTM 1020	TOF laser time interval	330–1000 m range, 15 cm accuracy, 20° swath	Optech Inc., (416) 661–5904
Range imaging (area scan)	Rangecam 7000	Laser or strobe triangulation	Uses standard CCD camera and light plane projector	Range Vision Inc., (604)473–9411
Range imaging (area scan)	LASAR	TOF, AM LIDAR	2–40 m range, 60° × 70° max field of view, 360,000 samples/s	Perceptron Inc., (810) 478–7710
Position tracking	OPTOTRAK	Active target triangulation	Up to 255 targets, submillimeter accuracy, 5000 3 DoF samples/s	Northern Digital Inc., (519) 884–5142
Position tracking	Flock of Birds	Magnetic field based	Up to 30 position/orientation targets, approx. 10 mm accuracy, 144 6-DoF samples/s	Ascension Technology Corp., (802) 860–6440

based on dc magnetic fields. The switching transient due to time-multiplexing does produce an eddy current effect, but it is allowed to die out before measurement is made. Details of the dc technique are available in Ref. [21].

An important difference between optical and magnetic tracking technologies is that the former require an unbroken line of sight to the targets while the latter do not. This gives magnetic trackers an advantage in some applications. On the other hand, the $1/R^3$ field distribution characteristic of magnetic tracking implies an extreme sensitivity loss with distance, whereas optical triangulation has a more benign $1/R$ characteristic. This, to some extent, explains why the volume of measurement and accuracy of optical triangulation systems is generally much better than for magnetic systems.

14.4 A Sampling of Commercial Ranging, Range-Imaging, and Motion-Tracking Products

Table 14.1 contains information collected from vendor literature. Be advised when comparing specifications that test conditions, standards, and interpretations can vary significantly. The specifications, therefore, should serve only as a rough guide.

References

1. R. Resnick and D. Halliday, *Physics (Part 1)*. John Wiley & Sons, New York, 1966, Ch. 4.
2. P. J. Besl, *Range Imaging Sensors*. General Motors Research Publication, GMR-6090, General Motors Research Laboratories, Warren, MI, March, 1988.
3. R. Resnick and D. Halliday, *Physics (Part 1)*. John Wiley & Sons, New York, 1966, Ch. 3.
4. D. F. McAllister (ed.), *Stereo Computer Graphics and Other True 3D Technologies*, Princeton University Press, Princeton, NJ, 1993. Ch. 4.
5. L. E. Kinsler and A. R. Frey, *Fundamentals of Acoustics*, 2nd edn., John Wiley & Sons, New York, 1962, Chs. 9, 15.
6. W. Diedrich, Foundations of reading sonar, *The In-Fisherman*, April–May, 42–56, 1996.
7. E. B. Blood, Device for quantitatively measuring the relative position and orientation of two bodies in the presence of metals utilizing direct current magnetic fields, U.S. Patent 4,945,305, July 31, 1990.
8. W. J. Steinway and C. R. Barrett, Development status of a stepped-frequency ground penetrating radar, in *Underground and Obscured Object Imaging and Detection*, *SPIE Proceedings*, Vol. 1942, Orlando, FL, April 1993, pp. 34–43.
9. J. Borenstein, H. R. Everett, and L. Feng, *Where am I? Sensors and Methods for Autonomous Mobile Robot Positioning*, 1995 edn., University of Michigan report for the United States Department of Energy Robotics Technology Development Program, Ann Arbor, MI, 1995. Ch. 3.
10. Hymarc Ltd., *Product Information*, Hyscan 3D Laser Digitizing Systems, Ottawa, Ontario, Canada, 1995.
11. Perceptron Inc., *Product Information*, TriCam Non-Contact Measurement Solutions, Farmington Hills, MI, 1995.
12. F. E. Goodwin, Frequency modulated laser radar, U.S. Patent 4,830,486, May 16, 1989.
13. F. Biais, M. Rioux, and J.-A. Beraldin, Practical considerations for a design of a high precision 3D laser scanner system, SPIE Vol. 959, Optomechanical and Electro-Optical Design of Industrial Systems, 1988.
14. D. K. Barton, *Radar System Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1964, Ch. 4.
15. A. V. Jelalian, *Laser Radar Systems*, Artech House, Boston, MA, 1992. Ch. 1.
16. E. S. Cameron, R. P. Srumski, and J. K. West, Lidar scanning system, U.S. Patent 5,006,721, April 9, 1991.

17. Acuity Research Inc., *Product Information*, Accurange 4000, Menlo Park, CA, 1995.
18. R. R. Clark, Scanning rangefinder with range to frequency conversion, U.S. Patent 5,309,212, May 3, 1994.
19. Products, Northern Digital Inc., NDI, http://www.ndigital.com/products.php (accessed on March 14, 2013).
20. F. H. Raab, E. B. Blood, T. O. Steiner, and H. R. Jones, Magnetic position and orientation tracking system, *IEEE Transactions on Aerospace Electronic Systems*, AES-15(5), September 1979.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

15

Altitude Measurement

15.1 Pressure Altitude Measurement **15**-1
15.2 Radar Altimeter **15**-2
15.3 Height Measurement **15**-2
15.4 Ground-Based Height Estimation **15**-3
Navigation Accuracy Measurement System • Height Estimation with Single Air Traffic Control Radar • Height Estimation with One SSR and One Omniradar • Dual-Synchronized Autonomous Monitoring System • Height Measurement by Quadrilateration • Height Estimation with SSRs and Pseudoranges • Height Measurement with SSRs and Range Differences (RDHMU)
15.5 Onboard–Derived Height Estimation **15**-12
Height Measurement with Distance Measuring Equipment • Estimation of Vertical Position with the Global Positioning System
15.6 Special Topics **15**-13
Inherent Bias • Existence and Uniqueness of Position Fix
References **15**-15

Dimitris E. Manolakis
Alexander Technological Education Institute of Thessaloniki

Altitude can be determined via barometric altimeters or it can be measured via electronic radio wave systems. The indication of the first equipment is referred to as pressure altitude, or simply altitude, whereas that of the second category is referred to as geometric height or simply height or true altitude. Pressure altitude is derived onboard, whereas the height can be derived either onboard or in ground stations.

15.1 Pressure Altitude Measurement

Actually, the altitude information is an atmospheric pressure measurement transformed to altitude indication through a formula expressing the pressure/altitude relationship. Thus, a pressure altimeter or barometric altimeter is a barometer that is provided with a nonlinear calibration in order to transform it to altitude indication, through the following form:

$$z = kT \log\left(\frac{P_0}{P}\right) \tag{15.1}$$

where
P is the pressure at altitude z
P_0 is the pressure at sea level
T is the absolute temperature

Constant k expresses the effect of the molar mass of the air and the acceleration of gravity according to a mathematical model defined by the International Standard Atmosphere (ISA). It is important to note that this type of altimeter relies on "density altitude" and consequently can shift its indication by tens or hundreds of meters because of abrupt changes in air pressure although the altitude actually remains the same.

Pressure altimeters are used in aircrafts, meteorological stations, mountain climbers, etc.

15.2 Radar Altimeter

A more direct measurement of the altitude is via onboard radar. The altitude z is derived by measuring the time T that the signal needs to travel from the board to the surface and back after its reflection:

$$z = \frac{Tc}{2} \tag{15.2}$$

where c is the velocity of the signal. Radar altimeters are used in commercial and military aircrafts to measure height above ground level during landing. These altimeters are part of terrain avoidance warning systems as well as and of terrain-following systems permitting fighter aircraft to fly at very low altitude.

15.3 Height Measurement

The altitude information at air traffic control (ATC) centers is based on pressure altitude measurement that is derived onboard, and the aircraft transponder sends it after an appropriate interrogation transmitted by secondary surveillance radar (SSR). When a flight level (FL) is cleared for an aircraft, it actually means that the pilot must keep flying on an isobaric surface. However, the barometric altimetry system may present systematic errors (biases) that are different for each airplane and that significantly affect safety. In addition, the atmospheric conditions rarely match the standard atmosphere, so true and indicated altitudes are not equal. For example, if the temperature is about 10°C colder than standard, then indicated altitude is approximately 4% higher than true altitude. This means an error of 600 ft at 15,000 ft flying level that is a substantial divergence if flying on a cold day. The vertical separation minimum (VSM) at FL lower than 2900 ft (FL290) is 1000 ft, whereas for FL higher than FL290, it is 2000 ft due to the deterioration of the performance of pressure altimeters at high FL. A program started in 1982 from ICAO to assess the feasibility of reducing the VSM from 2000 to 1000 ft for levels between FL290 and FL410. The benefit is the increase of the capacity of the airspace especially at levels where they operate most cost effectively [1]. Thus, the altimetry system performance as well as the aircraft height-keeping performance should be monitored by independent radar systems in order to grant approval to fly at reduced VSM [2]. Between 1997 and 2011, RVSM was implemented all over the world for aircrafts with certified altimeters and autopilots.

Radar or satellite systems determine the position of an object through algorithms that are fed with range, or range difference (RD), or range sum, or range and bearing measurements, and they estimate the object position vector employing appropriate techniques such as triangulation or trilateration. The radar measurements are contaminated by two kinds of errors: random and systematic errors. The effect of random errors can be reduced by the use of appropriate noise rejection filters, such as a Kalman filter. The second kind of error is usually removed by calibrating the instrument if this is possible; otherwise, suitable algorithms must be invented to anticipate it.

The estimation may be derived in ground stations or on board the aircraft according to where the data acquisition and processing is performed. In the latter case, the vertical position estimation has to be downlinked to the appropriate ATC center. Also, the estimation may be performed off-line or on-line. Ground-based methods or systems are the navigation accuracy measurement system (NAMS), the height estimation method with a single ATC radar, the method with one SSR and one omnidirectional radar,

the dual-synchronized autonomous height monitoring system (DAMS), and methods that use multiple SSRs and estimate the height by quadrilateration, or by the use of pseudorange measurements, or by the use of RD measurements. Onboard height measurement methods derive their estimates by trilateration using the distance measuring equipment (DME) or the Global Positioning System (GPS) signals.

15.4 Ground-Based Height Estimation

The radars used to derive the original measurements are either primary or SSR. Primary radar sends a signal and scans for the arrival of its reflection. The range to the object reflecting the signal is derived from the time elapsed between transmission and reception. With secondary radar, the radar sends an interrogation to aircraft—to all aircraft or to a selected one—and the appropriate aircraft sends a reply via its transponder. The range to the aircraft is computed from the time elapsed between the signal transmission and the signal arrival, taking into account the nominal delay time of the transponder. Most of the methods estimating the aircraft height make use of the SSR equipment because it is cost-effective, the transponder reply signal is stronger than that reflected to a primary radar, and the system can operate more reliably in dense traffic areas.

Any systematic errors in the primary radar and in the ground equipment of the SSR can be corrected by calibration. However, the problem encountered with SSR is that it involves the transponder delay time in the range measurement process. Thus, any systematic error in the transponder delay time causes range bias errors that are different for each aircraft, and thus, suitable methods must be used to anticipate for it in the subsequent measurement data processing.

15.4.1 Navigation Accuracy Measurement System

Nagaoka has proposed an off-line height estimation system [3,4]. It is composed of a primary marine radar located under an air lane and measures range R and depression angle β. Figure 15.1 shows the geometry of the system. The antenna rotates about a vertical axis and scans the area above it with rate equal to 1 rotation per 3 s. The principle exploited to derive the height estimate is that the range varies as the aircraft passes through the data acquisition area. The rate of change of range is mainly a function of the flight height z and secondarily of depression angle. It is easily derived from Figure 15.1 that the relation between the aforementioned quantities and the position x along the x-axis at time t is

$$R(t) = \sqrt{x^2(t) + R'(t)^2} = \sqrt{x^2(t) + \frac{z^2}{\cos^2 \beta(t)}} \tag{15.3}$$

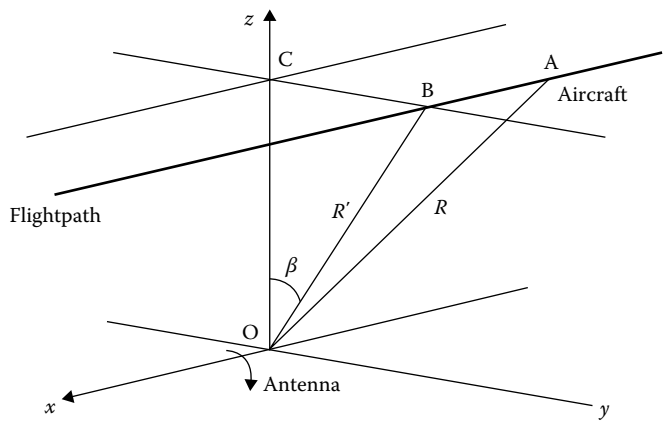


FIGURE 15.1 Geometry of the NAMS. BA = x, CB = y, OC = z, $R' = z/\cos(\beta)$.

Let x_0 denote the position of the aircraft at time t_0. Assuming the aircraft flies in straight and level flight, this means that the velocity V_x, the depression angle, and the height z remain constant during the data collection period. The aforementioned quantities and the range measurement at time t_t are related through Equation 15.4:

$$R_i = \sqrt{[x_0 + V_x(t_i - t_0)]^2 + \frac{z^2}{\cos^2 \beta}} = f(t_i, \beta, q) = f_i(q) \tag{15.4}$$

where $\mathbf{q}$ is the unknown quantities vector, $\mathbf{q} = [x_0, V_x, Z]^T$.

Measurements of R_i and β are collected at times $t_0, t_v, \ldots, t_n$, and their set is briefly expressed as a vector function of the unknown quantities with the following matrix equation:

$$\mathbf{R} = \mathbf{f}(\mathbf{q}) \tag{15.5}$$

where

$$\mathbf{R} = \begin{bmatrix} R_0 \\ R_1 \\ . \\ R_n \end{bmatrix} \quad \mathbf{f} = \begin{bmatrix} f_0 \\ f_1 \\ . \\ f_n \end{bmatrix}$$

Equation 15.5 is nonlinear. Thus, a nonlinear least squares method, such as the Gauss–Newton iterative method, must be used to estimate the unknown vector. Let $\mathbf{q_k}$ be the estimate at the kth iteration. Then, the next estimate is

$$\hat{\mathbf{q}}_{\mathbf{k+1}} = \hat{\mathbf{q}}_{\mathbf{K}} + (\mathbf{F}^{\mathbf{T}}\mathbf{F})^{-1}\mathbf{F}^{\mathbf{T}}[\mathbf{R} - \mathbf{f}(\hat{\mathbf{q}}_{\mathbf{K}})] \tag{15.6}$$

where $\mathbf{F}$ is the partial derivatives (Jacobian) matrix; that is,

$$\mathbf{F} = \frac{\vartheta f}{\vartheta q} = \begin{bmatrix} \frac{\vartheta f_0}{\vartheta x_0} & \frac{\vartheta f_0}{\vartheta V_x} & \frac{\vartheta f_0}{\vartheta z} \\ . & . & . \\ . & . & . \\ \frac{\vartheta f_n}{\vartheta x_0} & \frac{\vartheta f_n}{\vartheta V_x} & \frac{\vartheta f_n}{\vartheta z} \end{bmatrix} \tag{15.7}$$

Assuming typical aircraft velocities and FL and taking into account the antenna fan beam and rotation rate, it is determined that the observation time is approximately 1 min and the number of measurements is approximately 20. The standard deviation (SD) of the height estimation errors is σ_z = 10 m when the SD of the measurement errors of range and depression angle are σ_R = 10 m and σ_β = 0.08°, respectively.

The system has been developed and operates at the Electronic Navigation Institute of Tokyo, Japan. To estimate and anticipate for systematic errors in the depression angle and range measurements, it has been proposed to perform 40 experimental flights above the radar range. The basic assumption of the NAMS system (i.e., that the aircraft flies at constant height) cannot be validated by the NAMS itself. Thus, an SSR is required to confirm that the flight is indeed performed at the same level via the mode C interrogations [4].

15.4.2 Height Estimation with Single Air Traffic Control Radar

Nagaoka at the Electronic Navigation Institute of Tokyo investigated the possibility to derive height estimation with a very simple and inexpensive method [5]. The idea is to use single radar normally used for ATC. The radar antenna rotates around a horizontal axis once every 4 s and measures range R and bearing β, as shown in Figure 15.2.

Let (x, y, z) be the 3-D position of the aircraft. The bearing angle is defined as

$$\beta = \cos^{-1}\left(\frac{y}{\sqrt{x^2 + y^2}}\right) \tag{15.8}$$

Thus, the relation between the measurements and the unknown quantities is

$$R = \left(z^2 + x^2 + y^2\right)^{1/2} = \left(z^2 + \frac{y^2}{\cos^2\beta}\right)^{1/2} \tag{15.9}$$

The measurement principle is similar to that of NMAS, that is, that during the time the aircraft passes through the data acquisition area, the range varies as a function of both the horizontal position and the height. When straight and level flight is assumed, the unknown quantities are the height z, the y-axis initial position y_0, and the velocity V_y. The expression relating the measurements R_i and β_i and the unknown quantities vector $\mathbf{q} = [y_0, V_y, z]^T$ at time t_i is

$$R_i = \left(z^2 + \frac{\left[y_0 + V_y\left(t_i - t_0\right)\right]^2}{\cos^2\beta_i}\right) = f\left(t_i, \beta, \mathbf{q}\right) = f_i(\mathbf{q}) \tag{15.10}$$

The measurements are collected as the aircraft flies close to the radar at a distance of 30 km. Approximately, the data acquisition area is 15 km, the observation time is 1 min, and the total number of scans is 15. The total set of measurements is expressed as a nonlinear matrix function in the form of Equation 15.5; that is, $\mathbf{R} = \mathbf{f}(\mathbf{q})$, where $\mathbf{R}$ denotes the measurement vector. The unknown vector is estimated using a nonlinear least squares method like that of the NAMS system expressed by Equation 15.6.

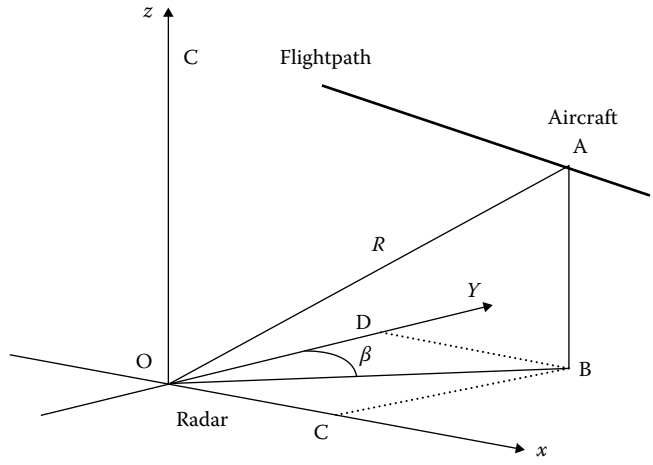


FIGURE 15.2 Geometric configuration of the height measurement system with a single ATC radar. OC = x, OD = y, AB = z, and $\beta = \widehat{\text{BOD}}$.

The system performance is evaluated with Monte Carlo simulations. The best accuracy achieved is σ_z = 100 m under the assumption that the SD errors of range and bearing measurements are σ_R = 100 m and σ_β = 0.06°. The method can use either primary or secondary radar. However, in the latter case, the method has to include systematic delays by the transponder.

15.4.3 Height Estimation with One SSR and One Omniradar

Manolakis et al. investigated the possibility to estimate the aircraft flight height with a system that consists of one standard SSR plus a secondary radar equipped with an omnidirectional antenna located far away from the SSR; hence, this system could be referred to as one SSR and one omnidirectional radar (SSROR) [6]. Figure 15.3 shows a typical geometric configuration of the system. The SSR measures bearing θ and slant range R_s, which is transformed to horizontal range R with correcting lookup tables. The omniradar measures range R_a. Let (x, y, z), (x_a, y_a, z_a), and (0, 0, 0) be the coordinates of aircraft, omniradar, and SSR, respectively. Then, from Figure 15.3, it follows that the theoretical height function is

$$z = f(R_a, x, y) = z_a + [R_a^2 - (x - x_a)^2 - (y - y_a)^2]^{1/2} \tag{15.11}$$

$$x = R\sin\theta \tag{15.12}$$

$$y = R\cos\theta \tag{15.13}$$

However, the height measurement obtained through the preceding formula will be biased because of the transponder delay systematic deviation from its nominal value that affects the range measurements. Since this bias cannot be removed by calibration, a suitable algorithm has to be applied to anticipate for it. The approach proposed is to augment the unknown vector by incorporating the bias term. Let b denote the bias in range measurements. The biased measurements R_b and R_{ab} of R and R_a, respectively, are expressed as

$$R_b = R + b\cos\phi \tag{15.14}$$

$$R_{ab} = R_a + b \tag{15.15}$$

where ϕ is the elevation angle. Let s denote the squared height:

$$s = (z - z_a)^2 \tag{15.16}$$

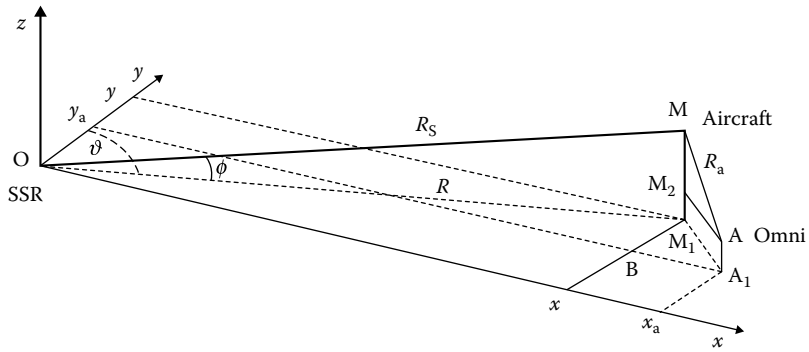


FIGURE 15.3 Geometry of the height estimation system with SSROR. The distance x_a of the radars typically ranges from 80 km up to 160 km. $M_1M = z$.

and s_b denote the corresponding quantity derived from (biased) range measurements. Then, after some manipulations of the preceding equations, the following relation is obtained:

$$s_b = s + ba \tag{15.17}$$

$$a = 2[R_a - \cos\phi(R - x_a \sin\theta - y_a \cos\theta)] \tag{15.18}$$

The term a is the bias multiplying factor determined by the relative geometry of the system. Equation 15.17 is a linear relation between measurement s_b and the unknown quantities s and b. The effect of the bias varies as the aircraft passes through the surveillance area. Its form is mainly determined by the flight height. Consequently, it is possible to estimate both the height and the bias by collecting data during the period the aircraft remains in the surveillance area. The measurement equation at time t_i is

$$s_{bi} = s_i + ba_i + e_i \tag{15.19}$$

where e_i represents the effect of the random measurement errors; hence, it could be referred to as the equation error. Assuming level flight, $s_i = s$, and the set of collected data is expressed with the following linear matrix equation:

$$\mathbf{s_b} = \mathbf{Aq} + \mathbf{e} \tag{15.20}$$

$$\mathbf{s_b} = \begin{bmatrix} s_{b0} \\ s_{b1} \\ . \\ s_{bn} \end{bmatrix} \quad \mathbf{A} = \begin{bmatrix} 1 & a_0 \\ 1 & a_1 \\ . & . \\ 1 & a_n \end{bmatrix} \quad \mathbf{e} = \begin{bmatrix} e_0 \\ e_1 \\ . \\ e_n \end{bmatrix} \quad \mathbf{q} = \begin{bmatrix} s \\ b \end{bmatrix}$$

The best estimate $\hat{\mathbf{q}}$ of $\mathbf{q}$, minimizing the weighted sum of the squared errors $\mathbf{e}^T\mathbf{We}$, is

$$\hat{\mathbf{q}} = (\mathbf{A}^T\mathbf{WA})^{-1}\mathbf{A}^T\mathbf{Ws_b} \tag{15.21}$$

where $\mathbf{W}$ is the weighting matrix defined as the inverse of the equation error covariance matrix:

$$\mathbf{W}^{-1} = \mathrm{E}\left\{\mathbf{ee}^T\right\} \tag{15.22}$$

The notation $E\{\}$ stands for the expected value operation. Thus, the estimate of the geometric height is

$$\hat{z} = \sqrt{\hat{q}_1} = \sqrt{\hat{s}} \tag{15.23}$$

In the case of nonlevel flights, an augmented unknown vector should be used that comprises bias b, initial height Z_0, and vertical velocity V_z. In this case, however, the measurement equation is nonlinear in terms of the unknown quantities; hence, a nonlinear least squares iterative algorithm must be employed.

The system performance is a function of (1) the range and bearing measurement SD errors, σ_R and σ_θ, respectively; (2) the aircraft velocity, which affects the number of scans; (3) the omniradar scan rate; and (4) geometric terms such as the FL and the distance between the radars. By assuming that $\sigma_R = 70$ m and $\sigma_\theta = 0.08°$, the SD of the height estimation error will be between 50 and 100 m when the values of the other factors lie in reasonable ranges.

15.4.4 Dual-Synchronized Autonomous Monitoring System

The research for this system has been conducted in the National Aerospace Laboratory NLR, the Netherlands [7]. The system is composed of two primary marine radars. Each of the radars has a rotating antenna. The antennas scan, with different directions, the same volume—that is an area 3 km × 10 km above the location site. The primary surveillance area is 3 km × 3 km. The two antennas (A and B) are separated by 2.5 m, mechanically synchronized, and scan once every 2.5 s. Their data extractors produce measurements of slant ranges and elevation angles, (R_A, ϕ_A) and (R_B, ϕ_B), respectively, for each radar. The tracking software derives estimation of the aircraft position and trajectory in off-line mode. At each scan, a combination of four measurements $(R_A, \phi_A, R_B, \phi_B)$ is available to be used in the estimation of the unknown vector $\mathbf{q} = [x_0, y_0, z_0, V_x, V_y, V_z]^T$. A local Cartesian frame is used to perform the calculations. Finally, 16 equations are available to be solved for the 6 unknowns with a weighted least squares method. The weight of each measurement is the measured amplitude of the radar pulse. The maths of the tracker is not presented in [7]. However, one approach could be the following. If a coordinate system is defined such that radar A is at the origin (0, 0, 0) and radar B is at $(x_B, 0, 0)$, then the measurements obtained at scan time t_i can be expressed as

$$R_{Ai} = \sqrt{x_i^2 + y_i^2 + z_i^2} = \sqrt{[x_0 + V_x(t_i - t_0)]^2 + [y_0 + V_y(t_i - t_0)]^2 + [z_0 + V_z(t_i - t_0)]^2} \tag{15.24a}$$

$$\phi_{Ai} = \tan^{-1}\left[\frac{z_i}{\sqrt{x_i^2 + y_i^2}}\right] = \tan^{-1}\left[\frac{z_0 + V_z(t - t_i)}{\sqrt{[x_0 + V_x(t - t_i)]^2 + [y_0 + V_y(t - t_i)]^2}}\right] \tag{15.24b}$$

$$\begin{aligned} R_{Bi} &= \sqrt{(x_i - x_B)^2 + y_i^2 + z_i^2} \\ &= \sqrt{[x_0 + V_x(t_i - t_0) - x_B]^2 + [y_0 + V_y(t_i - t_0)]^2 + [z_0 + V_z(t_i - t_0)]^2} \end{aligned} \tag{15.24c}$$

$$\phi_{Bi} = \tan^{-1}\left[\frac{z_i}{\sqrt{(x_i - x_B)^2 + y_i^2}}\right] = \tan^{-1}\left[\frac{z_0 + V_z(t - t_i)}{\sqrt{[x_0 + V_x(t - t_i)]^2 + [y_0 + V_y(t - t_i)]^2}}\right] \tag{15.24d}$$

Notice that x_B has a small value (2.5 m) compared to the magnitude of the other quantities; consequently, it can be neglected. The preceding set of equations at four different times yields 16 equations that can be solved with a nonlinear least squares method such as the Gauss–Newton iterative method presented in Equation 14.6.

The SD of the height estimation error will be less than 15 m in the primary surveillance area and 30 m at the edges of the area when the SD of range and elevation angle measurements are smaller than 10 m and 0.1°, respectively.

15.4.5 Height Measurement by Quadrilateration

Rice proposed a system consisting of four synchronized receiving SSR stations S_i, $i = 0, 1, 2, 3$, that use SSR transmissions from the aircraft transponder and estimate the height by quadrilateration [8]. Figure 15.4 shows a typical configuration of systems composed of N SSRs. One of them is an active station, which means that this station has both an interrogator and a receiver. Let (x, y, z) and (x_i, y_i, z_i) denote the Cartesian coordinates of aircraft and station S_i, respectively. Also let R_i denote the range from station S_i to aircraft, c denote the velocity of the light, and T_s denote the time of signal transmission

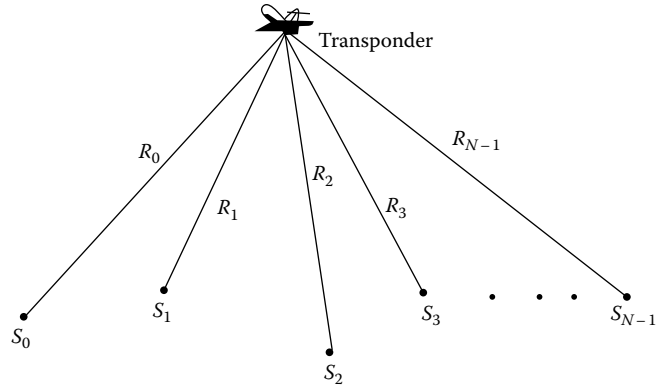


FIGURE 15.4 Typical configuration of the height estimation systems that are based on N SSR stations. One of the stations is active, that is, it both transmits the interrogations and receives the replies, whereas the other stations are receivers only.

from aircraft transponder. The stations measure the time of arrival (TOA) T_i, $i = 0, 1, 2, 3$, of the aircraft transponder signal at each site. The following relations hold:

$$T_i = T_s + \frac{R_i}{c} \quad i = 0, 1, 2, 3 \tag{15.25}$$

$$R_i^2 = (x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2 \quad i = 0, 1, 2, 3 \tag{15.26}$$

The preceding system of eight equations can be solved for the unknown quantities. The unknown quantities used by Rice are $(R_0, R_1, R_2, R_3, x, y, z, T_s)^T$. However, an equivalent approach is to substitute for R_i in Equation 15.25, which becomes

$$T_i = T_s + \frac{1}{c}\sqrt{(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2} = f_i(\mathbf{q}) \quad i = 0, 1, 2, 3 \tag{15.27}$$

where $\mathbf{q} = [x, y, z, T_s]^T$ is the unknown vector. Thus, there are four nonlinear equations to be solved for $\mathbf{q}$. One suitable method, for example, is the Newton–Raphson method, which iteratively approximates the solution via the following formula:

$$\mathbf{q_{k+1}} = \mathbf{q_k} + \mathbf{F}(\mathbf{q_k})^{-1}[\mathbf{T} - \mathbf{f}(\mathbf{q_k})] \tag{15.28}$$

where $\mathbf{T} = [T_0, T_1, T_2, T_3]^T$ is the measurement vector and $\mathbf{F}$ is the Jacobian matrix:

$$\mathbf{F} = \frac{\vartheta f}{\vartheta q} = \begin{bmatrix} \frac{\vartheta f_0}{\vartheta x} & \frac{\vartheta f_0}{\vartheta y} & \frac{\vartheta f_0}{\vartheta z} & \frac{\vartheta f_0}{\vartheta T_s} \\ . & . & . & . \\ \frac{\vartheta f_3}{\vartheta x} & \frac{\vartheta f_3}{\vartheta y} & \frac{\vartheta f_3}{\vartheta z} & \frac{\vartheta f_3}{\vartheta T_s} \end{bmatrix} \tag{15.29}$$

Notice that the time of interrogation transmission, as well as the transponder nominal delay time, is not involved in the measurements. The measured quantities are only the TOAs at the station sites. Thus, the height estimate is not affected by any transponder bias.

The theoretical and experimental research has been conducted at the GEC-Marconi Research Center, U.K. The optimum station arrangement is to locate the three of them equispaced on a circle and the

fourth in the middle. The typical circle radius is 35 km. The same magnitude holds for the measurement range. The method may be implemented in on-line or off-line mode. In the first case, there must be transmitters at the stations to transmit their measurements of TOA to the height monitoring center. The vertical dilution of precision (VDOP) is a performance index defined by the ratio:

$$\text{VDOP} = \frac{\sigma_z}{c\sigma_{\text{rte}}} \tag{15.30}$$

where σ_{rte} is the SD of the relative timing errors. The VDOP expresses the effect of the relative geometry to system performance. The VDOP of this system achieves a typical value of 3.

15.4.6 Height Estimation with SSRs and Pseudoranges

This approach has been investigated by Nagaoka at Electronic Navigation Institute of Tokyo [9]. The system consists of N SSR receiving stations like the previous system; see again Figure 15.5. Again one of them, say S_0, must be active to transmit interrogations to aircraft. The time of interrogation transmission, T_t, and the time of signal arrivals at the receiving stations T_i, $i = 0, 1, \ldots, N-1$, are measured. Thus, N pseudorange measurements r_i are obtained where $r_i = c(T_i - T_t)$. Let T_D denote the transponder delay and D denote the distance corresponding to this delay, $D = cT_D$. Then, for each pseudorange measurement r_i, the following relation holds:

$$\begin{aligned} r_i &= D + R_i + R_0 + = D + \sqrt{(x-x_i)^2 + (y-y_i)^2 + (z-z_i)^2} + \sqrt{(x-x_0)^2 + (y-y_0)^2 + (z-z_0)^2} \\ &= f_i(\mathbf{q}) \quad i = 0, 1, ..., N-1 \end{aligned} \tag{15.31}$$

where $\mathbf{q} = [x, y, z, D]^{\text{T}}$ is the unknown vector. The set of N measurements $\boldsymbol{\rho} = [\rho_0, \rho_1, \ldots, \rho_{N-1}]^{\text{T}}$, $N \geq 4$, and the unknown vector are related through Equation 15.30:

$$\mathbf{r} = \mathbf{f}(\mathbf{q}) \tag{15.32}$$

The unknown vector $\mathbf{q}$ can be obtained from the solution of Equation 15.32 with a nonlinear weighted least squares method. Thus, the best estimate of $\mathbf{q}$ is iteratively calculated as

$$\hat{\mathbf{q}}_{\mathbf{k+1}} = \hat{\mathbf{q}}_{\mathbf{k}} + (\mathbf{F}^{\mathbf{T}}\mathbf{F})^{-1}\mathbf{F}^{\mathbf{T}}[\mathbf{r} - \mathbf{f}(\hat{\mathbf{q}}_{\mathbf{k}})] \tag{15.33}$$

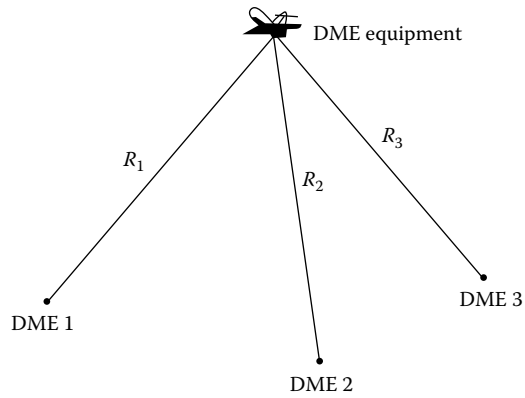


FIGURE 15.5 Configuration of the onboard height estimation system that utilizes the distance measurements derived from the DME.

where **F** is the Jacobian matrix

$$\mathbf{F} = \frac{\vartheta \mathbf{f}}{\vartheta \mathbf{q}} = \begin{bmatrix} \frac{\vartheta f_0}{\vartheta x} & \frac{\vartheta f_0}{\vartheta y} & \frac{\vartheta f_0}{\vartheta z} & \frac{\vartheta f_0}{\vartheta D} \\ . & . & . & . \\ \frac{\vartheta f_{N-1}}{\vartheta x} & \frac{\vartheta f_{N-1}}{\vartheta y} & \frac{\vartheta f_{N-1}}{\vartheta z} & \frac{\vartheta f_{N-1}}{\vartheta D} \end{bmatrix} \tag{15.34}$$

The estimate of **q** is free of the transponder delay systematic error because the estimation is based not on the nominal delay, but on the actual delay time, which is one of the parameters to be estimated, whereas the rest of the parameters are the aircraft 3-D position coordinates x, y, z.

The station arrangement proposed by Nagaoka, when there are four stations, is an equilateral triangle formed by the three stations, whereas the fourth station is located in the center. The VDOP, defined as the ratio σ_z/σ_R (where σ_R is the observation error) has a typical value of 4 when the aircraft is above the center at a height equal to the baseline radius. The VDOP increases as the aircraft flies higher and longer and as the baseline radius becomes smaller.

15.4.7 Height Measurement with SSRs and Range Differences (RDHMU)

This approach has been proposed in [10,11]. The system consists of $N-1$ receiving SSR stations S_i, $i = 1, N-1$, and one station, say S_0, which is both receiver and interrogator, like that in Figure 15.4. The stations receive the reply and the time difference of arrival (TDOA) between a reference station, say S_0, and station S_i is measured. A set of $N-1$ TDOA or, equivalently, RD measurements, is collected at each time the transponder sends a reply signal. The height estimation derived from this set of measurements is not affected by any transponder delay systematic error since this error is inherently subtracted from the measurements used. This system could be referred to as RD height monitoring unit (RDHMU). The systems that derive the position fix based on this kind of measurement are known as TDOA or hyperbolic systems.

Let τ_i denote the TDOA between stations S_i and S_0, and d_i denote the corresponding RD measurement, $d_i = c\tau_i$. The following relation holds:

$$\begin{aligned} d_i &= R_i - R_0 = \sqrt{(x-x_i)^2 + (y-y_i)^2 + (z-z_i)^2} - \sqrt{(x-x_0)^2 + (y-y_0)^2 + (z-z_0)^2} \\ &= f_i(\mathbf{q}) \quad i = 1, 2, ..., N-1 \end{aligned} \tag{15.35}$$

where $\mathbf{q} = [x, y, z]^T$ is the unknown aircraft position vector. The vector of RD measurements $\mathbf{d} = [d_1, d_2, \ldots, d_m)$ is expressed as

$$\mathbf{d} = \mathbf{f}(\mathbf{q}) \tag{15.36}$$

A commonly employed method to solve for **q** in this nonlinear equation is the Taylor series method or equivalently the Gauss–Newton iterative method. The best estimate of **q** is iteratively approximated as

$$\hat{\mathbf{q}}_{\mathbf{k}+1} = \hat{\mathbf{q}}_{\mathbf{k}} + (\mathbf{F}^{\mathbf{T}}\mathbf{F})^{-1}\mathbf{F}^{\mathbf{T}}[\mathbf{d} - \mathbf{f}(\hat{\mathbf{q}}_{\mathbf{k}})] \tag{15.37}$$

where **F** is the Jacobian matrix:

$$\mathbf{F} = \frac{\vartheta \mathbf{f}}{\vartheta \mathbf{q}} = \begin{bmatrix} \frac{\vartheta f_1}{\vartheta x} & \frac{\vartheta f_1}{\vartheta y} & \frac{\vartheta f_1}{\vartheta z} \\ \frac{\vartheta f_{N-1}}{\vartheta x} & \frac{\vartheta f_{N-1}}{\vartheta y} & \frac{\vartheta f_{N-1}}{\vartheta z} \end{bmatrix} \tag{15.38}$$

In the case of four stations, the best arrangement is an equilateral triangle with the fourth station in the center. The SD of height estimation error σ_z will be 15 m when the baseline radius is 6 km, the flying height is 9 km, and σ_{TDOA} is 10 ns.

Work on proof of principles and system development of an HMU based on the concept of TDOA measurement of SSR signals has been conducted by Roke Manor Research Ltd., U.K. [12].

15.5 Onboard–Derived Height Estimation

15.5.1 Height Measurement with Distance Measuring Equipment

With this approach, the geometric height is derived onboard the aircraft using DME [13,14]. Using the DME interrogation equipment, the distance from three DME ground stations is measured onboard (see Figure 15.5). The three stations are located under an airway. The height is then computed from the range measurement vector $\mathbf{R} = [R_1, R_2, R_3]^{\text{T}}$ by trilateration. An exact and efficient solution of the nonlinear measurement equation was derived in [14]. Specifically, the height is computed from the closed form:

$$z = g(\mathbf{R}) = \frac{-b(\mathbf{R}) + \sqrt{d(\mathbf{R})}}{2a} \tag{15.39}$$

where $b(\mathbf{R})$ and $d(\mathbf{R})$ are the following simple polynomial-type functions:

$$b(\mathbf{R}) = b_0 + b_1 R_1^2 + b_2 R_2^2 + b_3 R_3^2 \tag{15.40}$$

$$d(\mathbf{R}) = d_{00} + d_{01}R_1^2 + d_{02}R_2^2 + d_{03}R_3^2 + d_{11}R_1^4 + d_{22}R_2^4 + d_{33}R_3^4 + d_{12}R_1^2R_2^2 + d_{13}R_1^2R_3^2 + d_{23}R_2^2R_3^2 \tag{15.41}$$

The coefficients a, b_i, d_{ij} are analytically defined in the appendix of [15]. An important aspect of these coefficients is that they are completely defined by the ground stations' coordinates (x_i, y_i, z_i), which are fixed. Thus, the coefficients are calculated only once at the moment the aircraft enters the data acquisition area. Then, every time a new set of range measurements is available, the height is computed from the aforementioned equations using the range measurements and the stored coefficients. Define the ratio σ_z/σ_R as the VDOP of this technique, where σ_R is the SD of the ranging error. The VDOP is 1 in the case where the stations form an equilateral triangle inscribed in a circle with 10 km radius and the aircraft is above the triangle center at a height of 8 km.

15.5.2 Estimation of Vertical Position with the Global Positioning System

The height could be also determined with GPS receivers with four or more satellites. However, the vertical position determined using GPS alone is not precise. Thus, an augmentation method is required such as differential GPS. The VDOP, defined as $\sigma_z/\sigma_{\text{D}}$, depends on the geometry that varies continuously, even in the case of a fixed receiver, because the satellites are not geostationary but move in such orbits as to complete a rotation in 12 h. The precision is very bad when all the available satellites happen to be close to the horizon. The world mean value of VDOP is about 2 [15]. However typical VDOP values range from 1.5 to 7, depending on the area of the receiver and on the time of day. The details for position determination with satellite systems such as GPS, GLONASS, and Galileo are presented in a subsequent specific section of this chapter.

15.6 Special Topics

The performance analysis of the various parameter estimation systems is usually restricted to the variance analysis, and the estimation error is usually assumed to have zero mean value. However, it is proven that in all of the aforementioned systems, the estimation error does not have zero mean value due to the nonlinearity of the measurement equation. Another important aspect is that there are cases where it is not possible to obtain a solution due to the relative geometry of aircraft vs. stations that leads to large errors. In these cases, the successive iterations applied to solve the system of nonlinear equations may not converge. Also, even after convergence, the solutions need not necessarily be "the best" or the "correct ones."

15.6.1 Inherent Bias

The nonlinearity of the systems, joined with the measurement random errors, causes inherently biased estimations although the measured quantities are unbiased. For example, take the case of height estimation with DME measurements. The function $g(\mathbf{R})$ in Equation 15.39 determines that the height z is nonlinear. In addition, range measurements will be contaminated by additive zero mean value random errors. Let $\mathbf{R}_m$ denote the noisy measurement vector and z_m denote the height measurement derived from $g(\mathbf{R}_m)$, that is, from the measurement function when it is fed with noisy measurements. Extending $g(\mathbf{R}_m)$ in a Taylor series around the actual values of ranges R_1, R_2, R_3 up to second-order terms and taking the expected values, it is derived that the expected value of z_m will not be the actual value z, but it will differ by an amount b_z, which is called inherent bias. Specifically, for the DME case, the inherent bias is evaluated as

$$b_z = E[z_m] - z = \frac{1}{2}\left[\frac{\partial^2 g}{\partial R_1^2}\sigma_{R_1}^2 + \frac{\partial^2 g}{\partial R_2^2}\sigma_{R_2}^2 + \frac{\partial^2 g}{\partial R_3^2}\sigma_{R_3}^2\right] \tag{15.42}$$

Figure 15.6 shows the inherent bias generated in the case of the DME system. The inherent bias becomes larger as the magnitude of the measurement errors and the system nonlinearity becomes stronger. This bias error is inherently generated in all position estimation algorithms and must be taken into account when precise position estimation is required.

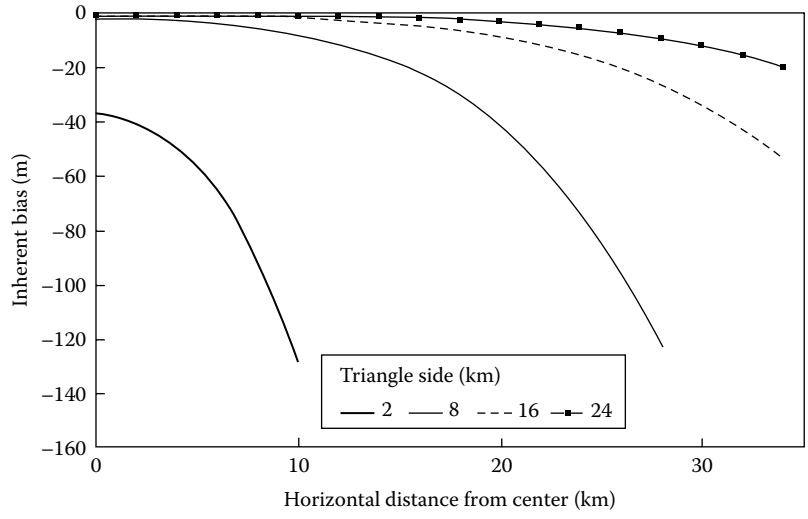


FIGURE 15.6 Bias inherently generated by the height estimation algorithm of the system based on DME measurements. The stations' sites form an equilateral triangle. The inherent bias is shown as function of the horizontal distance from the center for various magnitudes of the triangle side. The flight height is $z = 9$ km, and the SD of the distance measurement error is $s_R = 90$ m.

15.6.2 Existence and Uniqueness of Position Fix

In some systems, there are singular cases for which it is not possible to achieve a position fix. This fact has been reported by Abel and Chafee for the GPS system in [16], where it is shown that for some satellites/aircraft relative geometries, it is not possible to solve the relevant equation or there is more than one solution. Also, for the RDHMU, it has been shown that there are some station arrangements for which it is not possible to derive height estimation when the aircraft is at specific areas [11]. For example, in the case of four stations, when the quadrilateral defined by the stations is inscribed in a circle, it is not possible to estimate the height when the aircraft is above the center. Also, when the quadrilateral is a rectangle or symmetric trapezoid, it is not possible to derive a position fix when the aircraft is above the line that passes from the middle of the parallel sides. From a mathematical point of view, this singularity is expressed by the singularity of the Jacobian matrix; consequently, this matrix cannot be inverted as is required by the relevant position estimation algorithm. The algorithm in this case diverges from the actual height and finally collapses. Notice that height estimation is achieved everywhere except at this singular point. However, when the aircraft is close to the singular region, although a position fix is achieved, it is not actually reliable since it is affected by large errors; for example, the VDOP could be larger than 600 in regions close to the singular region.

Table 15.1 presents the institutes and companies that either investigate and develop prototype height monitoring units or provide relevant systems in the market.

TABLE 15.1 Height Monitoring Systems and Companies/Institutes that Develop and Provide Them

System	Company/Institute	
NAMS	Electronic Navigation Research Institute	
	Ministry of Transport, 6-38-1 Shinkawa, Mitaka, Tokyo, 181, Japan. Tel: +81 422 413171, Fax: 81-422-413176	
DAMS height monitoring unit	National Aerospace Laboratory NLR Anthony Fokkerweg 2, 1059 CM Amsterdam, the Netherlands. Tel: +31 (0)20 511 3113, Fax: +31 (0)20 511 3210	
SSR and quadrilateration technique	GEC-Marconi Electronics Ltd. Marconi Research Laboratories, West Hanningfield Road, Great Baddow, Chelmsford Essex, England	
SSR and TDOA technique (RDHMU)	Roke Manor Research Ltd. Roke Manor, Romsey, Hampshire S051 OZN, U.K. Tel: +44(0)794 833000, Fax: +44(0)794 833433	
GPS providers	Trimble Navigation	Trimble Navigation Europe Ltd.
	585 North Mary Avenue, Sunnyvale, CA 94086. Tel: (408) 730-2900	79-81 High Str., West Mailing, Kent ME19 6NA, U.K. Tel: +44(0)732 849242, Fax: +44(0)732 847437
	Rockwell International	Rockwell Semiconductor System
	Digital Communication Division, 4311 Jamboree Road, Newport Beach, CA 92660-3095. Tel: (714) 221-4600, Fax: (714) 221-6375	Berkshire Court, Western Road, Bracknell, Berkshire RG12 IRE, England. Tel: +44(0)1344 48644, Fax: +44(0)1344 48655

References

1. ICAO doc 9574, *Manual on Implementation of a 300 m (1000 ft) Vertical Separation Minimum between FL 290 and FL 410 Inclusive*, 2nd edn., 2002, ICAO, Montreal, retrieved from http://www.skybrary.aero/bookshelf/books/1311.pdf (accessed on March 14, 2013).
2. FAA, Reduced vertical separation minimum (RVSM), http://www.faa.gov/about/office_org/headquarters_offices/ato/service_units/enroute/rvsm/ (accessed on March 14, 2013).
3. S. Nagaoka, E. Yoshioka, and P. T. Muto, A simple radar for navigation accuracy measurements, *J. Navigation*, 34(3), 462–469, 1981.
4. S. Nagaoka, Possibility of detecting a non-level-flight aircraft by the navigation accuracy measurement system (NAMS), *ICAO, Review of the General Concept of Separation Panel, 7th Meeting*, Montreal, Canada, RGCSP-WP/180, October 30, 1990.
5. S. Nagaoka, Height estimation of a cruising aircraft via a radar for air traffic control, *Electronics and Communications in Japan, Part 1 (Communications), USA*, 71(11), 95–105, 1988.
6. D. E. Manolakis, C. C. Lefas, G. S. Stavrakakis, and C. M. Rekkas, Computation of aircraft geometric height under radar surveillance, *IEEE Transactions on Aerospace and Electronic Systems*, 28(1), 241–248, 1992.
7. J. Brugman, J. Verpoorte, and A. I. L. Willekens, DAMS height monitoring unit-phase one, Report CR 92328 C, NLR, the Netherlands, 1992.
8. D. E. Rice, Height measurement by quadrilateration, *The Marconi Review*, XLVI(228), 1–17, 1983.
9. S. Nagaoka, Possibility of geometric height measurement by using secondary surveillance radars, *ICAO, Review of the General Concept of Separation Panel, 7th Meeting*, Montreal, RGCSP-WP/181, October 30, 1990.
10. D. E. Manolakis and C. C. Lefas, Aircraft geometric height computation using secondary surveillance radar range differences, *IEEE Proceedings Radar, Sonar and Navigation*, 141(2), 119–124, 1994.
11. D. E. Manolakis and C. C. Lefas, Station arrangement effects on ground referenced height computation by using time differences, *Navigation*, *Journal of the Institute of Navigation*, 42(2), 409–420, 1995.
12. L. G. Hopkins, D. Sherry, and D. C. Rickard, Geometric Height Monitor Unit (HMU) Programme—Final Report on Phase 1, Proof of Principles, Roke Manor Research Ltd., Report No. 72/91/R161IU, Roke Manor, U.K., 1991.
13. D. E. Manolakis, Efficient solution and performance analysis of 3-D position estimation by trilateration, *IEEE Transactions on Aerospace and Electronic Systems*, 32(4), 1239–1248, 1996.
14. D. E. Manolakis and A. I. Dounis, Advances in aircraft height computation using distance measuring equipment, *IEEE Proceedings Radar, Sonar and Navigation*, 143(1), 47–52, 1996.
15. J. L. Leva, Relationship between navigation vertical error, VDOP, and pseudorange error in GPS, *IEEE Transactions on Aerospace and Electronic Systems*, 30(4), 1138–1142, 1994.
16. J. S. Abel and J. W. Chaffee, Existence and uniqueness of GPS solutions, *IEEE Transactions on Aerospace and Electronic Systems*, 27(6), 952–956, 1991.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

16

Attitude Measurement

16.1 Introduction to Attitude Sensors for Ships, Aircraft, and Crane Lifters **16**-1
16.2 Attitude Sensors for Spacecraft Applications **16**-3
16.3 Automatic On-Line Attitude Measurement for Ships and Crane Lifters **16**-5
16.4 Attitude Measurement for Ships **16**-5
Location of Sensors and Outputs • Dynamics of Attitude Signals
16.5 Attitude Measurement for Crane Lifters **16**-8
Dynamics of Attitude Signals • Sensor Outputs and On-Line Attitude Measurement
16.6 Aircraft Attitude Determination **16**-11
Vertical and Directional Gyro Analysis • Inertial Measurement Units
16.7 Spacecraft Attitude Determination **16**-15
Attitude Determination Methodology
16.8 PALADS **16**-17
References **16**-18

Mark A. Stedham
University of Alabama at Huntsville

Partha P. Banerjee
University of Alabama at Huntsville

Seiji Nishifuji
Yamaguchi University

Shogo Tanaka
Yamaguchi University

In many practical situations, it is important to determine and measure the attitude of a particular vehicle, such as a ship, an airplane, a piece of mechanical equipment such as a crane lifter, or a spacecraft. For this reason, many attitude sensors have been developed with advanced computer and semiconductor technologies. This section first introduces the various attitude sensors with an explanation of their operating principles and then presents several methodologies for attitude measurement and determination, including ships and crane lifters, aircraft, and spacecraft applications.

16.1 Introduction to Attitude Sensors for Ships, Aircraft, and Crane Lifters

There are many types of gyroscopes that, corresponding to the physical measurement mechanisms used, may be classified as two-axis *freedom gyro* and single-axis freedom gyro using precession, *vibratory gyro* using Coriolis' force, and *optic gyro* using Sagnac effect. Among them, the two-axis freedom gyro has the longest history. It consists of a high-speed rotating rotor around a spin axis supported by two orthogonal axes. This type of gyro is generally classified as either a *free gyro*, a *vertical gyro* (VG), or a *directional gyro* (DG).

The single-axis freedom gyro has only one output axis in addition to the spin axis. Depending on the specifications (in which) the gyro is designed, there are two types of gyros, the *rate gyro* and the *rate-integrating gyro*. Related to these rotating-type gyros is another type of gyro known as the *electrostatic gyro*, which makes use of a high-speed rotating sphere in a vacuum cavity. Because of its resistance-free property, the electrostatic gyro has the highest accuracy among existing gyros. There are also rotorless gyros.

TABLE 16.1 Performance of Different Types of Gyros

Type of the Gyro	Degrees of Freedom	Quantities to Be Detected	Accuracy (° h^{-1})
Free gyro	2	Angle	1
VG	2	Declination from horizontal plane	1
DG	2	Shift from reference direction	1
Rate gyro	1	Angular velocity	10
Rate-integrating gyro	1	Angle	0.001–1
Ring laser gyro	1	Angular velocity	0.003
Fiber-optic gyro	1	Angular velocity	0.01
Electrostatic gyro	2	Angle	0.00001–0.01

The first one is a vibratory gyro that uses Coriolis' force as the measurement principle. The second type is an optical one. Among optical gyros, there are two types: the *ring laser gyro* and the *fiber-optic gyro*. Both rely on the Sagnac effect in their measurement mechanisms. The performance of gyros is evaluated by their drift rates, and the performance of various gyros is shown in Table 16.1, for reference, with their primary usages.

Recently, with the development of computer technology, many types of three-axis gyros have been developed that can measure not only the tilt angles but also the angular velocities and the accelerations along the three axes by combining several gyros and accelerometers. Accelerometers are often coupled with gyros to provide flight and ship navigation systems as well as attitude sensors for dynamic objects such as crane lifters. Examples include the *attitude and heading reference system* (AHRS), *inertial navigation system* (INS), *inertial measurement unit* (IMU), and *gyro compass* (GC), as well as the VGs and DGs discussed previously [1].

The principle of a servo-type accelerometer is explained later (see Figure 16.1). As soon as the shift of the beam caused by the acceleration α is detected by the deflection pickup, the current i is generated by the servo amplifier, which produces a torque to keep the beam at the principal axis of the sensor. Since the torque and the current that generates the torque are proportional to α, the acceleration can be measured using the current. The measurement process forms a closed-loop system, so that the sensor is not only robust to disturbances but also achieves high measurement accuracy (see Table 16.2).

Similarly, an inclinometer is another inertial sensor that measures tilt angle to provide attitude information (see Figure 16.2). The principle of servo-type inclinometers is the same as that of the servo-type accelerometer, except that the beam in the accelerometer is replaced by a pendulum suspended from the supporting point in the sensor. When the sensor is placed on the inclined static inclined surface of tilt

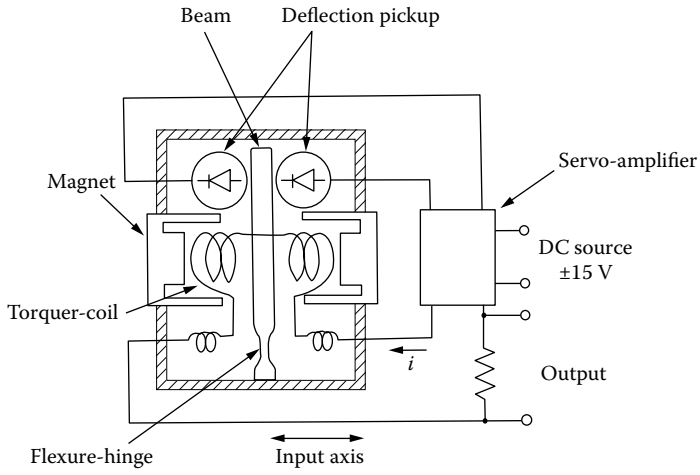


FIGURE 16.1 Servo-type accelerometer.

TABLE 16.2 Specification of a Servo-Type Accelerometer

Measurement range	±5 *g*
Resolution	Less than 5 μ*g* (dc)
Sensitivity	2 V g^{-1}
Output resistance	560 Ω
Torque current	3.5 mA g^{-1}
Case alignment	Less than ±1°
Frequency response	450 Hz (+3 dB)
Temperature range	−25 to +70 °C
Power source	±15 V (DC)
Consumption current	Less than 15 mA
Size	28.4 mm × 24.5 mm
Mass	46 g (including the cable 10 g)

Note: *g*, gravitational acceleration (according to the type TA-25D-05 by TOKIMEC).

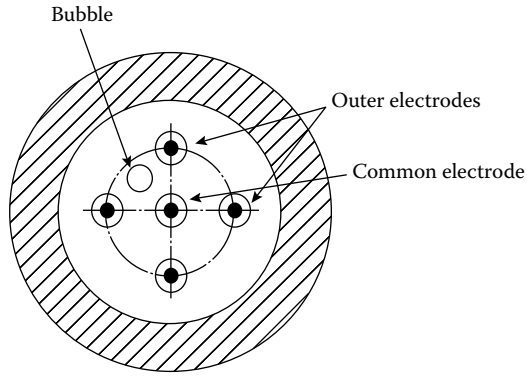


FIGURE 16.2 Dielectric-type inclinometer (front view).

angle β, the pendulum takes the angle β against the principal axis of the sensor, assuming the sensor has no force other than gravity acting on it. The sensor can, however, generate a torque $T_c = mg_l \sin\beta \cong mg_l\beta$ to keep the pendulum at the principal axis, then the tilt angle β can be accurately measured using the torque (and consequently the current producing the torque), where m and l are the pendulum mass and length of the pendulum to its mass center, respectively. One must note, however, that such a sensor is essentially designed to measure the tilt angles of static inclined surfaces. Thus, when applied to dynamic inclined surfaces, the accelerations will affect the torque, making the sensor unreliable. An intelligent attitude sensing system that overcomes such difficulty will be introduced later. Although application is limited to static inclined surfaces with minute tilt angles, a dielectric-type inclinometer employing electrodes and a bubble kept in an electrolyte can achieve high accuracies on the order of 10^{-4}°.

16.2 Attitude Sensors for Spacecraft Applications

Attitude measurement for spacecraft usually requires two or more sensors for detecting the reference sources needed to satisfy attitude requirements. The choice of which sensors to employ is primarily influenced by the direction the spacecraft is usually pointing as well as the accuracy requirements for attitude determination [2]. Table 16.3 summarizes some performance parameters for these sensors as well as typical manufacturers.

IMUs generally consist of gyroscopes coupled with accelerometers, which together measure both rotational and translational motion. These IMUs may be either gimbal mounted (movement about

TABLE 16.3 Spacecraft Attitude Determination Sensors

Sensor	Accuracy	Mass (kg)	Typical Vendors
IMU	$1–5 \times 10^{-6}\,g$	3–25	Northrop Grumman, Bendix, Kearfott. Honeywell, Hamilton, Standard, Litton, Teledyne
Sun sensor	10^{-2} to 3°	0.5–2	Adcole, TRW, Ball Aerospace
Horizon sensor	10^{-1} to 1°	2–5	Barnes, Ithaco, Lockheed Martin, Lockheed Barnes
Star sensor	10^{-3} to 10^{-2}°	3–7	Ball Aerospace, Bendix, Honeywell, Hughes
Magnetometer	0.5–3°	∼1	Schonstedt, Develco

Source: Adapted from Larson, W. J. and Wertz, J. R., eds., *Space Mission Analysis and Design*, Microcosm, Inc., Torrance, CA, 1992, p. 360.

a gimbal point, independent of the spacecraft) or a strapdown system (rigidly mounted to the spacecraft body), where expansive software is used to convert sensor outputs into reference frame measurements. IMUs tend to suffer gyro drift and other bias errors and, when used for spacecraft attitude measurements, are often used with one or more of the sensors discussed as follows.

Sun sensors detect the visible light from the sun, measuring the angle between the sun's radiation and the detector's photocell. The sun is a commonly chosen attitude reference source since it is by far the visually brightest object in the sky, having a total radiation per unit area of 1353 W m^{-2} at Earth distances [3]. Also, it is generally accepted as a valid point source for most attitude applications, having an angular radius of 0.25° at Earth distances. Increased measurement accuracy can be obtained by determining its centroid. Even though sun sensors are quite accurate (0.01°–3.0°), they do require clear fields of view and sometimes suffer periods of eclipse from both the Earth and the moon [4]. Also, sensitive equipment (such as imaging devices) must be protected from the powerful radiation of direct sunlight. When the sun is available, the angle between it and the sensor's primary axis is referred to as the *sun angle*.

For spacecraft in near-Earth orbits, the Earth is the second brightest object in the sky and covers as much as 40% of the sky. Earth *horizon sensors* detect the interface between the Earth's edge (or limb) and the space background. Horizon sensors can detect either of the Earth's visible limb (albedo sensor), infrared limb, or airglow. The infrared limb is the edge between the warm Earth and the cold space background. The airglow is a region of the atmosphere around the Earth that is visible to the spacecraft when it is on the night side of the Earth. Accuracies for horizon sensors are in the 0.1°–1.0° range. Increased accuracy requires Earth oblate spheroid modeling [4]. Some problems associated with albedo detection include the distortion effects of the Earth's atmosphere, falsely identifying the day/night terminator crossing as the true Earth limb, and the considerable variability of the Earth's albedo in the visible spectrum (varies from land, sea, ice).

Most sensors used to detect the Earth's horizon are scanning sensors with narrow fields of view that measure the time between horizon crossings. In general, two horizon crossings occur per sensor scan period: one crossing when the sensor scans from the space background onto the Earth, followed by a second crossing when the sensor scans from the Earth back to space. The combination of horizon crossing times, scan rate, and spacecraft altitude allows for the computation of the Earth's apparent *angular radius*. The apparent angular radius will be smaller than the real (or physical) angular radius if the spacecraft is tilted away from the Earth nadir vector. The nadir vector is defined as the vector connecting the center of the spacecraft to the center of the Earth. To see this effect, one needs to compute the Earth's physical radius p, which for a given spacecraft altitude h (in kilometers) is given by $\rho = \sin^{-1}[(6371)/(6371 + h)]$.

If the spacecraft horizon sensor is pointing exactly *nadir*, then the apparent angular radius as measured by the sensor will agree with the physical radius given by the aforementioned relation for ρ. However, if the horizon sensor is pointed away from nadir, the horizon crossing times will be smaller than when pointing exactly nadir. This results in an apparent angular radius that is smaller than the physical radius by an amount proportional to the angle between the sensor axis and the nadir vector. This angle is referred to as the *nadir angle*.

Star sensors are used when extreme accuracy requirements are necessary. This high degree of sensor accuracy (0.003°–0.01°) can be attributed mainly to the point source nature and precise fixed location

of stars in space. Star sensors may be categorized as either star trackers or star mappers. A star tracker utilizes a wide field of view in order to search for a given star of specific brightness. A star mapper is similar to a tracker, except that it scans over many stars, recording their relative positions and angular separations. By comparing the recorded data with that from a *star catalog* (database), exact spacecraft orientation can be obtained. The angle between the star line of sight and the sensor's primary axis is referred to as the *star angle*.

The accuracy of star sensors is obtained with higher costs, however. Star sensors are generally heavier and consume more power than other types of attitude sensors. In addition, star sensors are quite sensitive to stray light sources such as sunlight reflected from the spacecraft or the Earth and sunlight scattered from dust particles and jet exhausts [4]. Most rely on optical shielding to reduce the effects of stray light.

Magnetic sensors (called *magnetometers*) measure both the magnitude and direction of the Earth's magnetic field. The difference in orientation between the measured field and the true field translates into attitude determination. Magnetometer accuracies (0.5°–3.0°) are usually less than the other sensor types because of the uncertainty in the Earth's true field, which tends to change or shift over time. In addition, the Earth's magnetic field decreases with increasing altitude and magnetometers are generally limited to altitudes of about 6000 km. For this reason, magnetometers are often used with one of the other sensor types already discussed for improved measurement accuracy [2].

16.3 Automatic On-Line Attitude Measurement for Ships and Crane Lifters

For on-line attitude measurement for ships and crane lifters, the first thing that comes to mind is to use gyros. However, because they often suffer from drifts, accurate attitude measurements might not be achieved using the gyros. Accordingly, one uses on-line attitude measurement systems that do not utilize gyros but servo-type accelerometers and inclinometers. The philosophy of the measurement systems introduced here is to make the best use of the system dynamics of the object and the sensors and to apply Kalman filters or adaptive filters to achieve high measurement accuracy.

16.4 Attitude Measurement for Ships

On-line accurate measurement of a ship's attitude is extremely important in exact search of the seabed patterns with sonars [5,6]. It is also required by high-performance ships like hovercrafts from the viewpoint of suppressing swings by the waves. The measurement of a ship's attitude can usually be reduced to that of the heaving, rolling, and pitching of the ship. For such a measurement, a heave sensor has been used, whose output is given by double integration of the output of an accelerometer vertically directed with a gyroscope. However, since the initial values of heaving displacement and its velocity are unknown, the output will contain a bias that increases with time, and the accuracy of the sensor deteriorates considerably. From this viewpoint, one introduces a strapdown-type on-line measurement system that adequately processes the outputs of the two servo-type inclinometers and one accelerometer mounted on the ship [7].

16.4.1 Location of Sensors and Outputs

The two servo-type inclinometers and one servo-type accelerometer are located on the deck (at the point A) of vertical distance L from O, the intersection of rolling and pitching axes (see Figure 16.3). The two inclinometers are set in such a way that the rolling and the pitching angles are measured respectively. The accelerometer is set upward to the deck to obtain the information on the heaving. Because inclinometers were originally developed for the measurement of the tilt angles of static inclined surfaces, the rigid pendulum inside the sensor is considerably affected by the ship's acceleration other than the gravitational one.

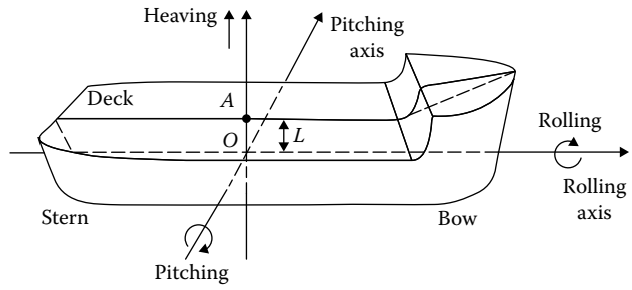


FIGURE 16.3 Location of sensors.

Applying Lagrange's equations of motion [8,9] to rigid pendulums and calculating the torques to keep their deflections from the principal axes almost zero yield the sensor outputs [7]:

$$z_1(t) = \theta(t) - \frac{L}{g}\ddot{\theta}(t) + v_1(t) \tag{16.1}$$

$$z_2(t) = p(t) - \frac{L}{g}\ddot{p}(t) + v_2(t) \tag{16.2}$$

where $z_1(t)$, $z_2(t)$, $\theta(t)$, $p(t)$, and g denote, respectively, the outputs of the two inclinometers, the rolling and the pitching angles, and the gravitational acceleration ($v_1(t)$ and $v_2(t)$: noises of the outputs, including the approximation errors in deriving the outputs).

On the other hand, the accelerometer output is expressed as

$$z_3(t) = (g + \alpha(t))\cos\theta(t)\cos p(t) + v_3(t) \tag{16.3}$$

where $\alpha(t)$ and $v_3(t)$ represent, respectively, the heaving acceleration and the accelerometer noise.

16.4.2 Dynamics of Attitude Signals

It is well known that each of the heaving, rolling, and pitching in inshore seas has two dominant waves in a short interval. That is, a sinusoidal wave of long periodic length (in the range of 6–10 s) and a sinusoidal wave of short periodic length (in the range of 2–3 s) [10–12]. Thus, one models each of the signals in a short interval by a composite wave of the two dominant sinusoidal waves. For the heaving (in a short interval), the displacement is modeled by

$$x(t) = a_1 \sin(\omega_1 t + \phi_1) + a_2 \sin(\omega_2 + \phi_2) \tag{16.4}$$

with the parameters $\{a_i\}$, $\{\phi_i\}$, and $\{\omega_i\}$ unknown. From the 4th-order differential equation satisfied by the $x(t)$, we obtain the linear dynamic equation [7]:

$$\dot{x}(t) = Ax(t), \quad A \equiv \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -\omega_1^2\omega_2^2 & 0 & -(\omega_1^2 + \omega_2^2) & 0 \end{bmatrix} \tag{16.5}$$

where $x(t) = (x_1, x_2, x_3, x_4)^{\mathrm{T}}$ ($x_n = d^{n-1}x/dt^{n-1}$ ($n = 1, \ldots, 4$)). On the other hand, the rolling and pitching angles can be modeled by

$$x(t) = a_1 \sin(\omega_1 t + \phi_1) + a_2 \sin(\omega_2 t + \phi_2) + b \tag{16.6}$$

because there are usually some biases associated with them. From the 5th-order differential equation that Equation 16.6 satisfies, we get the similar state variable representation of the model as Equation 16.5. In practice, the heaving, rolling, and pitching signals have many nondominant sinusoidal waves in addition to the dominant ones. Therefore, Equation 16.5 is modified by introducing a white Gaussian noise $w(t)$ with zero mean and adequate variance σ^2 as follows:

$$\dot{x}(t) = Ax(t) + \Gamma w(t) \tag{16.7}$$

where $\Gamma = (0, 1, 0, 0)^{\mathrm{T}}$ for the heaving and $\Gamma = (0, 1, 0, 0, 0)^{\mathrm{T}}$ for the rolling and pitching. The higher the order of the models, the better the measurement accuracy will be. If we consider the on-line measurement of the signals, Equation 16.7 will be sufficient.

16.4.2.1 On-Line Attitude Measurement

The observation Equations 16.1 and 16.2 are expressed using their own state vector $x(t)$. The observation equations in a discretized form are

$$y_k = Hx_k + v_k \tag{16.8}$$

where $H = (1, 0, -L/g, 0, 0)$ and y_k, x_k, and v_k, respectively, denote $y(t)$, $x(t)$, and $v(t)$ of the corresponding signals at the kth sampling instant [7,9]. The discretized form of the dynamic Equation 16.7 is

$$x_{k+1} = Fx_k + w_k \tag{16.9}$$

where

$$F = \Phi(t)\Big|_{t=\Delta T}, \quad \Phi(t) \equiv L^{-1}\{(sI - A)^{-1}\} \tag{16.10}$$

Here, L^{-1} and ΔT, respectively, denote the inverse Laplace transformation and the sampling period. The discretized transition noise w_k becomes a white Gaussian noise with zero mean and covariance:

$$w = \sigma^2 \int_0^{\Delta T} \Phi(\Delta T - \tau)\Gamma\Gamma^{\mathrm{T}}\Phi^{\mathrm{T}}(\Delta T - \tau)d\tau \tag{16.11}$$

The measurement of the rolling and pitching can thus be reduced to the state estimation of the linear discrete dynamic systems (Equations 16.8 and 16.9), if the angular frequencies ω_1 and ω_2 are given and v_k is assumed to have a white Gaussian property. The state estimation is achieved by a Kalman filter [7,13]. However, difficulties in implementing the filter are that the exact values of the two angular frequencies are a priori unknown and also time variant. To overcome the difficulty, adequate candidates $\{(\omega_1^i, \omega_2^i); 1 \le i \le M\}$ for the parameters $\{\omega_1\omega_2\}$ are set and a bank of Kalman filters is used. Then, the final estimate is obtained as the conditional expectation of the state estimate as follows:

$$\hat{x}_{k/k}^0 = \sum_{i=1}^{M} p_k^i \hat{x}_{k/k}^i \tag{16.12}$$

where $\hat{X}_{k/k}^i$ represents the state estimate $\hat{X}_{k/k}$ for the ith candidate $\Omega_i = (\omega_1^i, \omega_2^i)$ and p_k^i denotes the conditional posteriori probability of the ith candidate calculated based on the Bayesian theorem:

$$p_k^i = \frac{p(y_k/\Omega_i, Y^{k-1})p_{k-1}^i}{\sum_{j=1}^{M} p_{k-1}^j p(y_k/\Omega_j), Y^{k-1}} \tag{16.13}$$

Here, $p(y_k/\Omega_i Y^{k-1})$ represents the conditional Gaussian probability density function of y_k under Ω_i and $y^{k-1} = \{y_j; j \le k-1\}$, whose mean and variance are calculated recursively [7].

The proposed measurement system can adaptively and automatically select the most appropriate candidate versus time. It thus enables an accurate on-line measurement of the rolling and pitching whose dominant angular frequencies vary with time. The first, second, and third components of the final estimate $\hat{X}^0_{k/k}$ represent the estimates of the displacement, velocity, and acceleration, respectively. The proposed system thus has an advantage in that it can measure not only the displacements but also the velocities and the accelerations of the three signals. In order to improve the measurement accuracy of the rolling and pitching, one should place the inclinometers near the intersection O of the rolling and the pitching axes.

Finally, the dynamics of the heaving is given by Equation 16.7 similar to that of the rolling and pitching. Substituting the estimates $\hat{\theta}(t)$ and $\hat{p}(t)$ obtained earlier into Equation 16.3 and subtracting the effect of the gravitational acceleration, one can derive a linear observation equation for $\alpha(t)$:

$$\begin{aligned} y_k &= [z_3(t) - g\cos\hat{\theta}(t)\cos\hat{p}(t)]\big|_{t=k\triangle T} \\ &= H_k x_k + \nu_k \end{aligned} \tag{16.14}$$

where

$$H_k = [0, 0, \cos\hat{\theta}(t)\cos\hat{p}(t), 0]\big|_{t=k\triangle T}$$

$$\nu_k = \nu_3(k\Delta T)$$

Thus, the on-line measurement of the heaving is also realized by executing the same procedure as described before. The locations of the rolling and pitching axes were assumed to be known; however, even when they are unknown, the attitude measurement system described previously is effective, if we introduce the candidates on the location of the axes adding to the angular frequencies.

16.5 Attitude Measurement for Crane Lifters

16.5.1 Dynamics of Attitude Signals

An illustrative diagram of a crane lifter system is shown in Figure 16.4. One of the easiest ways to measure the attitude of the lifter is to set up a high-resolution camera on the bottom of the trolley and to track a mark on the top of the lifter. However, it increases the cost and also the difficulty in maintenance. Furthermore, sometimes the scheme does not work because of shadows and light reflection. As previously mentioned, for gyros not offering sufficiently accurate measurement, a high-sensitivity servo-type accelerometer is used to extract the attitude signals. When setting up the sensor on the lifter, however, there is a secondary swing signal adding to the primary one, due to the free suspension of the lifter and the structure of the lifter. Despite its small amplitude, the secondary one has a higher frequency and for this reason has a large magnitude on the sensor output. The important signal for practical applications, such as the attitude control of the lifter, is the primary one, which has larger amplitude with a lower angular frequency of $\omega = \sqrt{g/l}$ (g, the gravitational acceleration; l, the wire length from the primary supporting point to the center of gravity of the pulley). If we try to attenuate the secondary swing signal by passing the output through a low-pass filter, the phase lag is also introduced into the primary swing signal and the signal can no longer be used for the accurate attitude control of the lifter.

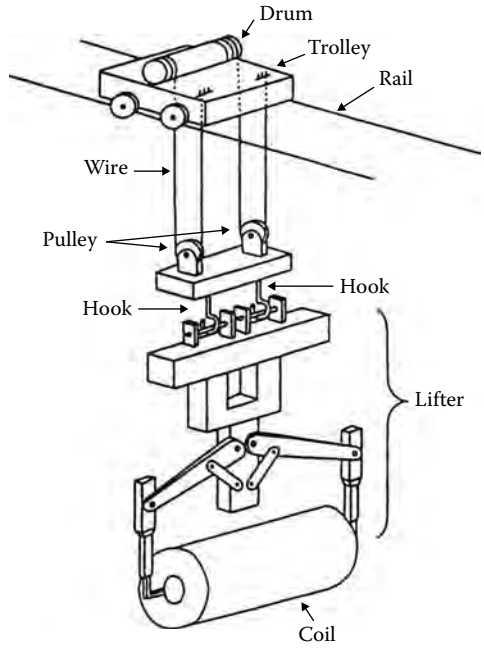


FIGURE 16.4 A crane lifter system.

For the aforementioned reasons, we introduce an autonomous measurement system that measures both the primary and the secondary swings by modeling the lifter system with a double pendulum and applying a Kalman filter to it [14]. The dynamics of the trolley-lifter system is derived using Lagrange's equations of motion [8,9]:

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{x}}\right)-\frac{\partial T}{\partial x}+\frac{\partial V}{\partial x}=u-z\dot{x} \tag{16.15}$$

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{\theta}_1}\right)-\frac{\partial T}{\partial \theta_i}+\frac{\partial V}{\partial \theta_i}=0(i=1,2) \tag{16.16}$$

where

T and V represent, respectively, the kinetic and the potential energies of the trolley-lifter system

θ_1, θ_2 denote, respectively, the angles that the primary and the secondary pendulums take against the vertical line

The other variables x, u, and a represent the location of the trolley, the driving force, and the coefficient of friction between the trolley and the rail, respectively. Considering that θ_i, θ_i, $(1 \le i \le 2)$ are small, the dynamic equation of the trolley-lifter system can be expressed as [14]

$$\dot{x}(t)=Ax(t)+bu(t) \tag{16.17}$$

where $\mathbf{x}(t)$ is the state vector $x(t)=(x, \dot{x}, \theta_1, \dot{\theta}_1, \theta_2, \dot{\theta}_2)^{\mathrm{T}}$. Taking into account the approximation errors in deriving Equation 16.17, air resistance, friction in the wires, and microscopic swings at the other

connection points, it is reasonable to introduce white Gaussian noises $w(t)$ $(1 \le i \le 3)$ with zero mean and appropriate variances to the dynamic Equation 16.17 as in Equation 16.5 as follows [14]:

$$\dot{x}(t) = Ax(t) + bu(t) + \Gamma w(t) \tag{16.18a}$$

where

$$\Gamma = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}^{\mathrm{T}}, \quad \omega(t) = [\omega_1(t), \omega_2(t), \omega_3(t)]^{\mathrm{T}} \tag{16.18b}$$

16.5.2 Sensor Outputs and On-Line Attitude Measurement

When a servo-type accelerometer is set up on the lifter (in the direction of the swing) at the place of the distance ℓ'_2 from the secondary supporting point, the output of the sensor becomes [14]

$$y(t) \cong -\ddot{x} - \ell'_1 \ddot{\theta}_1 - \ell'_2 \ddot{\theta}_2 - g\theta_2 \tag{16.19}$$

where ℓ'_1 is the distance between the primary and the secondary supporting points (see Figure 16.5). Substitution of Equation 16.17 into Equation 16.19 yields an output expressed in terms of the state vector $\mathbf{x}(t)$, as in Equation 16.8. Using a rotary encoder to measure the location and the velocity of the trolley, and then combining these three sensor outputs with the dynamic Equation 16.18 and applying a Kalman filter, enables the state vector to be estimated on-line. Using this approach, both angular displacement and velocity of the deflections θ_1, θ_2 of the two pendulums can be measured exactly.

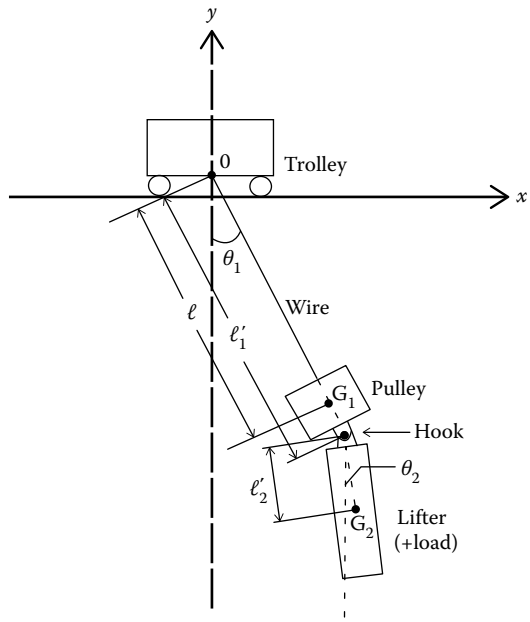


FIGURE 16.5 Dynamics of a trolley-lifter system.

16.6 Aircraft Attitude Determination

The determination of aircraft attitude requires the measurement of angles about three independent body axes. These angles are the roll, pitch, and yaw angles. There are two primary means employed today for measuring these angles; the first method uses VGs to measure the roll and pitch angles and a DG to measure the yaw angle. The second method, more commonly used today, employs an IMU for full three-axis attitude determination coupled with a baro-altimeter to correct for vertical drift errors in the IMU. Both methods are described as follows.

16.6.1 Vertical and Directional Gyro Analysis

A VG is a two degree-of-freedom gyro with its spin axis mounted nominally vertical. It employs two specific force sensors mounted nominally horizontal on the inner gimbal. The two angles measured by the VG—roll and pitch—require nearly identical analyses [1]. Consider the situation shown in Figure 16.6, depicting an aircraft with a roll angle of ϕ with respect to the local vertical. The sensed roll angle ϕ_s is given by the difference in the actual roll angle and the gyro roll drift error ϕ_Δ:

$$\phi_s = \phi - \phi_\Delta \tag{16.20}$$

In order to compensate for this drift error, gyros employ a specific force sensor such as an electrolytic bubble device, which senses drift error. This correction device senses the angular difference between the specific force vector **f** acting on the aircraft roll axis and the gyro axis, as shown in Figure 16.6. Thus,

$$\phi_B = \tan^{-1}\left[\left(\frac{f_\phi}{f_v}\right) - \phi_\Delta\right] \cong \left(\frac{f_\phi}{f_v}\right) - \phi_\Delta \tag{16.21}$$

where

f_ϕ is the side horizontal component of **f**

f_v is the force of gravity is the vertical component

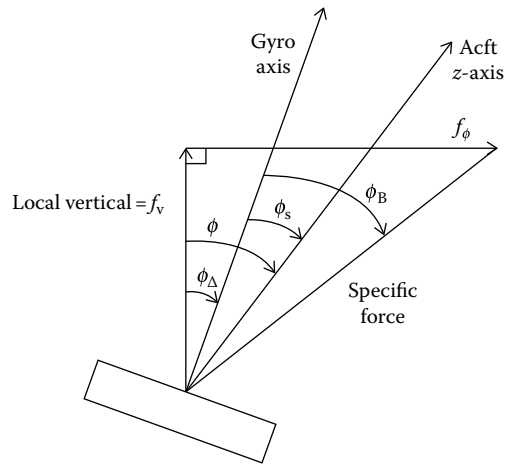


FIGURE 16.6 VG analysis.

A similar analysis for the pitch angle θ yields

$$\theta_s = \theta - \theta_\Delta \tag{16.22}$$

$$\theta_B = \tan^{-1}\left[\frac{f_0}{f_v} - \theta_\Delta\right] \cong \frac{f_\theta}{f_v} - \theta_\Delta \tag{16.23}$$

where f_θ is the back horizontal component of **f**. Next, define the gyro angular momentum vector by

$$H_{VG} = [J_x \dot{\phi}_\Delta, J_y \dot{\theta}_\Delta, -h] \tag{16.24}$$

where

J_x and J_y are the sensor moments of inertia
h is the gyro spin angular momentum

In addition, define the inner gimbal axes angular velocity vector as

$$\omega_{VG} = [\dot{\phi}_\Delta, \dot{\theta}_\Delta, 0] \tag{16.25}$$

Finally, define the gimbal torque vector by

$$Q_{VG} = [Q_{cx} + Q_{dx}, Q_{cy} + Q_{dy}, 0] \tag{16.26}$$

where

$$Q_{cx} = \text{gimbal roll control torque} = -k_c\theta_B \tag{16.27a}$$

$$Q_{cy} = \text{gimbal pitch control torque} = k_c\phi_B \tag{16.27b}$$

$$Q_{dx} = \text{gimbal roll disturbance torque} = -k_d(\dot{\phi}_\Delta - \dot{\phi}) + \text{random torques} \tag{16.27c}$$

$$Q_{dy} = \text{gimbal pitch disturbance torque} = k_d(\dot{\theta}_\Delta - \dot{\theta}) + \text{random torques} \tag{16.27d}$$

and the k_c and k_d are constant scaling factors related to each torque component.

Using the vectors defined in Equations 16.24 through 16.26, the gyro equations of motion are given by

$$\frac{\partial}{\partial t}(H_{VG}) + (\omega_{VG} \times H_{VG}) = Q_{VG} \tag{16.28}$$

Taking the Laplace transform of the expansion of Equation 16.28, with the assumption that $J_x = J_y = J$, yields the following gyro equations of motion in the Laplace domain:

$$\begin{bmatrix} J_x s^2 + k_d s & -(hs + k_c) \\ hs + k_c & J_y s^2 + k_d s \end{bmatrix} \begin{bmatrix} \phi_\Delta(s) \\ \theta_\Delta(s) \end{bmatrix} \equiv \begin{bmatrix} -k_c\theta_B + k_d s\phi(s) + \text{random torques} \\ (k_c/g) f_\phi(s) + k_d s\theta(s) + \text{random torques} \end{bmatrix} \tag{16.29}$$

For normal gyro operation, $J_x \cong J_y \cong 0$ and $k_d/h \ll 1$, so these factors may be ignored in Equation 16.29. Thus, solving for the desired roll and pitch angles under these assumptions gives [1]

$$\phi_s = \begin{cases} \phi & \omega \gg k_c/h \\ \phi - f_\phi/g & \omega \ll k_c/h \end{cases} \tag{16.30}$$

$$\theta_s = \begin{cases} \theta & \omega \gg k_c/h \\ \theta - f_\theta/g & \omega \ll k_c/h \end{cases} \tag{16.31}$$

A DG is a two degree-of-freedom gyro with its spin axis mounted nominally horizontal and pointing in the direction of magnetic north. It employs a single specific force sensor mounted on the inner gimbal [1]. The DG measures the third required aircraft angle, yaw, generally denoted by ψ. The sensed yaw angle ψ_s is given by the difference in the actual yaw angle ψ (angle between the aircraft z-axis and true north) and the gyro heading angle drift error ψ_Δ (angle between the gyro axis and true north):

$$\psi_s = \psi - \psi_\Delta \tag{16.32}$$

Define the gyro angular momentum vector by

$$H_{DG} = [J_y \dot{\theta}_\Delta, J_z \dot{\psi}_\Delta, -h] \tag{16.33}$$

and the inner gimbal axes angular velocity vector as

$$\omega_{DG} = [\dot{\theta}_\Delta, \dot{\psi}_\Delta, -h] \tag{16.34}$$

and the gimbal torque vector as

$$Q_{DG} = [Q_{cy} + Q_{dy}, Q_{cz} + Q_{dz}, 0] \tag{16.35}$$

Here, the torque vector components are given by

$$Q_{cy} = k_c(M_\Delta - \psi_\Delta) \tag{16.36a}$$

$$Q_{cz} = -k_c\theta_B \tag{16.36b}$$

$$Q_{dy} = -k_d(\dot{\theta}_\Delta - \dot{\theta}) + \text{random torques} \tag{16.36c}$$

$$Q_{dz} = -k_d(\dot{\psi}_\Delta - \dot{\psi}) + \text{random torques} \tag{16.36d}$$

where M_Δ = magnetic compass heading error (from true north). Therefore, the DG equations of motion are given in Laplace domain as

$$\begin{bmatrix} J_y s^2 + k_d s & -(hs + k_c) \\ hs + k_c & J_z s^2 + k_d s \end{bmatrix} \begin{bmatrix} \theta_\Delta(s) \\ \psi_\Delta(s) \end{bmatrix} \cong \begin{bmatrix} k_c M_\Delta(s) + k_d s\theta(s) + \text{random torques} \\ -k_c\theta_B + k_d s\psi(s) + \text{random torques} \end{bmatrix} \tag{16.37}$$

The desired yaw angle measurement for the DG is thus given as [1]

$$\Psi_s = \begin{cases} \Psi & \omega \gg k_c/h \\ \Psi - M_\Delta & \omega \ll k_c/h \end{cases} \tag{16.38}$$

As indicated in Table 16.1, the accuracies of both VGs and DGs are approximately 1°. An improvement of over two orders of magnitude can be obtained through the use of IMUs, which are described next.

16.6.2 Inertial Measurement Units

IMUs consist of gyroscopes and accelerometers that together provide full three-axis attitude measurements. Most are mounted on stable gimbaled platforms that remain locally horizontal via torquing devices. An IMU aboard an aircraft cannot measure exactly the local vertical due to the fact that the specific force acting on the aircraft has a horizontal component due to vehicle motion. In addition, since the vehicle is moving with respect to the inertial reference frame, the Earth's magnetic pole cannot be determined precisely [1].

These problems (errors) are minimized by aligning the IMU to be exactly horizontal and north pointing while the aircraft is stationary. Once platform motion begins, the IMU may be constantly realigned by sensing changes in the direction of vertical and north and then applying appropriate torques to the platform to keep it properly aligned. This realignment is accomplished by integrating the two orthogonal accelerometer outputs to determine the components of horizontal velocity. These data, combined with the Earth's rotation rate, yield the desired rates of change in local vertical and true north at the vehicle's current latitude and longitude. Performing a second integration of the sensor outputs yields an estimate of relative position.

Analysis in Bryson [1] has shown that the pitch angle (variation in platform horizontal position) is given by the IMU sensor output as

$$\theta(t) = \frac{-\varepsilon}{\omega_s}\sin(\omega_s t) - \frac{b}{g} \tag{16.39}$$

where

ε is the gyro drift rate error

b is the specific force sensor error

$\omega_s \equiv$ Schuler frequency $= \sqrt{g/R}$, [g = force of gravity, R = Earth's radius]

Thus, the platform root-mean-square (rms) pitch angle becomes

$$\theta_{rms} = \left[\frac{1}{2}\left(\frac{\varepsilon}{\omega_s}\right)^2 + \left(\frac{b}{g}\right)^2\right]^{\frac{1}{2}} \tag{16.40}$$

Using typical values for $\varepsilon(\cong 0.05°\text{h}^{-1})$, $\omega_s \cong (0.71°\text{h}^{-1})$, and $b(\cong 0.01)$ yields an rms pitch angle error of $\theta_{rms} = 0.01°$. Thus, it is apparent that under normal operating conditions, the IMU provides a two order-of-magnitude improvement in sensor accuracy when compared to the VG and DG.

16.7 Spacecraft Attitude Determination

Most spacecraft attitude determination techniques rely upon finding the orientation of a single axis in space (e.g., the spacecraft z-axis) plus the spacecraft rotation about this axis. This provides a full three-axis attitude solution. In order to achieve this, reference sources that are external to the spacecraft must be used. Specifically, full three-axis spacecraft attitude determination requires at least two external vector measurements. Commonly used reference sources for these external vector measurements include the sun, Earth, moon, stars, planets, and the Earth's magnetic field. In addition, IMUs are also used to provide the necessary attitude measurements.

16.7.1 Attitude Determination Methodology

The first step in attitude determination is to determine the angles between the spacecraft's primary axis and the two (or more) attitude reference sources. For example, suppose a particular spacecraft is using the sun and the Earth for attitude reference. The two angles in this case are referred to as the sun angle β_s and the nadir angle Γ_N. Since the orientation of even a single spacecraft axis is unknown at this point, these angles establish two *cones* along which the attitude vector **A** must lie. Since the attitude vector must lie on both cones, it must lie along the intersection between the two cones [4] (see Figure 16.7). The two vectors, notably $\mathbf{A}_1$ and $\mathbf{A}_2$, resulting from the intersection of these two cones may be determined by the following method derived by Grubin [15]. Let **S** represent the sun vector, **E** the spacecraft nadir vector, and **A** the desired attitude vector, each defined in Cartesian space as follows:

$$\mathbf{S} = (\mathbf{s}_x, \mathbf{s}_y, \mathbf{s}_z) \tag{16.41}$$

$$\mathbf{E} = (\mathbf{E}_X, \mathbf{E}_y, \mathbf{E}_z) \tag{16.42}$$

$$\mathbf{A} = (\mathbf{A}_x, \mathbf{A}_y, \mathbf{A}_z) \tag{16.43}$$

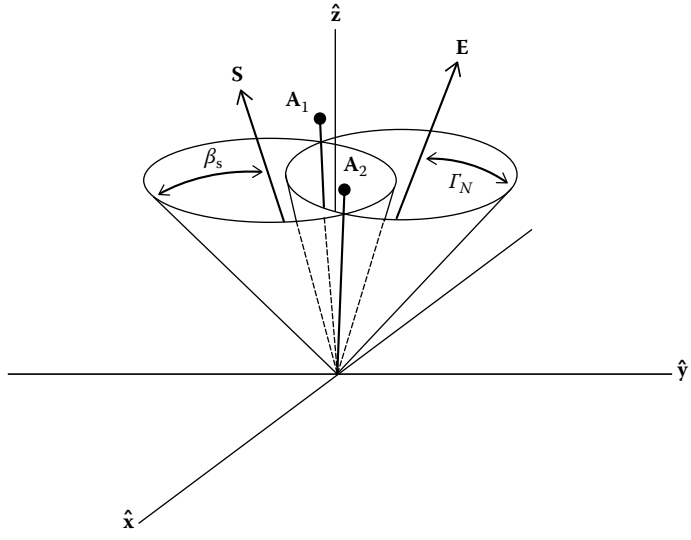


FIGURE 16.7 Relationship between reference vectors and single-axis attitude cones.

Let the vectors **S**, **E**, and **N** define a set of base unit vectors with

$$\mathbf{N} = \frac{\mathbf{S} \times \mathbf{E}}{|\mathbf{S} \times \mathbf{E}|} = (\mathbf{N}_x, \mathbf{N}_y, \mathbf{N}_z) \tag{16.44}$$

If we introduce a proper set of scaling factors as follows:

$$I_x = \frac{[\cos\beta_s - (\mathbf{S} \cdot \mathbf{E})\cos\Gamma_N]}{1 - (\mathbf{S} \cdot \mathbf{E})^2} \tag{16.45a}$$

$$I_y = \frac{[\cos\Gamma_N - (\mathbf{S} \cdot \mathbf{E})\cos\beta_s]}{1 - (\mathbf{S} \cdot \mathbf{E})^2} \tag{16.45b}$$

$$I_z = \sqrt{1 - I_x \cos\beta_s - I_y \cos\Gamma_N} \tag{16.45c}$$

then the two possible attitude vectors **A** and $\mathbf{A}_2$ are found to be

$$\mathbf{A}_{1,2} = [(I_x\mathbf{S}_x + I_y\mathbf{E}_y \pm I_z\mathbf{N}_x), (I_x\mathbf{S}_y + I_y\mathbf{E} \pm I_z\mathbf{N}_y), (I_x\mathbf{S}_z + I_y\mathbf{E}_z \pm I_z\mathbf{N}_z)] \tag{16.46}$$

In Equations 16.44 through 16.46, $\mathbf{S} \times \mathbf{E}$ represents the Cartesian vector product, and $\mathbf{S} \times \mathbf{E}$ represents the Cartesian scalar product. The radicand in Equation 16.45c may be negative, thus producing imaginary values for I_z whenever the two cones do not intersect. Such occurrences are usually attributed to sensor error or random noise fluctuations. In this case, one can add a predetermined sensor bias to both sensors in order to "broaden" the cone angles, thus forcing the cones to intersect.

Note that for most applications involving spacecraft attitude determination, the principal coordinate system used is the *celestial sphere* coordinate system. This coordinate system has the *z*-axis aligned with the Earth's polar axis and the *x*-axis aligned with the intersection of the Earth's equatorial plane and the Earth's orbital plane around the sun (i.e., aligned with the *vernal equinox*). In this coordinate system, all vectors are considered unit vectors and the two principal measurements describing a vector's position are the *right-ascension* and *declination angles*, denoted Ω and Δ, respectively. Thus, the sun vector **s** and the Earth nadir vector **E** used in Equations 16.41 and 16.42 will in general be given as right-ascension and declination angles that can be converted to Cartesian coordinates via the following set of transformations:

$$x = \cos(\Omega)\cos(\Delta); \quad y = \sin(\Omega)\cos(\Delta); \quad z = \sin(\Delta) \tag{16.47a}$$

$$\Omega = \tan^{-1}\left(\frac{y}{x}\right); \quad \Delta = \sin^{-1}(z) \tag{16.47b}$$

The final step in measuring three-axis attitude is to determine which attitude solution is correct, $\mathbf{A}_1$ or $\mathbf{A}_2$, and then measure the rotation about this axis. The two ambiguous attitude solutions may be resolved by comparison with a priori attitude information, if available, or through the use of *trace averaging* [4]. Trace averaging is a method of plotting each attitude solution on a right-ascension versus declination plot and choosing the area of greatest concentration as the correct solution, as demonstrated in Figure 16.8. Since the attitude is assumed to change more slowly than the attitude sensor's sample rate, over short time intervals, the data for the correct solution usually form a "cluster" near the correct attitude; the data for the incorrect solution are usually much more scattered.

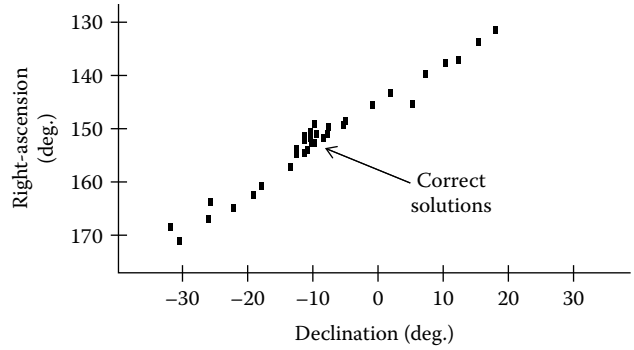


FIGURE 16.8 Method of trace averaging for resolving ambiguous attitude solutions.

Once the correct attitude vector has been obtained, the orientation of the remaining two orthogonal axes may be found by measuring the rotation, or phase angle, of the spacecraft about the preferred axis. Any sensor measurement that provides this phase angle may be used. An example of this technique is provided by the panoramic annular lens attitude determination system (PALADS), described in the next section. This imaging system uses a unique "3D" lens that provides simultaneous detection of two (or more) reference sources [16]. This information, combined with the orientation of the single axis, uniquely determines three-axis attitude.

The three angles derived earlier, which are commonly referred to as *Euler angles*, define the orientation of the three spacecraft axes with respect to the chosen reference frame. A more formal treatment of the attitude solution usually requires specifying the components of the 3×3 attitude matrix **A**. Each component of the attitude matrix defines the angular relationship between a given spacecraft axis and a reference frame axis. Various methods exist for computing the attitude matrix **A** (see Ref. [4]); the preferred method depends on the particular application at hand.

16.8 PALADS

The primary component of PALADS is the panoramic annular lens (PAL), a single-element lens made from a high index of refraction glass with certain portions of the lens coated with a mirrored surface. Hence, it relies on both refraction and reflection in forming an image (Figure 16.9). The lens is unique in that it images a 3D object field onto a 2D image plane, whereas a "normal" lens is capable of only imaging 2D object space onto an image plane. The combination of high index of refraction glass and mirrored surfaces provides the PAL with a field of view extending from approximately 65° to 110° from the optical axis. This 45° field of view covers the entire 360° surrounding the optical axis [17]. Any ray originating from outside the 45° field of view will not form a part of the image. The PAL may be attached to any

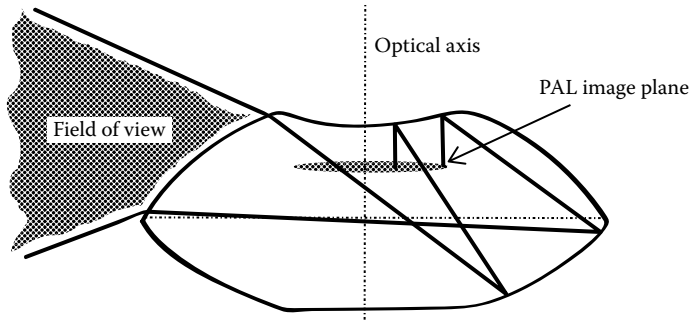


FIGURE 16.9 PAL lens ray diagram.

high-quality imaging system using an appropriate *transfer lens*. As currently configured, the PALADS imaging system utilizes a Sony XC-73 charged-couple device (CCD), a black-and-white video camera coupled to the PAL via a Nikon *f*/1.4 transfer lens.

The hemispherical view provided by PALADS allows for single-sensor detection of multiple attitude reference sources, such as the Earth and the sun or moon. The position of each reference source in the image plane translates into a unique azimuth elevation angle between the PAL's optical axis and the reference source. Since the PAL has a 360° field of view surrounding the optical axis, it may detect several reference sources simultaneously. The data points associated with each source are extracted from the image plane using digital image processing techniques. Thus, it is easy to see how a single image from PALADS (containing two or more reference sources) provides the necessary angle data to determine three-axis spacecraft attitude.

References

1. A. E. Bryson, *Control of Spacecraft and Aircraft*, Chapter 10, Princeton, NJ: Princeton University Press, 1994.
2. W. J. Larson and J. R. Wertz (eds.), *Space Mission Analysis and Design*, Chapter 11, Torrance, CA: Microcosm, Inc., 1992.
3. NASA Technical Memorandum NASA TM X-64757, *Terrestrial Environment (Climatic) Criteria Guidelines for Use in Aerospace Vehicle Development (1973 Revision)*, Marshall Space Flight Center, AL, 1973.
4. J. R. Wertz (ed.), *Spacecraft Attitude Determination and Control*, Chapters 11 and 12, Dordrecht, the Netherlands: Reidel Publishing Company, 1980.
5. R. D. Angelari, A deterministic and random error model for a multibeam hydrographic sonar system, *Proceedings of the OCEANS'78: The Ocean Challenge*, Washington, DC, 1978, pp. 48–53.
6. C. de Moustier, T. Hylas, and J. C. Phillips, Modifications and improvements to the Sea Beam system on board R/V Thomas Washington, *Proceedings of the OCEANS'88: A Partnership of Marine Interests*, Baltimore, MD, 1988, pp. 372–378.
7. S. Tanaka and S. Nishifuji, Automatic on-line measurement of ship's attitude by use of a servo-type accelerometer and inclinometers, *IEEE Trans. Instrum. Meas*., 45, 209–217, 1996.
8. D. G. Shultz and J. L. Melsa, *State Functions and Linear Control Systems*, New York: McGraw-Hill, 1967.
9. Y. Takahashi, M. J. Rabins, and D. M. Auslander, *Control and Dynamic Systems*, Reading, MA: Addison-Wesley, 1971.
10. S. Tanaka and S. Nishifuji, On-line sensing system of dynamic ship's attitude by use of servo-type accelerometers, *IEEE J. Oceanic Eng*., 20, 339–346, 1995.
11. S. Tanaka, On automatic attitude measurement system for ships using servo-type accelerometers (in Japanese), *Trans. SICE*, 27, 861–869, 1991.
12. D. E. Cartwright and M. S. Longuet-Higgins, The statistical distribution of the maxima of a random function, *Proc. R. Soc. Lond, Ser. A*, 237, 212–232, 1956.
13. R. E. Kaiman, A new approach to linear filtering and prediction problems, *Trans. ASME, J. Basic Eng*., 82, 35–45, 1960.
14. S. Tanaka, S. Kouno, and H. Hayashi, Automatic measurement and control of attitude for crane lifters (in Japanese), *Trans. SICE*, 32(1), 97–105, 1996.
15. C. Grubin, Simple algorithm for intersecting two conical surfaces, *J. Spacecraft Rockets*, 14(4), 251–252, 1977.
16. M. A. Stedham and P. P. Banerjee, The panoramic annular lens attitude determination system, *SPIE Proceedings, Space Guidance, Control, and Tracking II*, Orlando, FL, April 17–18, pp. 108–117, 1995.
17. J. A. Gilbert, D. R. Matthys, and P. Greguss, Optical measurements through panoramic imaging systems, *Proceedings of the International Conference on Hologram Interferometry and Speckle Metrology*, Baltimore, MD, November 4–7, 1990.

17

Inertial Navigation

17.1 Principles of Inertial Navigation **17**-1
17.2 Gyroscopes **17**-4
Mechanical Gyroscopes • Flywheel Gyroscopes • Optical Gyroscopes
17.3 Accelerometers **17**-8
17.4 Errors and Stabilization in Inertial Navigation Systems **17**-9
Errors • Stabilization
17.5 Vehicular Inertial Navigation **17**-10
17.6 Applications of Inertial Navigation **17**-11
Aircraft • Underwater • Robotics
Bibliography **17**-13
Partial List of Manufacturers/Suppliers **17**-13

Halit Eren
Curtin University

17.1 Principles of Inertial Navigation

The original meaning of the word *navigation* is "ship driving." In ancient times when sailing boats were used, navigation was a process of steering the ship in accordance with some means of directional information, and adjusting the sails to control the speed of the boat. The objective was to bring the vessel from location A to location B safely. At present, navigation is a combination of science and technology. No longer is the term limited to the control of a ship on the sea surface; it is applied to land, air, sea surface, underwater, and space.

The concept of inertial-navigator mechanization was first suggested by Schuler in Germany in 1923. His suggested navigation system was based on an Earth-radius pendulum. However, the first inertial guidance system based on acceleration was suggested by Boykow in 1938. The German A-4 rocket, toward the end of World War II, used an inertial guidance system based on flight instrument type gyroscopes for attitude control and stabilization. In this system, body-mounted gyro-pendulum-integrating accelerometers were used to determine the velocity along the trajectory. The first fully operational inertial autonavigator system in the United States was the XN-1 developed in 1950 to guide C-47 rocket. Presently, inertial navigation systems (INSs) are well developed theoretically and technologically. They find diverse applications, allowing the choice of appropriate navigation devices, depending on cost, accuracy, human interface, global coverage, time delay, and autonomy.

Inertial navigation is a technique using a self-contained system to measure a vehicle's movement and determine how far it has moved from its starting point. Acceleration is a vector quantity involving magnitude and direction. A single accelerometer measures magnitude but not direction. Typically, it measures the component of acceleration along a predetermined line or direction. The direction information is usually supplied by gyroscopes that provide a reference frame for the accelerometers. Unlike other positional methods that rely on external references, an INS is compact and self-contained, as it is not required to communicate to any other stations or other references. This property enables the craft to navigate in an unknown territory.

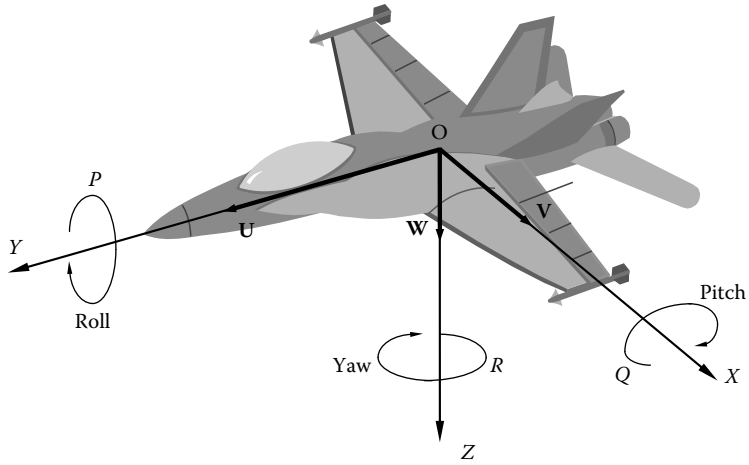


FIGURE 17.1 Inertial navigation is for monitoring the movement of an object such as a vehicle, rocket, ship, aircraft, and robot with respect to a reference axis. Vehicles such as aircrafts and marine vessels have their own local axes, known as the roll, the pitch, and the yaw.

Inertial navigation can also be described as a process of directing the movement of a vehicle, rocket, ship, aircraft, robot, and so on, from one point to another with respect to a reference axis. The vehicle's current position can be determined from "dead reckoning" with respect to a known initial starting reference position. On the Earth's surface, the conventional reference will be north, east, and down. This is referred to as the *Earth's fixed axes*. A vehicle such as an aircraft or a marine vessel will have its own *local axes:* roll, pitch, and yaw, as shown in Figure 17.1.

The inertial sensors of the INS can be mounted in such a way that they stay leveled and pointing in a fixed direction. This system relies on a set of gimbals and sensors attached on three axes to monitor the angles at all times. This type of INS is based on a *navigational platform*. A sketch of a three-axis platform is shown in Figure 17.2. Another type of INS is the *strapdown system* that eliminates the use of gimbals. In this case, the gyros and accelerometers are mounted to the structure of the vehicle. The measurements received are made in reference to the local axes of roll, pitch, and yaw. The gyros measure the movement of angles in the three axes in a short time interval of few milliseconds. The computer then uses this information to resolve the accelerometer outputs into the navigation axes. A schematic block diagram of the strapdown system is shown in Figure 17.3.

The controlling action is based on the sensing components of acceleration of the vehicle in known spatial directions, by instruments that mechanize Newtonian laws of motion. The first and second integrations of the sensed acceleration determine velocity and position, respectively. A typical INS includes a set of gyros, a set of accelerometers, and appropriate signal processing units. Although the principle of the systems may be simple, the fabrication of a practical system demands a sophisticated technological base. The system accuracy is independent of altitude, terrain, and other physical variables but is limited almost purely by the accuracy of its own components. Traditional INSs mainly rely on mechanical gyros and accelerometers, but today there are many different types available, such as optical gyroscopes, piezoelectric vibrating gyroscopes, and active and passive resonating gyroscopes. Also, micromachined gyroscopes and accelerometers are making an important impact on modern INSs. A brief description and operational principles of gyroscopes and accelerometers suitable for inertial navigation are as follows.

Major advances in INS over the years include the development of the *electrostatic gyro* (ESG) and the laser gyro. In ESG, the rotor spins at a speed above 200,000 rpm in a near-vacuum environment. The rotor is suspended by an electrostatic field; thus, it is free from bearing friction and

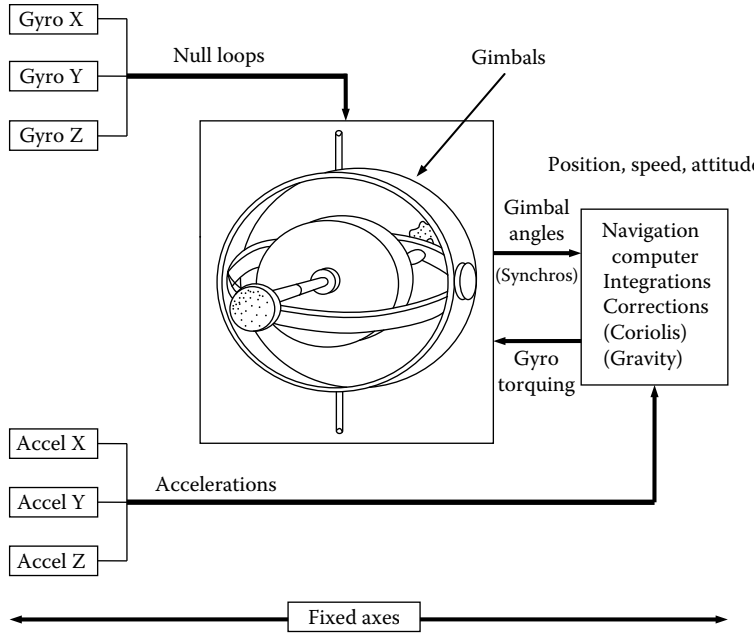


FIGURE 17.2 Some INSs are based on a navigational platform. The sensors are mounted in such a way to stay leveled at all times and pointing in a fixed direction. This system uses a set of gimbals and sensors attached on three axes in the *x*, *y*, and *z* directions to continuously monitor the angles and accelerations. The navigation computer makes corrections for coriolis forces, gravity, and other adverse effects.

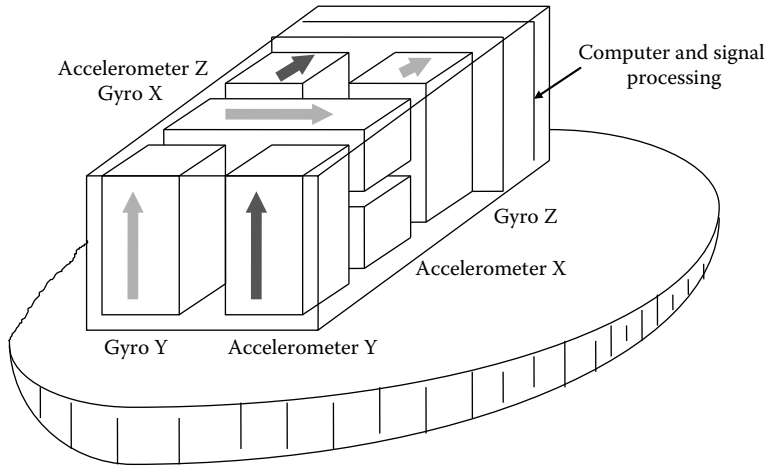


FIGURE 17.3 Strapdown gyros and accelerometers are mounted rigidly on the structure of the vehicle, and the measurements are referenced to the local axes of roll, pitch, and yaw. The gyros measure the movement of angles in the three axes in short time intervals to be processed by the computer. This information is used, together with the accelerometer outputs, for predicting navigation axes, thus eliminating the need for gimbals.

other random torques due to mechanical supports. Hence, its operation results in a superior performance compared to others, closely resembling the performance of a theoretical gyro. Although no system can claim to reach perfection, an ESG requires less frequent updates compared to other mechanical gyros. Gyroscopes are an important element of inertial navigation; therefore, they will be described further.

17.2 Gyroscopes

There are two broad categories: (1) mechanical gyroscopes and (2) optical gyroscopes. Within both of these categories, there are many different types available. Only the few basic types will be described to illustrate the operating principles; detailed information may be found in the references listed at the end of this chapter.

17.2.1 Mechanical Gyroscopes

The first mechanical gyroscope was built by Foucault in 1852, as a gimbaled wheel that stayed fixed in space due to angular momentum while the platform rotated around it. They operate on the basis of *conservation of angular momentum* by sensing the change in direction of an angular momentum. There are many different types, which are the following:

1. *Single-degree-of-freedom gyroscopes* include the rate, rate integrating, a spinning rotor flywheel, electron, and particle gyros.
2. *Two-degree-of-freedom gyroscopes* incorporate the external gimbal types, two-axis-floated, spherical free-rotor and electrically suspended, gas-bearing free-rotor gyros.
3. *Vibrating gyroscopes* include the tuning fork, vibrating string, vibrating shell, hemispherical resonating, and vibrating cylinder gyros.
4. *Continuous-linear-momentum gyroscopes* incorporate a steady stream of fluid, plasma, or electrons, which tends to maintain its established velocity vector as the platform turns. One typical example is based on a differential pair of hot-wire anemometers to detect the apparent lateral displacement of the flowing air column.

The operating principle of all mechanical gyroscopes is based on the conservation of angular momentum, as shown in Figure 17.4. The angular momentum is important since it provides an axis of reference. From Newton's second law, the angular momentum of a body remains unchanged unless it is acted upon by a torque. The rate of change of angular momentum is equal to the magnitude of the torque. It can be expressed, in vectorial form, as

$$T = \frac{\mathrm{d}H}{\mathrm{d}t} \tag{17.1}$$

where H is the angular momentum (=inertia × angular velocity, $I\omega$).

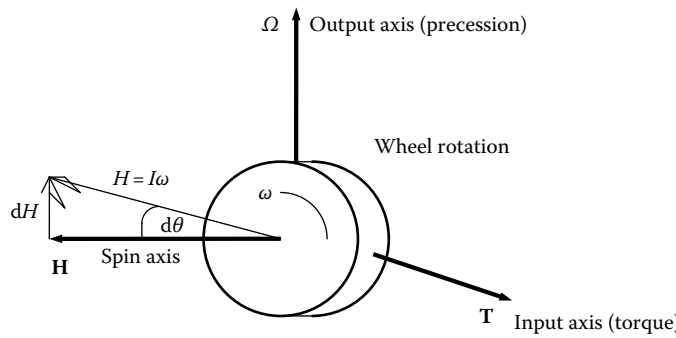


FIGURE 17.4 The operational principle of gyroscopes is based on the angular momentum of a carefully constructed rotating body. The angular momentum stabilizes the system and it remains unchanged unless acted upon by a torque. If the torque is orthogonal to the spin axis, it cannot change the velocity, but it can change the direction in the same direction as the torque. The spin axis always tries to align with the external torque.

If a torque acts about the axis of rotation, it changes the angular velocity by

$$T = \frac{I\mathrm{d}\omega}{\mathrm{d}t} = I\alpha \tag{17.2}$$

where

I is the inertia about the spin axis
α is the angular acceleration

If the torque is orthogonal to the spinning axis, it cannot change the magnitude of the angular velocity vector, but it can change direction in the same direction as torque T; then

$$\mathrm{d}H = H\mathrm{d}\theta \tag{17.3}$$

where θ is the angle of rotation.

Therefore, from Equations 17.1 and 17.4,

$$T = \frac{\mathrm{d}H}{\mathrm{d}t} = \frac{H\mathrm{d}\theta}{\mathrm{d}t} = H\Omega \tag{17.4}$$

where Ω is the precession rate or the angular velocity of the spinning wheel about the axis normal to the plane of the spin and the input torque. Generally, the spin axis tries to align with the external input torque.

These equations can be elaborated to describe the operating principles of mechanical gyros by taking into account the angular momentum in x, y, and z directions, nutation, coriolis accelerations, directions of other influencing torques, and linear forces. Here, the operation of the well-known *flywheel gyroscope* will be described as the basis for further discussions on INSs.

17.2.2 Flywheel Gyroscopes

Flywheel gyroscopes are based on an electrically driven rotor suspended in a pair of precision low-friction bearings at both ends of the rotor axle. An example of a double-axis flywheel gyro is shown in Figure 17.5. The rotor bearings are supported by a circular ring known as an *inner gimbal ring*, which in turn pivots on a second set of bearings that is attached to the *outer gimbal ring*. The pivoting action of the inner gimbal defines the horizontal axis of the gyro, which is perpendicular to the spin axis of the rotor. The outer gimbal ring is attached to the instrument frame by a third set of bearings that defines the vertical axis of the gyro that is perpendicular to both the horizontal axis and the spin axis. This type of suspension has the property of always preserving the predetermined spin-axis direction in inertial space. Equations governing the two-degree-of-freedom gyroscope can be written using Equations 17.1 through 17.4. The torque with respect to an inertial reference frame can be expressed as

$$T = \dot{H}_{\mathrm{t}} \tag{17.5}$$

If the Earth is taken as a moving reference frame, then

$$\dot{H}_{\mathrm{t}} = \dot{H}_{\mathrm{E}} + \omega_{\mathrm{IE}}H \tag{17.6}$$

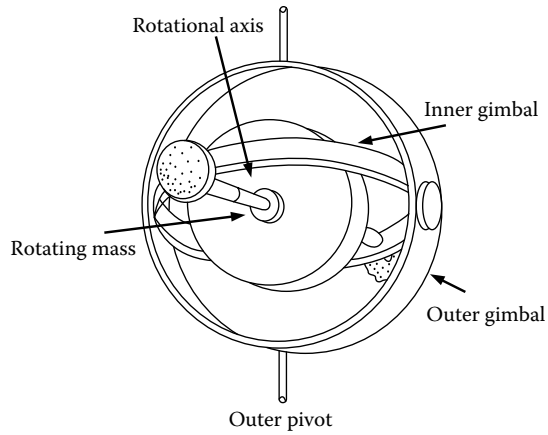


FIGURE 17.5 In a double-axis flywheel gyro, an electrically driven rotor is suspended by a pair of precision low-friction bearings at the rotor axle. The rotor bearings are supported by a circular inner gimbal ring. The inner gimbal ring in turn pivots on a second set of bearings attached to an outer gimbal ring. The pivoting action of the inner gimbal defines the horizontal axis of the gyro, which is perpendicular to the spin axis of the rotor. The outer gimbal ring is attached to the instrument frame by a third set of bearings.

If the gyroscope itself is mounted on a vehicle (e.g., aircraft) that is moving with respect to the Earth, then

$$\dot{H}_E = \dot{H}_B + \omega_{EB} H \tag{17.7}$$

The case of the gyroscope can be mounted on a platform so that it can rotate relative to the platform; then

$$\dot{H}_B = \dot{H}_C + \omega_{BC} H \tag{17.8}$$

Finally, the inner gimbal can rotate relative to the case; hence,

$$\dot{H}_C = \dot{H}_G + \omega_{GC} H \tag{17.9}$$

Substituting Equations 17.7 and 17.9 into Equation 17.4 yields

$$T = \dot{H}_G(\omega_{GC} + \omega_{BC} + \omega_{EB} + \omega_{IE})H \tag{17.10}$$

But

$$(\omega_{GC} + \omega_{BC} + \omega_{EB} + \omega_{IE}) = \omega_{IG} \tag{17.11}$$

Therefore,

$$T = \dot{H}_G + \omega_{IG} H \tag{17.12}$$

By carefully constructing the gyroscope and maintaining the spin velocity of the rotor constant, H_G can be made to be zero. Thus, the law of gyroscopes can be written as

$$T = \omega_{IG} H \tag{17.13}$$

This means that if an external torque T is applied to the gyroscope, the inner gimbal will precess with respect to the inertial frame with a velocity ω such that Equation 17.13 is satisfied.

In most designs (e.g., rate gyros), the gimbal is hermetically sealed in a liquid and liquid is floated in the case, to unload the gimbal bearings and to provide viscous damping. A pick-off senses gimbal deflection by means of position transducers, and it controls a servo system, with a servomotor driving the case to maintain pick-off null.

17.2.3 Optical Gyroscopes

Optical gyroscopes are based on the inertial properties of light instead of Newton's law of motion. They operate on the Sagnac effect, which produces interferometer fringe shift against the rotation rate. In this case, two light waves circulate in opposite directions around a path of radius R, beginning at source S. A typical arrangement for the illustration of operation principles is shown in Figure 17.6. When the gyro is stationary, the two beams arrive at the detector at the same time, and no phase difference will be recorded. Assume that the source is rotating with a velocity CO so that light traveling in the opposite direction to rotation returns to the source sooner than that traveling in the same direction. Thus, any rotation of the system about the spin axis causes the distance covered by the beam traveling in the direction of rotation to lengthen and the distance traveled by the beam in the opposite direction to shorten. The two beams interfere to form a fringe pattern and the fringe position may be recorded, or the phase differences of the two beams may be sensed. This phase difference is directional and proportional to the angular velocity. Usually, photodetectors are used to measure the phase shift.

Optical gyros can be categorized to be either passive or active and resonant or nonresonant. In passive gyrosensors, the Sagnac phase is measured by some external means, whereas in active gyros, the Sagnac phase causes a frequency change internal to the gyro that is directly proportional to the rotation rate.

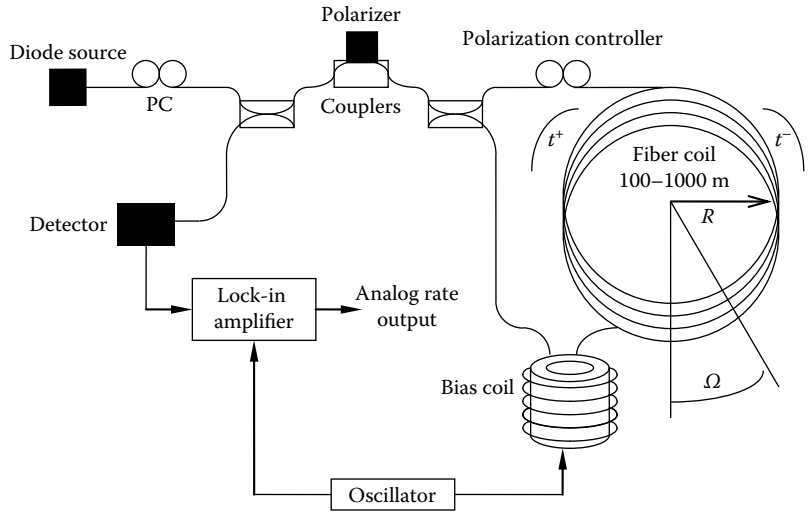


FIGURE 17.6 A typical fiber-optic gyroscope is based on the inertial properties of light, making use of the Sagnac effect. The Sagnac effect describes interferometer fringe shift against rotation rate. Two light waves circulate in opposite directions around a path of radius R, beginning at source S. When the gyro is stationary, the two beams arrive at the detector at the same time, and no phase difference is recorded. If the optical path is rotating with a velocity, the light traveling in the opposite direction to rotation returns to the source sooner than that traveling in the same direction. The two beams interfere to form a fringe pattern and the fringe position may be recorded, or the phase differences of the two beams may be sensed by photodetectors.

The Sagnac interferometer is the basis of the *interferometric fiber-optic gyro* (IFOG). A typical fiber-optic gyroscope is shown in Figure 17.6. However, the most widely used gyro is the active resonant *ring-laser gyro* (RLG), which is applied extensively in aircraft navigation. Two different types of resonant passive gyros, the *resonant fiber-optic gyro* (RFOG) and the *micro-optic gyro* (MOG), are lower-cost devices commonly used and comparable to RLGs.

17.3 Accelerometers

In inertial navigation, the absolute acceleration is measured in terms of three mutually perpendicular components of the total acceleration vector. Integrating these acceleration signals twice gives the displacement from an initial known starting location. Details of the acceleration and accelerometers are given elsewhere in this book (e.g., acceleration, vibration, and shock); therefore, accelerometers as applied to inertial navigations will be discussed here.

Accelerometers are made from three basic elements: proof mass, suspension mechanism, and pick-off mechanism. Some accelerometers require electric or magnetic force generators and appropriate servo loops. Accelerometers measure not only real vehicular acceleration but also respond to gravitational reaction forces. Acceleration due to gravity is a function of position—in particular, latitude and altitude—and is compensated by computers.

The most commonly used accelerometer in navigation systems is based on the pendulous types. They can be classified as

1. Generic pendulous accelerometer
2. Q-flex-type accelerometers
3. Micromachined accelerometers (A typical example of a modern micromachined accelerometer is given in Figure 17.7.)

Accelerations in the three axes are measured by suitably positioned accelerometers. Since accelerometers contain errors, the readings must be compensated by removing fixed biases or by applying scaling factors. The errors may be functions of operating temperature, vibration, or shock. Measurement of time must be precise as it is squared within the integration process for position determination. The Earth's rotation must also be considered, and gravitational effects must be compensated appropriately.

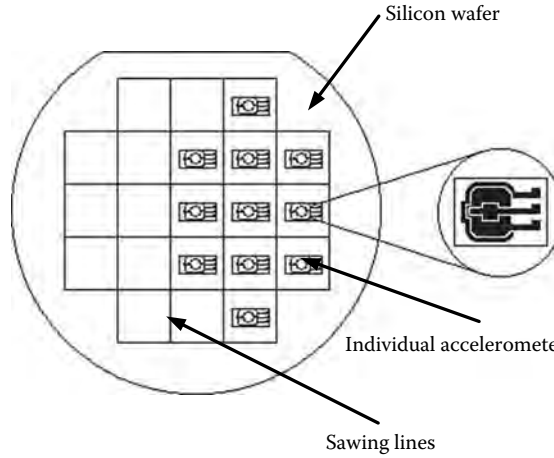


FIGURE 17.7 In modern micromachined accelerometers, multiple accelerometers can be mounted on a single chip to sense accelerations in x, y, and z directions. The primary signal conditioning is also provided in the same chip. The output from the chip is usually read in the digital form.

17.4 Errors and Stabilization in Inertial Navigation Systems

17.4.1 Errors

In general, inertial navigation is an initial value process in which the location of the navigating object is determined by adding distances moved in known directions. Any errors in the system cause misrepresentation of the desired location by being off target. The major disadvantage of an inertial guidance system is that its errors tend to grow with time. These errors in the deduced location are due to a number of reasons, including imperfect knowledge of the starting conditions, errors in computation, and mainly errors generated by gyros and accelerometers.

If the error buildup with time becomes too large, external aids (e.g., the global positioning system [GPS]) may be used to reset or update the system. Optimal use of the data from external aids must account for the geometry of the update and also for the accuracy of the update relative to the accuracy of the inertial system. The Kalman filter, for example, is one of the computational procedures frequently applied for optimally combining data from different sources.

Errors can broadly be classified as

1. *System heading error*: A misalignment angle in the heading of an object traveling with a velocity can cause serious errors. For example, a vehicle traveling with a velocity of 500 km h^{-1} in the same direction with 0.1° initial heading error will be off the target by approximately 873 m at the end of 1 h travel.
2. *Scale error*: Error in scaling can accumulate. In order to minimize scale errors, a scale factor is used. The scale factor is the ratio between changes in the input and output signals. It simply translates the gyro output (counts per second in the case of RLG) into a corresponding angle rotation. Some instruments may have different scale factors for positive and negative inputs, known as *scale factor asymmetry*. (Scale factors are measured in h^{-1}, mA^{-1}, h^{-1} Hz^{-1}, or g Hz^{-1}.)
3. *Nonlinearity and composite errors*: In most cases, scale factors are not constant, but they can have second- or higher-order terms relating the output signals to the input. Statistical techniques can be employed to minimize these errors.
4. *Bias errors*: Zero offset or bias error is due to existence of some level of output signal for a zero input. Bias errors exist in accelerometers, gyros, and tilt misalignments.
5. *Random drift and random walk errors*: In some cases, the outputs of the devices can change due to disturbances inside the sensors, such as ball bearing noise in mechanical gyros. These disturbances may be related to temperature changes, aging, etc. White noise in optical gyros can cause a long-term accumulation in angle error known as the *random walk*.
6. *Dead band, threshold, resolution, and hysteresis errors*: These errors can be related to inherent operation of accelerometers and gyros. They can be due to stiction, minimum input required for an output, minimum measurable outputs, and nonrepeatability of variations in the output versus variations in the input.

Note that this list is by no means exhaustive. For those readers with further interest, the detailed error analysis can be found in the bibliography.

17.4.2 Stabilization

The inertial navigation sensors must maintain angles within specified limits in spite of the disturbances imposed by the moving object. Accuracy requirements demand that the system must provide reliable and stable information in spite of vibrations and other disturbing factors. One way of achieving stabilization is by using a stabilized platform. These platforms are designed to maintain accelerometers and gyros accurately leveled and oriented in the azimuth direction. In some cases, the platform is driven around its axis by servo amplifiers and electric motors. Usually, outputs of doubly integrating

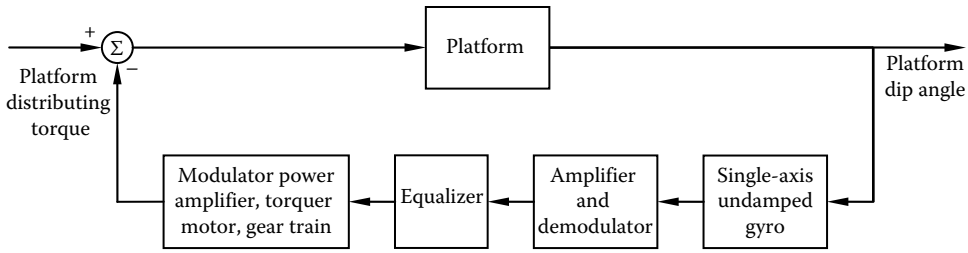


FIGURE 17.8 Stabilization is obtained using platforms designed to maintain accelerometers and gyros leveled accurately and properly oriented in the azimuth direction. In some cases, the platform is driven around its axis by servo amplifiers and electric motors. Sensitive pick-offs on the gyroscopes that fed error signals are used to maintain the desired stability of the platform in the presence of disturbances.

accelerometers are used directly to control the level-axis gyroscope precession rates. Sensitive pick-offs on the gyroscopes that fed error signals are used to maintain a desired stable platform in the face of disturbing torques. The operation of a typical system, in block diagram form, is shown in Figure 17.8.

Unlike platform models, in a strapdown system, gyroscopes and accelerometers are rigidly mounted to the vehicle structure so that they move with the vehicle. The accelerometers and gyroscopes are manufactured to measure accelerations and angles up to the maximum expected values. As the vehicle travels, the measured values are frequently transmitted to a computer. The computer uses these values to resolve the readings into the navigation axis sets and make deductions on the body axis sets.

17.5 Vehicular Inertial Navigation

In modern vehicular navigation, computerized maps and mobile communication equipment are integrated together with inertial and/or other electronic navigation systems. In recent years, in the wake of low-cost GPS systems, the vehicular navigation system has found a wide range of applications from consumer product to self-navigating systems.

Automobile navigation systems are based on dead-reckoning, map-matching, satellite-positioning, and other navigational technologies. Map intelligent systems achieve high relative accuracy by matching dead-reckoned paths with road geometry encoded in a computerized map. This is also used to perform other functions such as vehicle routing and geocoding. Satellite-based navigation systems achieve high absolute accuracy with the support of dead-reckoning augmentation. The capabilities and functions of automobile navigation systems depend on

- Choosing the necessary technology
- Integrating the overall system
- Resolving driver interface
- Providing map data basis
- Coordinating mobile communications

Digital maps and mobile data communications combine together for full usefulness and effectiveness. The navigation systems are greatly enhanced in conjunction with stored digital maps combined with effective communications.

The usefulness of a navigation system is related to the accuracy in position determination. There are a number of methods available with varying accuracy; these include the following:

Dead reckoning: Dead reckoning is the process of determining vehicle location relative to an initial position by integrating measured increments and directions of travel. The devices include the odometer, the differential odometer, and a magnetic compass. Gyros and inertial systems prove to have limited applications in harsh automotive environments. Although dead-reckoning systems suffer from error accumulation,

they are still used in INSs, particularly in robotics and vehicular applications. Even the most precise navigation system requires periodic re-initialization and continuous calibrations by computers.

Most of the dead-reckoning equipment commercially available is sufficiently robust to support map matching when operating in a defined road network. However, a good dead-reckoning accuracy is required to achieve re-initialization through map matching upon returning to the road network after off-road operations.

Radiolocation: In radiolocation, the GPS is used extensively.

Map matching: Artificial intelligence concepts are applied to match dead-reckoned vehicle paths, which are stored in computers. In map matching, sensed mathematical features of the vehicle paths are continuously associated with those encoded in a map database. Thus, a vehicle's dead-reckoned location can be initialized automatically at every turn to prevent accumulation of dead-reckoning errors.

The first application of map-matching technology was in the automatic route control system (ARCS), which used a differential odometer for dead reckoning. In another system, the Etak map-matching system, a solid-state flux gate compass, is used as well as a differential odometer to dead-reckon paths for matching with digitized maps and aerial photographs.

In a map-matching system, as long as the streets and road connectivities are accurately defined, the process identifies position relative to the road network as visually perceived by the automatic system or by the driver of the vehicle.

Proximity beacon: This uses strategically located short-range transmitters, and reception of their location-coded signal infers the receiving vehicle's instantaneous location. There are several variations of the proximity approach; some versions involve two-way communications with the equipped vehicle. Typically, the driver enters the destination code on the vehicle panel, for automatic transmission to the roadside unit, as the vehicle approaches an instrumented intersection. The roadside unit, which can be networked with a traffic management system, analyzes the destination code and transmits route instructions to the display on the vehicle panel.

The approach to the interface between an onboard navigation system and a vehicle operator must take into account ergonomics and safety considerations as well as functional requirements. As a result of intensive research, especially in the aerospace industry, display of information for the operator is a well-developed area. In a well-known European system, Philips' CARiN, a color CRT map display, is used to show vehicle location relative to the surroundings. Many other systems use short visual messages, symbolic graphics, and voice.

Data communications in modern navigation systems provide current updates (road additions, closures, detours, etc.) for onboard map databases. It also provides real-time information on traffic conditions for onboard route generation.

17.6 Applications of Inertial Navigation

17.6.1 Aircraft

The primary navigation aid for civil aircraft flying in the airspace of most of the developed countries is the very-high-frequency omnirange (VOR)/distance-measuring equipment (DME) system. The VOR and the DME enable onboard determination of an aircraft's bearing relative to north at the fixed ground station and slant range from the station, respectively.

Many commercial aircraft are still equipped with precision INSs not only for navigation purposes but also for stabilization of the aircraft at all times. Suitably located gyroscopes and accelerometers give signals to control the stability of the aircraft, as illustrated in Figure 17.9.

Many aircraft INSs utilize a gyro-stabilized platform on which the accelerometers are mounted. The platform is aligned before take-off to the desired orientation. Due to alignment errors and in-flight

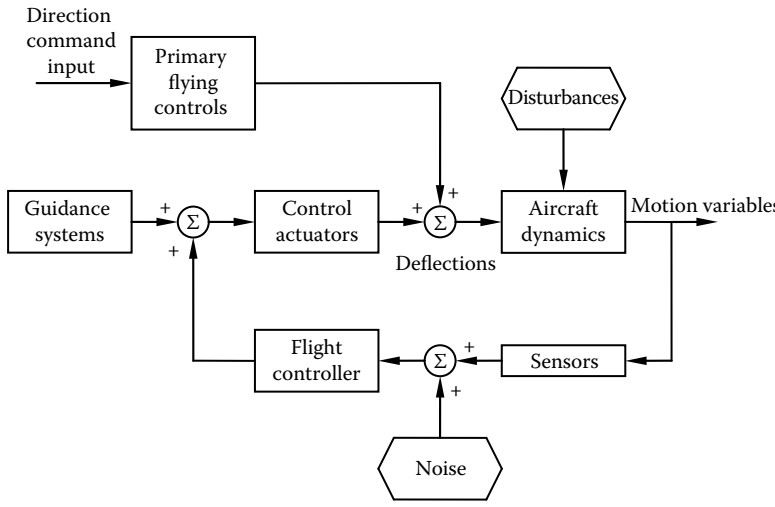


FIGURE 17.9 In many commercial aircraft, suitably located gyroscopes and accelerometers give signals for controlling the stability of the aircraft. Due to various instrumental errors and in-flight random disturbances such as gyro drift, scale factor errors, and accelerometer bias errors, the errors in the desired orientation increase with time. These accumulated errors need to be compensated periodically by information from the GPS systems.

random disturbances such as gyro drift, scale factor errors, and accelerometer bias errors, the errors in the desired orientation increase with time. The errors need to be compensated periodically by external information such as radar and GPS navigation systems.

GPS systems give all vehicles on or near the Earth unprecedented navigation accuracy. Most of the airlines are equipped with confined GPS receivers. Airborne receivers are designed that combine navaids operating in a common radio bands.

17.6.2 Underwater

The Ship's Inertial Navigational System (SINS) was originally developed for precision position finding required by ballistic missile submarines in the late 1950s and the early 1960s. The first deployment was onboard U.S. George Washington in 1960, and SINS are used today in submarines, aircraft carriers, and other surface warships. As the cost and size are continually decreasing, the system is also deployed in naval as well as merchant vessels. Another development of INS for underwater application is in the area of the autonomous underwater vehicle (AUV). In this section, a number of such products are described.

AUVs are used extensively for military and civilian purposes. Application examples are mapping, surveillance, ocean exploration, survey, and mining, all of which require precise position determination. The desired features of such systems are low power, high accuracy, small volume, light weight, and low cost. Two typical examples were given as follows:

KN-6050 underwater navigators are based on Kearfott's Monolithic RLG INS system. It is assisted by a Doppler Velocity Log (DVL) and GPS position inputs. They find applications in AUVs as well as surface unmanned vehicles (SUV) and remotely operated vehicles (RUV). It requires a multitude of inputs, such as depth sensor, position inputs, and speed and sound sensors. Navigation outputs are position, altitude and depth, ground speed, and heading. These devices provide vehicle stabilization and control by monitoring angular rates, linear accelerations, roll, and pitch.

The second example is the HUGIN 1000 AUV, developed by Konsberg, Norway, suitable for military applications as well as ocean research and environmental monitoring. HUGIN 1000 uses hydroacoustic-aided INS, which combines acoustic and inertial positioning principles. It also uses various types of transponders.

17.6.3 Robotics

Autonomous mobile robots also use INS extensively as a self-contained, independent navigation system. Typical applications are mining, unknown terrain exploration, and off-line path planning. There are many commercially available INSs suitable for cost-effective utilization in the navigation of robots. Some of these are gyrocompasses, rate gyros, gyrochip, piezoelectric vibrating gyros, RLGs, interferometric, and other types of fiber-optic gyros.

The Honeywell Modular Azimuth Position System (MAPS): The MAP-H is a typical example of an INS for land-based vehicles. It consists of a dynamic reference unit (DRU) that provides processed information from the inertial sensor assembly, a control display unit (CDU) that is used for human–machine interface, and a vehicle motion sensor (VMS) that monitors the vehicle's directional and distance information. The inertial sensor assembly comprises three Honeywell ring-laser gyros and three accelerometers mounted to measure movements in three local axes. The inertial processor translates the information to the navigation processor that resolves the vehicle movement information from the VMS. The CDU provides mode selection, data display, waypoint information, and general operator interface.

Bibliography

Aggarwall, P., Syed, Z., Noureldin, A., El-Sheimy, N., *MEMS-Based Integrated Navigation*, Artech House, Norwood, MA, 2010.

Bijker, J., Kalman filter configurations for a low-cost loosely, integrated inertial navigation system on an airship, *Control Engineering Practice*, 16(12), 1509–1518, 2008.

Bird, J., Arden, D., Indoor navigation with foot mounted strapdown inertial navigation and magnetic sensors, *IEEE Wireless Communications*, 18(2), 35–42, 2011.

Hablani, H. B., Autonomous inertial relative navigation with sight-line stabilized integrated sensors for spacecraft randevous, *Journal of Guidance, Control, and Dynamics*, 32(1), 172–183, 2009.

Lawrence, A., *Modern Inertial Navigation Technology: Navigation, Guidance, and Control*, 2nd edn., Springer, London, U.K., 1998.

Qin, F. J., Xu, J. N., Jiang, S., A new scheme of gyroscope free inertial navigation system using 9 accelerometers, *2009 International Workshop on Intelligent Systems and Applications*, ISA 2009, Wuhan, China, May 2009.

Sonmez, T., Bingol, H. H., Modelling and simulation of a terrain aided inertial navigation algorithm, for land vehicles, *IEEE Position, Location and Navigation Symposium*, Monterey, Canada, pp. 1046–1052, 2008.

Titterton, D. H., Weston, J. L., *Strapdown Inertial Navigation Technology*, 2nd edn., The Institution of Electrical Engineering, Stevenage, U.K., 2004.

Woodman, O. J., An introduction to inertial navigation, Technical Report No. 696, University of Cambridge, Cambridge, U.K., 2007. Available at http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-696.pdf (accessed on March 3, 2012).

Partial List of Manufacturers/Suppliers

ACUTRONICUSA, Inc., 700 Waterfront Dr., Pittsburgh, PA 15222-4742, Phone: 412-926-1200, Fax: 412-697-8111, www.acutronic.com

Astronautics Corp. of America, 4115 N. Teutonia Ave., Milwaukee, WI 53209-6731, Phone: 414-449-4000, Fax: 414-447-8231, www.astronautics.com

Automatic Equipment Mfg. Co./BlueOx, 1 Mill Rd, P.O. Box P, Pender, NE 68047, Phone: 402-385-3051, Fax: 402-385-3360, www.aemfg.com

Boeing Co., 1215 N. 2200 W., Salt Lake City, UT 84116-4128, Phone: 801-537-6400, www.boeing.com

Condor Pacific Industries, Inc., 31829 La Tienda Dr., Westlake Village, CA 91362-4011, Phone: 818-865-3000, Fax: 818-865-3010, www.condorpacific.com

Crossbow Technology, Inc., 4145 N. 1st St., San Jose, CA 95134-1504, Phone: 408-965-3300, Fax: 408-324-4840, www.xbow.com

Davis Instruments Corp., 3465 Diablo Ave., Hayward, CA 94545, Phone: 510-732-9229, 800-678-3669 (toll free), Fax: 510-670-0589, www.davisnet.com

Electron Energy Corp., 924 Links Ave., Landisville, PA 17538-1615, Phone: 717-898-2294, 800-824-2735 (toll free), Fax: 717-898-0660, www.electronenergy.com

Frog Navigation Systems, 691 N. Squirrel Rd Ste 115, Auburn Hills, MI 48326, Phone: 248-373-7200, Fax: 248-373-7400, www.frogusa.com

Garmin Corporation, 1200 E. 151st St., Olathe, KS 66062-3426, Phone: 913-397-8200, 800-525-6726 (toll free), Fax: 913-397-8282, www.garmin.com/garmin/cms/site/us

Hitachi Cable America, Inc., Optical Div., 3031 Tisch Way, Ste. 150, San Jose, CA 95128, Phone: 408-260-2630, Fax: 408-260-2756, www.hitachi-cable.com

Honeywell International, Inc., 101 Columbia Rd., Morristown, NJ 07962-1057, Phone: 973-455-2000, 800-707-4555 (toll free), Fax: 973-455-4807, honeywell.com/Pages/Home.aspx

Inertial Labs, 13800 Coppermine Rd., Suite 300, Herndon, VA 20171, Phone: 703-880-4222, Fax: 703-991-5378, www.inertiallabs.com

Jae Electronics, Inc., 142 Technology Dr. W. # 100, Irvine, CA 92618-2430, Phone: 949-753-2600, Fax: 949-753-2699, www.jae.com

Kearfott Guidance & Navigation Corp., 1150 McBride Ave., Little Falls, NJ 07424, Phone: 973-785-6000, Fax: 973-785-5400, www.kearfott.com/

Litton Guidance & Control Systems, 5500 Canoga Ave., Woodland Hills, CA 91367, Phone: 818-715-4040, Fax: 818-715-4467, http://www.littoncorp.com/

Lockheed Martin, 6417 Deere Rd., Syracuse, NY 13206, Phone: 315-456-3660, www.lockheedmartin.com/

Mi-Tech Metals, Inc., 4701 Massachusetts Ave., Indianapolis, IN 46218-3144, Phone: 317-549-4290, 800-624-1895 (toll free), Fax: 317-549-4295, www.mi-techmetals.com

Mil-Spec Industries Corp., 10 Mineola Ave., Roslyn Heights, NY 11577, Phone: 516-625-5787, Fax: 516-625-0988, www.mil-spec-industries.com

Navhouse Co., 10 Loring Drive, Bolton, ON L7E1J9, Phone: 905-857-8102, Fax: 905-857-8104, www.navhouse.com

Navico, Inc., 1010 S.E. Everett Mall Way Ste. 101, Everett, WA 98208, Phone: 603-324-2042, Fax: 603-324-2043, http://www.navico.com/en/

Northrop Grumman Corporation, 755 Dedham St., Canton, MA 02021, Phone: 866-646-2879 (toll free), http://www.northropgrumman.com/

Northstar, 30 Sudbury Rd., Acton, MA 01720, Phone: 978-897-6600, 800-628-4487 (toll free), Fax: 978-897-8264, www.northstarnav.com/

Precision Navigation, Inc., 5464-T Skylane Blvd., Santa Rosa, CA 95403-1084, Phone: 707-566-2260, Fax: 707-566-2260, www.pnicorp.com/

Rockwell International Corp., Aerospace & Defense, Autonetics & Missile Systems Div., 3370-T Miraloma Ave., Anaheim, CA 92803, Phone: 714-762-8111, Fax: 714-762-1243, http://www.rockwellautomation.com/

Select Controls, Inc., 45 Knickerbocker Ave., Bohemia, NY 11716-3110, Phone: 866-929-5312 (toll free), Fax: 631-567-9013, www.select-controls.com

Southern Avionics Company, P.O. Box 5345, Beaumont, TX 77726-5345, Phone: 409-842-1717, 800-648-6158 (toll free), Fax: 409-842-2987, www.southernavionics.com

Systron Donner Inertial, 2700 Systron Dr., Concord, CA 94518, Phone: 925-671-6400, Fax: 925-671-6647, www.systron.com

Trimble Navigation, Ltd., 936 Stewart Dr., Sunnyvale, CA 94088, Phone: 408-481-8000, www.trimble.com

Warren-Knight Instrument Co., 2045 Bennett Dr., Philadelphia, PA 19116-3019, Phone: 215-464-9300, Fax: 215-464-9303, www.warrenknight.com

18

Level Measurement

18.1 Measurements Using the Effects of Density **18**-2
Displacer • Float • Pressure Gages • Balance Method
18.2 Time-of-Flight Measurements **18**-6
Basic Principle • Ultrasonic • Microwaves • Laser/Light
18.3 Commonly Used Evaluation Methods **18**-11
Frequency-Modulated Continuous-Wave RADAR • Time-of-Flight through Product
18.4 Level Measurements by Detecting Physical Properties **18**-13
Electrical Properties • Capacitive • Conductive • Radiation Attenuation • Radiometric • Ultrasonic Switch • Microwave Switch • Photoelectric Barrier • Thermal and Mechanical • Thermal • Viscosity
18.5 Level Measuring Instruments **18**-17
References **18**-17
Further Readings **18**-18
Partial List of Manufacturers and Suppliers **18**-18

Detlef Brumbi
Krohne Messtechnik GmbH

Level is defined as the filling height of a liquid or bulk material, for example, in a tank or reservoir. Generally, the position of the surface is measured relative to a reference plane, usually the tank bottom. If the product's surface is not flat (e.g., with foam, waves, turbulences, or with coarse-grained bulk material), level usually is defined as the average height of a bounded area.

Various classic and modern methods exist to measure product level in process and storage tanks in the chemical, petrochemical, pharmaceutical, water, and food industries, in mobile tanks on vehicles and ships, but also in natural reservoirs like seas, dams, lakes, and oceans. Typical tank heights are approximately between 0.5 and 40 m.

Two different tasks can be distinguished: (1) continuous level measurements (level indication, LI) and (2) level switches (LSs) (e.g., to detect an alarm limit to prevent overfilling). Figure 18.1 shows the principal operational modes of level measurement. Every continuous system can also be used as a programmable switch. Many level devices are mounted on top of the tank and measure primarily the distance d between their mounting position and the product's surface. The level L is then calculated, defining the tank height h as constant, as shown in Figure 18.1 and expressed as:

$$L = h - d \tag{18.1}$$

The following examples describe primarily the measurement of liquids, but most of the methods can also be applied to solids (bulk material). The emphasis of this chapter will be general information about the measurement principles. The focus of the descriptions is on the methods most commonly practiced; other principles are mentioned less comprehensively. Readers interested in more detailed discussions may refer to [1–4].

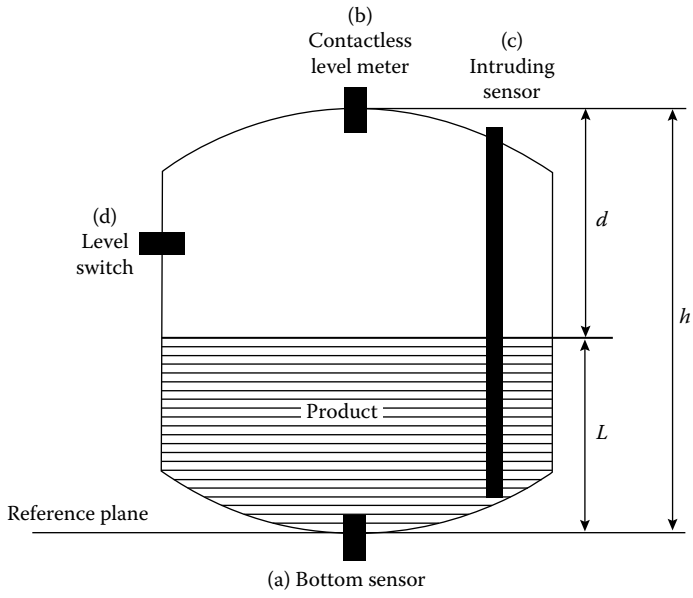


FIGURE 18.1 Representation of a tank with a liquid or solid material (hatched area), the product to be measured. The level sensor can be mounted (a) contacting product at the bottom, (b) as a contactless instrument on top, (c) as an intrusive sensor, or (d) at the sides as an LS.

18.1 Measurements Using the Effects of Density

All methods described in this chapter have in common that the product in the tank has an effect due to its density ρ, (1) producing buoyancy to a solid submerged into the liquid or (2) executing a force due to its weight.

18.1.1 Displacer

Displacers measure the buoyancy of a solid body that is partially submerged in the liquid. The change in weight is measured. Figure 18.2 illustrates the parameters used for these calculations. The cross section A of the body is assumed to be constant over its length b. The weight of force F_G due to gravity g and mass m is

$$F_G = gm = gAb\rho_D \tag{18.2}$$

The buoyant force F_B accounts for the partial length L_d that is submerged with the remainder of the body in the atmosphere:

$$F_B = gAL_d\rho_L + gA(b - L_d)\rho_A \tag{18.3}$$

Combining Equations 18.2 and 18.3 gives the resulting force to be measured by an appropriate method as follows:

$$F_R = F_G - F_B \tag{18.4}$$

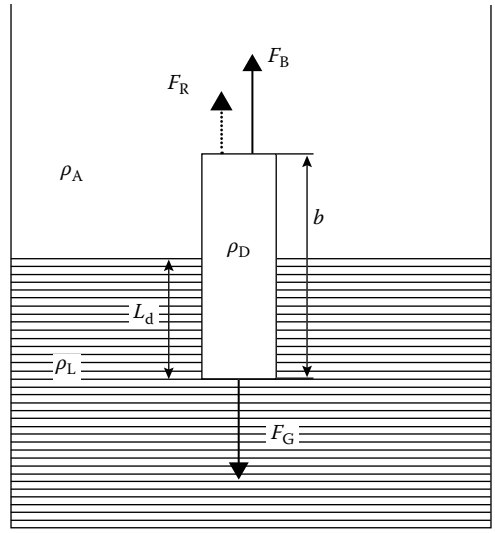


FIGURE 18.2 Quantities of a solid body immersed into a liquid. The forces F can be calculated from Equations 18.2 through 18.4. ρ = density, b = length of the body, and L_d = dipped length.

The result for level L_d, related to the lower edge of the displacer, is

$$L_d = \frac{b(\rho_D - \rho_A) - F_R/gA}{\rho_L - \rho_A} \tag{18.5}$$

The density of the body should be higher than the density of the liquid; otherwise, the measurement operating range is limited (until the displacer floats on the liquid). In another version, a servo gage moves the displacer up and down to detect the interface between the atmosphere and a liquid, or between two different liquids, by measuring the change in buoyancy. Figure 18.3 shows a special configuration, in which a small ball with volume V is mounted to a thin wire driven by a stepping motor and put into resonant vibration. The resulting force F can be measured from the resonating frequency f of the wire between points A and B:

$$F = \rho_w A_w 4 f^2 l^2 \tag{18.6}$$

where

l is the length of the wire between the points A and B
ρ_w is the density of the wire
A_w is the cross-sectional area of the wire

And the surrounding density ρ_L can be calculated as follows:

$$F = gV(\rho_D - \rho_L) \Leftrightarrow \rho_L = \rho_D - \frac{F}{gV} \tag{18.7}$$

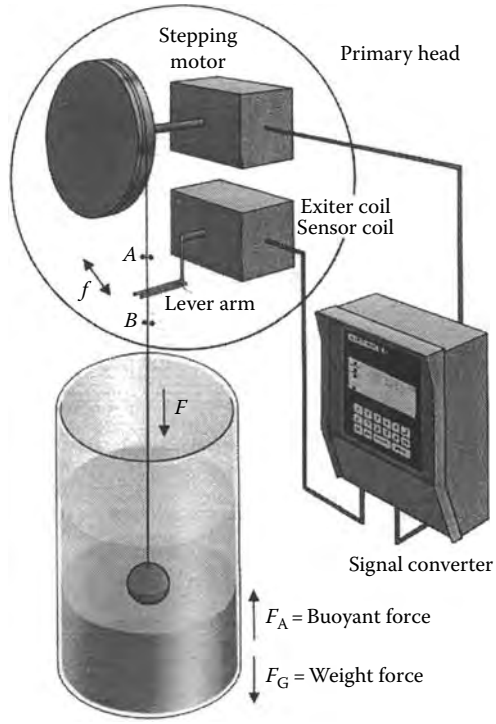


FIGURE 18.3 Level, interface, and density sensor using the effects of buoyancy. A stepping motor drives the small ball attached to the thin wire to different heights in the liquid or to the interfaces. The resulting force F as a difference between weight force and buoyant force is measured from the resonant frequency of the wire-ball system. The lever arm excites the wire into oscillation and a sensor coil coupled to the lever arm measures its frequency. The signal converter controls the stepping motor and calculates the measured values. (From Messtechnik, K., *Technical Data Sheets of Level Measurement Products*, Krohne, Duisburg, Germany, 1996.)

18.1.2 Float

Floats are similar to displacers but are swimming on the liquid's surface due to the buoyancy. Hence, the density of the float must be lower than the density of the liquid. Figure 18.4a shows the principle of operation. The position of the float is (1) observed visually or (2) transferred to an external display or to an angle transmitter. In general, the float is coupled to the transmitter magnetically. Figure 18.4b shows an LS, using a reed relay magnetically coupled with the float. Also, a magnetostrictive linear sensor may determine the position of the float.

If the float is very flat, it is called a "sensing plate." This plate is mechanically guided, for example, by a servo control, on the surface until uplift is detected. For solids, specially shaped perpendicular floats are helpful.

18.1.3 Pressure Gages

A hydrostatic pressure p, caused by the weight of the product, is present at the bottom of a tank, in addition to the atmospheric pressure p_0:

$$p = p_0 + g\rho_L L \Leftrightarrow L = \frac{p - p_0}{g\rho_L} \tag{18.8}$$

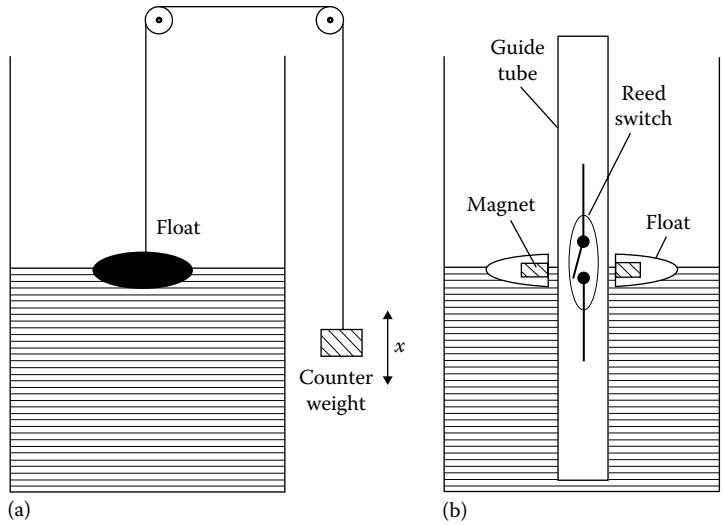


FIGURE 18.4 Principle of operation for float level meters. (a) *A* counterweight balances the float that swims on the liquid's surface. Its position represents the level; (b) the float contains a magnet that contacts a reed switch inside a guide tube. Using a bistable relay, this system is used as an LS. One can also insert multiple relays into the tube to achieve different switching points for quasicontinuous operation.

Pressure gages at the bottom of the tank measure this pressure. In process tanks with varying atmospheric pressure, a differential pressure measurement is achieved by measuring the difference between the pressure at the bottom and that at the top of the tank, above the liquid. Figure 18.5a shows such a configuration with a differential pressure sensor. Because measurement by hydrostatic pressure is proportional to the density, level errors result if density changes; see Equation 18.8. Primary pressure gaging is a mass measurement. Figure 18.5b shows a vertical arrangement with three sensors; the measurements of p_1 and p_2 are used to compensate for the influence of density ρ_L and to calculate the level:

$$\rho_L = \frac{p_2 - p_1}{gl} \Rightarrow L = \frac{p_2 - p_0}{p_2 - p_1} l \tag{18.9}$$

A system of this configuration is often called "hydrostatic tank gaging" (HTG). Figure 18.5c shows a further arrangement, called "bubble tube," in which the bottom pressure is transmitted to the top of the tank. This is often used for level gaging if the sensor cannot be mounted at the bottom of the tank. It requires a tank with pressure equalization due to the steady insertion of inert gas.

18.1.4 Balance Method

Here simply the weight F of the complete tank is measured, dependent on the level L:

$$F = F_0 + gAL\rho_L \tag{18.10}$$

where

F_0 is the weight of the empty tank
A is the cross-sectional area

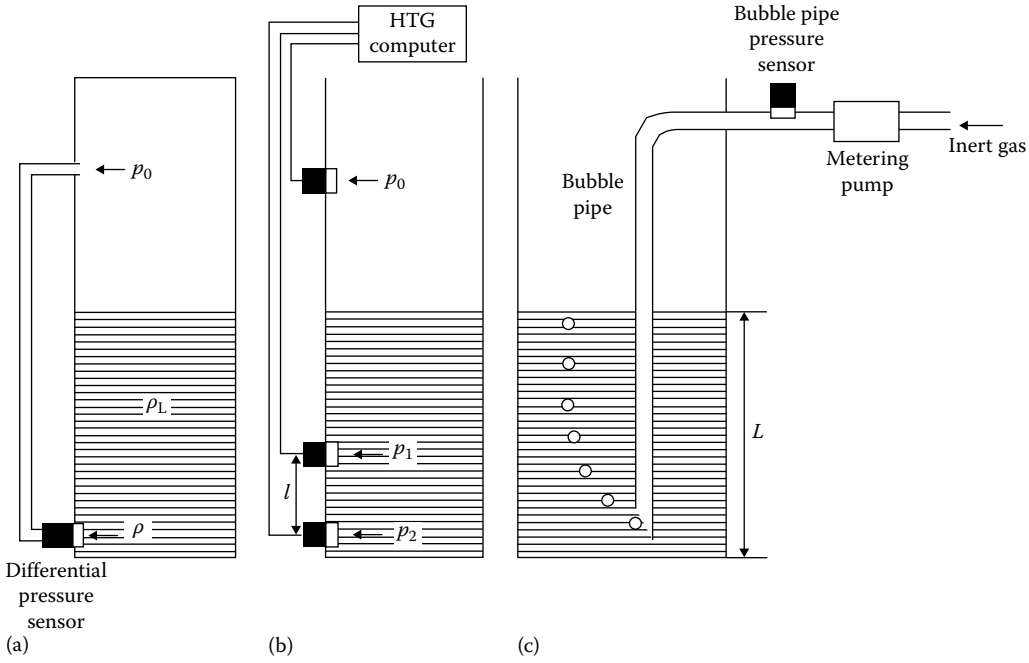


FIGURE 18.5 Level gaging by hydrostatic pressure measurement. The bottom pressure p is proportional to level. (a) The atmospheric pressure p_0 can be taken into consideration by a differential measurement. The low side of the differential pressure sensor is connected via a thin pipe to the top of the tank. (b) A differential measurement within the liquid is called "hydrostatic tank gaging" and can be used to compensate for errors due to density variations of the liquid. The signals from all three sensors are evaluated by a computer. (c) With a so-called bubble tube, the sensor can be mounted on the top of the tank: an inert gas is injected into the tube such that bubbles of gas escape from the end of the tube. The flow rate of the gas is constant, so the head pressure in the system can be measured at the inlet of the pipe.

which is assumed to be constant throughout the tank height. In order to measure the weight force correctly, it is necessary to isolate the complete tank mechanically. For precise measurements, the buoyancy in air must be taken into consideration:

$$F = F_0 + gAL(\rho_L - \rho_A) \Leftrightarrow L = \frac{F - F_0}{gA(\rho_L - \rho_A)} \tag{18.11}$$

This method has severe disadvantages when the tank is not inside a building. Outside, wind forces and the weight of snow and rain can cause errors.

18.2 Time-of-Flight Measurements

An indirect measurement of level is evaluating the time of flight of a wave propagating through the atmosphere above the liquid or solid. This is primarily a distance measurement; the level can then be calculated accordingly. The increasing demand of industry for nonintrusive continuous level gaging systems has been instrumental in accelerating the development of technologies using time-of-flight measurements [5].

18.2.1 Basic Principle

Although different types of physical waves (acoustic or electromagnetic) are applied, the principle of all these methods is the same: a modulated signal is emitted as a wave toward the product, reflected at its surface and received by a sensor, which in many cases is the same (e.g., the ultrasonic piezoelectric transducer or the RADAR antenna). Figure 18.6 demonstrates the principle of operation. The measuring system evaluates the time of flight t of the signal:

$$t = \frac{2d}{v} \tag{18.12}$$

where v is the propagation velocity of the waves.

One can generate an unmodulated pulse, a modulated burst as in Figure 18.6b, or special forms. Table 18.1 lists the main properties of the three preferred types of waves, used for time-of-flight level gaging.

The very short time spans of only a few nanoseconds for RADAR and laser measurement techniques require the use of time expansion by sampling as discussed later.

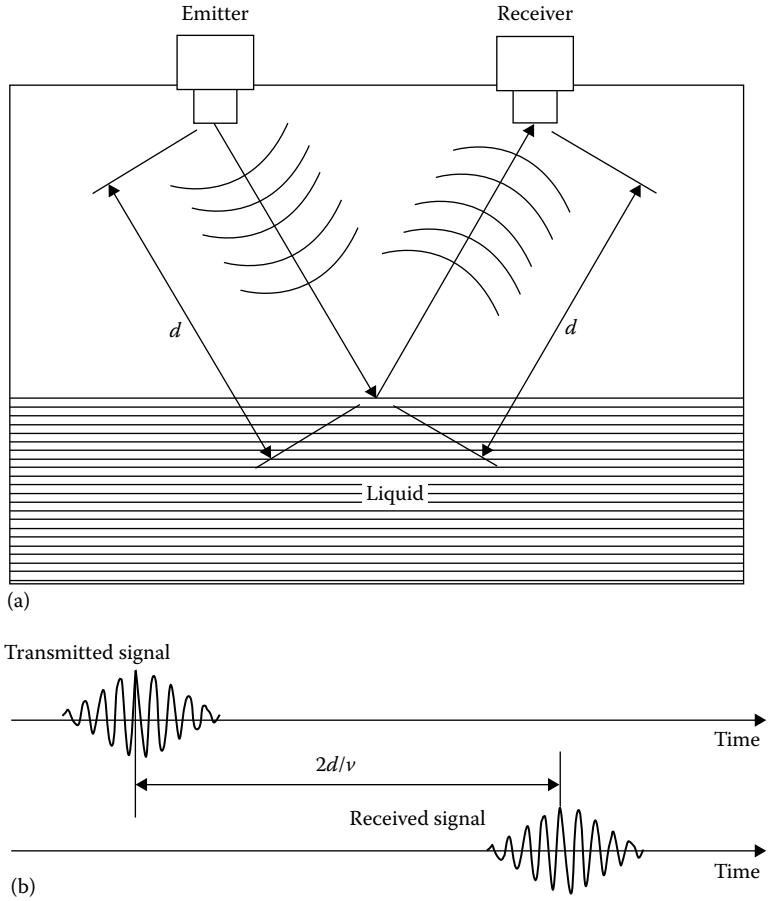


FIGURE 18.6 (a) Representation of time-of-flight measurements. The emitter couples a wave (ultrasonic or electromagnetic) into the atmosphere that propagates the wave toward the liquid. Its surface reflects the wave and a sensor receives it. (b) Due to the propagation velocity v, a time delay is measured between emission and receipt of the signal. This example is characterized by a modulated burst. The time scale is arbitrary.

TABLE 18.1 Properties of the Wave Types for Time-of-Flight Measuring

Principle	Wave Velocity	Avg. Carrier Frequency	Wavelength	Avg. Burst Time
Ultrasonic	340 m s^{-1}	50 kHz	7 mm	1 ms
RADAR	300,000 km s^{-1}	10 GHz	3 cm	1 ns
Laser	300,000 km s^{-1}	300 THz	1 µm	1 ns

18.2.2 Ultrasonic

Ultrasonic waves are longitudinal acoustic waves with frequencies above 20 kHz. Ultrasonic waves need a propagation medium, which for level measurements is the atmosphere above the product being measured. Sound propagates with a velocity of about 340 m s^{-1} in air, but this value is highly dependent on temperature and composition of the gas and also on its pressure. In vacuum, ultrasonic waves cannot propagate. In practice, the reflection ratio is nearly 100% at the product's surface (e.g., at transitions gas/liquid or gas/solid). Piezoelectric transducers are utilized as emitter and detector for ultrasonic waves, a membrane coupling it to the atmosphere. The sensor is installed as in Figure 18.1b, the signal form is as in Figure 18.6b. Level gaging is, in principle, also possible with audible sound 16 Hz–20 kHz or infrasonic waves less than 16 Hz.

Another procedure is to propagate the waves within the liquid by a sensor mounted at the bottom of the tank. The velocity of sound in the liquid must be known, considering the dependence on temperature and type of liquid. This method is similar to an echo sounder on ships for measuring the water depth.

18.2.3 Microwaves

Microwaves are generally understood to be electromagnetic waves with frequencies above 2 GHz and wavelengths of less than 15 cm. For technical purposes, microwave frequencies are used up to a maximum of 120 GHz; in practice, the range around 10 GHz (X-band) is preferred.

The usually applied time-of-flight measurements with microwaves are RADAR based [6,7]. The term "RADAR" is generally understood to mean a method by means of which short electromagnetic waves are used to detect distant objects and determine their location and movement. It is an acronym from radio detection and ranging. Figure 18.7 shows preferred antenna forms. They are usually combined with a compact sensor, as in Figure 18.8. For level measuring systems, a small radiation angle is desirable in order to avoid interfering reflections from the tank wall or tank internals as much as possible. The larger the aperture area, the smaller the radiation angle and the higher the antenna gain. The power balance is given by the general RADAR equation:

$$P_R = \frac{P_T G_T R G_R}{D^2} \tag{18.13}$$

where

P_R is the received power
P_T is the transmitted power
G_T is the transmitting antenna gain
R is the reflection factor of target
G_R is the receiving antenna gain
D^2 is the propagation loss to and from the surface, due to power density decrease and atmospheric influences

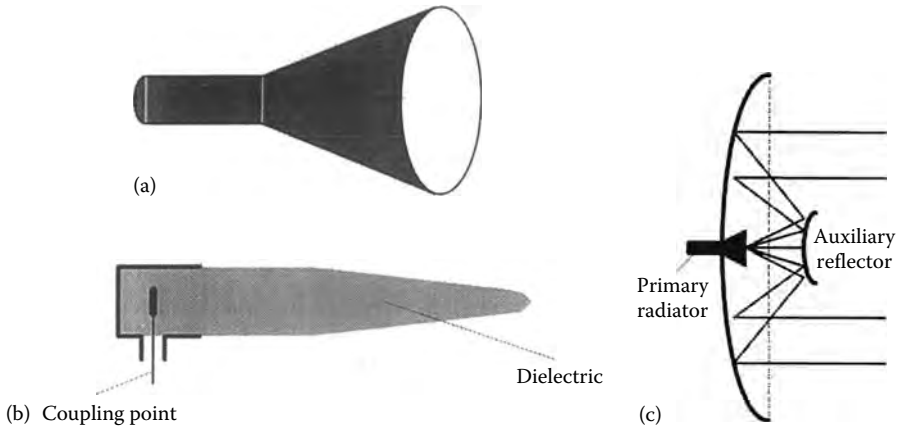


FIGURE 18.7 Practical antenna forms used for RADAR level instruments: (a) conical horn antenna, (b) dielectric rod antenna, and (c) parabolic mirror with a small antenna as primary radiator and an auxiliary reflector giving a very small beam angle (so-called Cassegrain model).



FIGURE 18.8 Design of a compact industrial level RADAR system. The converter above the flange includes the complete microwave circuitry, signal-processing stages, microprocessor control, display, power supply, and output signal. (From Messtechnik, K., *Technical Data Sheets of Level Measurement Products*, Krohne, Duisburg, Germany, 1996.)

The reflection factor R of the product's surface is dependent on the dielectric permittivity ε_r of the liquid or bulk material:

$$R = \frac{\left(\sqrt{\varepsilon_r} - 1\right)^2}{\left(\sqrt{\varepsilon_r} + 1\right)^2} \tag{18.14}$$

In level measurement situations, the reflecting area is so large that it intersects the beam cross section completely; therefore, D^2 is approximately proportional with distance d^2. Thus also, the received power decreases proportionately with d^2, as derived in [6]:

$$P_R \propto \frac{1}{d^2} \tag{18.15}$$

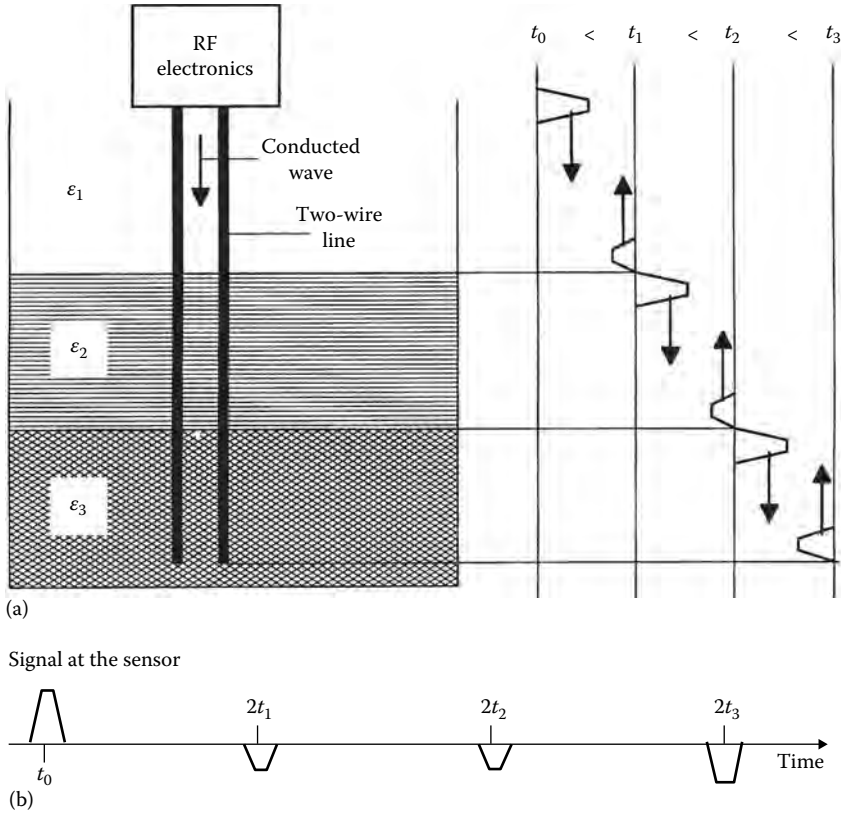


FIGURE 18.9 Principle of operation of a wire-conducting high-frequency level measurement system. (a) An electrical pulse is generated (time t_0) and a two-wire line guides the electromagnetic wave. At every position where the surrounding permittivity e changes, a part of the wave is sharply reflected (time t_1) back to the sensor. The wave propagates along the entire line and is reflected a second time (t_2) at the interface between the two liquids and a third time at the end of the line. (b) The signal delay times ($2t_1$, $2t_2$, and $2t_3$) represent the positions of the interfaces with respect to the end of the line, which can be used as a reference. The signal polarity is inverted due to the reflection from lower to higher permittivity. The time scale is arbitrary.

This is not the case if the waves propagate within an electromagnetic waveguide, like a vertical tube dipping into the liquid, called a stilling well. Here, the propagation is nearly without losses.

An alternative method using electromagnetic waves is to propagate them in a cable. Figure 18.9a illustrates the operation with a cable dipped into the liquid or bulk material. In this case, the dielectric permittivity of the surrounding medium changes; part of the wave is reflected. This method can be applied to interface measurements too. Figure 18.9 shows the signals in an application with a two-phase product. This method is called "time domain reflectometry" (TDR).

18.2.4 Laser/Light

Lasers and light-emitting diodes produce electromagnetic waves of very short wavelength (less than 2 μm), which can also be used for time-of-flight measurements, similar to the described microwave methods. Preferred laser signals are (1) short pulses of less than 1 ns duration or (2) lasers with amplitude-modulated intensity with frequencies of some megahertz. For more details about laser operation and interferometry methods, refer to Section 6.5 of this handbook.

Laser systems are very precise and can achieve accuracies better than 1 mm. Because the laser beam is very narrow, such level measurement systems can be installed without influence of tank internals. Some practical disadvantages of laser level measurement are as follows: (1) It functions as does your eye to see the product's surface and therefore fails if dust, smoke, etc. are present; (2) it is sensitive to dirt on the optical sensors; and (3) the equipment is expensive.

18.3 Commonly Used Evaluation Methods

Due to the great benefits of contactless time-of-flight measurement, some typical methods have been evaluated for level gaging within the last few years, mainly in RADAR techniques [6].

18.3.1 Frequency-Modulated Continuous-Wave RADAR

Because the flight times in typical level applications are very short (a resolution of 1 mm requires a 7 ps time resolution), it is difficult to evaluate information directly in the time domain. By modulation of the microwave signals, the time delay is transformed into the frequency domain, obtaining low-frequency signals.

Therefore, frequency-modulated continuous-wave (FMCW) RADAR has been established as the dominant technique. FMCW RADAR utilizes a linearly frequency-modulated microwave signal; the transmission frequency rises linearly in a time interval T. The frequency difference in this interval is called the frequency sweep F.

Figure 18.10 shows the principle of FMCW RADAR. Due to the time delay during signal propagation, the transmitted frequency changes such that the difference between the momentary transmitted frequency and the received frequency, a low-frequency signal, is obtained. The frequency f of that signal (typically up to a few kilohertz) is proportional to the reflector distance d (see Figure 18.6); in this method, therefore, the delay t is transformed into a frequency f:

$$f = \frac{F}{T}t = \frac{F}{T}\frac{2d}{c} \Leftrightarrow d = \frac{fcT}{2F} \tag{18.16}$$

In Equation 18.16, c is the speed of light and F/T is the sweep velocity; see Figure 18.10.

Figure 18.11 shows a basic circuit block diagram of an FMCW RADAR system. Because the resultant signal frequencies are low, further signal processing is technically simple and very accurate. Normally, evaluation is by means of digital signal processing.

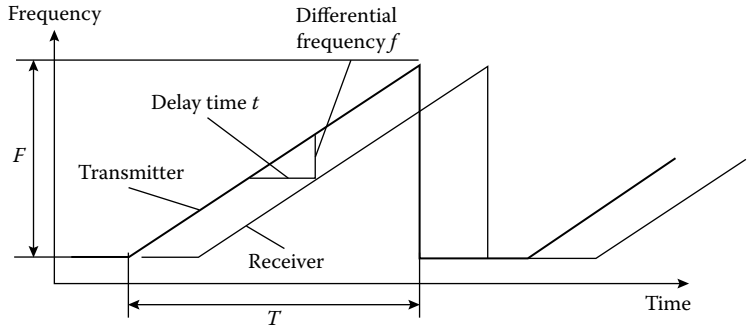


FIGURE 18.10 Operation characteristics of FMCW RADAR. The frequency of the transmitter changes linearly by time in an interval (sweep). The received signal has the same form but is time delayed. At every point of the sweep, the differential frequency is constant and proportional to the time delay. Time and frequency scales are arbitrary.

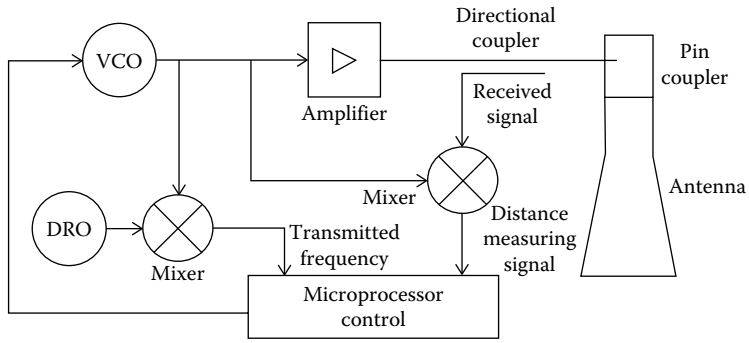


FIGURE 18.11 Basic circuit block diagram of an FMCW RADAR system: a microprocessor controls a voltage-controlled oscillator (VCO), such that the desired frequency sweep is obtained. This signal is amplified and fed into the transmitting antenna. The instantaneous frequency must be measured in order to ensure good sweep linearity. This is accomplished by counting the frequency after it has been mixed with the known frequency of a dielectric resonance oscillator (DRO). A directional coupler decouples the received signal, which is mixed with the transmission signal and processed by the microprocessor.

18.3.2 Time-of-Flight through Product

Alternatively, the propagation time of the waves through a weakly absorbing liquid or bulk material of low permittivity ε_r can be measured, as well as the propagation through the air. In cases where the reflection from the interface between air and the upper surface of the product is poor, part of the signal travels through the liquid and is reflected a second time at the tank bottom or at an interface between two liquids (e.g., oil on water).

Figure 18.12 demonstrates this technique. The evaluation is done in the following four steps:

1. Where microwaves in the tank atmosphere of height d are propagated at the speed of light c, microwaves in the medium (relative permittivity = ε_r, height L) are propagated at a slower velocity v.
2. Hence, the reflection r_2 from the tank bottom appears to be shifted downward, and the apparent tank height h_v is greater than the true height h.
3. The transit time in the medium is $t_1 = L/v$, where for the same distance in an empty tank would be $t_0 = L/c$. The ratio of apparent "thickness layer" ($h_v - d$) to true filling height ($h - d$) therefore corresponds to the ratio of the wave propagation rates:

$$\frac{h_v - d}{h - d} = \frac{c}{v} = \sqrt{\varepsilon_r} \tag{18.17}$$

4. When ε_r, h, and h_v are known, distance d and, from that, filling height L can be calculated exactly as follows:

$$L = h - d = \frac{h_v - h}{\sqrt{\varepsilon_r} - 1} \tag{18.18}$$

By this method, a direct level measurement—not a distance measurement—is attained. It can even be applied when signal r_1 from the surface of the medium is no longer measurable. The evaluation of the tank bottom reflection signal is known as "tank bottom tracing" and is used in the RADAR level system offered by Krohne in Figure 18.8.

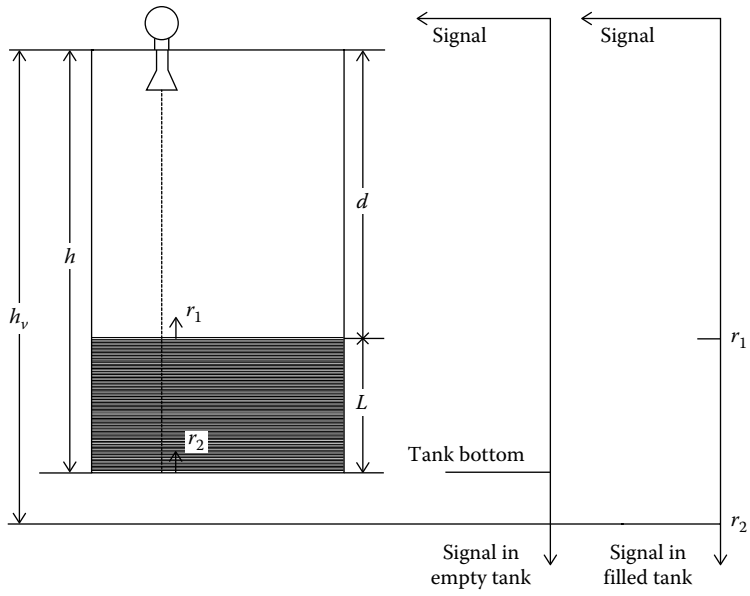


FIGURE 18.12 Representation of time-of-flight measurement through liquid: the wave is reflected once (r_1) at the product's surface and a second time (r_2) at the tank bottom. Due to the reduced wave velocity within the liquid, the reflection r_2 appears below the geometric position of the bottom. From that shift, the level can be calculated; see Equations 18.17 and 18.18.

18.4 Level Measurements by Detecting Physical Properties

To measure level, one can detect physical parameters that are significantly different between the atmosphere and the product; for example, conductivity, viscosity, or attenuation of any type of radiation. Two types are possible: (1) continuous measurement with an integral sensor or (2) switching by a point measurement when the sensor comes in contact with the product.

18.4.1 Electrical Properties

The sensor must be in direct or indirect contact with the product to detect its electrical properties. For continuous measurement, only part of the intrusive sensor must be in contact with the product to detect the difference in dielectric permittivity or conductivity.

18.4.2 Capacitive

In most applications, a rod electrode is arranged vertically in the tank. The electrode can be (1) noninsulated if the liquid is nonconductive or (2) insulated. The metallic vessel acts as a reference electrode.

The result depends on the permittivity ε_2 of the product. Figure 18.13a shows an electrode concentrically mounted on a cylindrical tank. For such a rotationally symmetrical configuration, the capacitance C of an insulated electrode changes with level L according to

$$C = \frac{2\pi\varepsilon_0 L}{1/\varepsilon_1 \ln d_2/d_1 + 1/\varepsilon_2 \ln d_3/d_2} \Leftrightarrow L = \frac{C\left(1/\varepsilon_1 \ln d_2/d_1 + 1/\varepsilon_2 \ln d_3/d_2\right)}{2\pi\varepsilon_0} \tag{18.19}$$

where

ε_0 is the dielectric constant of vacuum (8.85×10^{-12} $V^{-1}m^{-1}$)

ε_1 and ε_2 are the relative permittivities of the insulation material and the liquid, respectively

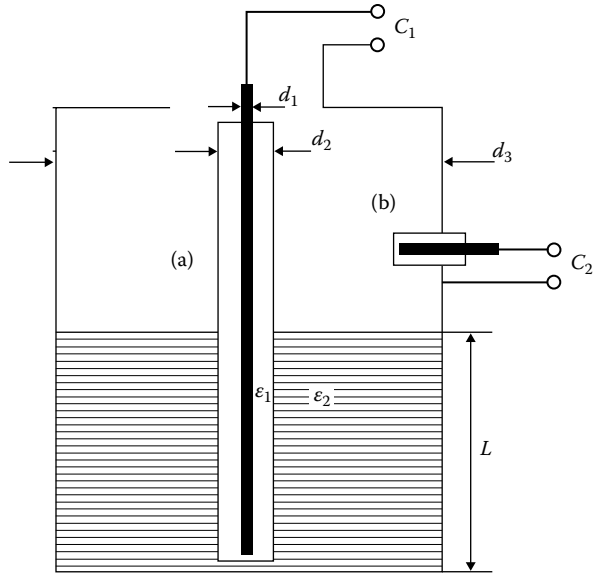


FIGURE 18.13 Principle of operation for a capacitance-type level device. (a) An insulated electrode protrudes into the liquid. The capacitance between the inner conductor and the tank walls is measured. (b) As a capacitance LS, the electrode can be mounted at the appropriate position directly into the tank wall.

If the liquid itself is highly conductive, Equation 18.19 simplifies to

$$C = \frac{2\pi\varepsilon_0\varepsilon_1 L}{\ln d_2/d_1} \Leftrightarrow L = \frac{\ln d_2/d_1}{2\pi\varepsilon_0\varepsilon_1} \tag{18.20}$$

If the electrode is not insulated, the following equation is valid:

$$C = \frac{2\pi\varepsilon_0\varepsilon_2 L}{\ln d_3/d_1} \Leftrightarrow L = \frac{\ln d_3/d_1}{2\pi\varepsilon_0\varepsilon_2} \tag{18.21}$$

When arranged horizontally, as in Figure 18.13b, a capacitive sensor can act as an LS.

For a more precise measurement of conductive liquids, a method measuring the complex impedance is helpful.

18.4.3 Conductive

The resistance of the liquid between two electrodes is measured with (1) a strip line with two parallel electrodes similar to Figure 18.9a or (2) a rod electrode with the metal vessel as the reference electrode, similar to Figure 18.13a without insulator.

18.4.4 Radiation Attenuation

All radiation (e.g., gamma rays, ultrasonic waves, electromagnetic waves) is attenuated to some degree in any medium. In general, attenuation in liquids or bulk materials is higher than in gases. This effect is used to measure level or limits, without direct contact of the sensor.

18.4.5 Radiometric

The intensity I of gamma rays is attenuated by the liquid according to its damping factor α:

$$I = I_0 \exp(-\alpha d) \tag{18.22}$$

The source can be a radioactive material Co-60 or Cs-137, having half-lives of 5.23 years and 29.9 years, respectively. Emitter and sensor may take the form of (1) a point emitting the rays radially in all directions, (2) a rod emitting radially from a line, or (3) an array consisting of several point emitters in a row. Any combination of point/rod/array emitter with point/rod/array detector is possible. Figure 18.14 shows two different configurations. Radiation protection regulations must be considered. A real-time clock in the system must compensate for the decrease of intensity (dose rate) I by time t according to the half-life T_H of the applied material:

$$I = I_0 2^{-t/T_H} \tag{18.23}$$

For more information about radiation detection methods, refer to Chapter 66. Plastic scintillators and counting tubes are preferred for radiometric level gaging. The level–intensity characteristic is nonlinear, so the equipment should be calibrated in place. Mengelkamp [8] describes the radiometric techniques in more detail.

18.4.6 Ultrasonic Switch

A short ultrasonic transmission path can be used to detect products that dampen sonar waves. For instance, this method is applicable for the detection of slurries or to determine the interface between two different liquids. When combined with a servo system, the vertical profile of ultrasonic attenuation can be measured. Another application uses a noncontact sensor mounted on the outside of the vessel. It measures the acoustic impedance through the vessel wall that changes if liquid or gas is present behind the wall.

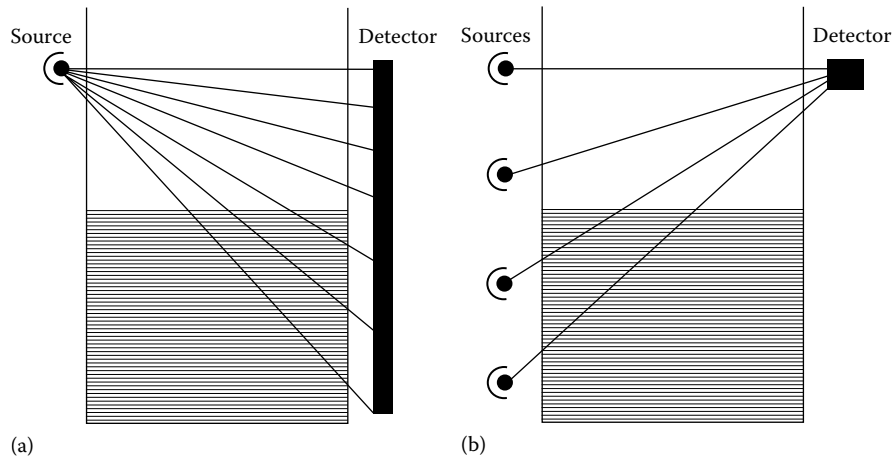


FIGURE 18.14 Representation of a radiometric continuous level gage. The rays are emitted by a radioactive source, propagate through the tank walls, and are damped by the liquid. (a) A point source is combined with a rod detector (e.g., scintillator rod); (b) a source array is combined with a point detector.

18.4.7 Microwave Switch

Liquids and solids dampen microwaves in many cases, sometimes absorbing them completely. A simple unmodulated microwave source and an accompanying receiver are sufficient for level switching.

18.4.8 Photoelectric Barrier

A photoelectric barrier can act as an LS for liquids that are not transparent, as well as most solids. But in closed nontransparent tanks, the coupling of the photoelectric components to the tank will not be possible in most cases.

18.4.9 Thermal and Mechanical

For some special applications, level sensors utilize the different heat dissipation properties and viscosities of the media.

18.4.10 Thermal

A self-heated resistor with a high temperature coefficient is immersed into the liquid. Heat dissipation causes the temperature to drop somewhat in the region where the liquid covers the sensor. Therefore, the resistance change is nearly linear with the level. This method is often used in automotive applications. In some applications, with heated liquids (e.g., chemical reaction vessels), a simple temperature sensor can be used as an LS by emitting a signal when the liquid contacts the sensor and heats it.

18.4.11 Viscosity

The effect of viscosity, which is significantly higher for liquids than for gases, dampens the movement of a body. These level sensors measure the degree of damping of a vibrating fork when dipped in a liquid. Normally, it is only used as a point LS. Figure 18.15 shows such a "tuning fork," named according to the



FIGURE 18.15 Design of a vibrating LS. The switch reacts to product viscosity changes, which dampens the vibration of the paddles. (From Messtechnik, K., *Technical Data Sheets of Level Measurement Products*, Krohne, Duisburg, Germany, 1996.)

TABLE 18.2 Priority Level Measurement Techniques and Prices for Industrial Applications

Technique	Application Range	Attainable Accuracy (mm)	Average Price
Displacer/float	Continuous, liquids	1	$2,000
Float	Switch, liquids	10	$500
Pressure	Continuous, liquids	10	$2,000
Ultrasonic	Continuous, liquids, solids	5	$2,500
RADAR	Continuous, liquids, solids	1	$3,000
TDR	Continuous, liquids, solids	3	$2,500
Laser	Continuous, liquids, solids	0.1	$10,000
Radiometric	Continuous, liquids, solids	10	$5,000
Capacitive	Switch, liquids	10	$500
Viscosity	Switch, liquids	10	$500

typical form with two or three vibrating paddles. The integrated electronics evaluate the power loss or the frequency shift of the mechanical resonance system. For solids, a sensor with a rotating paddle that stops when contacting the product is useful.

18.5 Level Measuring Instruments

Table 18.2 lists the most common techniques for level gaging with corresponding application range, accuracy, and average prices. Pricing as shown is for typical industrial applications and can vary widely, depending on materials of construction, measuring range, accuracy, approval requirements, etc. Partial list of manufacturers and suppliers of level instruments is provided at the end of this chapter. Most of these companies manufacture various instruments, using the different technologies shown in Table 18.2. This list is not comprehensive; additionally, some smaller companies provide equipment for specialized level measurements.

References

1. D.M. Considine, *Process Instruments and Control Handbook*, 2nd edn., New York: McGraw-Hill, 1974.
2. E.B. Jones, *Instrument Technology*: *Measurement of Pressure, Level and Temperature*, Vol. 1, London, U.K.: Butterworth & Co., 1974.
3. Verein Deutscher Ingenieure, Verband Deutscher Elektrotechniker (VDI/VDE), *Füllstandmessung von Flüssigkeiten und Feststoffen* (*Level Measurement of Liquids and Solids*), VDI/VDE 3519, Part 1, Berlin, Germany: Beuth, 1984.
4. K.W. Bonfig (ed.), *Technische Füllstandsmessung und Grenzstandskontrolle*, Ehningen, Germany: Expert, 1990.
5. Harold R. Raemer, *Radar System Principles*, Boca Raton, FL: CRC Press, 1997.
6. D. Brumbi, *Fundamentals of Radar Techniques for Level Gauging*, Duisburg, Germany: Krohne, 1995.
7. D. Brumbi, Measuring process and storage tank level with radar technology, *Int. Radar Conf. IEEE*, 256–260, 1995.
8. William L. Wolfe, *Introduction to Radimetry*, Bellingham, Washington, DC: SPIE, 1998.
9. K. Messtechnik, *Technical Data Sheets of Level Measurement Products*, Duisburg, Germany: Krohne, 1996.

Further Readings

Technical Reference-Level Measurement, OMEGA, http://www.omega.com/toc_asp/sectionSC.asp?section=K&book=Green (accessed on June 14, 2012).

Process Instrumentation, Level Measurements with Level Measuring Instruments, SIEMENS, http://www.automation.siemens.com/w1/automation-technology-level-measurement-with-level-measuring-instruments-18693.htm (accessed on July 18, 2012).

Partial List of Manufacturers and Suppliers

Berthold Systems, Inc., 101 Corporation Drive, Aliquippa, PA 15001, Tel: (508) 378–1900

Bindicator, 150 Venture Blvd, Spartanburg, SC 29306, Phone: 864-574-8060, www.bindicator.com

Danfoss Automatic Controls, 7941 Corporate Dr., Baltimore, MD 21236, Phone: 410-931-8250, Fax: 410-931-8256, www.danfoss.com/North_America/

Drexelbrook Engineering Company, 205 Keith Valley Road, Horsham, PA 19044, Tel: (215) 674–1234, www.gilsoneng.com

Dwyer Instruments, Inc., Hwy. 212 at 12, P.O. Box 373, Michigan City, IN 46361, Phone: 219-879-8000, 800-872-9141 (toll free), Fax: 219-872-9057, www.dwyer-inst.com.

Endress & Hauser Instruments, 2350 Endress Place, Greenwood, IN 46143, Tel: (317) 535–7138, www.us.endress.com

The Foxboro Company, 500 N. Bedford St., P.O. Box 500, East Bridgewater, MA 02333, Phone: 508-543-8750, 800-321-0322 (toll free), iom.invensys.com

Honeywell Enraf, 2000 Northfield Ct, Roswell, GA 30076, Phone: 770-475-1900, Fax: 770-475-1717, hpsweb.honeywell.com/Cultures/en-US/Prod…

Krohne America Inc., 7 Dearborn Road, Peabody, MA 01960, Tel: (508) 535–1720, http://krohne.com/

Miltronics LTD-Sherrex Systems Division, 4130-93rd St., Edmonton, AB T6E5P5, www.sherrex.com

MTS Systems Corporation, Sensors Division, 1001 Sheldon Drive, Cary, NC 27513, Tel: (919) 677–0100, www.mts.com

Rosemount, Inc., 21785 Hamburg Ave., Lakeville, MN 55044–9035, Phone: 952-985-6200, 877-985-6221 (toll free), Fax: 952-985-6205, www.rosemount-office.com

Saab Tank Control, 10700 Hammerley Blvd., Houston, TX 7743, Tel: (713) 722–9199, Fax: (713) 722–9115, www.rosemount-tankradar.com/home_index.a…

Siemens Corp., 300 New Jersey Avenue, N.W., Suite 1000, Washington, DC 20004–2611, Phone: 202-434-4800, 800-SIEMENS (toll free), Fax: 202-347-4015, www.usa.siemens.com/entry/en/

TN Technologies, Inc., 2555 North Interstate 35, P.O. Box 800, Round Rock, TX 78680, Tel: (512) 388–9100, www.tn-technologies.com

Yokogawa Filed Instruments Division, 2 Dart Rd., Newnan, GA 30265, Phone: 770-254-0400

Fax: 770-254-4913, www.yokogawa.com/us/ia/fieldinstruments/…

19

Area Measurement

Charles B. Coulbourn
Los Angeles Scientific Instrumentation Co.

Wolfgang P. Buerner
Los Angeles Scientific Instrumentation Co.

19.1 Theory **19**-1
Planimeter • Digitizer • Tablet Digitizer • Sonic Digitizers • Arm Digitizer • Grid Overlay
19.2 Equipment and Experiment **19**-7
19.3 Evaluation **19**-10
Defining Terms **19**-11
References **19**-11
Further Information **19**-11

One must often measure the area of enclosed regions on plan-size drawings. These areas might be either regular or irregular in shape and describe one of the following:

- Areas enclosed by map contours
- Cross section of the diastolic and systolic volumes of heart cavities
- Farmland or forestland shown in aerial photographs
- Cross sections of proposed and existing roads
- Quantities of materials used in clothing manufacture
- Scientific measurements
- Swimming pools
- Quantities of ground cover

The tools for this type of measurement include planimeters, digitizer–computer setups, digitizers with built-in area measuring capability, and grid overlay transparencies.

19.1 Theory

19.1.1 Planimeter

A planimeter is a mechanical integrator that consists of a bar (tracer arm), a measuring wheel with its axis parallel to the bar, and a mechanism that constrains the movement of one end of the bar to a fixed track (Figure 19.1). The opposite end of the bar is equipped with a pointer for tracing the outline of an area. The measuring wheel (Figure 19.2) is calibrated with 1000 or more equal divisions per revolution. Each division equals one count. It accumulates counts, P, according to

$$P = \frac{K}{\pi D} \int \sin\phi \, ds \tag{19.1}$$

where

K is the number of counts per revolution of the measuring wheel
D is the diameter of the measuring wheel
ϕ is the angle between the measuring wheel axis and the direction of travel
s is the traced path

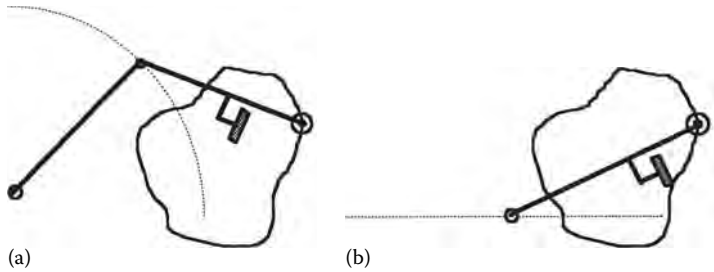


FIGURE 19.1 Constrained end of a polar planimeter (a) follows a circular path; the constrained end of a linear planimeter (b) follows a straight line path.

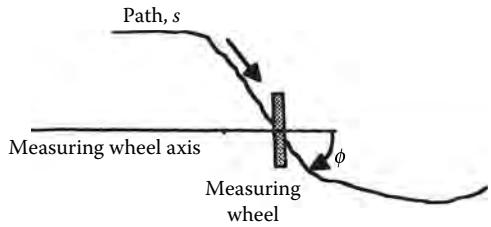


FIGURE 19.2 Rotation of a measuring wheel is proportional to the product of distance moved and the sine of the angle between the wheel axis and the direction of travel.

The size of an area, A, traced is

$$A = \frac{P}{K} \times \pi D \times L \tag{19.2}$$

where

L is the length of bar

P is the accumulated counts (Equation 19.1)

Figure 19.3 shows how a basic wheel and bar mechanism determines the area of a parallelogram. The traced path is along the sloped line; however, the wheel registers an amount that is a function of the product of the distance traveled and the sine of the angle between the direction of travel and the axis of the measuring wheel (Equation 19.1). This is the altitude of the parallelogram. The product of the altitude (wheel reading converted to distance) and base (bar length) is the area.

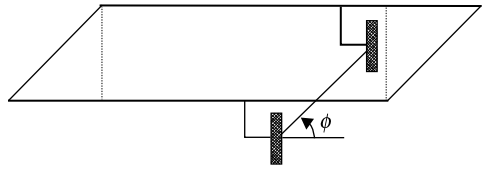


FIGURE 19.3 Area of this parallelogram is proportional to the product of tracer arm length and measuring wheel revolutions.

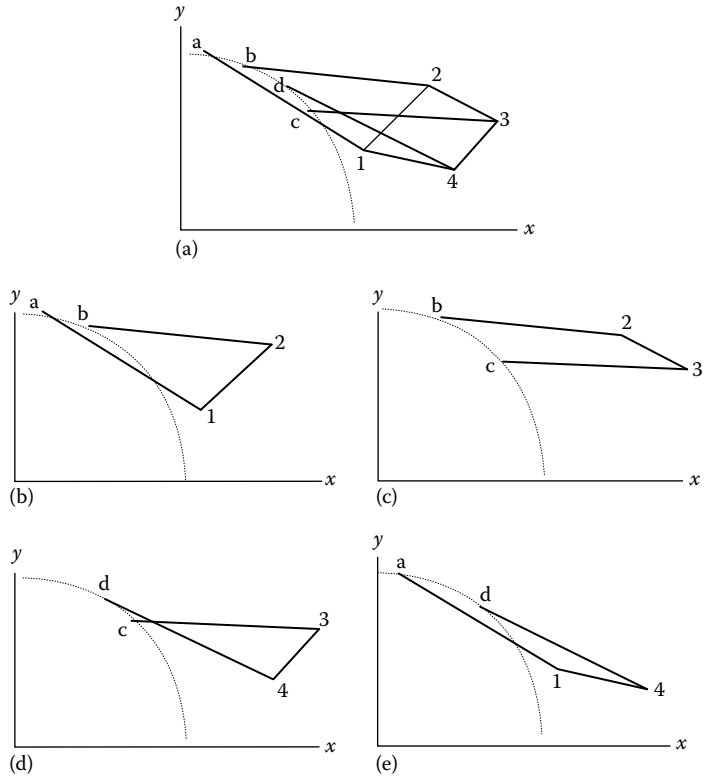


FIGURE 19.4 This schematic shows a planimeter pointing to each junction of a four-sided figure being traced in (a) and at the ends of each of the segments in (b) through (e). The constrained end of the tracer arm follows a circle.

Figure 19.4a illustrates the operation of a planimeter when the area of a four-sided figure is measured. Figure 19.4b–e shows the initial and final positions of the bar as each side of the figure is traced. Applying the general expression for the area under a curve, $A = \int f(x)dx$, for each of these partial areas gives

$$A_a = \left(\int_a^1 + \int_1^2 + \int_2^b + \int_b^a \right) f(x)dx \tag{19.3}$$

$$A_b = \left(\int_b^2 + \int_2^3 + \int_3^c + \int_c^b \right) f(x)dx \tag{19.4}$$

$$A_c = \left(\int_c^3 + \int_3^4 + \int_4^d + \int_d^c \right) f(x)dx \tag{19.5}$$

$$A_d = \left(\int_d^4 + \int_4^1 + \int_1^a + \int_a^d \right) f(x)dx \tag{19.6}$$

where A_a, A_b, A_c, and A_d are the four partial areas.

The total area of the figure is the sum of the four partial areas. Combining the terms of these partial areas and rearranging them so that those defining the area traced by the left end of the bar are in one group, those defining the area traced by the other end of the bar are in a second group, and those remaining are in a third group, results in the following:

$$A = \left\{ \left[\int_{b}^{a} + \int_{c}^{b} + \int_{d}^{c} + \int_{a}^{d} \right] f(x)dx + \left[\int_{a}^{1} + \int_{2}^{b} + \int_{b}^{2} + \int_{3}^{c} + \int_{c}^{3} + \int_{4}^{d} + \int_{d}^{4} + \int_{1}^{a} \right] f(x)dx \right.$$

$$\left. + \left[\int_{1}^{2} + \int_{2}^{3} + \int_{3}^{4} + \int_{4}^{1} \right] f(x)dx \right\} \tag{19.7}$$

The first four integrals describe the area traced by the left end of the bar. Since this end runs along an arc, it necessarily encloses an area equal to zero. The final four integrals describe the area traced by the right end of the bar. This is the four-sided figure. The remaining eight integrals cancel out since $\int_{a}^{1} + \int_{1}^{a} = 0$, etc. Thus, the total area equals the area traced by the right end of the bar. Note that the same reasoning applies to figures of any number of sides and of any shape.

19.1.2 Digitizer

A digitizer converts a physical location on a map to digital code representing the (x, y) coordinates of the location. The digital code is normally converted to a standard ASCII or binary format and transmitted to a computer where computations are made to determine such things as area or length. Certain digitizers can also compute areas and lengths without the use of a computer.

Area, A, can be computed using the coordinate pairs that define the area boundary:

$$A = \frac{1}{2}(y_1 + y_2)(x_2 - x_1) + \frac{1}{2}(y_2 + y_3)(x_3 - x_2) + \cdots + \frac{1}{2}(y_{n-1} + y_n)(x_n - x_{n-1})$$

$$+ \frac{1}{2}(y_n + y_1)(x_1 - x_n) \tag{19.8}$$

where

x_1, x_2, x_3, etc., are the sequentially measured x-coordinates along the boundary
y_1, y_2, y_3, etc., are the corresponding y-coordinates

All digitized coordinate pairs must be used in the computations since the Δx intervals of these data are typically unequal.

In Equation 19.8, the final term deserves special consideration because it contains both the first and the last coordinate pairs of the series. When computations are performed from the data stored in a computer, the information is available for computing the final term so that no special actions are necessary. However, when the computations are performed by an embedded microprocessor with limited memory, normally each term is computed as the coordinates are read. In this case, one of the two actions must be taken. First, the first coordinate pair is saved, and then, once the measurement has been completed, a key is pressed to initiate computation of the final term. This is called "closing the area." Second, if the first coordinate pair is not saved, to prevent an error, the final point digitized must coincide with the first point digitized so that $x_n = x_1$, $y_n = y_1$, and the final term is zero.

The popular types of digitizers in use today are tablets, sonic digitizers, and arm digitizers. Probably the most popular type is the tablet.

19.1.3 Tablet Digitizer

Tablet digitizers consist of a pointer and a work surface containing embedded wires configured as a grid. The horizontal wires are parallel and spaced by about 12 mm. The vertical wires are also parallel and spaced the same. Different sensing techniques are used to locate the pointer position relative to the grid wires.

One sensing technique employs grid wires made from magnetostrictive material that has the property of changing shape very rapidly when subjected to a magnetic field. Each set of grid wires, the horizontal and the vertical, is independently energized by a send wire that lies perpendicular to that set of wires [1]. A pulse transmitted over the send wire has a magnetic field that changes the shape of each magnetostrictive wire, causing a strain wave to propagate down the wire. Coincidentally, the pulse starts a counter. A coil in the pointer senses the strain wave and sends a signal to stop the counter. The counter reading is thus proportional to the propagation time of the strain wave. The product of the propagation time and velocity of the strain wave equals the physical distance between send line and pointer detector. The velocity of the strain waves is slow enough so that any errors in time measurement can be made acceptable.

A second sensing technique uses a grid made of conductive wires and a pointer that emits a signal on the order of 57.6 kHz. The vertical grid wires are sensed to determine the amplitude distribution of signal induced in each wire. The point of maximum signal strength determines the location of the pointer along the horizontal axis. The horizontal grid wires are likewise sensed to find the point of maximum signal strength that determines the pointer location along the vertical axis. Coupling between the pointer and the grid can be by either electromagnetic or electrostatic means.

Signals can also be applied to the grid wires and the pointer used as a receiver. In this case, the signal to each grid wire must be coded or else applied sequentially, first to one set of grid wires and then to the other. The amplitude profiles of the signals received by the pointer are determined to locate the pointer along each axis.

19.1.4 Sonic Digitizers

Sonic digitizers consist of a pointer and two or more microphones. The pointer, used to identify the point to be digitized, typically contains a spark gap that periodically emits a pulse of sonic energy [2]. The microphones in certain cases are mounted in a bar that is located along the top of the drawing area. In other cases, they are mounted in an L frame and located along the top and one side of the drawing area.

The microphones receive the sonic pulse emitted by the pointer. The time taken for the pulse to travel from the transmitter to each receiver is measured, and the slant distance is computed from the product of this elapsed time and the sonic velocity. The x- and y-coordinates of the pointer are computed from the slant ranges and the locations of the microphones.

Ambiguities exist since a single set of slant ranges describe two points; however, the ambiguities can be resolved either by using additional microphones or by ensuring that the ambiguous points are outside the work area. With the microphones aligned along the top of the work area, the ambiguous points are outside the work area; with them aligned along the top and the side, the ambiguities have to be resolved with additional microphones.

The sound-wave velocity is slow enough so that errors in time measurements can be made acceptable.

19.1.5 Arm Digitizer

An arm digitizer consists of a base, two arms, two rotary encoders, and a pointer. The *base* must be anchored at a point removed from the work area, normally at the top of the user's desk. One end of the arm, called the *base arm*, swivels about a vertical axis at the base. The other end of the base arm

hinges to one end of a second arm, called the *tracer arm*, about a vertical axis. On the other end of the tracer arm is the *pointer*. One of the encoders, the *base encoder*, detects the angle between the base and the base arm. The other encoder, the *arm encoder*, detects the angle between the base arm and the tracer arm.

To simplify implementation, each arm is made the same length, and for sufficient accuracy, each encoder provides in excess of 36,000 counts per revolution. The encoder output signals consist of dual square waves with a 90° phase relationship. The leading and trailing edges of each square wave are counted and the 90° phase relationship provides count direction.

Since each encoder is a relative rather than absolute counting device, the count registers of each encoder must be initialized at a known angle. The count register of the arm encoder is set to the angle that results when the pointer is moved to a precisely known location. This location is called "home." The count register for the base encoder is initialized when the rotation of the *x*–*y*-coordinate system is set. Thereafter, the encoders add to or subtract from their count registers as the pointer is moved.

With the arm lengths and the angles precisely known, the x and y positions of the pointer can be computed. However, x and y are relative to a coordinate system whose origin is at the axis of the base and whose rotation is unknown. The user must therefore select an origin, usually near the lower left corner of the drawing area, and the direction of the x-axis, usually parallel to the bottom edge of the drawing area. This information is sufficient for the processor to compute the correct translation and rotation or the coordinate system.

19.1.6 Grid Overlay

A grid overlay is simply a transparent sheet onto which a grid has been drawn. To use it, place it atop the drawing or photo of the area to be measured (Figure 19.5). Then count the number of squares that lie within the boundary of the area. Squares that are at least half enclosed should be counted. The unknown area equals the product of the number of blocks counted and the area of each block. The accuracy achieved is dependent on grid size, precision of the grid dimensions, and counting accuracy.

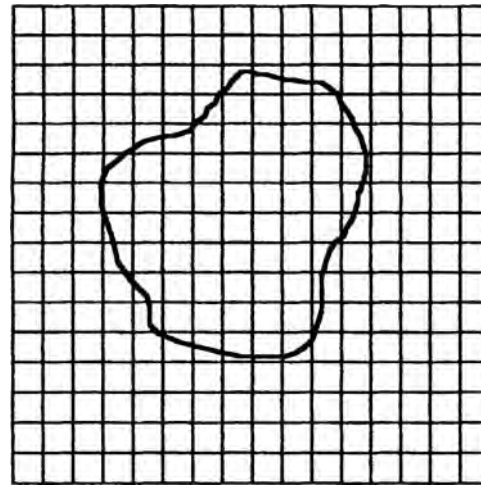


FIGURE 19.5 Grid overlays offer a simple and readily available method for measuring areas.

19.2 Equipment and Experiment

Many different types of planimeters and digitizers are manufactured in the United States, Europe, and Japan. A representative sample of these instruments is described in this section.

One of the simplest and least costly area measuring devices is the manual polar planimeter (Figure 19.6). It consists of a weight, two arms, a measuring wheel, and a pointer. The *weight* secures one end of a pole arm, allowing the other end to rotate along a fixed arc. The rotating end of the pole arm attaches to one end of a *tracer arm* and constrains its movement to the arc. At the other end of the tracer arm is a *pointer* used for tracing the periphery of an unknown area. The *measuring wheel* is located in the box at one end of the tracer arm. The location of the measuring wheel is not critical; however, its axle must be parallel to the tracer arm.

The length of both arms of the planimeter shown in Figure 19.6 can be adjusted. The length of the pole arm has absolutely no effect on measurement accuracy and is adjustable only for convenience. The effective length of the tracer arm directly affects the reading; a shorter arm results in a larger reading. By adjusting the tracer arm length, one can achieve a very limited range of scaling; however, this is usually not done. Rather, the arm length is adjusted according to the general size of areas to be measured: a shorter arm for smaller areas and a longer arm for larger areas. A shorter arm results in a greater number of counts per unit area, which is needed for smaller areas. Scaling is usually done by multiplying the result by an appropriate value.

One zeros the measuring wheel of the planimeter shown in Figure 19.6 by turning a small knurled wheel attached to the measuring wheel axle. On other models, one pushes a plunger to zero the wheel. The latter method is easier but is sensitive to misalignment.

A planimeter pointer is usually a lens with a small circle engraved in the center, although some planimeter models use a needle as the pointer. The best type of pointer is a matter of personal preference although lens pointers are much more popular.

Figure 19.7 shows two electronic planimeters with digital readouts. With these planimeters, one can measure length as well as area. To measure length, snap out the measuring wheel housing, attach an auxiliary handle, and roll the wheel along the line to be measured. Extremely high accuracy can be achieved.

The upper planimeter in Figure 19.7 is a polar type. It consists of the same parts as the model in Figure 19.6, except that the measuring wheel is attached to a small optical encoder and the weight is packaged in the digital readout. The encoder provides two square-wave outputs that are in phase quadrature, that is, one is 90° out of phase with the other. Both outputs are fed to a processor that counts the pulses and uses the phase difference to determine count direction.

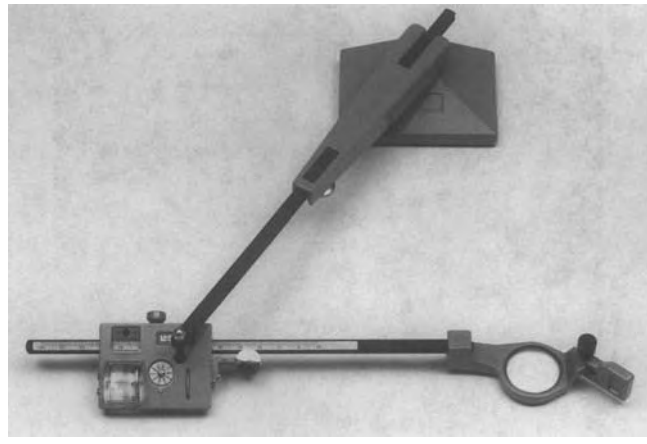


FIGURE 19.6 Mechanical planimeters are normally preferred when occasional use is required.

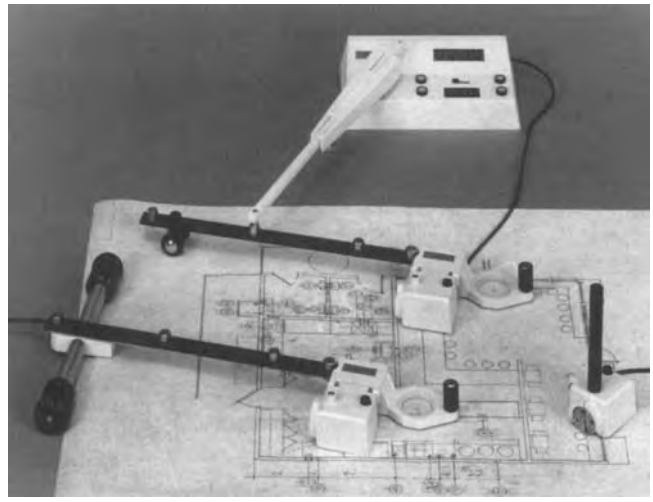


FIGURE 19.7 Electronic planimeters are easier to use and read and are preferred especially when frequently used.

The processor has an electronic scale feature that translates the planimeter reading to real measurement units such as square feet, square meters, acres, or hectares. One can transmit processor data to a computer using an auxiliary interface unit (not shown).

The planimeter at the bottom of Figure 19.7 is a linear model since the path traveled by the constrained end of the tracer arm is a straight line. The straight line path is maintained by a rigid connection between the two carriage wheels and their axle. Other linear planimeters use an actual rail to guide the constrained end of the tracer arm along a straight line path.

The polar planimeter shown at the top of Figure 19.8 and the linear planimeter shown at the bottom of Figure 19.8 are both compact battery-operated models. The measuring wheel is built into the processor, which is attached to the pole arm of the planimeter. The effective length of the tracer arm for both instruments extends from the axis of the constrained end to the pointer, which for these instruments is a small circle engraved in the center of a lens. These planimeters provide electronic scaling and averaging of multiple readings. They cannot be used to measure length.

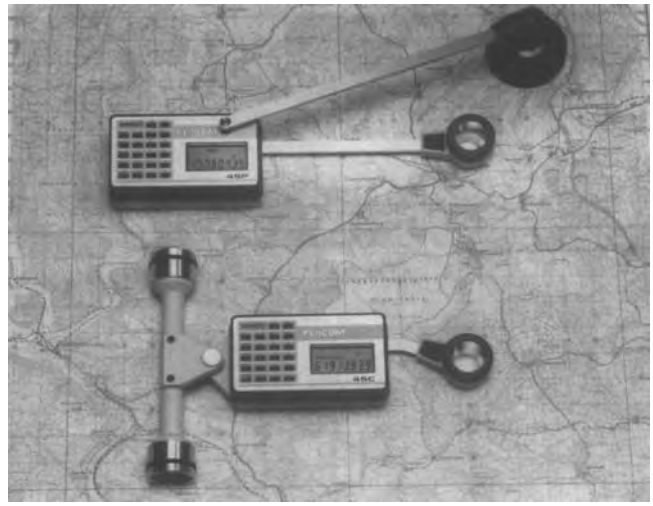


FIGURE 19.8 Battery-operated planimeters have the advantages of electronic planimeters and are portable.

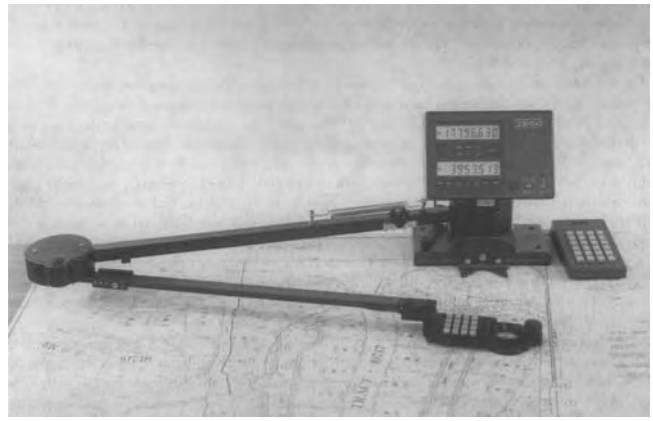


FIGURE 19.9 Arm digitizer lets one measure area, length, and coordinates and transmits the displayed data to a computer; it requires minimum desk space.

Figure 19.9 shows an arm digitizer that can be used either as a stand-alone area and length measuring device or to digitize a map or drawing. When operating as a digitizer, the arm digitizer displays the (x, y) coordinates as well as transmits them to a computer. The digitizer has a built-in interface and can transmit using any of over 24 different ASCII and binary codes, each with a choice of parameters. It can be set to measure x- and y-coordinates in either English or metric units.

Three other modes, in addition to the digitizer mode, are available for computing and displaying either area and length and area and item count or item count and length. Measurements can also be made in either English or metric units. Any displayed item can be transmitted to a computer through the built-in interface.

The arm digitizer has other features that are especially useful when it is used in a nondigitizer mode. It provides independent scaling in each axis, empirical scale factor determination, key-press programming, four-function calculating, unit conversion, and averaging. Also, the arm digitizer can be folded out of the way when not in use.

Arm digitizers can be used on transparent surfaces as well as on normal opaque work surfaces. Thus, they are also useful for digitizing or measuring areas on rear-projected maps or drawings.

Tablet digitizers, like the one shown in Figure 19.10, are popular instruments used for digitizing maps and drawings and as extended keypads for certain software. When used as an extended keypad, a template with figures of key functions is placed somewhere in the work area. Pointing to a function

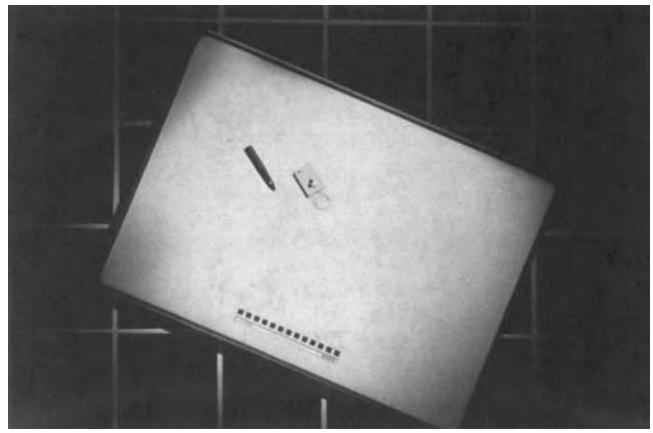


FIGURE 19.10 Popular pad digitizer includes a lightweight puck or pen-type pointer for selecting points whose coordinates are transmitted to a computer for processing.

TABLE 19.1 Companies That Make Planimeters

Gebruder Haff Gmbh Tiroler Strasse 5, D-87459 Pfronten, Germany, Tel: 49-8363-9122-0
Koizumi Sokki Mfg. Co., Ltd. 1–132, Midori-Cho, Nagaoka-Shi, Niigata 940–21, Japan, Tel: (0)258-27-1102
Lasico Inc. 2451 Riverside Drive, Los Angeles, CA 90039, Tel: (213) 662–2128
Sokkia Corporation 9111 Barton St., Overland Park, KS, P.O. Box 2934, Tel: (800) 476–5542

TABLE 19.2 Companies That Make Digitizers

Altek Corporation 12210 Plum Orchard St., Silver Spring, MD 20904, Tel: (301) 572–2555
Calcomp Technology Inc. 2411 West La Palma Ave., Anaheim, CA 92801-2589, Tel: (800) 445–6515
GTCO Corporation 7125 Riverwood Drive, Columbia, MD 21046, Tel: (800) 344–4723
Lasico Inc. 2451 Riverside Drive, Los Angeles, CA 90039, Tel: (213) 662–2128
Numonics Corporation 101 Commerce Drive, Box 1005, Montgomeryville, PA 18936, Tel: (215) 362–2766
Wacom Technology Corporation 501 S.E. Columbia Shores Blvd., Suite 300, Vancouver, WA 98661, Tel: (360) 750–8882

and clicking a cursor key selects it. Tablets feature a built-in interface and one or more binary and ASCII codes. Communication parameters on some tablets are controlled by dip switches and on others by computer software. Pointers are lightweight and include between 1 and 16 cursor keys. Tablets are available in sizes ranging from that of a notebook up to 122 cm × 168 cm (48 in. × 66 in.) or more. Models are available with transparent work surfaces and can be used with rear-projected maps and drawings.

Tablet digitizers have the advantage of lightweight pointers, but the larger tablets have the disadvantages of occupying a significant amount of floor space and being relatively difficult to move around. One manufacturer has overcome this disadvantage by designing a digitizer tablet that rolls up.

Sonic digitizers, like tablets, are intended for digitizing maps and drawings and for providing an extended keypad for certain software packages. They feature a built-in computer interface with a number of ASCII and binary codes, with a choice of parameters, which make them suitable for use with many different software packages. A major advantage enjoyed by sonic digitizers is their portability and the fact that they operate well with a transparent work surface required for rear-projected drawings and maps.

Table 19.1 provides a list of manufacturers of planimeters, and Table 19.2 lists many of the manufacturers of graphic digitizers.

19.3 Evaluation

Each of the area measuring devices described in this section is excellent and has been thoroughly proven by use. However, some of the devices are more suited to certain tasks and operating environments than others.

To measure an area that is smaller than an equivalent circle of about 2 cm in diameter, a digitizer is probably the best choice. For areas of this magnitude, the resolution element of planimeters starts to become a significant portion of the total area. The resolution element of most digitizers is significantly less than that of a planimeter, and any measurement is always plus or minus a resolution element.

For measuring areas that are equivalent to a circle between 2 and 55 cm in diameter, either a planimeter or a digitizer will provide excellent results.

When measuring areas larger than an equivalent circle of about 55 cm in diameter, one can still use a planimeter; however, one must subdivide the large area into smaller areas and then individually measure each of the smaller areas. A digitizer can measure significantly larger areas, but at some point, it too will reach a limit. Then, one can use the same technique of subdividing the large area into smaller areas that the digitizer can handle.

Area measuring instruments of the future will undoubtedly make even greater use of microprocessors to provide more features such as incorporation of slope and tilt correction, statistical operations, and determination of centroids, moment, etc. The instruments should become mechanically simpler and more reliable. Features such as conservation of office space and portability will be emphasized. The ultimate area measuring device will consist of a detached cursor for pointing and a small calculator-like device for operating on the results, displaying them, storing them, and sending them to a computer.

Scanners and associated software will also impact the field of area measurement, particularly as their coverage increases and their price decreases.

Defining Terms

Digitizer: A device to convert data or an image to digital form. The digitizers discussed here convert images to digital form and are categorized as graphic digitizers.

Linear planimeter: A planimeter with a tracer arm whose constrained end follows a straight line.

Measuring wheel: The planimeter wheel whose degree of rotation is directly proportional to the area.

Planimeter: A mechanical integrator used for measuring the area of enclosed regions on maps, plans, etc.

Pointer: The part of a planimeter or digitizer that is used to follow the line being traced.

Pole arm: One of the two bars comprising a polar planimeter. One end of the pole arm is fixed and the other end is free to rotate. The length of the pole arm has no effect on the planimeter reading.

Polar planimeter: A planimeter with a tracer arm whose constrained end follows a circle.

Resolution element: The smallest elemental area that can be discerned. When referred to in connection with area measurement, it is an area with a value of 1.

Tracer arm: The bar of a planimeter that is attached to the measuring wheel. One end of the tracer arm is constrained to a fixed path, while the other end traces the perimeter of an enclosed region whose area is being measured. The length of the pole arm is indirectly proportional to the planimeter reading.

References

1. P. A. Santi, J. Fryhofer, and G. Hansen, Electronic planimetry, *Byte*, 5(3), 113–122, March 1980.
2. D. Dayton Reuter, *Computer Graphics for Architects, Engineers, and Environmental Designers*, New York: Allworth Press, 2002.

Further Information

Sullivan, N., *Area, Distance and Volume*, Marshall Cavendish Corporation, New York, 2007.
Willers, F. A., *Mathematische Instrumente*, Munchen und Berlin, Germany: Verlag von R. Oldenbourg, 1943.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

20

Volume Measurement

20.1 Plethysmography Theory **20**-2
Air/Water Plethysmography or Chamber Plethysmography • Electrical Plethysmography
20.2 Numerical Integration with Imaging **20**-7
20.3 Indicator Dilution Methods **20**-11
Thermodilution • Radionuclide Techniques • First-Pass Method • Dynamic Recording Method • Gas Dilution
20.4 Water Displacement Volumetry **20**-13
20.5 Equipment and Experiments **20**-15
20.6 Evaluation **20**-15
References **20**-16
Further Information **20**-18
Partial List of Manufacturers and Suppliers **20**-18

René G. Aarnink
University Hospital Nijmegen

Hessel Wijkstra
University Hospital Nijmegen

For simple geometric shapes, volume measurements can be performed analytically by measuring the dimensions of the object in question and using the appropriate formula for that shape. Volume formulae are found in geometry and calculus textbooks as well as in reference books such as CRC Press's *Handbook of Mathematical Science* or other references.

Volume can also be measured by fluid displacement. The object whose volume is to be measured is placed in a container filled with fluid, and the initial and final volumes are measured. The object's volume is equal to the final volume minus the initial volume. This technique is especially useful for irregularly shaped objects.

Fluids can also be used to measure volume of cavities within objects by filling the cavity entirely with a fluid and then measuring the volume of the fluid after it is removed from the cavity.

The remainder of this chapter is dedicated to the more specific problem of volume measurements in medical applications.

Quantitative volume information can be of importance to the clinician in the diagnosis of a variety of diseases or abnormalities. It can also improve the understanding of the physiology of the patient. Information on volume may be used in various applications such as cardiac monitoring, diagnosis of prostate diseases, follow-up during organ transplantation, surgery for tumor resection, blood flow measurements, plastic surgery, follow-up of preterm infants, and sports performance analysis. Because of this wide spectrum of applications, various techniques for volume measurements have been developed, some of which are useful in determining the amount of blood flow to the organ (dynamic), while others are used to obtain the size of the object (static). The techniques can be invasive or noninvasive and are based on either direct or indirect measurements. Each technique has its own advantages and disadvantages, and the application determines the selection of volume measurement method.

One of the earliest techniques to measure (changes of) body volume was *plethysmography*, originally developed by Glisson (1622) and Swammerdam (1737) to demonstrate isovolumetric contraction of

isolated muscle. The measuring technique consists of surrounding the organ or tissue with a rigid box filled with water or air. The displacement of the fluid or the change in air pressure indicates the volume changes of the tissue due to arterial inflow. Two major types of plethysmography exist, and these can be distinguished by the technique used to measure the volume change. These are *volume plethysmography* (direct-measurement displacement plethysmography including water and air types) and *electrical plethysmography* (strain-gage, inductive, and impedance plethysmographs). The physical condition in which the measurements should be performed determines the plethysmographic method chosen.

Advances in medical imaging have provided new possibilities for noninvasively extracting quantitatively useful diagnostic information. Images are constructed on a grid of small picture elements (pixels) that reflect the intensity of the image in the array occupied by the pixel. Most medical images represent a 2D projection of a 3D object.

Currently, the most commonly used medical imaging modalities are ultrasound, nuclear magnetic resonance imaging, x-rays, and x-ray computer tomography. *Ultrasound imaging* is based on the transmission and reflection of high-frequency acoustic waves. Waves whose frequencies lie well above the hearing range can be transmitted through biological tissue and will be reflected if they cross a boundary between media of different acoustic properties. These reflected signals can be reconverted to electrical signals and displayed to obtain a 2D section. *Magnetic resonance imaging* is based on the involvement of the interaction between magnetic moment (or spin) of nuclei and a magnetic field. The proton spins are excited by an external radio-frequency signal and the return to an equilibrium distribution is used to construct cross-sectional images. *Computer tomography* or *CT scanning* uses a rotating source of x-rays. The x-ray source and detector are scanned across a section of the object of interest, and the transmission measurements are used to reconstruct an image of the object.

These imaging techniques display cross sections of views that can be used to estimate the size of specific components. One way to estimate the volume of internal objects is to make certain assumptions concerning the shape and to apply formulae to estimate the volume with dimensions of the object such as length, height, and width. A more accurate technique is *step-section planimetry*, a clinical application of numerical integration. During this procedure, cross sections of the object are recorded with a certain (fixed) interval. The area of the object is determined in every section, and the total volume is calculated by multiplying the contribution of each section with the interval and summarizing all contributions.

The volume of a fluid-filled region can be calculated if a known quantity of indicator is added to the fluid and the concentration measured after it has been dispersed uniformly throughout the fluid. The selection of the indicator used to measure the volume depends on the application and can be based on temperature, color, or radioactivity.

Finally, volume (changes) can also be performed directly using *water displacement volumetry*, a sensitive but time-consuming method to measure the volume of an extremity. This method is not suitable for patients in the immediate postoperative period.

20.1 Plethysmography Theory

Fluid inflow can be measured by blocking the outflow and then measuring the change in volume due to the inflow over a defined time interval. Plethysmography enables the volume change to be determined in a reasonably accurate manner. A direct-volume displacement plethysmograph uses a rigid chamber usually filled with water, into which the limb or limb segment is placed. This type of plethysmograph, sometimes called *chamber plethysmography*, can be used in two ways. It can be used to measure the sequence of pulsations proportional to the individual volume changes with each heartbeat (arterial plethysmography or rheography). Also, the total amount of blood flowing into the limb or digit can be measured by venous occlusion: by inflating the occluding cuff placed upstream of the limb or digit just above the venous pressure 5–8 kPa (40–60 mm Hg), arterial blood can enter the region but venous blood is unable to leave. The result is that the limb or digit increases its volume with each heartbeat by the volume entering during that beat.

20.1.1 Air/Water Plethysmography or Chamber Plethysmography

Air plethysmography uses the relation between the volume change of a gas in a closed system and the corresponding pressure and temperature [1]. The relation between pressure and volume is described by Boyle's law, which can be written as

$$P_i V_i = P_f V_f \tag{20.1}$$

where P_i and V_i are the initial pressure and volume, respectively, and P_f and V_f are the final pressure and volume, respectively.

Measured at constant temperature, the displacement of the fluid or the compressing of the air is a direct measure of the blood flow or original volume. The air plethysmograph uses the change in pressure that occurs in a cuff wrapped around the segment of interest due to the change in volume. By inflating the cuff to about 60 mm Hg, the arterial inflow causes small increases in pressure. These small changes in pressure over the cardiac cycle can be monitored (see Figure 20.1). The measurement of blood flow is achieved by comparing the pressure changes to changes caused by removing known amounts of air from the system. The second measurement uses volume changes at various pressures between systolic and diastolic pressures, and the peak deflection is compared with the deflection caused by removal of known amounts of air from the system. In segmental plethysmography, two cuffs are used to measure the volume changes in a segment of a limb. Venous occlusion is established by the first cuff, while the second is inflated to a pressure that will exclude blood flow from those parts that should not be included in the measurement.

The technique can also be used for whole-body plethysmography [2], a common technique used to measure residual volume, functional residual capacity (FRC), and total lung volume, the parameters usually determined during pulmonary function testing [3]. In body plethysmography, the patient sits inside an airtight box and inhales or exhales to a particular volume (usually the FRC, the volume that remains after a normal exhalation). Then, a shutter closes the breathing tube, while the subject tries to breathe through the closed tube. This causes the chest volume to expand and decompress the air in the lungs. The increase in chest volume slightly reduces the air volume of the box, thereby increasing the pressure. First, the change in volume of the chest is quantified using Boyle's law with the initial pressure and volume of the box and the pressure in the box after expansion. The change in volume of the box is

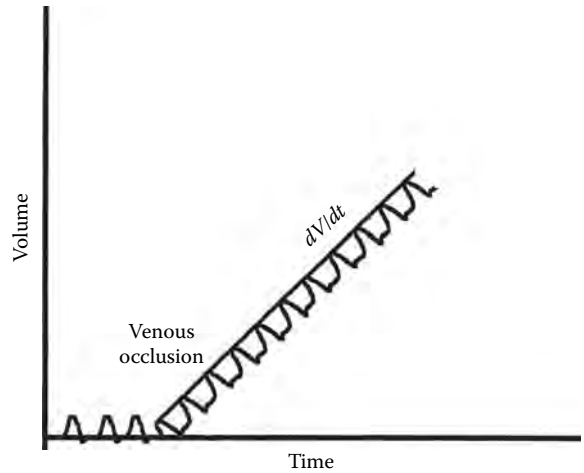


FIGURE 20.1 Typical recording from air plethysmograph during venous occlusion, revealing the changes in volume over time caused by arterial inflow.

equal to the change in volume of the chest. A second measurement method is based on the initial volume of the chest (unknown) and the initial pressure at the mouth, and the inspiratory volume (the unknown chest volume and the change in volume obtained in the first measurement) together with the pressure at the mouth during the inspiratory effort. By solving Boyle's law for the unknown volume, the original volume of gas present in the lungs when the shutter was closed is obtained, which is normally the volume present at the end of a normal exhalation or FRC.

Alternative methods for measurement of volume changes with plethysmography have been introduced to overcome the disadvantages of chamber plethysmography such as cost, complexity, and awkwardness of use. Important alternatives are elastic-resistance strain-gage plethysmography, impedance plethysmography, inductive plethysmography, and photoelectric plethysmography.

20.1.2 Electrical Plethysmography

20.1.2.1 Strain-Gage Plethysmography

In 1953, Whitney described the elastic-resistance strain-gage plethysmograph [4]. The strain-gage instrument quantifies the change in resistance of the gage as it is stretched due to the enlargement of the object (e.g., limb segment). The gage is typically a small elastic tube filled with mercury or an electrolyte or conductive paste. The ends of the tube are sealed with copper plugs or electrodes that make contact with the conductive column. These plugs are connected to a Wheatstone bridge circuit.

To illustrate the principle of the strain-gage method, a strain gage of length l_0 is placed around the limb segment, a circular cross section with radius r_0. The length of the strain gage can thus also be expressed as $l_0 = 2\pi r_0$. Expansion of the limb gives a new length $l_1 = 2\pi r_1$, with r_1 the new radius of the limb. The increase in length of the strain gage is thus $\delta l = 2\pi(r_1 - r_0)$, while the change in cross-sectional area of the limb δA is $\pi(r_1^2 - r_0^2)$. This can also be expressed as

$$\delta A = \pi[2r_0(r_1 - r_0) + (r_1 - r_0)^2] \tag{20.2}$$

Since $r_1 - r_0$ is usually small, $(r_1 - r_0)^2$ can be neglected. Consequently, δA can be written as $2\pi r_0(r_1 - r_0)$, which, on dividing by $A = \pi(r_0)^2$, gives

$$\frac{\delta A}{A} = 2\left(\frac{r_1 - r_0}{r_0}\right) \tag{20.3}$$

But since $\delta V/V = \delta A/A$ and $\delta l/l = \delta r/r$,

$$\frac{\delta V}{V} = 2\left(\frac{l_1 - l_0}{l_0}\right) \tag{20.4}$$

Thus, the percentage increase in volume can be obtained by measuring the initial gage length and the change in length. In practice, this change in length of the strain gage is recorded over a relatively short period to overcome the problem of back pressure that builds up in the limb or digit because of the venous occlusion, and the measurement is usually expressed as milliliters per minute per 100 g of tissue or volume flow rate = $2(\delta l/l_0) \times (100/t)$, where δl is the increase in gage length during time t. To measure $\delta l/t$, the gage must be calibrated by removing it from the limb and stretching it on a measuring jig until the recorder returns to the zero level; this is the length l_0. The gage can then be stretched accurately by typical amounts, to calibrate the output and calculate the volume flow rate.

20.1.2.2 Impedance Plethysmography

The physical principle of impedance plethysmography is the variation in electrical impedance of a tissue segment over the cardiac cycle due to a changing conductance. The change is caused by arterial inflow to the segment while the venous drainage is occluded.

If a segment of limb is assumed to be cylindrical and of length l and volume V, the impedance Z of that part can be written as

$$z = \frac{\rho l^2}{V} \tag{20.5}$$

where ρ is the specific resistivity of the tissue forming the cylinder.

Blood entering the tissue segment during the cardiac cycle introduces a parallel impedance Z_b described as $Z_b = Z_0 Z_1/(Z_0 - Z_1)$ with Z_0 the initial impedance and Z_1 the new value. A small change in volume δV is related to a small change in resistance δZ, such that

$$\delta Z = Z_0 - Z_1 = \rho l^2 \left(\frac{1}{V_0} - \frac{1}{V_1} \right) = \rho l^2 \frac{(V_1 - V_0)}{V_1\ V_0} \approx -\rho l^2 \frac{\delta V}{V^2} = -Z \frac{\delta V}{V} \tag{20.6}$$

when very small changes in volume are assumed [5]. Thus, δV can be written as

$$\delta V = \rho \left(\frac{1}{Z_0^2} \right) \delta Z \tag{20.7}$$

and these parameters Z_0 and δZ can be obtained from the impedance plethysmogram. If it is assumed that $\rho = 150\ \Omega$ cm, being the specific resistivity of soft tissue, the change in volume can be determined.

Two major techniques have been used to measure tissue impedance. One method uses two electrodes to measure both the voltage drop across the electrodes and the current flow. The main disadvantage of this technique is that the current distribution in the vicinity of the electrodes is not known, meaning the exact volume of the interrogated tissue is also unknown. In 1974, Kubicek et al. [6] described the use of four electrodes to measure tissue impedance: the outer electrodes supply a small-amplitude, high-frequency AC signal, while the inner two are used to measure the potential difference between two points on the same surface (see Figures 20.2 and 20.3).

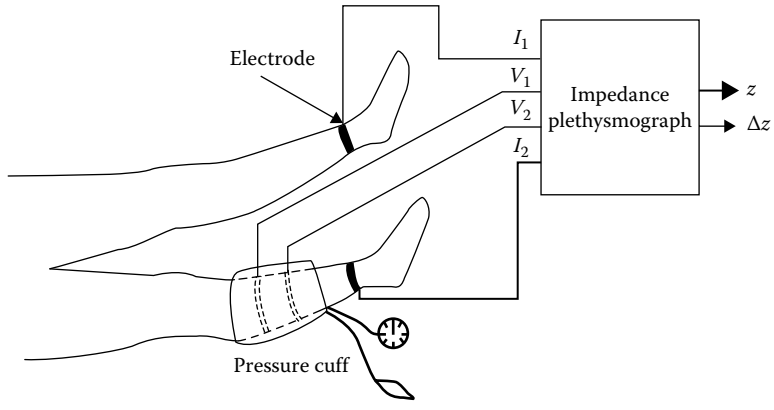


FIGURE 20.2 Electrode positioning for impedance plethysmography used for early detection of peripheral atherosclerosis: the outer two supply current, while the inner two measure voltage. The combination of these two is used to obtain the changes in impedance, and these changes are then used to obtain the changes in volume of the leg. (From Shankar, R. and Webster, J. G., *IEEE Trans. Biomed. Eng.*, 38, 62, 1993. With permission.)

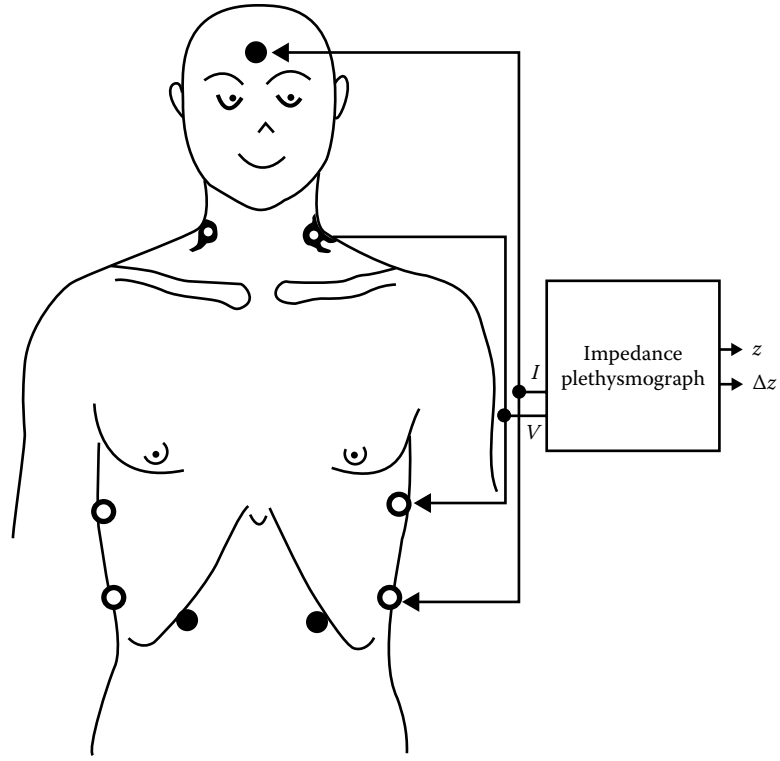


FIGURE 20.3 A second example of electrode positioning for impedance plethysmography here used to estimate stroke volume: the outer two supply current, while the inner two measure voltage. (From Verschoor, N. et al., *Physiol. Meas.*, 17, 29, 1996. With permission.)

20.1.2.3 Inductive Plethysmography

For measurement of volume changes during respiration, a system based on inductive plethysmography has been developed; the respiratory inductive plethysmograph was introduced in 1978. During inductive plethysmography, two coils are attached—one placed around the rib cage (RC) and the other around the abdomen. During respiration, the cross-sectional diameter of the coils changes, leading to a change in the inductances of the coils. By converting these inductances into proportional voltages, a measure for the changes in volume during respiration can be obtained. The conversion is achieved by using oscillators whose frequencies depend on a fixed internal capacitor and the inductance of each coil. These oscillator frequencies are converted to proportional voltages, voltages that are then recorded. After calibration on the patient, the system can be used for respiratory volume measurement [9].

Two general methods have been introduced for calibration of the respiratory inductive plethysmograph. These calibration methods are based on the following equation:

$$V_T = V_{RC} + V_{AB} \tag{20.8}$$

where V_{RC} and V_{AB} are the contributions of the RC and abdominal compartments (AB) to the tidal volume V_T. Calibration can be performed by determining the calibration constant in an isovolume measurement or by regression methods of differences in inductance changes due to a different body position or a different state of sleep [10]. Once it is calibrated, the respiratory inductive plethysmograph can be used for noninvasive monitoring of changes of thoracic volume and respiratory patterns.

20.2 Numerical Integration with Imaging

For diagnostic purposes in health care, a number of imaging modalities are currently in use. Webb nicely reviewed the scientific basis and physical principles of medical imaging [11].

With medical imaging modalities, diameters of objects can be determined in a noninvasive manner. By applying these diameters in an appropriate formula for volume calculation, the size of the object can be estimated. For example, in a clinical setting, the prostate volume V_p can be approximated with an ellipsoid volume calculation:

$$V_p = \frac{\pi}{6} HWL \tag{20.9}$$

where

H is the height
W is the width
L is the length of the prostate as illustrated in Figure 20.4

Variations have been introduced to find the best formulae to estimate the prostate volume [12,13]. Also, for the determination of ventricular volumes, formulae have been derived, based on the following variation of the formula for prostate volume measurements: (1) The volume is obtained with the same formula as for prostate volume measurements, (2) the two short axes are equal, (3) the long axis is twice the short axis, and (4) the internal diameter is used for all axes.

By displaying cross-sectional images of a 3D object (e.g., an organ), volume calculations can also be performed by integrating areas of the organ over a sequence of medical images as illustrated in Figure 20.5. For ultrasound images, applications have been reported in the literature on prostate, heart, bladder, and kidney volume. The application of integration techniques involves the application of segmentation algorithms on sequences of images, either in transverse direction with fixed intersection distance or in the radial with images at a fixed intersection angle.

For a volume defined by the function $f(x, y, z)$, the volume can be calculated from

$$V = \int_{x=x_1}^{x_2} \int_{y=y_1}^{y_2} \int_{z=z_1}^{z_2} f(x,y,z)dzdydx \tag{20.10}$$

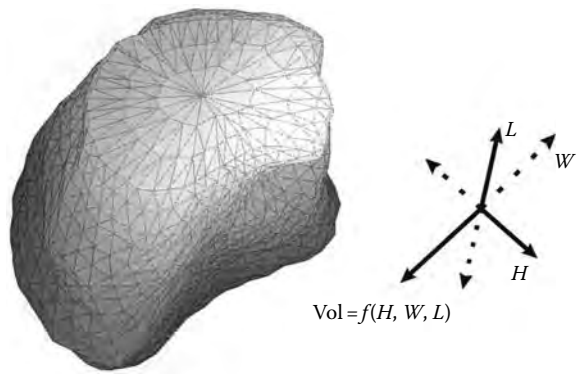


FIGURE 20.4 Volume estimate of a prostate, obtained from 3D reconstruction of cross-sectional images, using the dimensions in three planes: the length, height, and width in a volume formula describing an ellipsoid shape.

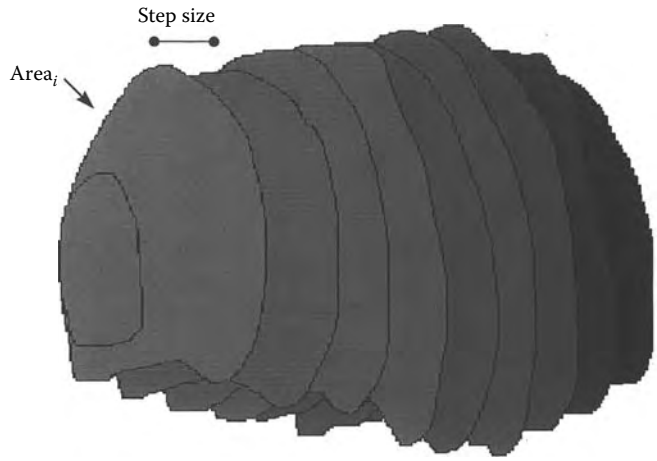


FIGURE 20.5 Outlines of cross-sectional images of the prostate used for planimetric volumetry of the prostate: transverse cross sections obtained with a step size of 4 mm have been outlined by an expert observer, and the contribution of each section to the total volume is obtained by determining the area enclosed by the prostate contour.

By applying numerical integration, the volume calculation using an integral is approximated by a linear combination of the values of the integrand. For reasons of simplicity, this approximation is first illustrated for a 1D case:

$$\int_{x=a}^{b} f(x)dx \approx w_1 f(x_1) + w_2 f(x_2) + \cdots + w_{nf} f(x_n) \tag{20.11}$$

In this equation, $x_1, x_2, \ldots, x_n$ are n points chosen in the interval of integration $[a, b]$, and the numbers $w_1, w_2, \ldots, w_n$ are n weights corresponding to these points. This approximation can also be expressed in the so-called Riemann sum:

$$\int_{x=a}^{b} f(x)dx \approx w_1 f(x_1) + w_2 f(x_2) + \cdots + w_n f(x_n) = \sum_{i=1}^{n} f(\xi_i)(x_i - x_{i-1}) \tag{20.12}$$

For volume calculation, the Riemann sum needs to be extended to a 3D case:

$$V = \int_{x=x_1}^{x_2} \int_{y=y_1}^{y_2} \int_{z=z_1}^{z_2} f(x,y,z)dzdydx \approx \sum f(p_{ijk})(x_{i+1} - x_i)(y_{j+1} - y_j)(z_{k+1} - z_k) \tag{20.13}$$

This 3D case can be reduced by assuming that $(x_{i+1} - x_i)(y_{i+1} - y_i)f(P_{ijk})$ represents the surface S_k of a 2D section at position z_k, while $(z_{k+1} - z_k)$ is constant for every k and equals h, the intersection distance:

The volume is then determined by summarizing all contributions after multiplying them with the step size:

$$V \approx h \sum S(z_k) = h \sum_{k=0}^{n-1} S(a + kh) \tag{20.14}$$

where

a represents the position of the first section

n is the number of sections

h is the step size, given by $h = (b - a)/n$, where b is the position of the last section

In general, three different forms of the approximation can be described, depending on the position of the section. The first form is known as the rectangular rule and is presented in Equation 20.14. The midpoint rule is described by

$$V \approx h\sum S(z_k) = h\sum_{k=0}^{n-1} S\left[a + \left(k + \frac{1}{2}h\right)\right] \tag{20.15}$$

The trapezoidal rule is similar to the rectangular rule, except that it uses the average of the right-hand and left-hand Riemann sum:

$$T_n(f) = h\left[\frac{f(a)}{2} + f(a+h) + f(a+2h) + \cdots + f(a+(n-1)h) + \frac{f(b)}{2}\right] \tag{20.16}$$

The trapezoidal and midpoint rules are exact for a linear function and converge at least as fast as n^{-2}, if the integrand has a continuous second derivative [14].

A number of parameters are associated with the accuracy of the approximation of the volume integral with a Riemann sum; first of all, the number of cross sections n is important as the sum converges with n^{-2}: the more sections, the more accurate the results. Furthermore, the selection of the position of the first section is important. While the interval of integration is mostly well defined in theoretical analyses, for in vivo measurements, the first section is often less well defined. This means not only that the selection of this first section is random with the first possible section and this section increased with the step size but also that sections can be missed. The last effect that may influence the numerical results is the effect of taking oblique sections, cross sections that are not taken perpendicular to the axis along which the intersection distance is measured. All the aforementioned assume that every section S_k can be exactly determined.

The influence of the parameters associated with these kinds of measurements can be tested in a theoretical model. In a mathematical model with known exact volume or surface, the parameters—including step size, selection of the first section, and the scan angle—can be tested. Figure 20.6 shows a model used to model the prostate in two dimensions. It consists of a spheroid function with a length of the long axis of 50 mm and a length of the short axis of 40 mm. In mathematical form, this function can be described as

$$y = f(x) = y_m \pm y_0\sqrt{1 - \left(\frac{x - x_m}{x_0}\right)^2} \tag{20.17}$$

where

x_0 is half the diameter of the ellipse in x-direction (25 mm)
y_0 is half the diameter of the ellipse in y-direction (20 mm)
(x_m, y_m) is the center point of the ellipse

The analytical solution of the integral to obtain the surface of this model is $\pi x_0 y_0$.

Figure 20.6 also reveals the possible influence of the parameters h and α. Figure 20.7 shows the results for varying the step size h between 4 and 16 mm corresponding to between 13 and

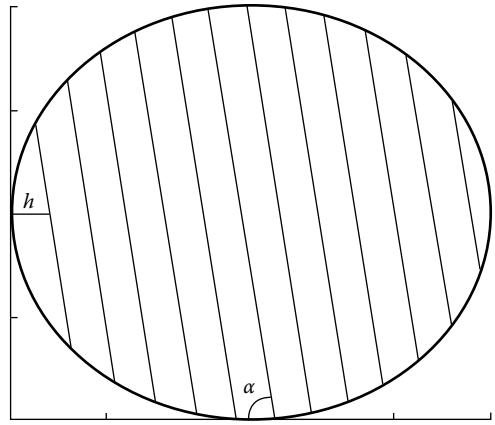


FIGURE 20.6 An ellipsoid-shaped model of the prostate that was used for theoretical analysis of planimetric volumetry. Indicated are the intersection distance h and the scan angle α with the axis of the probe. (From Aarnink, R.G. et al., *Physiol. Meas*., 16, 141, 1995. With permission.)

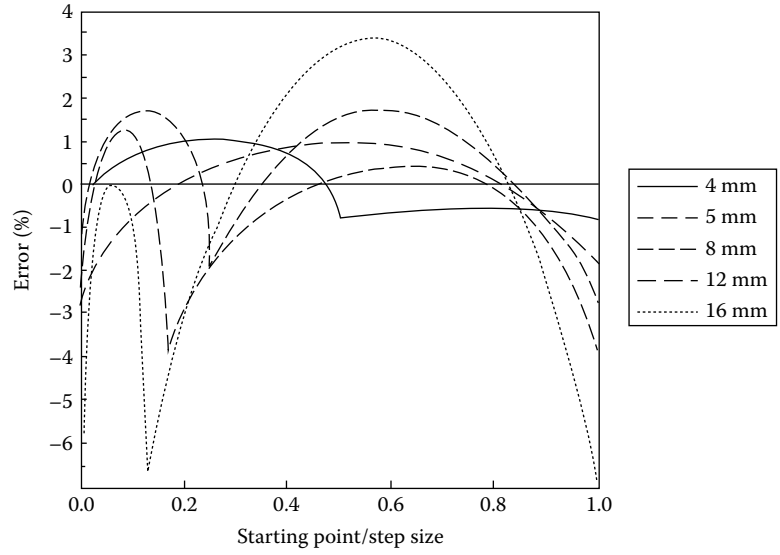


FIGURE 20.7 The errors in surface estimates by numerical integration for different step sizes h presented as percentage of the exact solution, as a function of the position of the first section, located between 0 and h. (From Aarnink, R.G. et al., *Physiol Meas*., 16, 141, 1995. With permission.)

3 sections used for numerical integration as a function of the position of the first section. The first section is selected (at random) between the first coordinate of the function and this coordinate plus the step size h. The sections were taken perpendicular to the long axis. From this study, which is fully described by Aarnink et al. [15], it was concluded that if the ratio of step size to the longest diameter is 6 or more, then the error in volume estimation is less than 5%. Although clinical application might introduce additional errors such as movement artifacts, this rule provides a good indication of the intersection distance that should be used for good approximations with numerical integration.

20.3 Indicator Dilution Methods

The principle of the indicator dilution theory to determine gas or fluid volume was originally developed by Stewart and Hamilton. It is based on the following concept: if the concentration of an indicator that is uniformly dispersed in an unknown volume is determined, and the volume of the indicator is known, the unknown volume can be determined. Assuming a single inflow and single outflow model, all input will eventually emerge through the output channel, and the volumetric flow rate can be used to identify the volume flow. It can be described with two equations, of which the second is used to determine the volume using the result of the first equation [16]:

$$F = \frac{I_i}{\int_0^\infty C(t)dt} \tag{20.18}$$

$$\text{Vol} = Ft_m \tag{20.19}$$

where

F is the flow rate
I_i is the total quantity of injected indicator
$C(t)$ is the concentration of the contrast as function of time
Vol is the volume flow
t_m is the mean transit time

The integrated value of the concentration over time can be obtained from the concentration-time curve (Figure 20.8) using planimetry and an approximation described by the Riemann sum expression (Equation 20.13). Care must be taken to remove the influence of the effect of recirculation of the indicator. An indicator is suitable for use in the dilution method if it can be detected and measured and remains within the circulation during at least its first circuit, without doing harm to the object. Subsequent removal from the circulation is an advantage. Dyes such as Coomassie blue and indocyanine green were initially used, using the peak spectral absorption at certain wavelengths to detect the indicator. Also, radioisotopes such as *I*-labeled human serum albumin have been used. Currently, the use of cooled saline has become routine in measurement of cardiac output in the intensive care unit.

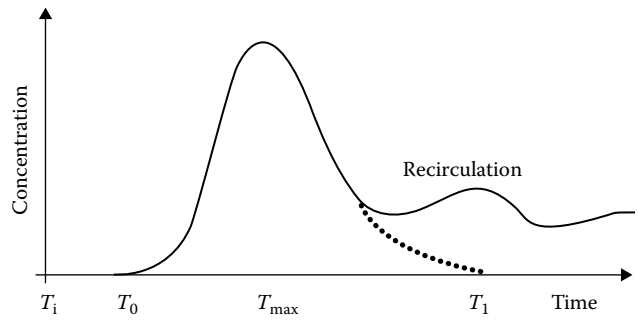


FIGURE 20.8 The time–concentration curve of the indicator is used to obtain the unknown volume of the fluid. The area under the time–concentration curve is needed to estimate the volumetric flow rate, and this flow rate can be used to estimate the volume flow. For accurate determination of the flow rate, it is necessary to extrapolate the first-pass response before the area under the curve is estimated.

20.3.1 Thermodilution

To perform cardiac output measurements, a cold bolus (with a temperature gradient of at least 10 K) is injected into the right atrium, and the resulting change in temperature of the blood in the pulmonary artery is detected using a rapidly responding thermistor. After recording the temperature–time curve, the blood flow is calculated using a modified Stewart–Hamilton equation. This equation can be described as follows [17]:

$$V_i SW_i SH_i (T_b - T_i) = F \int T_b(dt) SW_b SH_b \tag{20.20}$$

The terms on the left represent the cooling of the blood (with temperature T_b), as caused by the injection of a cold bolus with volume V_i, temperature T_i, specific weight SW_i, and specific heat SH_i. The same amount of "indicator" must appear in the blood (with specific weight and heat SW_b and SH_b, respectively) downstream to the point of injection, where it is detected in terms of the time course of the temperature T_b in the flow F by means of a thermistor. The flow can be described as

$$F = \frac{T_b - T_i}{\int T_b(dt)} K \tag{20.21}$$

where K (the so-called calculation constant) is described as

$$K = V_i \frac{SW_i - SH_i}{SW_b - SH_b} C \tag{20.22}$$

This constant is introduced to the computer by the user. T_i is usually measured at the proximal end of the injection line; therefore, a "correction factor" C must be introduced to correct for the estimated losses of cold saline in the catheter, due to the dead space volume. Furthermore, the warming effect of the injected fluid during passing through the blood must be corrected; this warming effect depends on the speed of injection, the length of immersion of the catheter, and the temperature gradient. This warming effect reduces the actual amount of injected fluid, so the correction factor C is less than 1 [17].

20.3.2 Radionuclide Techniques

Radionuclide imaging is a common noninvasive technique used in the evaluation of cardiac function and disease. It can be compared to x-ray radiology, but now the radiation emanating from inside the human body is used to construct the image. Radionuclide imaging has proven useful to obtain ejection fraction (EF) and left ventricular volume. As in thermodilution, a small volume of indicator is injected peripherally; in this case, radioisotopes (normally indicated as radiopharmaceuticals) are used. The radioactive decay of radioisotopes leads to the emission of alpha, beta, gamma, and x-radiation, depending on the radionuclide used. For in vivo imaging, only gamma or x-radiation can be detected with detectors external to the body, while the minimal amount of energy of the emitted photons should be greater than 50 keV. Using a photon counter such as a gamma camera, equipped with a collimator, the radioactivity can be recorded. Two techniques are commonly used: the first-pass method and the dynamic recording method [11].

20.3.3 First-Pass Method

To inject a radionuclide bolus into the blood system, a catheter is inserted in a vein, for example, an external jugular or an antecubital vein. The camera used to detect the radioactive bolus is usually positioned in the left anterior oblique position to obtain optimal right and left ventricular separation and is tilted slightly in an attempt to separate left atrial activity from that of the left ventricle. First, the background radiation is obtained by counting the background emissions. Then, a region of interest is determined

over the left ventricle and a time–activity curve is generated by counting the photon emissions over time. The radioactivity count over time usually reveals peaks at different moments; the first peak occurs when the radioactivity in the right ventricle is counted, the second peak is attributed to left ventricular activity. More peaks can occur during recirculation of the radioactivity. After correction for background emissions, the time–intensity curve is evaluated to determine the EF of the heart, which is given by

$$\mathrm{EF} = \frac{c_\mathrm{d} - c_\mathrm{s}}{c_\mathrm{d} - c_\mathrm{b}} \tag{20.23}$$

where

c_d is the end diastolic count
c_s is the end systolic count
c_b is the background count [18]

Using a dynamic cardiac phantom, the accuracy of the EF measurement and LV volume estimation by radionuclide imaging has been determined [18]. The count-based method was found to give accurate results for evaluating cardiac EF and volume measurements.

20.3.4 Dynamic Recording Method

During dynamic recording, the data acquisition is triggered by another signal, usually the ECG, to obtain the gated activity during ventricular systole and diastole. The contraction of the left ventricle of the heart is used to align the acquisition data from different cardiac cycles. During dynamic recording, information about the cardiac function is averaged over many heartbeats during a period when the radiopharmaceuticals are uniformly distributed in the blood pool. Estimations of end systolic and end diastolic volumes are made and the EF is calculated (see preceding text). Also, other parameters can be obtained, for example, the amplitude and phase of the ejection from the left ventricle. These parameters may show malfunction in the cardiac cycle. These dynamic recordings may also be applied to the brain, lungs, liver, kidney, and vascular systems, and they can be important to judge the function of these organs (e.g., after kidney transplantation).

20.3.5 Gas Dilution

Gas dilution is a method to determine lung volumes because standard techniques for lung measurements measure only the inhaled or exhaled air as a function of time (spirometry) and cannot be used to assess the absolute lung volume. The subject is connected to a spirometer that contains a known concentration of gas (e.g., helium). The subject is then asked to breathe for several minutes, to equalize the concentration of helium in the lung and the spirometer. Using the law of conservation of matter, the volume of the lung can be calculated. Since the total amount of helium is the same before and after measurement, the fractional concentration times the volume before equals the fractional concentration times the volume after: $C_1, \times V_1, = C_2 \times (V_1, + V_2)$. The volume after the measurement can be extracted from this equation, and by subtracting the volume of the spirometer, the lung volume is calculated.

20.4 Water Displacement Volumetry

The measurement of the volume of an extremity such as arm or leg can be measured using a water tank, which is illustrated schematically in Figure 20.9. The advantage of water displacement volumetry is the possibility for direct measurement of objects with an irregular form. In the clinical situation, volume determination of the leg can be valuable for monitoring the severity of edema or hematoma after surgery or trauma. A setup for water displacement can be developed by the user, and an example of such a system is described by Kaulesar Sukul et al. [19]. A special tank for water displacement volumetry of

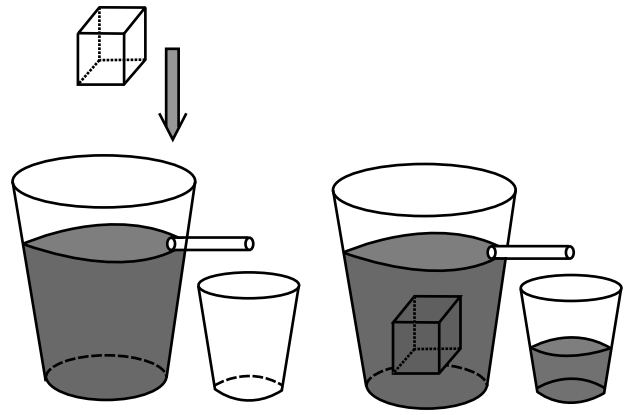


FIGURE 20.9 A schematic overview of a volume-measuring method using water displacement: the desired volume of the object is represented by the overflow volume in the small tank after lowering the object into the large tank. This technique is especially useful for irregular-shaped objects.

the leg was constructed consisting of two overflow tubes. The tank was filled to the lower overflow tube, and the overflow tube was subsequently closed. The patient was then asked to lower his leg into the tank, and the amount of overflow of the upper tube was measured. The total volume was then calculated by measuring the fluid volume delivered to the upper tube and the volume difference between the lower and upper overflow tube, the so-called reserve volume. The volume of ankle and foot was then measured by filling the tank to the upper overflow tube and measuring the amount of water in the cylinder when the foot and ankle were immersed in the water.

Disadvantages of water displacement volumetry are hygiene problems; it is time-consuming and not suitable for measurements of the volume of extremities of patients in the immediate postoperative period. Therefore, alternatives have been sought, with the disk model method as a promising one. The calculation of the volume of the leg is performed by dividing the leg into disks of thickness h (e.g., 3 cm) as illustrated in Figure 20.10. The total volume is equal to the sum of the individual disk volumes:

$$V = \sum_{i=1}^{n} \frac{C_i^2}{4\pi} h = \sum_{i=1}^{n} \pi r_i^2 h = h \sum_{i=1}^{n} \pi r_i^2 \tag{20.24}$$

where C_i is the circumference of the disk at position i with assumed radius of r_i.

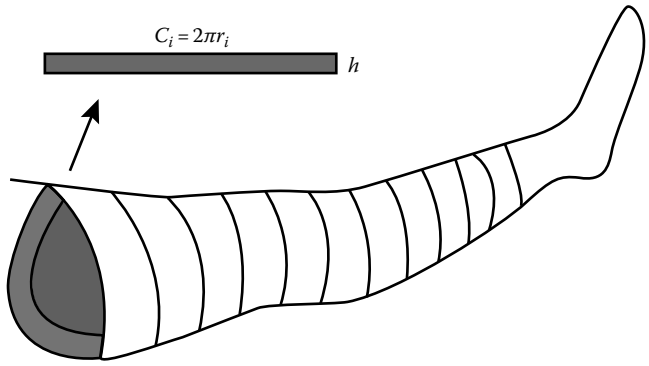


FIGURE 20.10 Illustration of the disk model to measure leg volume. The leg is divided in different sections with a fixed intersection distance; the circumference of each section is obtained. This circumference is then used to describe the leg as perfect circles with an assumed radius calculated from the circumference. This radius is then used to calculate the contribution of that section to the volume and the total volume is obtained by summation.

A study to compare water displacement volumetry with the results of the disk model method indicated that both methods give similar results. Consequently, because of the ease of application, the disk model method is the method of choice to measure volumes of extremities. Assuming the length of a leg to be 75 cm, a ratio between length and step size (3 cm was proposed in [19]) of 25 is obtained. According to a study by Aarnink et al. [15], this ratio can be reduced while still obtaining an accuracy of >95%. A disk height of 10 cm will be sufficient to measure the leg volume with an accuracy of at least 95%. However, it might be necessary to use a small interdisk section to obtain a higher accuracy, which might be needed to accurately monitor the edema changes in volume.

20.5 Equipment and Experiments

Table 20.1 summarizes the different methods for volume measurement and their applications. The table serves as a starting point to evaluate the available equipment for volume measurement.

20.6 Evaluation

Plethysmography is an easy and noninvasive method to obtain knowledge for assessing vascular diseases, cardiac output disorders, or pulmonary dysfunctions. Systems for direct measurement of volume changes have been developed, but physical properties have also been introduced as an indirect measure of the change in volume. Each system has its own advantages and disadvantages.

For *chamber plethysmography*, the water-filled type is more stable with respect to temperature change, but thermal problems may occur if the water temperature is significantly different from that of the limb segment. Furthermore, the hydrostatic effect of the water may alter the blood flow. Also, the rubber sleeve between the limb segment and the water may influence the release of sweat. It is a cumbersome measurement that is not very useful during or after exercise. The *air displacement* type introduces problems of drift because of its high coefficient of expansion, although self-compensating systems

TABLE 20.1 Volume-Measuring Techniques, Applications, and Equipment for Different Applications

Technique	Application	Companies	Products	Price (U.S.$)
Spirometry	Lung volume	Nellcor Puritan Bennett	Renaissance	
		Morgan Science Spirometrics		
		CDX Corporation	Spiro 110S	
Whole-body plethysmography	Lung volume	Morgan Science ACI Medical Inc.		25,000.00
Gas dilution	Lung volume	Equilibrated Biosystems Inc. Melville		
Thermodilution	Heart	Abbott Critical Care System Baxter		
		American Edwards Laboratories		
Strain-gage plethysmography	Cardiac output	Parks Medical Electronics		
Impedance plethysmography	Perfusion studies	Ambulatory monitoring systems Vitalog		
		RJL systems Detroit Codman and Shurtleff Inc. Randolph		
		Electrodiagnostic Instrument Inc. Burbank		
Inductive Plethysmograph	Lung volume	SensorMedics BV	RespiTrace plus SomnoStar PT	15,000.00
Radionuclide imaging	Heart, peripheral organs	Schering		

have been proposed. Also, the thermal behavior of the body plethysmograph may produce variations in pressure inside the airtight box. Two sources of temperature variation can be mentioned: the chest volume variation induced by breathing produces heat and a temperature gradient may exist between the patient and the internal temperature of the box [20].

In *impedance plethysmography*, the changes in impedance in tissue are used, which are primarily due to changes in the conductivity of the current path with each pulsation of blood. Several theories attempt to explain the actual cause of these changes in tissue impedance. One explanation is that blood filling of a segment of the body lowers the impedance of that segment. A second theory is that the increase in diameter due to additional blood in a segment of the body increases the cross-sectional area of the segment's conductivity path and thereby lowers the resistance of the path. A third explanation is based on the principle of pressure changes on the electrodes that occur with each blood pulsation and uses the changes in the impedance of the skin–electrode interface. The main difficulty with the procedure is the problem of relating the output resistance to any absolute volume measurement. Detection of the presence of arterial pulsations, measurement of pulse rate, and determination of time of arrival of a pulse at any given point in the peripheral circulation can all be satisfactorily handled by impedance plethysmography. Also, the impedance plethysmograph can measure time-variant changes in blood volume. A problem with impedance plethysmography may be the sensitivity to movement of the object. Research is being conducted to reduce the influences of these movement artifacts, either by different electrode configuration, electrode location, or using multiple sensors or different frequencies [21].

A low-cost *inductive plethysmograph* was designed by Cohen et al. to obtain a noninvasive measure of lung ventilation [22]. This plethysmograph indirectly monitors ventilation by measuring the cross-sectional area of the chest and abdomen. They attached commercially available elastic bands containing wire around the chest and abdomen and determined their inductances by measuring the frequency of an inductive-controlled oscillator.

New devices for plethysmographic measurements are under development, using different properties to obtain the quantitative knowledge required. For example, *acoustic plethysmography* measures body volume by determining changes in the resonant frequency of a Helmholtz resonator. A Helmholtz resonator consists of an enclosed volume of gas connected to its surroundings through a single opening. The gas can be forced to resonate acoustically by imposing periodic pressure fluctuations of the opening. The resonator frequency is inversely proportional to the square root of the volume of air in the resonator. An object placed in the resonator reduces the volume of air remaining in the resonator by its own volume, causing an increase in the resonator frequency. From a study to obtain density values of preterm infants, it was concluded that the acoustic plethysmograph can be used to measure total body volume of preterm infants [23].

In addition, volume measurements with diagnostic imaging modalities are well established in the clinical environment, for example, prostate ultrasonography or echocardiography. Besides important parameters such as step size and first step selection, accurate determination of the surface in different sections is important. While predominantly performed manually now, several attempts for automated detection of the surface have been reported [24–27]. Automatic detection of the surface in each section should be possible and would enable a system for automated measurement of the prostate, heart, liver, etc. Further research will indicate the usefulness of such a system in a clinical setting.

References

1. A. J. Comerota, R. N. Harada, A. R. Eze, and M. L. Katz, Air plethysmography: A clinical review, *Int. Angiol.*, 14, 45–52, 1995.
2. A. B. Dubois, S. J. Botello, G. N. Beddell, R. Marshall, and J. H. Comroe, A rapid plethysmographic method for measuring thoracic gas volume: A comparison with a nitrogen washout method for measuring functional residual capacity in normal subjects, *J. Clin. Invest.*, 35, 322–326, 1956.
3. J. B. West, *Respiratory Physiology—The Essentials*, Baltimore, MD: Williams and Wilkins, 1987.

4. J. R. Whitney, The measurement of volume changes in human limbs, *J. Physiol*., 121, 1–27, 1953.
5. J. Nyboer, S. Bagno, and L. F. Nims, The impedance plethysmograph: An electrical volume recorder, National Research Council, Committee on Aviation Medicine, Report No. 149, Washington, DC, 1943.
6. W. G. Kubicek, F. J. Kottke, M. V. Ramos, R. P. Patterson, D. A. Witsoe, J. W. Labree, W. Remole, T. E. Layman, H. Schoening, and J. T. Garamala, The Minnesota impedance cardiograph—Theory and applications, *Biomed. Eng*., 9, 410–416, 1974.
7. R. Shankar and J. G. Webster, Noninvasive measurement of compliance of human leg arteries, *IEEE Trans. Biomed. Eng*., 38, 62–67, 1993.
8. N. Verschoor, H. H. Woltjer, B. J. M. van der Meer, and P. M. J. M. de Vries, The lowering of stroke volume measured by means of impedance cardiography during endexpiratory breath holding, *Physiol. Meas*., 17, 29–35, 1996.
9. M. A. Cohn, H. Watson, R. Weisshaut, F. Stott, and M. A. Sackner, A transducer for non-invasive monitoring of respiration, in *ISAM 1977, Proceedings of Second International Symposium on Ambulatory Monitoring*, London, U.K.: Academic Press, 1978, pp. 119–128.
10. J. A. Adams, Respiratory inductive plethysmography, in J. Stocks, P. D. Sly, R. S. Tepper, and W. J. Morgan (eds.), *Infant Respiratory Function Testing*, New York: Wiley-Liss, 1996, pp. 139–164.
11. S. Webb, *The Physics of Medical Imaging*, Bristol, U.K.: IOP Publishing, 1988, pp. 204–221.
12. M. K. Terris and T. A. Stamey, Determination of prostate volume by transrectal ultrasound, *J. Urol*., 145, 984–987, 1991.
13. R. G. Aarnink, J. J. M. C. H. de la Rosette, F. M. J. Debruyne, and H. Wijkstra, Formula-derived prostate volume determination, *Eur. Urol*., 29, 399–402, 1996.
14. P. J. Davis and P. Rabinowitz, *Methods of Numerical Integration*, San Diego, CA: Academic Press, 1975, pp. 40–43.
15. R. G. Aarnink, R. J. B. Giesen, J. J. M. C. H. de la Rosette, A. L. Huynen, F. M. J. Debruyne, and H. Wijkstra, Planimetric volumetry of the prostate: How accurate is it? *Physiol. Meas*., 16, 141–150, 1995.
16. E. D. Trautman and R. S. Newbower, The development of indicator-dilution techniques, *IEEE Trans. Biomed. Eng*., 31, 800–807, 1984.
17. A. Rubini, D. Del Monte, V. Catena, I. Ittar, M. Cesaro, D. Soranzo, G. Rattazzi, and G. L. Alatti, Cardiac output measurement by the thermodilution method: An in vitro test of accuracy of three commercially available automatic cardiac output computers, *Intensive Care Med*., 21, 154–158, 1995.
18. S. Jang, R. J. Jaszczak, F. Li, J. F. Debatin, S. N. Nadel, A. J. Evans, K. L. Greer, and R. E. Coleman, Cardiac ejection fraction and volume measurements using dynamic cardiac phantoms and radionuclide imaging, *IEEE Trans. Nucl. Sci*., 41, 2845–2849, 1994.
19. D. M. K. S. Kaulesar Sukul, P. T. den Hoed, E. J. Johannes, R. van Dolder, and E. Benda, Direct and indirect methods for the quantification of leg volume: Comparison between water displacement volumetry, the disk model method and the frustum sign model method, using the correlation coefficient and the limits of agreement, *J. Biomed. Eng*., 15, 477–480, 1993.
20. P. Saucez, M. Remy, C. Renotte, and M. Mauroy, Thermal behavior of the constant volume body plethysmograph, *IEEE Trans. Biomed. Eng*., 42, 269–277, 1995.
21. J. Rosell, K. P. Cohen, and J. G. Webster, Reduction of motion artifacts using a two-frequency impedance plethysmograph and adaptive filtering, *IEEE Trans Biomed. Eng*., 42, 1044–1048, 1995.
22. K. P. Cohen, D. Panescu, J. H. Booske, J. G. Webster, and W. L. Tompkins, Design of an inductive plethysmograph for ventilation measurement, *Physiol. Meas*., 15, 217–229, 1994.
23. O. S. Valerio Jimenez, J. K. Moon, C. L. Jensen, F. A. Vohra, and H. P. Sheng, Pre-term infant volume measurements by acoustic plethysmography, *J. Biomed. Eng*., 15, 91–98, 1993.
24. R. G. Aarnink, R. J. B. Giesen, A. L. Huynen, J. J. M. C. H. de la Rosette, F. M. J. Debruyne, and H. Wijkstra, A practical clinical method for contour determination in ultrasonographic prostate images, *Ultrasound Med. Biol*., 20, 705–717, 1994.

25. C. H. Chu, E. J. Delp, and A. J. Buda, Detecting left ventricular endocardial and epicardial boundaries by digital two-dimensional echocardiography, *IEEE Trans. Med. Imag.*, 7, 81–90 1988.
26. J. Feng, W. C. Lin, and C. T. Chen, Epicardial boundary detection using fuzzy reasoning, *IEEE Trans. Med. Int.*, 10, 187–199, 1991.
27. S. Lobregt and M. A. Viergever, A discrete contour model, *IEEE Trans. Med. Imag.*, 14, 12–24, 1995.

Further Information

Anonymous, AARC clinical practice guideline: Static lung volume, *Respir. Care*, 39, 830–836, 1994.
Anonymous, AARC clinical practice guideline: Body plethysmography, *Respir. Care*, 39, 1184–1190, 1994.
Bernstein, E. F. (ed.), *Noninvasive Diagnostic Techniques in Vascular Disease*, 3rd edn., St. Louis, MO: Mosby, 1985.
Davis, P. J. and Rabinowitz, P., *Methods of Numerical Integration*, London, U.K.: Academic Press, 1975.
Feigenbaum, H., *Echocardiography*, 5th edn., Philadelphia, PA: Lee & Febiger, 1993.
McDicken, W. N., *Diagnostic Ultrasonics: Principles and Use of Instruments*, 3rd edn., London, U.K.: Crosby Lockwood Staples, 1991.
Nyboer, J., *Electrical Impedance Plethysmography*, Springfield, IL: Charles Thomas, 1959.
The Science Notebook, *Exploring Measurement Volume*, http://science-notebook.com/measure02-vol.html, accessed on August 6, 2012.
Webb, S., *The Physics of Medical Imaging*, Bristol, U.K.: IOP Publishing, 1988.
West, J. B., *Respiratory Physiology—The Essentials*, Baltimore, MD: Williams and Wilkins, 1987.

Partial List of Manufacturers and Suppliers

BrandTech Scientific, Inc., 11 Bokum Rd., Essex, CT 06426, Phone: 860-767-2562, Fax: 860-767-2563, www.brandtech.com
Emerson Process Management, 310 E. University Dr., McKinney, TX 75069, Phone: 972-548-3574, 800-558-5853 (toll free), Fax: 972-542-6433, www2.emersonprocess.com/en-US/brands/fis…
Floyd Bell, Inc., 720 Dearborn Park Lane, Columbus, OH 43085, Phone: 614-294-4000, 888-356-9323 (toll free), Fax: 614-291-0823, www.floydbell.com
Fluid Metering, Inc., 5 Aerial Way, Suite 500, Syosset, NY 11791, Phone: 516-922-6050, 800-747-5858 (toll free), Fax: 516-624-8261, www.fmipump.com/
Seraphin Test Measure/Pemberton, 30 Indel Ave., Rancocas, NJ 08073-0227, Phone: 609-267-0922, 800-573-6322 (toll free), Fax: 609-261-2546, 800-373-6322 (toll free), www.seraphinusa.com
Sick, Inc., 6900 W. 110th St., Bloomington, MN 55438, Phone: 952-941-6780, 800-325-7425 (toll free), Fax: 952-941-9287, www.sick.com/us/en-us/home/Pages/Homepag…
Simtech, USA, 47-A Runway Rd., Bristol Industrial Park, Levittown, PA 19057, Phone: 215-547-0444, 877-777-2467 (toll free), Fax: 215-547-9129, www.simtechusa.com
Vindum Engineering. Inc., 1 Woodview Court, San Ramon, CA 94582, Phone: 925-275-0633, Fax: 925-275-9697, www.vindum.com

21

Tilt Measurement

Adam Chrzanowski
University of New Brunswick

James M. Secord
University of New Brunswick

21.1 Tiltmeters or Inclinometers **21**-1
21.2 Geodetic Leveling **21**-4
21.3 Hydrostatic Leveling **21**-5
21.4 Suspended and Inverted Plumb Lines **21**-7
21.5 Integration of Observations **21**-9
References **21**-9
Further Readings **21**-11
Partial List of Vendors and Suppliers **21**-11

If the relative position of two points is described by the 3D orientation of a line joining them, then, in general, *tilt* is the angular amount that the orientation has changed, in a vertical plane, from a previous direction or from a reference direction. If the original or reference orientation is nearly horizontal, then the term "tilt" is usually used. If it is nearly vertical, then the change is often regarded as "inclination." Here, "tilt" will refer to either. The two points can be separated by a discernable amount, the base, or the tilt can be measured at a point with the reference orientation being defined by the direction of the force of gravity at that point. Thus, the same instrument that measures tilt at a point can be called either a *tiltmeter* or an *inclinometer* (or clinometer), depending on the interpretation of the results. The instrument used to measure a series of tilts along any vertical orientation is often called an inclinometer (e.g., [1]).

Angular tilt is directly related to the linear amount of change subtending the length of the base. Consequently, angular tilt does not have to be measured directly but can be derived from the mechanical or other measurement of this linear change if the length of the base is known.
Therefore, the following discussion has been subdivided into

1. Tiltmeters or inclinometers (angular tilt at a point or over a limited, relatively small base length)
2. Geodetic leveling (tilt derived from a height difference over an extended base of virtually limitless length)
3. Hydrostatic leveling (tilt derived from a height difference over an extended base of limited length)
4. Suspended and inverted pendula or plumb lines (inclination from a difference in horizontal relative position over a vertical base or height difference)

21.1 Tiltmeters or Inclinometers

Considering the basic principle of operation, tiltmeters may be divided into liquid (including spirit bubble type), vertical pendulum, or horizontal pendulum. Dunnicliff [1] provides a comprehensive review of tiltmeters and inclinometers according to the method in which the tilt is sensed (i.e., mechanical, with accelerometer transducer, with vibrating wire transducer, or with electrolytic transducer).

The sensitivity of tilt falls into two distinct groups: geodetic or geophysical special tiltmeters with a resolution of 10^{-8} rad (0.002 in.) or even 10^{-9} rad and engineering tiltmeters with resolutions from 0.1 in. to several seconds of arc, depending on the required range of tilt to be measured.

The first group includes instruments that are used mainly for geophysical studies of Earth tide phenomena and tectonic movements, for example, the Verbaander–Melchior [2] and the Zöllner [3,4] horizontal pendulum tiltmeters and the Rockwell Model TM-1 [5] liquid-bubble type. This category of instrument requires extremely stable mounting and a controlled environment. There are very few engineering projects where such sensitive instruments are required. However, deformation measurements of underground excavations for the storage of nuclear waste may be one of the few possible applications. An example is a mercury tiltmeter (Model 300) developed for that purpose by the Auckland Nuclear Accessory Co. Ltd. in New Zealand. In this instrument, the change in capacitance between electrodes and a mercury surface is proportional to the tilt. This tiltmeter, with a total range of 15 in., is claimed to give a resolution of 10^{-9} rad (0.0002 in.), which corresponds to a relative vertical displacement of only 6×10^{-7} mm over its base length of 587 mm.

In the second group, there are many models of liquid or pendulum tiltmeters of reasonable price ($2000–$5000) that satisfy most needs in engineering studies. Apart from a spirit level or level vial by itself, the simplest form of tiltmeter is a base that is tens of centimeters long and leveled by centering the bubble in the vial by direct viewing or by an optical coincidence viewing of the two ends of the bubble. Direct viewing gives a resolution of 1/5 of a vial division (usually 2 mm), which typically has a sensitivity of 10–30 in. per division. Coincidence viewing increases the setting accuracy to 0.03 of the sensitivity of the vial. The discrepancy from horizontal between the two measurement points can be determined by a dial gage or micrometer that has a resolution of 0.0005 in. or 0.02 mm. Huggenberger AG claims a sensitivity of 0.3 in. over a range of ±21 ft for their clinometer with a 100 mm base and coincidence centering of the bubble in the level vial. The clinometer can be attached to 1 m bases for either horizontal or vertical measurements.

If the vial is filled with an electrolytic liquid, the centering of the bubble can be done electrically. An example is the Electrolevel (by the British Aircraft Corp.), which uses the spirit bubble principle [6] and in which the movement of the bubble is sensed by three electrodes. A tilt produces a change in differential resistivity between the electrodes that is measured by means of a Wheatstone bridge. A resolution of 0.25 in. is obtained over a total range of a few minutes of arc. Many other liquid types of tiltmeters with various ranges (up to 30°) are available from various companies. Holzhausen [7] and Egan and Holzhausen [8] discuss the application of electrolytic tiltmeters (resolution of 2 in. over a range of ±1°, manufactured by Applied Geomechanics) in the monitoring of various hydroelectric power dams in the United States.

The Rank Organization in the United Kingdom [9] makes a liquid-dampened pendulum-type electronic level, the Talyvel, which gives an accuracy of ±0.5 in. over a total range of ±8 ft. A similar transducer of the pendulum type is used in the Niveltronic tiltmeter (range of ±150 in. with an accuracy of ±0.2 in.) produced by TESA S.A. in Switzerland. Of particular popularity are servo-accelerometer tiltmeters with horizontal pendula. They offer ruggedness and durability and can operate in low temperatures. The output voltage is proportional to the sine of the angle of tilt. Schaevitz Engineering produces such a servo-accelerometer that employs a small-mass horizontal paddle (pendulum), which tries to move in the direction of tilt, due to the force of gravity. Any resultant motion is converted by position sensors to a signal input to the electronic amplifier whose current output is applied to the torque motor. This develops a torque that is equal and opposite to the original. The torque motor current produces a voltage output that is proportional to the sine of the angle of tilt.

The typical output voltage range for tiltmeters is ±5 V, which corresponds to the maximum range of tilt and readily allows for serial interfacing. The angular resolution of a tiltmeter depends on its range of tilt since a larger range would result in more angular tilt per unit voltage so a higher-resolution tiltmeter would have a smaller range of measurable tilt. Typically, the resolution is 0.02% of the range (volts) [10].

There are many factors affecting the accuracy of tilt sensing, not just the resolution of the readout. A temperature change produces dimensional changes in the mechanical components and changes in the viscosity of the liquid in electrolytic tiltmeters and of the dampening oil in pendulum-type tiltmeters.

Also, electric characteristics can alter with temperature changes. Drifts in tilt indications and fluctuations of the readout may also occur. Compensation for the effects of temperature changes can be incorporated in the construction of an instrument but at an increased cost. An alternative is to design a linear reaction by the instrument to the effects of temperature and to apply simple calibration corrections.

In less expensive models, compensation for the aforementioned sources of error is not very sophisticated, and such tiltmeters may show nonlinear output in reaction to changes in temperature and erratic drifts or other behavior that would be difficult to predict without testing. Consequently, very thorough testing and calibration are required even when the accuracy requirement is not very high [11]. Testing should investigate, at least, the linearity of output in reaction to induced tilts over the instrument's full range (±) and to changes in temperature. Some suggestions for testing and calibrating inclinometers, among other instruments, are given in Dunnicliff [1]. It is further emphasized that regular and up-to-date calibration is important in order to ensure continuity in the fidelity of the data being gathered. In most cases, the behavior being investigated changes with time and incorrect data cannot be recaptured.

Compensators for vertical circle readings in precision theodolites work on the same principle as some engineering tiltmeters. The liquid compensator of the Kern E2 electronic theodolite [12] gave a repeatability of better than 0.3 in. over a range of ±150 in. and was incorporated in their tiltmeter, NIVEL 20, in 1989. The same compensation system has been used in the currently available Leica TC2002 precision electronic total station [13]. Consequently, the theodolite may also be used as a tiltmeter, in some applications, giving the same accuracy as the Electrolevel, for example.

Tiltmeters have a wide range of applications. A series of tiltmeters, if arranged along a terrain profile in a mining area, may replace geodetic leveling in the determination of ground subsidence [11] as shown in Figure 21.1. For example, the subsidence (i.e., the variation from the previous or original profile) of point 4 (Δh_4) with respect to point 1 may be estimated from the observed changes in tilt, from a base or original position, (α_i in radians) and known distances between the points as

$$\delta h_4 = s_1 \frac{(\alpha_1 + \alpha_2)}{2} + s_2 \frac{(\alpha_2 + \alpha_3)}{2} + s_3 \frac{(\alpha_3 + \alpha_4)}{2} \tag{21.1}$$

The fidelity of this method depends on the density of tilt measurements along the profile and the continuity of the profile (a constant slope of the terrain between measurement points is assumed). Similarly, deformation profiles of tall buildings may be determined by placing a series of tiltmeters at different levels of the structure [14]. Also, changes in borehole profiles can be created in a similar manner. The absolute profile of a borehole can be generated by considering the horizontal displacement in the direction of the orientation of the inclinometer (usually controlled by guide grooves in the borehole casing) for the *i*th position, as it traverses a borehole with observation of α_i at a depth s_i. However, this would require calibration of the inclinometer to correct its output to show zero in its vertical position since the α_i are tilts from the vertical rather than angular changes from an original inclination.

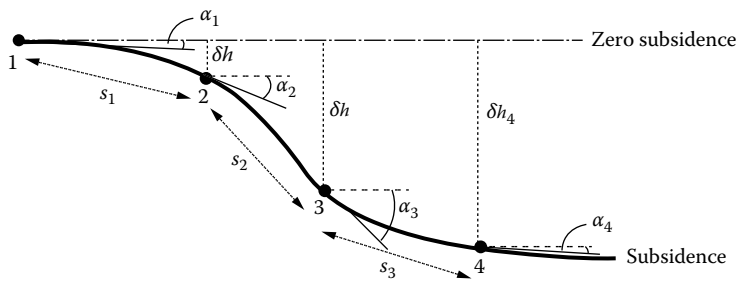


FIGURE 21.1 Ground subsidence derived from tilt measurements.

In geomechanical engineering, the most popular application of tiltmeters and borehole inclinometers is in slope stability studies and in monitoring earth-fill dams. Torpedo-shaped biaxial inclinometers are used to scan boreholes drilled to the depth of expected stable strata in the slope. By lowering the inclinometer on a cable with marked intervals and taking readings of the inclinometer at those intervals, a full profile of the borehole and its changes may be determined through repeated surveys, as mentioned earlier. Slope Indicator Co. (SINCO) and other producers of tiltmeters provide special borehole inclinometers (50 cm or 2 ft long) with guide wheels to control the orientation of the inclinometer. A special borehole casing (plastic or aluminum) with guiding grooves for the wheels is available. Usually, servo-accelerometer-type inclinometers are used with various ranges of inclination measurements, for example, ±6°, ±53°, or even ±90°. A 40 m deep borehole, if measured every 50 cm with an inclinometer having an accuracy of only ±100 in., should allow for the determination of linear lateral displacement of the collar of the borehole with an accuracy of ±2 mm.

In cases where there is difficult access to the monitored area or a need for continuous data acquisition or both, tiltmeters or borehole inclinometers can be left in place at the observing station with a telemetry monitoring system allowing for communication to the processing location. One example of a station setup of a telemetric monitoring of ground subsidence in a mining area near Sparwood, British Columbia, Canada, used a telemetry system developed for the Canadian Centre for Mining and Energy Technology (CANMET) by the University of New Brunswick [11,15]. Terra Technology biaxial servo-accelerometer tiltmeters of ±1° range were used in the study. The telemetry system could work with up to 256 field stations. Each station accepted up to six sensors (not only tiltmeters but any type of instrument with electric output, e.g., temperature, power level, or voltage). Another example is a fully automated borehole scanning system with a SINCO inclinometer and telemetric data acquisition that was also developed at the University of New Brunswick [16]. It has been used successfully in monitoring highwall stability at the Syncrude Canada Limited tar sand mining operation in northern Alberta.

21.2 Geodetic Leveling

Geodetic or differential leveling measures the height difference between two points using precision tilting levels, or precision automatic levels with compensators, with parallel plate micrometers and calibrated invar leveling staves or rods. Recent technology has provided digital automatic levels that use a CCD sensor in the instrument and bar codes, rather than linear graduations, on the staves [17]. Their ease of use and comparable precision have quickly made them rivals to the traditional optical instruments. In a single setup of the level, the height difference is the backsight rod reading minus the foresight rod reading. Any number of setup height differences can be combined to determine the height difference between two points of interest; however, the errors involved accumulate with the number of setups. With sight lengths limited to no more than 20 m, geodetic leveling can produce height differences with a precision of ±0.1 mm per setup, which is equivalent to a precision of ±0.5 in. in tilt. Although geodetic leveling is traditionally used to determine elevations, it is often used not only to monitor the settlement of sensitive structures but also to describe the tilt of components of a structure by determining the tilt between appropriate pairs of benchmarks (monumented in or on the structure) [18]. Since the level reference is created by an optical line of sight through a telescope (magnification up to 40×), a major source of systematic error is the effect of vertical atmospheric refraction. A vertical temperature gradient of 1°C m^{-1} across the line of sight would bend the line of sight to be in error by 0.4 mm at 30 m. Gradients of this magnitude are commonly encountered in industrial settings and are usually even more evident outdoors. Less effect is realized if the sight lengths are kept short, but this must be weighted against the accumulation of random error with each additional setup (shorter sight lengths would require more setups to traverse the same height difference). The errors that seem to be insignificant in a single setup (or in a few setups) become magnified in height differences involving a large number of setups (e.g., rod scale error and settlement of the instrument or rods). Such errors have become quite noticeable in the

monitoring of tectonic plate movement, and undetected systematic effects can be misleading. Further discussion on precision leveling and sources of error is available in Vanicek et al. [19] and in Schomacker and Berry [20].

Having the elevations or heights, h_1 and h_2, of two points or having measured, or determined, the height difference between them, $\Delta h_{12^t} = h_2 - h_1$, at a time t, means that, if $\delta\Delta h = \Delta h_{12^{t2}} - \Delta h_{12^{t1}}$, the tilt, T_{12}, can be calculated if the horizontal separation s_{12} is known, since $T_{12} = \delta\Delta h/s_{12}$. The separation does not have to be known as precisely as the height difference since the total random error is $\sigma_{T^2} = \sigma_{\delta\Delta h^2}/s^2 + \sigma_{s^2}(\delta\Delta h^2/s^4)$. As an example, for two points that are 60 m apart with a height difference of 0.5 m (extreme in most structural cases) with the height difference known to ±50 (μm ($\sigma_{\delta\Delta h}$) and the distance known to ±0.01 m (σ_s), the tilt would have a precision (σ_T) of ±0.3 in. Further, neither the measurement of the height difference nor the determination of the separation has to be done directly between the two points. The leveling can be done along whatever route is convenient and the separation can be obtained in a variety of ways, for example, inversing from coordinated values for the points [21].

21.3 Hydrostatic Leveling

If two connected containers (Figure 21.2) are partially filled with a liquid, then the heights h_1 and h_2 are related through the hydrostatic equation (Bernoulli's equation, as given in [22]):

$$h_1 + \frac{P_1}{(g_1\rho_1)} = h_2 + \frac{P_2}{(g_2\rho_2)} = c \tag{21.2}$$

where

P is the barometric pressure
g is the force of gravity
ρ is the density of the liquid, which is a function of temperature
c is a constant

The previously mentioned relationship has been employed in hydrostatic leveling, as shown schematically in Figure 21.3. The air tube connecting the two containers eliminates possible error due to different air pressures at two stations. The temperature of the liquid should also be maintained constant

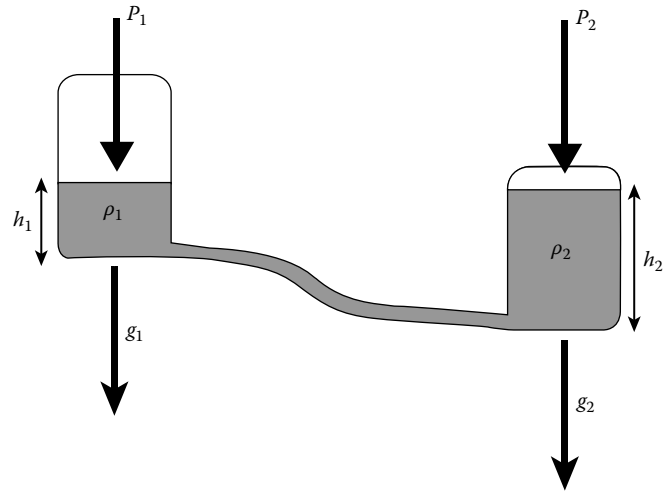


FIGURE 21.2 Hydrostatic equilibrium in connected vessels.

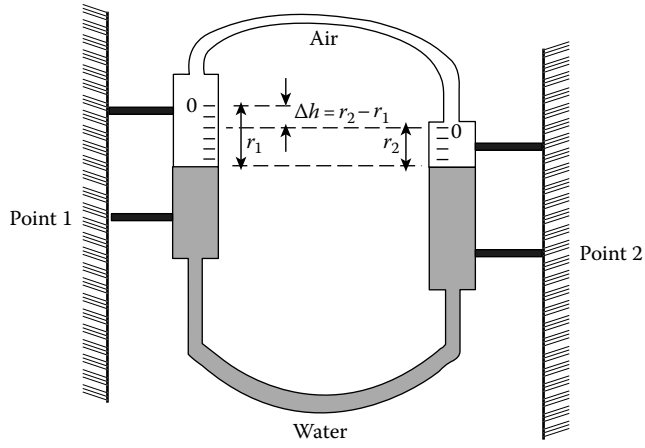


FIGURE 21.3 Hydrostatic leveling ($\Delta h_{12} = r_2 - r_1$).

because, for example, a difference of 1.2 °C between two containers may cause an error of 0.05 mm in a Δh determination for an average $h = 0.2$ m and $t = 20$ °C. Huggett et al. [23] devised a two-fluid tiltmeter to overcome the temperature effect by using two liquids with different temperature coefficients and claim a resolution of 10^{-8}–10^{-9} rad over a separation of up to 1 km. In a discussion of liquid level gages, Dunnicliff [1] emphasizes that care should be exercised to ensure that there is no discontinuity in the liquid since any gas (usually air, often entering when filling the tubing) in the liquid line will introduce an error in the level difference, especially in a vertical, more than in a horizontal, portion of tubing. He also mentions that the behavior of the liquid is influenced by the inside diameter and capillary effects of the tubing, while the outside diameter is likely what is quoted by manufacturers. Dunnicliff [1] also provides a comprehensive summary of the variety of liquid level gages.

Two examples of typical hydrostatic instruments used in precision leveling will be mentioned here. The ELWAAG 001, developed in Germany [24], is a fully automatic instrument with a traveling (by means of an electric stepping motor) sensor pin that closes the electric circuit on touching the surface of the liquid. A standard deviation of ±0.03 mm is claimed over distances of 40 m between the instruments [22]. Another automatic system, the Nivomatic Telenivelling System, is available from Telemac or Roctest Ltd. The Nivomatic uses inductance transducers that translate the up and down movements of its floats into electric signals (frequency changes in a resonant circuit). An accuracy of ±0.1 mm is claimed over a 24 m length. P & S Enterprises, Ltd. produces a Pellissier model H5 portable hydrostatic level/tiltmeter, for which they claim an accuracy of ±5 μm over a tube length of 14 m, for engineering and Earth tide measurements.

Hydrostatic levels may be used in a network formation of permanently installed instruments to monitor tilts in large structures. Robotti and Rossini [25] report on a DAG (automatic measuring device of levels and inclinations) network monitoring system available from SIS Geotecnica (Italy) that offers an accuracy of about ±0.01 mm using inductive transducers in the measurement of liquid levels. Various systems of double liquid (e.g., water and mercury) settlement gages based on the principle of hydrostatic leveling are used for monitoring power dams [26] with extended networks of connecting tubing.

Instruments with direct measurement of liquid levels are limited in their vertical range by the height of their containers. This problem may be overcome if liquid pressures are measured instead of the changes in elevation of the liquid levels. Pneumatic pressure cells or pressure transducer cells may be used. Both Dunnicliff [1] and Hanna [26] give numerous examples of various settlement gages based on that principle. Meier [27] mentions the application of a differential pressure tiltmeter in the monitoring of a concrete dam.

21.4 Suspended and Inverted Plumb Lines

Two types of mechanical plumbing are used in monitoring the stability of vertical structures:

1. Suspended pendula or plumb lines (Figure 21.4a)
2. Floating pendula or inverted, or reversed, plumb lines (Figure 21.4b)

Typical applications are in the monitoring of power dams and of the stability of reference survey pillars. Suspended pendula are also commonly used in mine orientation surveys and in monitoring the stability of mine shafts. Tilt, or inclination, is derived from differences in horizontal relative position combined with vertical separation in the same way as tilt is derived from geodetic leveling. So, similarly, the vertical separation between two reading tables or between a reading table and anchor point does not have to be known as precisely as the change in relative position. Two table readings, each ±0.02 mm, with a relative position difference of 100 mm and a vertical separation of 10 m, known to ±0.01 m, would result in a tilt precision of ±2 in.

Inverted plumb lines have become standard instrumentation in large dams (e.g., Hydro-Quebec uses them routinely). Their advantage over suspended plumb lines is in the possibility of monitoring the absolute displacements of structures with respect to deeply anchored points in the foundation rock, which may be considered as stable. In power dams, the depth of anchors must be 30 m or even more below the foundation in order to obtain absolute displacements of the surface points. The main problem with inverted plumb lines is the drilling of vertical boreholes so that the vertical wire of the plumb line would have freedom to remain straight and vertical. A special technique for drilling vertical holes has been developed at Hydro-Quebec [28].

Several types of recording devices that measure displacements of structural points with respect to vertical plumb lines are produced by different companies. The simplest are mechanical or electromechanical micrometers with which the plumb wire can be positioned with respect to reference lines of a recording (coordinating) table with an accuracy of ±0.2 mm or better. Traveling microscopes may give the same accuracy. Automatic sensing and recording is possible with a Telecoordinator from

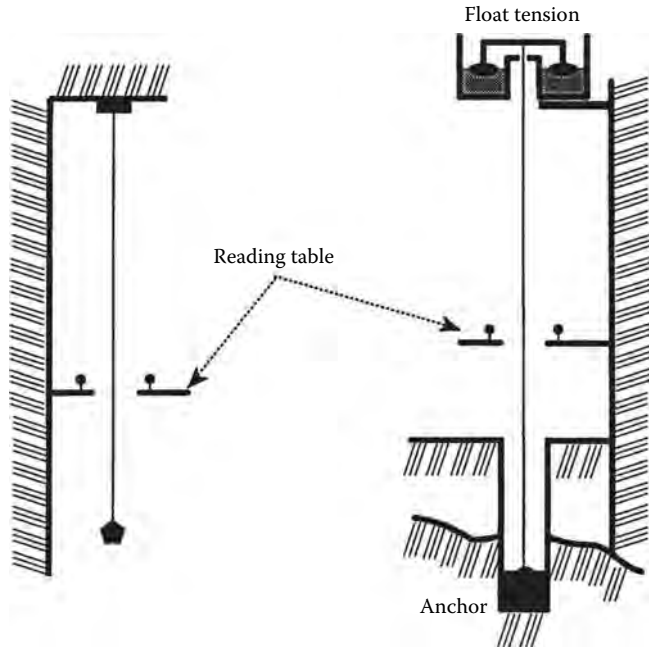


FIGURE 21.4 Inclination measurements with plumb lines. (a) Suspended pendulum, (b) Inverted pendulum.

Huggenberger AG in Switzerland. Telemac Co. (France) developed a system, Telependulum (marketed by Roctest), for continuous sensing of the position of the wire with remote reading and recording. A rigidly mounted reading table supports two pairs of induction-type proximity sensors arranged on two mutually perpendicular axes. A hollow cylinder is fixed on the pendulum wire at the appropriate level, passing through the center of the table and between the sensors. Changes in the width of the gap between the target cylinder and the sensors are detected by the corresponding changes in the induction effect. The system has a resolution of ±0.01 mm.

An interesting Automated Vision System has been developed by Spectron Engineering (Denver, Colorado). The system uses solid-state electronic cameras to image the plumb line with a resolution of about 3 μm over a range of about 75 mm. Several plumb lines at Glen Canyon dam and at Monticello dam, near Sacramento, California, have been using the system since 1982 [29].

Two sources of error, which may be often underestimated by the user, may strongly affect plumb line measurements:

1. The influence of air currents
2. The spiral shape of the wire

If the wire of a plumb line (Figure 21.5a), with pendulum mass Q, is exposed along a length h to an air current of speed v at a distance H from the anchor point, then the plumb line is deflected by an amount [30]:

$$e = f_o h \frac{H}{Q} \tag{21.3}$$

where f_o is the acting force of air current per unit length of the wire. The value of f_o may be calculated approximately from [30]:

$$f_o = 0.08 \frac{dv^2}{Q} \tag{21.4}$$

where

d is the diameter of the wire in millimeters
v is in meters per second
Q is in kilograms

As an example, if $H = 50$ m, $h = 5$ m, $d = 1$ mm, $Q = 20$ kg, and $v = 1$ m s^{-1} (only 3.6 km h^{-1}), then $e = 1$ mm.

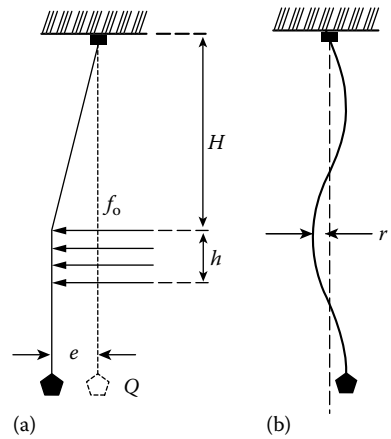


FIGURE 21.5 (a) Influence of air currents on a suspended plumb line. (b) Horizontal error due to the spiral shape of the wire.

The second source of error, which is usually underestimated in practice, is that the spiral shape (annealing) of the wire (Figure 21.5b) affects all wires unless they are specially straightened or suspended for a prolonged time (on the order of several months). If the wire changes its position (rotates) between two campaigns of measurements, then the recorded displacements could have a maximum error of 2*r*. The value of *r* can be calculated from [30]:

$$r = \frac{(\pi d^4 E)}{(64\ RQ)} \tag{21.5}$$

where

E is Young's modulus of elasticity (about 2×10^{11} Pa for steel)
R is the radius of the free spirals of the unloaded wire that, typically, is about 15 cm for wires up to 1.5 mm diameter
d and *Q* are as mentioned previously

For a plumb wire with $d = 1$ mm, $R = 15$ cm, and $Q = 196$ N (i.e., 20 kg), $r = 0.3$ mm.

If one plumb line cannot be established through all levels of a monitored structure, then a combination of suspended and inverted plumb lines may be used as long as they overlap at least at one level of the structure. At Hydro-Quebec, the drill holes of the plumb lines are also used for monitoring height changes (vertical extension) by installing tensioned invar wires [31].

21.5 Integration of Observations

The earlier discussion has considered using individual instruments. Because many investigations, using other instrumentation as well as the measurement of tilt, involve the repetition of measurements, often over a long term, the fidelity of the measurements and their being referenced to the original or initial observation are vital to the investigation. It is risky to expect consistent behavior of instrumentation, particularly in environments with dramatic variations (especially in temperature), and over a long period of time. Any conclusion relating to the behavior of a structure is only as good as the data used in the analysis of the behavior. Two ways to ensure reliability are possible. One is to make regular testing and calibration of a component of the observation regimen. The other is to analyze the observations as they are accumulated either each observable as a temporal series of repeated measurements or observations of different locations or types together in an integrated analysis. The analytical tools for integrated analyses, as well as for calibration testing and temporal series analysis, have been developed [21,32] and successfully implemented in several projects [15,18,33]. Proper calibration testing and correction, rigorous statistical analysis of trend in temporal series, and integrated analysis have proven to be valuable tools in the analysis of deformations and serve to enhance the monitoring of sensitive structures.

References

1. J. Dunnicliff, *Geotechnical Instrumentation for Monitoring Field Performance*, John Wiley & Sons, New York, 1988.
2. P. Melchior, *The Tides of the Planet Earth*, 2nd edn., Pergammon Press, Oxford, U.K., 1983.
3. W. Torge, *Geodesy*, Walter de Gruyter, New York, 1991.
4. M. van Ruymbeke, Sur un pendule horizontal équipé d'un capteur de déplacement a capacité variable, *Bull. Géodesique*, 50, 281–290, 1976.
5. G. L. Cooper and W. T. Schmars, Selected applications of a biaxial tiltmeter in the ground motion environment, *J. Spacecraft Rockets*, 11, 530–535, 1974.
6. M. A. R. Cooper, *Modern Theodolites and Levels*, Crosby Lockwood, London, U.K., 1971.
7. G. R. Holzhausen, *Low-Cost, Automated Detection of Precursors to Dam Failure: Coolidge Dam, Arizona*; *8th Annual Conference on Association of State Dam Safety Officials*, San Diego, CA, 1991.

8. N. H. Egan and G. R. Holzhausen, Evaluating structures using tilt (rotation) measurements, *Sensors Expo West Proceedings*, San Diego, CA, 1991.
9. W. Caspary and A. Geiger, Untersuchungen zur Leistungsfähigkeit elektronischer Neigungsmesser, *Schriftenreihe, Vermessungswesen HSBW*, 3, München, Germany, 1978.
10. A. Chrzanowski, Geotechnical and other non-geodetic methods in deformation measurements, in Y. Bock (ed.), *Proceedings of Deformation Measurements Workshop*, Massachusetts Institute of Technology, Boston, MA, 1986, pp. 112–153.
11. A. Chrzanowski, W. Faig, B. Kurz, and A. Makosinski, Development, installation and operation of a subsidence monitoring and telemetry system, Contract report submitted to CANMET, Calgary, Canada, November 1980.
12. Kern & Co. Ltd., E2 Instruction Manual, Kern & Co. Ltd., Aarau, Germany, 1984.
13. Leica AG, *Wild TC2002 User Manual*, Leica AG, Heerbrugg, Germany, 1993.
14. H. Kahmen, *Elektronische Messverfahren in der Geodäsie*, Herbert Wichmann, Karlsruhe, Germany, 1978.
15. A. Chrzanowski and M. Y. Fisekci, Application of tiltmeters with a remote data acquisition in monitoring mining subsidence, *Proceedings of 4th Canadian Symposium on Mine Surveying*, Banff, Alberta, Canada, June 1982.
16. A. Chrzanowski, A. Makosinski, A. Zielinski, and W. Faig, Highwall monitoring system, Contract report to Syncrude Canada Limited, Canada, April 1988.
17. H. Ingensand, Wild NA2002, NA3000, Technical paper digital levels, Leica AG, Heerbrugg, Germany, 1993.
18. A. Chrzanowski and J. M. Secord, The 1989 Integrated analysis of deformation measurements at the Mactaquac generating station, Contract Report to N.B. Power, May 1990.
19. P. Vanicek, R. O. Castle, and E. I. Balazs, Geodetic leveling and its applications, *Rev. Geophys. Space Phys*., 18(2), 505–524, 1980.
20. M. C. Schomaker and R. M. Berry, *Geodetic Levelling*, NOAA Manual NOS NGS 3, U.S. Department of Commerce, National Oceanic and Atmospheric Administration, National Ocean Survey, Rockville, MD, 1981.
21. J. M. Secord, Development of the automatic data management and the analysis of integrated deformation measurements, Ph.D. dissertation, Department of Geodesy and Geomatics Engineering Technical Report 176, University of New Brunswick, Fredericton, New Brunswick, Canada, 1995.
22. K. Schnädelbach, Neuere Verfahren zur präzisen Längen-und Höhenmessung, *Algemeine Vermessungs-Nachrichten*, 1980.
23. G. R. Huggett, L. E. Slater, and G. Pavlis, Precision leveling with a two-fluid tiltmeter, *Geophys. Res. Lett*., 3(12), 754–756, 1976.
24. H. Thierbach and W. Barth, Eine neue automatische Präzisionsschlauchwaage, *Z. Vermessungswesen*, 100, 470–478, 1976.
25. F. Robotti and T. Rossini, Analysis of differential settlements on monumental structures by means of the DAG automatic measuring device of levels and inclinations, in *Land Subsidence*, IAHA Publication No. 151, International Association of Hydrological Sciences, IAHS, Wallingford, Oxfordshire, UK, 1984.
26. T. H. Hanna, *Field Instrumentation in Geotechnical Engineering*, Trans Tech Publications, Clausthal-Zellerfeld, Germany1985.
27. E. Meier, A differential pressure tiltmeter for large-scale ground monitoring, *Water Power & Dam Constr*., 43(1), 38–40, 1991.
28. L. Dubreuil and R. Hamelin, Le forage vertical des trous de pendules inversés, *Proceedings of 2nd Canadian Symposium on Mining Surveying Rock Deformation Measurements*, Kingston, Ontario, Canada, 1974.
29. G. Kanegis, Automated vision system installed at Glen Canyon Dam, Spectron Engineering, n.d. [circa 1983].

30. A. Chrzanowski, E. Derenyi, and P. Wilson, Underground survey measurements—Research for progress, *Can. Min. Metall. Bull.*, June 1967.
31. B. Boyer and R. Hamelin, Monitoring survey: Recent developments in the use of inverted pendula, Report No. 4, Question 56, *15th Congress International Commission Large Dams,* Lausanne, Switzerland, 1985.
32. A. Chrzanowski, Y. Q. Chen, P. Romero, and J. Secord, Integration of geodetic and geotechnical deformation surveys in the geosciences, *Tectonophysics*, 130, 369–383, 1986.
33. A. Chrzanowski, Y. Q. Chen, J. Secord, and A. Szostak-Chrzanowski, Problems and solutions in the integrated monitoring and analysis of dam deformations, *CISM J. ACSGC*, 45(4), 547–560, 1991.

Further Readings

Leavitt, J., Sideris, A., and Bobrow, J. E., High bandwidth tilt measurement using low-cost sensors, *IEEE/ASME Trans. Mechatron*., 11(3), 320–327, 2006.

Miskam, A., Korakkotil, S., Zaidi, M., and Sidck, O., Development of a tilt measurement unit using micro-electromechanical system accelerometer, *J. Appl. Sci*., 9(13), 2451–2456, 2009.

Wang, J., Zhou, W., Lim, L. E. N., and Asundi, A. K., Strain and tilt measurement using multi-point diffraction strain sensor, *Opt. Laser Technol.*, 40(8), 1099–1103, 2008.

Yang, J., Zhao, Y., Yang, S. Y., Xu, Y., and Zhao, Y., Novel tilt measurement method based on self-demodulated FBG sensor and pendulum clinometer, *Proc. SPIE Int. Soc. Opt. Eng.*, 5634(1), 533–540, 2005.

Zhu, R., and Zheng, W., Tilt measurement research based on Moire fringe technique of SLR digital camera, *Appl. Mech. Mater*., 90–93, 2840–2843, 2011.

Partial List of Vendors and Suppliers

Applied Geomechanics Inc.
1336 Brommer Street, Santa Cruz, CA 95062

Eastman Whipstock GmbH
Gutenbergstrasse 3, 3005 Hannover-Westerfeld,
West Germany

Huggenberger AG
Hohlstrasse 176, CH-8040 Zürich, Switzerland

Leica AG
CH-9435 Heerbrugg, Switzerland

Maihak AG
Semperstrasse 38, D-2000 Hamburg 60, West Germany

RST Instruments Ltd.
1780 McLean Avenue, Port Coquitlam, B.C., Canada V3C 4K9

Serata Geomechanics, Inc.
4124 Lakeside Drive, Richmond, CA 94806

Soil Instruments Ltd.
Bell Lane, Uckfield, East Sussex TN22 1Q1, England

Solinst Canada Ltd.
2440 Industrial St., Burlington, Ontario, Canada L7P 1A5

Spectron Engineering
800 West 9th Avenue, Denver, CO 80204

Telemac
2 Rue Auguste Thomas, 92600 Asnieres, France

P & S Enterprises, Ltd.
240 South Monaco Pkwy, # 302, Denver, CO 80224

Auckland Nuclear Accessory Company Ltd.
P.O. Box 16066, Auckland, 3, New Zealand

Geotechnical Instruments Ltd.
Station House, Old Warwick Road, Leamington Spa,
Warwickshire CV31 3HR, England

IRAD GAGE
Etna Road, Lebanon, NH 03766

Measurement Devices Limited
11211 Richmond Avenue, Suite 106, Building B,
Houston, TX 77082

Roctest Ltée Ltd.
665 Pine, Montreal, P.Q., Canada J4P 2P4

Schaevitz Engineering
P.O. Box 505, Camden, NJ 08101

SINCO
3668 Albion Place N., Seattle, WA 98103

Solexperts AG
Postfach 274, CH-8034 Zürich, Switzerland

SIS Geotecnica
Via Roma, 15, 20090 Segrate (Mi), Italy

Spectron Glass and Electronics Inc.
595 Old Willets Path, Hauppauge, NY 11788

TerraMetrics
16027 West 5th Avenue, Golden, CO 80401

Edi Meier & Partner
8408 Winterthur, Switzerland



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

22

Proximity Sensing for Robotics

Ricardo E. Saad
Harmonics Lightwaves

Ben Benhabib
University of Toronto

A. Bonen
University of Toronto

K.C. Smith
University of Toronto

22.1 Proximity Definition **22**-1
22.2 Typical Sensor Characteristics **22**-2
22.3 Technologies for Proximity Sensing **22**-2
Electro-Optical Sensors • Phase Modulation • Amplitude Modulation • Geometrical Techniques • Time of Flight • Photothermal Effect • Capacitive Sensors • Ultrasonic Sensors • Magnetic Sensors
Acknowledgments **22**-15
References **22**-15
Further Readings **22**-16
Partial List of Vendors and Suppliers **22**-17

The objective of this chapter is to review the state of the art in proximity-sensing technologies for robotics. Special attention is paid to the sensing needs of robotic manipulators for grasping applications, in contrast to the needs of mobile robots for navigation purposes. For a review of the application of proximity sensing to mobile robots, the reader is referred to [1].

Robotic sensors can be categorized into three groups: medium-range (object recognition and gross position/orientation estimation) sensors, short-range (proximity) sensors, and contact sensors. Recent literature [2–6] suggests that robotic end effectors should be equipped with both short-range proximity and contact sensors.

Proximity sensors should be able to measure the position and orientation (pose) of an object's surface. The range must be sufficiently large to compensate for uncertainties in the medium-range pose-estimation process while maintaining sufficient accuracy to permit effective grasping of the object.

Transducers used by current proximity sensors vary in sophistication. Despite their great variety, however, these transducers and their accompanying electronic interface circuits (together comprising proximity sensor) cannot presently meet the stringent robustness requirements of most industrial robotic applications. Novel sensing algorithms and techniques still must be developed in order to improve on their current characteristics and, furthermore, to control both the sensing and grasping processes.

22.1 Proximity Definition

The term "proximity," quantified by "pose" in this chapter, refers to three geometrical parameters x, u, and v as shown in Figure 22.1, where

x is the translation from the origin of the sensor's reference coordinate frame, F_p, to a target point on the surface of the object measured along X_p. This target point defines the origin of the surface frame, F_r.

u is the *vertical* orientation of the object's surface, defined as a rotation around Y_p (of the translated frame), thereby specifying the new Z_r.

v is the *horizontal* orientation of the object's surface, defined as a rotation around Z_r, thereby specifying Y_r.

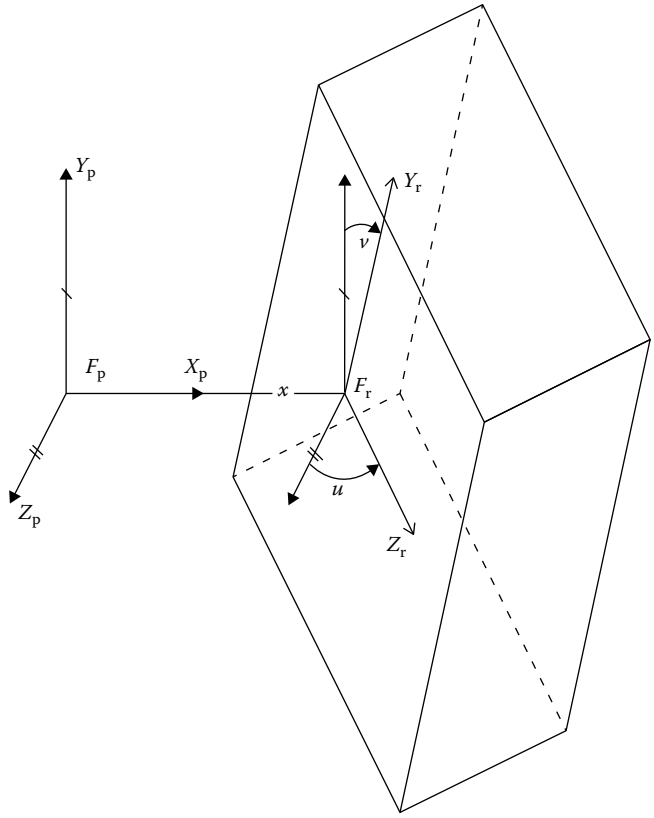


FIGURE 22.1 Proximity parameters.

22.2 Typical Sensor Characteristics

Conventionally, proximity sensors should be capable of measuring distances of up to 50 mm and two degree-of-freedom orientations equivalent to an overall inclination of up to ±30°. The intended principal application of the sensor is to act as a guide for the robot. Thus, it would be desirable to have higher sensitivity and accuracy as the gripper approaches the object, namely, when both the relative orientation and the distance approach near-zero values.

The signals received by the electronic interface circuit should be processed without limiting the required operating range of the sensor. The interface circuit should also minimize the effect of interference from the surroundings. It should therefore employ solutions to reduce background-noise interference and dynamic-range limitations.

The operation of the robot should not be slowed down by the sensor. Namely, a pose of the object should normally be estimated in 1–10 ms.

22.3 Technologies for Proximity Sensing

Proximity sensors have employed various transduction media, including sound waves, magnetic fields, electric fields, and light. Presently, electro-optical techniques seem to be the most appropriate for robotic-grasping applications. Such sensors are relatively small in size, have a large range of operation, and impose almost no restrictions on the object's material. However, recently, some new ultrasonic and capacitive proximity sensors have been fabricated directly as ICs, also showing the possibility of very-small-size proximity sensors based on these technologies [7,8].

Brief descriptions of the principles of the primary technologies used by proximity sensors are given in the following, with the main emphasis being on optical transducers. A survey of commercial proximity sensors capable of measuring distances can be found in [6].

22.3.1 Electro-Optical Sensors

Many proximity sensors use light, directly scattered from a target surface, to determine the distance and orientation of the target object from the gripper. The mechanism by which light is reflected can be explained by a model that specifies four different reflection phenomena. According to this model, light reflects from the surface primarily as a result of one or more of the following interactions:

1. *Single-surface reflection*: Light waves that reflect specularly a single time off a planar microfacet, whose dimensions are significantly larger than the wavelength
2. *Multiple-surface reflection*: Light waves that reflect specularly at least twice between multiple microfacets
3. *Reflection after penetration*: Light waves that penetrate into the material, refract, and then reflect back out as diffused light
4. *Corner reflection*: Light waves that diffract from interfaces with surface details about the same size or smaller than the wavelength (such as from corners of microfacets)

The primary phenomenon (1) usually exists in both dielectrics and metal. However, due to the high conductivity of metal surfaces, most of the light reflects specularly off the interface between the metal and the air, while the portion that penetrates into the metal surface is absorbed. Accordingly, the reflection intensity originating from internal refraction in metals is practically zero. In dielectrics, however, a large portion of the light penetrates into the surface and then reflects back out as diffused light (3). The secondary phenomena (2) and (4) exist both in metals and dielectrics and add to the diffused reflectance.

Common measurement techniques used in optical proximity sensing utilize one or more of the reflected components to determine the pose of the object in relation to the transducer.

22.3.2 Phase Modulation

A phase-modulated (PM) proximity sensor usually consists of two light sources and one or more photodetectors. The light sources are driven by modulated sinusoidal signals having a 90° phase relationship (Figure 22.2).

The emitter control voltages of the emitters, V_{em1} and V_{em2}, have amplitudes of a and b, respectively:

$$V_{em1}(t) = a \cdot \cos \omega t \tag{22.1}$$

$$V_{em2}(t) = b \cdot \sin \omega t \tag{22.2}$$

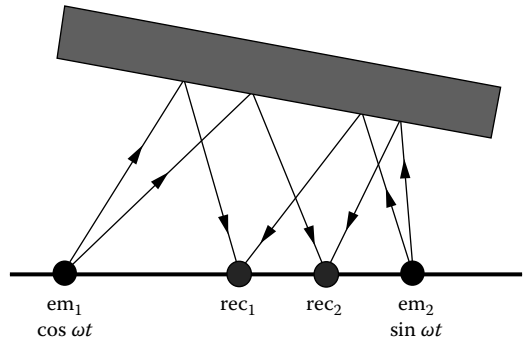


FIGURE 22.2 Basic PM proximity-sensor configuration.

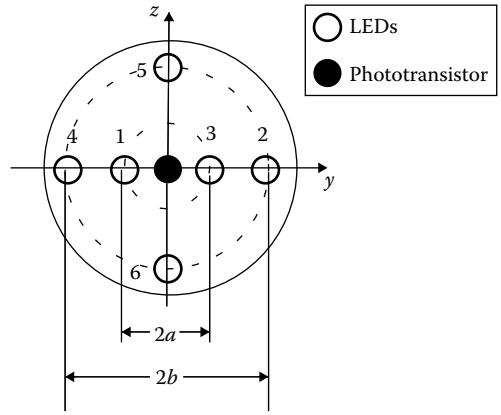


FIGURE 22.3 Sensor head reported in [9].

The signal detected by the receiver is a superposition of the two reflected signals, having corresponding attenuations of A and B:

$$V_{\text{rec}}(t) = A \cdot V_{\text{em1}} + B \cdot V_{\text{em2}} \tag{22.3}$$

The signal attenuation is a function of the geometrical and electrical parameters of the sensor, the reflectivity characteristics of the object's surface, and the surface's distance and orientation with respect to the sensor. The combined signal at the receiver is therefore

$$V_{\text{rec}}(t) = M \cdot \sin(\omega t + \phi) \tag{22.4}$$

where

M is the combined attenuation function
ϕ is the combined phase shift

Usually, only the phase information ϕ is used, and the amplitude is completely neglected or used only for verifying the likelihood of error and its potential magnitude.

A proximity sensor that uses this technique has been reported in [9]. Figure 22.3 shows a *sensor head*; it comprises six light sources (LEDs) and a photodetector (phototransistor). This sensor can measure the distance from the sensor's coordinate frame to the target point on the surface of the object (x), as well as the horizontal and vertical orientation of the object surface (u, v).

A simple model for the sensor was developed assuming that the light sources (LEDs) have low directivity, the photodetector (phototransistor) has high directivity, and the surface has diffused reflectivity. Figure 22.4 shows the basic configuration for the measurement of distance (x) and orientation (u and v). Table 22.1 shows the combinations of the driving signals in each LED needed for the measurement of distance and orientation.

For the measurement of the distance x, LED1 and LED3 are modulated by $K_1 \sin\omega t$ and LED2 and LED4 by $K_2 \cos\omega t$, respectively. The brightness detected by the photodetector can be calculated (using Lambert's law) to be

$$L_P = C\left(G_1 \frac{\cos(\alpha - v)}{a^2 + x^2}\right) + G_2 \frac{\cos(\beta - v)}{b^2 + x^2} + G_3 \frac{\cos(\alpha + v)}{a^2 + x^2} + G_4 \frac{\cos(\beta + v)}{b^2 + x^2} \tag{22.5}$$

where

G_i (i = 1, 2, 3, 4) are the intensities of the light sources
C is the reflection factor of the surface at point P

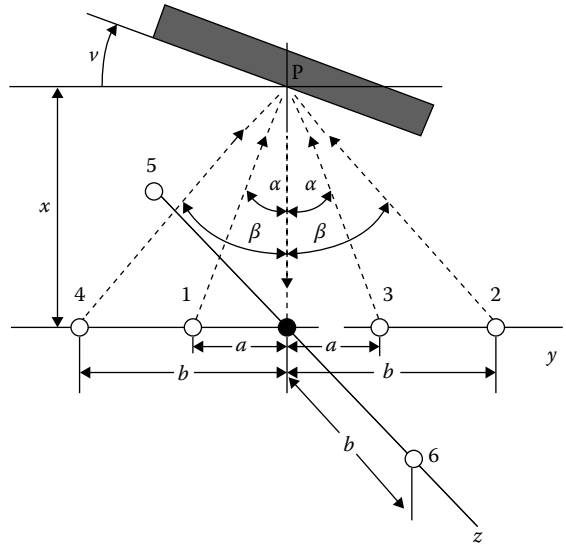


FIGURE 22.4 Measurement parameters for the sensor reported in [9].

TABLE 22.1 Combinations of the LED Driving Signals

Mode	LED1	LED2	LED3	LED4	LED5	LED6
Distance (x)	$\sin \omega t$	$\cos \omega t$	$\sin \omega t$	$\cos \omega t$	—	—
Orientation (u)	—	$\cos \omega t$	—	$\sin \omega t$	—	—
Orientation (v)	—	—	—	—	$\cos \omega t$	$\sin \omega t$

Considering that for this case $G_1 = G_3 = K_1 \sin\omega t$, $G_2 = G_4 = K_2 \cos\omega t$, $\cos\alpha = x/\sqrt{a^2 + x^2}$ and $\cos\beta = x/\sqrt{b^2 + x^2}$, Equation 22.5 can be rewritten as

$$L_P = 2Cx\left\{\frac{K_1}{(a^2 + x^2)^{3/2}}\sin\omega t + \frac{K_2}{(b^2 + x^2)^{3/2}}\cos\omega t\right\}\cos v = M\sin(\omega t + \phi_x) \tag{22.6}$$

The amplitude M and the phase shift ϕ_x are given by

$$M = 2Cx\left\{\frac{K_1^2}{(a^2 + x^2)^3} + \frac{K_2^2}{(b^2 + x^2)^3}\right\}^{1/2}\cos v \tag{22.7}$$

$$\phi_x = \tan^{-1}\left[\frac{K_2(a^2 + x^2)^{3/2}}{K_1(b^2 + x^2)^{3/2}}\right] \tag{22.8}$$

From Equation 22.8, it can be observed that the distance (x) can be obtained from ϕ_x. It is important to note that, in theory, the calculation is not affected by the reflection factor (C) of the surface.

Similarly, the orientation angles u and v can be obtained by driving the LEDs as indicated in Table 22.1. For example, v can be determined by modulating LED4 and LED2 by the signals $K_1\sin\omega t$ and $K_2\cos\omega t$, respectively. For this case, the phase shift associated with the brightness of the object at point P is given by

$$\phi_v = \tan^{-1}\left[\frac{K_2(x - b\tan v)}{K_1(x + b\tan v)}\right] \tag{22.9}$$

Note that, in order to recover v from Equation 22.9, x must be known. Accordingly, the distance x must be determined first. Correspondingly, the orientation angle (v) can be calculated from the new phase shift ϕ_v. The angle u can be calculated by modulating LED6 and LED5 with $K_1 \sin\omega t$ and $K_2 \cos\omega t$, respectively, and then determining the corresponding phase shift of the associate brightness at point P.

The pose-estimation results using this sensor were quite satisfactory and showed a good agreement between theory and experiment.

In [10], an experimental setup of a PM distance sensor, similar to the one in [9], was reported for investigating the effect of the geometrical and electronic parameters on the performance of the sensor. Optimal parameters were obtained for some targeted sensor-operation characteristics.

22.3.3 Amplitude Modulation

In amplitude-modulated (AM) sensors, the magnitude of the light reflected from a surface is utilized to determine the pose of the object.

AM transducers usually consist of one light source and several photodetectors (Figure 22.5). They were redesigned and optimized several times over the past decade to yield better measurement accuracy [11–14].

Many AM proximity sensors utilize optical fibers to illuminate and collect light from the surfaces of objects. The use of optical fibers, in a Y-guide configuration (Figure 22.6), facilitates the operation of sensitive low-noise circuitry in a shielded environment appropriately remote from the robot's electromagnetic interference sources.

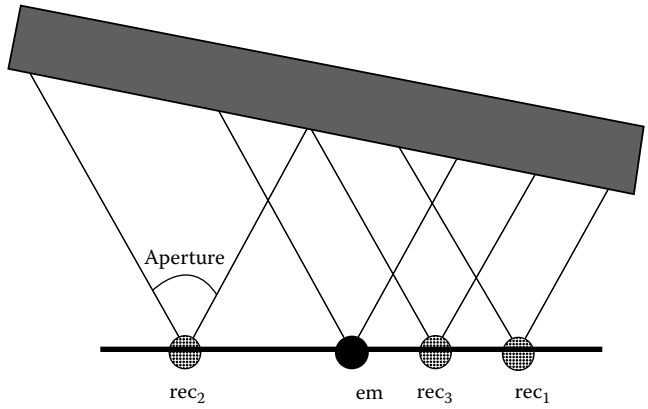


FIGURE 22.5 Bask AM proximity-sensor configuration.

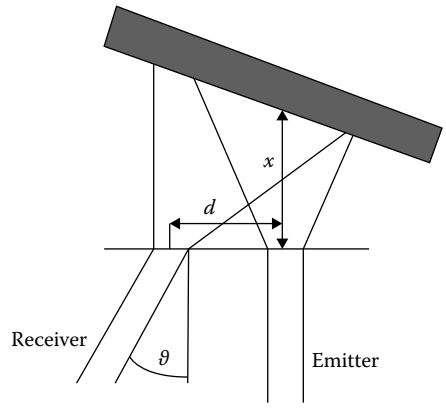


FIGURE 22.6 Y-guide transducer.

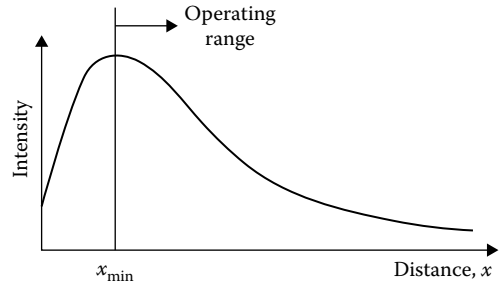


FIGURE 22.7 *Y*-guide response for distance measurement.

AM transducers primarily use variations of the basic *Y*-guide transducer. Two important parameters can be varied in the design of *Y*-guides: the distance, d, between the emitting and receiving fibers (referred to hereafter as the emitter and the receiver, respectively) and the inclination angle, ϑ, of the receiver fiber with respect to the transducer's surface. The emitter is usually placed perpendicular to the transducer's surface, due to symmetry requirements, as will be explained later in this section.

The collection of a sufficient amount of reflected light requires the use of relatively wide-diameter fibers, typically having a 0.3–2 mm core size. This requirement demands the use of relatively low-grade plastic fibers. Although attenuations of up to 1 dB m^{-1} are common in such plastic fibers, this loss rate is relatively insignificant for *Y*-guide applications because of the short length of the cables normally used. The numerical aperture (NA) of the plastic fibers, on the other hand, is an important parameter in the transducer design, since lenses are rarely used in conjunction with AM-type transducers. In particular, the acceptance angle of the fiber is fixed and given by $\alpha = 2\sin^{-1}$ NA.

For a *Y*-guide, the intensity of the light reflected from the surface is not monotonic function of the distance. Thus, the minimum operating distance of the transducer (x_{min}) is usually limited to a value that will guarantee monotonic response (Figure 22.7).

For the measurement of surface orientation, a symmetrical three-fiber constellation (Figure 22.8) can be used. In this *Y*-guide configuration, the emitter is at the center and the two receivers are positioned symmetrically on either side [12]. The light intensities detected by the receivers, for the transducer shown in Figure 22.8, are illustrated in Figure 22.9 as a function of the surface orientation.

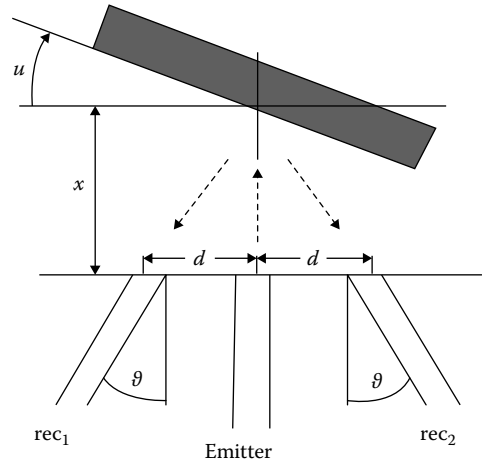


FIGURE 22.8 Typical receiver-pair constellation for orientation measurements.

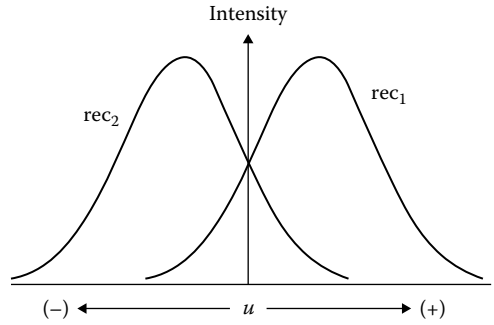


FIGURE 22.9 Light intensity detected by each receiver as a function of the surface orientation (u).

In the usual operating range of an AM transducer, the intensity of the light at the receiver is inversely related to the distance squared. As a result, it is conceptually possible to configure a transducer such that its sensitivity and accuracy will increase as the gripper nears the contact point, at which both the distance and the orientation of the object's surface are zero [10,12]. However, in practice, because of the limited dynamic range of the electronic transducer interface, a trade-off exists between the desired maximum accuracy near contact and the maximum range of operation. These and other considerations must be taken into account when establishing the geometrical features of the transducer.

Another important factor to take into account in the design of an AM sensor is the need to reduce, as much as possible, the effect of the variation in the emitting power of the light source, P_0, on the transducer's measurements. This normally leads to the employment of a pair of receivers. A normalized differential voltage (DV) estimation scheme, such as the following, is then applied to the pair of measurements:

$$\mathrm{DV} = \frac{V_{\mathrm{rec1}} - V_{\mathrm{rec2}}}{V_{\mathrm{rec1}} + V_{\mathrm{rec2}}} \tag{22.10}$$

where V_{rec1}, V_{rec2} are the voltages measured by receivers 1 and 2. However, in order to eliminate the effect of P_0 on DV, each receiver must linearly convert the light intensity to a corresponding voltage measurement.

In order to use a DV scheme for the measurement of distance, an asymmetrical transducer configuration can be used (Figure 22.10). However, one must note that, although orientation measurements are not affected by variations in distance, distance measurements are significantly affected by the orientation of the surface, for example [15].

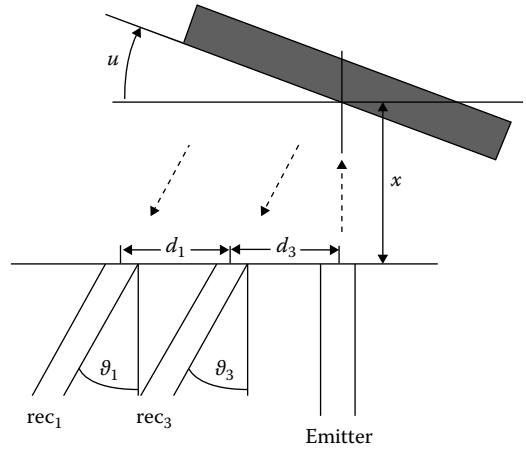


FIGURE 22.10 Asymmetrical receiver-pair constellation for distance measurements.

Accordingly, in using an AM proximity sensor with a DV scheme, the orientation is first approximated, and subsequently the distance is determined. The accuracies of the measured distance and orientation angle can be further improved by an iterative process.

Based on the earlier issues and observations, the outputs of the three receivers of the basic AM proximity sensor (Figure 22.5) can be paired for measuring both distance and orientation: the pair rec_1–rec_2 can be used for orientation measurement, while the pair rec_1–rec_3 can be used for distance measurement.

An experimental AM proximity sensor, capable of estimating the pose of an object with high accuracy, was reported in [14,16,17]. The transducer consists of one emitter, placed perpendicularly to the sensor head, and eight inclined receiver elements (Figure 22.11). The receivers of this transducer were paired for the specific measurements of distance, as well as of the vertical or horizontal orientation. However, the pose of the surface was determined with higher accuracy by using a polynomial fit technique (as opposed to the DV scheme described earlier) that provided relationships between the individual estimated parameters (x, u, and v) and all eight signals received.

The sensor presented in [14], and shown in Figure 22.11, operates in the range of 0–50 mm and ±20°. It can achieve an accuracy of 6.25 μm in distance and an accuracy of 0.02° in angular measurements in the near-contact region (0–6 mm range), using a general calibration-per-group strategy for different material groupings. This implies that the measured object's material belongs to a calibration group, which includes similar object surface characteristics, for example, machined metals. Better accuracies can be achieved using a calibration-per-surface strategy.

A similar configuration to the one shown in Figure 22.11 was reported earlier in [18] for the measurement of distances, where the orientations of each receiver pair relative to the emitter ϑ_1 and ϑ_3 were set at 10°. However, in this case, the apertures of the emitter and receiver were severely restricted by a collimating graded index (GRIN) lens. The emitter diameter was larger than that of the receivers in order

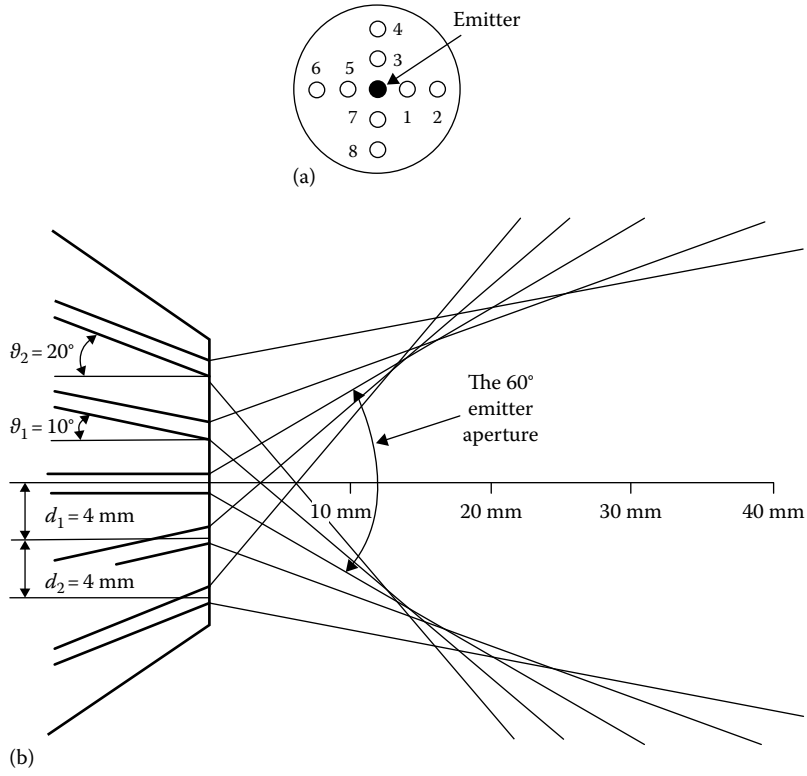


FIGURE 22.11 AM transducer design for the sensor reported in [14]: (a) top view, (b) front view.

to transmit more light. The measurements of the transducer were then processed in two phases: (1) The DVs of all the receiver pairs were processed independently to provide four distance estimations, and (2) the four distance estimations were then averaged to provide a more accurate estimate, eliminating adverse effects due to variations in surface orientation.

22.3.4 Geometrical Techniques

Proximity sensors based on geometrical techniques determine the pose of the object by examining the geometrical attributes of the reflected and incident light beams. Two of these techniques, one based on triangulation and the other based on the Gaussian lens law, are presented here.

Figure 22.12 shows the basic configuration of a proximity transducer for measuring distance (x) based on the triangulation technique [19,20]. The sensor head consists of a laser light source and a linear array of photodetectors (R_i with $i = 1, 2, \ldots, n$). A narrow light beam illuminates point P, and the receivers detect the reflected light from the illuminated point through a transmitting slit. The geometry of the ray trajectory provides the basic information for the estimation of the distance (x). While the light source illuminates the surface of the object, the photodetector array is scanned to detect the light path used for making the output signal maximum. The light path obtained by this scanning is called the effective light path [19]. This light path is the one indicated in Figure 22.12. The distance (x) can be determined by accurately detecting the position (y_i) and precisely measuring the dimensions (h) and (y_o):

$$\frac{x}{x+h} = \frac{y_o}{y_i} \tag{22.11}$$

or

$$x = \frac{y_o h}{y_i - y_o} \tag{22.12}$$

In [26], it is claimed that such a sensor has the following properties: (1) The influence of irregularities, reflectivity, and orientation of the object is negligible; (2) the distance measurement is not affected by illumination from the environment and luminance of the object (their influence is eliminated by comparison of two sensor signals obtained in successive on-and-off states of the light source); and (3) the sensor head is sufficiently small to be used in a robot hand.

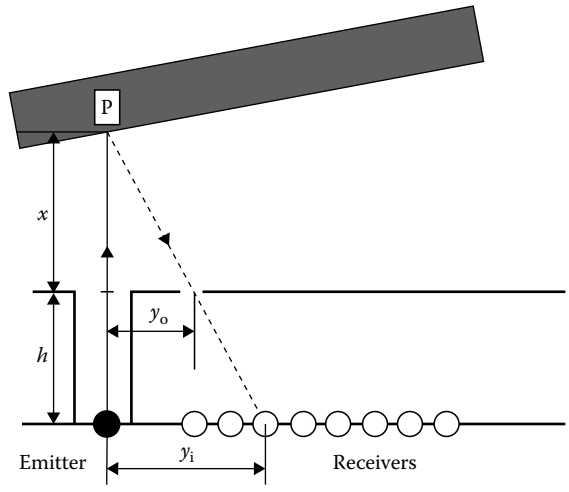


FIGURE 22.12 Basic principle of a proximity sensor for measuring distance based on triangulation.

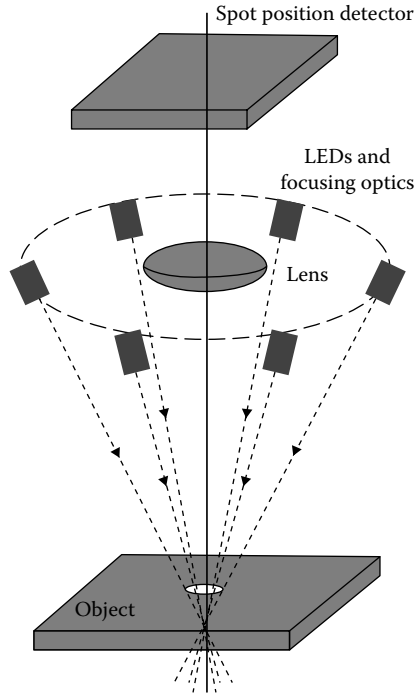


FIGURE 22.13 Multilight source proximity sensor reported in [21].

An experimental proximity-sensor configuration, based on triangulation and capable of measuring both distance and orientation, is shown in Figure 22.13 [21]. The sensor uses six infrared LEDs as light sources, an objective lens, and an area-array detector chip for detecting spot positions. The directions of the beams are aligned to form a fixed cone of light. This sensor is of a type called scanning emitter since each LED is sequentially pulsed to perform the measurements. As each LED is sequentially pulsed, the sensor IC detects the position of the spot projected by the reflected light beam from the object's surface. Knowledge of the spot's position, together with the camera's optics and the trajectory of each light beam, can be used to perform a triangulation to determine the 3D coordinates of each light spot on the target surface. A set of six 3D points are obtained for the six LEDs. Then, by fitting a plane to those points, the distance and orientation of the object's surface are approximated.

Another scanning-emitter-type proximity sensor was reported in [22]. In this case, a mechanical scanning system was utilized. One notes that inherent problems with sensors that use mechanical scanning devices include lower reliability and increased overall size.

Some recently reported triangulation sensors are sufficiently small in size to be mounted on a gripper [5]. However, they are still susceptible to errors due to distortion and separation of the light beam's reflection caused by surface irregularity and can also have blind spots as a result of discontinuities associated with the shape of the sensed object.

Another group of geometrical electro-optical proximity sensors are those based on the Gaussian lens law [23,24]. The basic configuration of such a transducer is shown in Figure 22.14. A light beam, collinear to the optical axis, forms a spot on the target's surface. The light scattered from the spot is collected by the lens. In Figure 22.14, PN represents the limiting ray that can be collected by the lens. The target distance (x) can be calculated in terms of the focal length of the lens (f) and the image position (w). Applying the Gaussian lens law, the distance x can be calculated as

$$x = \frac{fw}{w - f} \tag{22.13}$$

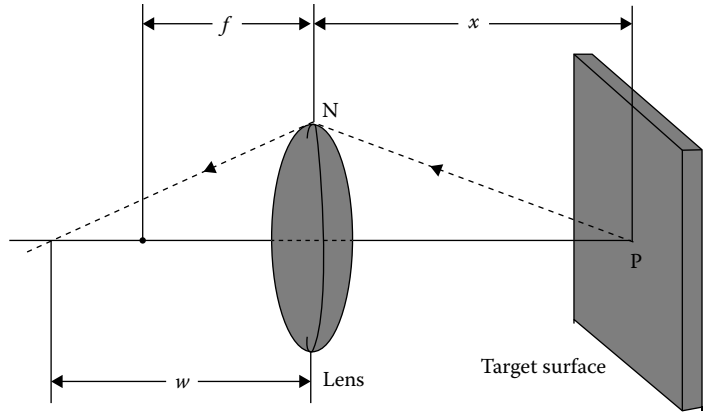


FIGURE 22.14 The principle of proximity sensing based on the Gaussian lens law reported in [7].

A proximity sensor (the HexEye) based on this lens principle is shown in Figure 22.15 [23,24]. The sensor consists of seven identical "sensing units" arranged in a hexagonal pattern. Each sensing unit in turn comprises four main parts: an objective lens, a conical mirror, six linear receiver arrays, and a laser diode light source.

The light beam generated by a light source forms a spot on the target surface, and the light flux scattered from the spot is collected by the objective lens and projected onto the receiver arrays in each sensing unit. The target distance is determined by the position, size, and shape of an image formed on individual receiver arrays of the unit.

The sensor operates in two modes, either distance or orientation. In the distance measurement mode, the seven light sources are activated one at a time to generate a light spot on the target surface, while all the sensing units receive the light flux scattered from the same light spot. The active unit determines the distance using the principle of the Gaussian lens law, while the nonactive units determine the distance based on triangulation. For the orientation measurement mode, the seven sensing units are grouped into orientation measurement units (OMUs). Each OMU comprises three neighboring sensing units around the center, resulting in a total of six possible OMUs. The local orientation of the surface is estimated by integrating the six orientation measurements.

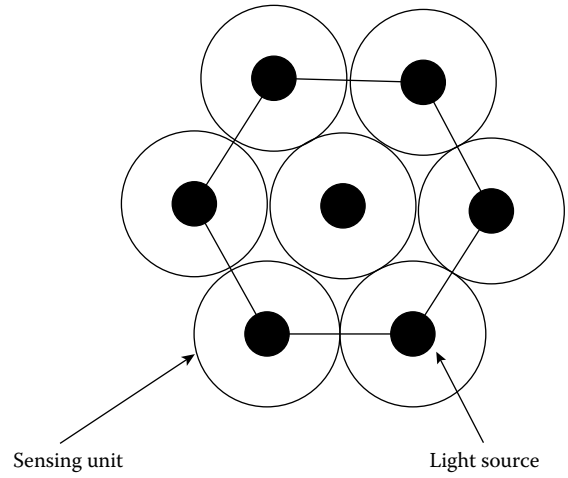


FIGURE 22.15 Configuration of HexEye (top view).

The mapping between the light distribution on the arrays of light detectors and the target distance is obtained through calibration. Using the light distribution (instead of light intensity as in the case of AM sensors), it is intended that the mapping be independent of surface properties such as color, material, diffusion, and reflectance factors. However, it has been shown that the mapping may be corrupted by "noise" from several sources, including reflection patterns and ambient light.

Using the Gaussian lens law has the following advantages over the triangulation principle: (1) The light source can be located at the center of the objective lens, allowing not only the sensor to be compact but also the amount of light flux input to the lens to be maximized, and (2) the sensitivity can be optimized for a certain range of measurement distance by controlling *f*.

22.3.5 Time of Flight

Time-of-flight-measuring electro-optical sensors are radar-type systems. However, unlike regular radar, which transmits a pulse of radio-frequency energy, these sensors normally use a modulated light beam. The distance to the target is extracted from the measured phase shift of the reflected light. Two problems associated with such sensors are difficulty in measuring short distances (which requires a very high modulation frequency) and the need for a mechanical scanning/switching system to get additional information (such as orientation) [5,25].

22.3.6 Photothermal Effect

The photothermal effect transducer uses a strong light beam directed toward the object's surface. The distance to the object is extracted from measurements of the thermal wave generated by the light absorbed by the object. The detection scheme and signal processing are similar to those used in an AM sensor. Since the shape of the thermal wave generated at the surface is surface texture independent, the photothermal sensor does not suffer from the surface robustness problem associated with AM sensors. However, the photothermal sensor is rather slow and limited to highly absorbing surfaces [26].

22.3.7 Capacitive Sensors

Capacitive sensors generate and measure changes in an electric field caused by either a dielectric or conducting object in their proximity.

There are basically two types of capacitive proximity sensor. One type uses the principle of a parallel plate capacitors, and the other uses the principle of fringing capacitances [8,27,28]. For the parallel plate-type proximity sensor, the transducer forms one plate and the object measured forms the other plate. The structure of a parallel plate-type proximity sensor and its typical response are shown in Figure 22.16 [8].

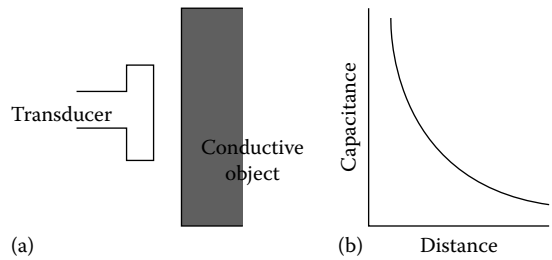


FIGURE 22.16 Capacitive proximity sensor based on the principle of parallel plates: (a) structure and (b) sensor response.

The parallel plate-type proximity sensor is widely applied in industry. However, this type of a sensor has three major limitations: (1) The object being measured must be conductive, (2) the inverse gap–capacitance relationship is highly nonlinear, and (3) the sensitivity drops significantly in the case of large gaps.

The second type of capacitive proximity sensor uses the principle of fringing capacitance [8]. The sensor has two "live" electrodes and the object being measured does not need to be part of the sensor system. The target object could be either conductive or nonconductive. However, the measurement of distances is affected by the type of object material. Therefore, separate calibrations must be carried out for different materials.

In [29], an innovative capacitive microsensor was presented. Using micromachining technology, the electrode thickness can be significantly reduced and the fringing effect increased when compared with other capacitive sensors. Consequently, this sensor yields a better sensitivity. An array of such transducers can be implemented to measure the distance and orientation of an object.

Proximity capacitive sensors have the following general advantages: (1) low energy consumption and (2) simple structure. The major disadvantages, however, are that they are influenced by external signals and a calibration-per-surface technique must be carried out, since their operation directly depends on the object's material.

22.3.8 Ultrasonic Sensors

The basic principle underlying ultrasonic ranging sensors is the measurement of the time required for a sound wave to travel from the emitter to the object's surface and return to the detector. By using several such emitters and detectors, one can obtain information about the distance and orientation of the surface.

In [19], a novel method is proposed to measure the orientation angles of an object's surface using the phase differences of reflected echoes. Figure 22.17a shows the configuration of a planar sensor head with n receivers, which are equally spaced and located on a circle of radius R around the transmitter T. A linear relationship exists between the difference in lengths, δ, of two reflecting paths for an adjacent pair of receivers and the object inclinations.

In Figure 22.17b, the relationship between the inclination of the target surface and δ is shown. It can be observed that the measuring range of the sensor can be enlarged with an increase in the number of receivers. In [30], it is also shown that the measuring range can be enlarged by reducing R. However, it was noted that measurements carried out with a small sensor are potentially less accurate.

Experimental results using a transducer with six (R = 30 mm) and eight (R = 20 mm) receivers were reported in [30]. With the six-receiver transducer, the measuring range of the orientation angles was ±15°,

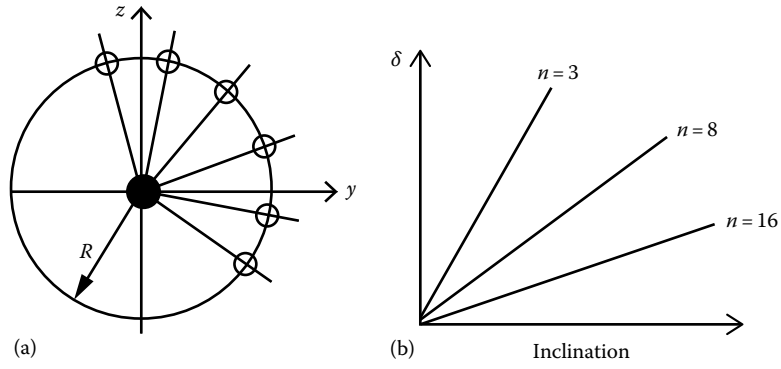


FIGURE 22.17 (a) Configuration of the ultrasonic sensor and (b) surface inclination versus δ.

while for the eight-receiver transducer, the maximum measuring range was ±30°. With the six-receiver transducer, the orientation angle could be determined with an accuracy of 0.5°, in the measuring range of ±15°, and 0.2° when the range was restricted to ±5°.

One of the major disadvantages of ultrasonic proximity sensors is that they are relatively large in size. However, implementing these sensors using micromachining could solve this problem. In [7], the generation and detection of ultrasound, for proximity sensing, were investigated using micromachined resonant membrane structures.

22.3.9 Magnetic Sensors

A magnetic-type sensor creates an alternating magnetic field, whose variation provides information about the object's position.

The simplest magnetic sensors are reed microswitches or Hall effect switches. However, the most commonly used sensors in robotics are based on the electromagnetic inductive principle, emphasizing eddy current generation. The basic principle consists of creating a magnetic field using appropriate coils around a core with high permeability and an oscillator with a frequency excitation high enough to minimize the penetration of the field inside a conductive material. The main problems with magnetic sensors are their high size/range ratio and difficulty in providing reliable distance measurements in varying magnetic environments.

Acknowledgments

The authors would like to thank Martin Bonert for careful review and critique of this chapter. They also acknowledge the financial support of the Natural Sciences and Engineering Research Council of Canada.

References

1. H. R. Everett, *Sensors for Mobile Robots: Theory and Application*, Natick, MA: A. K. Peters, Ltd., 1995.
2. B. Espiau, An overview of local environment sensing in robotics applications, *Sensors and Sensory Systems for Advanced Robots, NATO ASI Series*, F43, 125–151, 1988.
3. W. D. Koenigsberg, Noncontact distance sensor technology, *SPIE, Intelligent Robots*, 449, 519–531, 1988.
4. Å. Wernersson, B. Boberg, B. Nilsson, J. Nygårds, and T. Rydberg, On sensor feedback for gripping an object within prescribed posture tolerances, *Proceedings of IEEE, International Conference on Robotics and Automation*, Nice, France, 1992, pp. 1654–1660.
5. A. Bradshaw, Sensors for mobile robots, *Measurement and Control*, 23(2), 48–52, 1990.
6. R. Volpe and R. Ivlev, A survey and experimental evaluation of proximity sensors for space robotics, *Proceedings of IEEE International Conference on Robotics and Automation*, Vol. 4, San Diego, CA, 1994, pp. 3466–3473.
7. O. Brand, H. Baltes, and U. Baldenweg, Ultrasound-transducer using membrane resonators realized with bipolar IC technology, *Proceedings of IEEE Conference on Micro Electro Mechanical Systems*, Oiso, Japan, 1994, pp. 33–38.
8. R. C. Luo and Z. Chen, Modeling and implementation of an innovative micro proximity sensor using micromachining technology, *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, Yokohama, Japan, 1993, pp. 1709–1716.
9. R. Masuda, Multifunctional optical proximity sensor using phase modulation, *Journal of Robotic Systems*, 3(2), 137–147, 1986.
10. O. Partaatmadja, B. Benhabib, and A. A. Goldenberg, Analysis and design of a robotic distance sensor, *Journal of Robotic Systems*, 10, 427–445, 1993.

11. O. Partaatmadja, B. Benhabib, A. Sun, and A. A. Goldenberg, An electrooptical orientation sensor for robotics, *Proceedings of IEEE Transactions on Robotics and Automation*, 8, 111–119, 1992.
12. O. Partaatmadja, B. Benhabib, E. Kaizerman, and M. Q. Dai, A two-dimensional orientation sensor, *Journal of Robotic Systems*, 9, 365–383, 1992.
13. P. P. L. Regtien, Accurate optical proximity detector, *Proceedings of IEEE Conference on Instrumentation and Measurement Technology*, San Jose, CA, 1990, pp. 141–143.
14. A. Bonen, R. E. Saad, K. C. Smith, and B. Benhabib, Active-sensing via a novel robotic proximity sensor, *Proceedings of International Conference on Recent Advances in Mechatronics (ICRAM'95)*, Istanbul, Turkey, 1995, pp. 1053–1058.
15. Y. F. Li, Characteristics and signal processing of a proximity sensor, *Robotica*, 12, 335–341, 1994.
16. A. Bonen, R. E. Saad, K. C. Smith, and B. Benhabib, A novel calibration technique for electro-optical proximity sensors, *Proceedings of International Conference on Industrial Electronics, Control and Instrumentation (IECON'95)*, Orlando, FL, 1995, pp. 1190–1195.
17. A. Bonen, R. E. Saad, K. C. Smith, and B. Benhabib, A novel optoelectronic interface-circuit design for sensing applications, *IEEE Transactions on Instrumentation and Measurement*, 45, 580–584, 1996.
18. H. Bukow, Fiber optic distance sensor for robotic applications, *Proceedings of SME Conference, Sensors*, Detroit, MI, 1986, pp. MS86–MS938.
19. T. Okada, Development of an optical distance sensor for robots, *International Journal of Robotics Research*, 1, 3–14, 1982.
20. M. A. Kujoory, Real-time range and elevation finder, *Proceedings of IEEE*, 72(12), 1821–1822, 1984.
21. M. Fuhrman and T. Kanade, Optical proximity sensor using multiple cones of light for measuring surface shape, *Optical Engineering*, 23, 546–553, 1984.
22. T. Okada and U. Rembold, Proximity sensor using a spiral-shaped light-emitting mechanism, *IEEE Transactions on Robotics and Automation*, 7, 798–805, 1991.
23. S. Lee, Distributed optical proximity sensor system: HexEYE, *Proceedings of IEEE International Conference on Robotics and Automation*, Vol. 2, Nice, France, 1992, pp. 1567–1572.
24. S. Lee and J. Desai, Implementation and evaluation of HexEye: A distributed optical proximity sensor system, *Proceedings of IEEE International Conference on Robotics and Automation*, Vol. 3, Nagoya, Aichi, Japan, 1995, pp. 2353–2360.
25. S. Shinohara et al., Compact and high precision range finder with wide dynamic range using one sensor head, *Proceedings of IEEE Conference on Instrumentation and Measurement Technology,* Atlanta, GA, 1991, pp. 126–130.
26. M. Ito, K. Hane, F. Matsuda, and T. Goto, Proximity sensing technique using the photothermal effect, *Journal of the Japan Society of Precision Engineering*, 58, 139–144, 1992.
27. B. E. Noltingk, A novel proximity gauge, *Journal of Scientific Instruments, Series 2*, 2, 356–360, 1969.
28. B. E. Noltingk, A. E. T. Nye, and H. J. Turner, Theory and application of a proximity gauge using fringing capacitance, *Proceedings of ACTA IMEKO*, London, UK, 29, 537–549, 1976.
29. A. Moldoveanu, Inductive proximity sensors, fundamentals and standards, *Sensors*, 10(6), 11–14, 1993.
30. S. Nakajima and Y. Takahashi, An ultrasonic orientation sensor with distributed receivers, *Advanced Robotics*, 4, 151–168, 1990.

Further Readings

Everitt, B. S., Landau, S., Leese, M., and Stahl, D., *Measurement of Proximity*, Wiley Online library, http://onlinelibrary.wiley.com/doi/10.1002/9780470977811.ch3/summary, Published Online: January 25, 2011.

Hornung, M. R. and Brand, O., *Micromachined Ultrasound-Based Proximity Sensor*, Kluwer Academic Publisher, Boston, MA, 1999.

Mechanical Engineering Design Guide—Engineer's Handbook, *Sensors: Proximity, Mechanical Components*, http://engineershandbook.com/Components/proximitysensors.htm (accessed on June 24, 2012).

Partial List of Vendors and Suppliers

Autonics USA, 1353 Armour Blvd., Mundelein, IL 60060, Phone: 847-680-8160, 866-862-2602 (toll free), Fax: 847-680-8155, www.autonicsonline.com

Gems Sensors and Control, 1 Cowles Rd., Plainville, CT 06062-1198, Phone: 860-747-3000, 800-378-1600 (toll free), Fax: 860-747-4244, www.gemssensors.com

IFM Efector, Inc., 782 Springdale Dr., Exton, PA 19341-2843, Phone: 610-524-2000, 800-441-8246 (toll free), Fax: 800-329-0436, www.ifm.com/ifmus/web/home.htm

Kinequip. Inc., 365 Old Niagara Falls Blvd., Buffalo, NY 14228, Phone: 716-694-5000, 800-982-8894 (toll free), Fax: 716-694-5097, www.kinequip.com

Madison Co., 27 Business Park Dr., Branford, CT 06405, Phone: 203-488-4477, 800-466-5383 (toll free), Fax: 203-481-5036, madisonco.com/

Measurement Specialties, Inc., 1000 Lucas Way, Hampton, VA 23666, Phone: 757-766-1500, 800-745-8008 (toll free), Fax: 757-766-4297, www.meas-spec.com

Micro-Epsilon America, 8120 Brownleigh Dr., Raleigh, NC 27617, Phone: 919-787-9707, Fax: 919-787-9706, www.micro-epsilon.us

Schneider Electric, 1875 Founders Dr., Dayton, OH 45420, Phone: 937-514-7207, Fax: 937-258-5830, www.tesensors.com

Smith Systems, Inc., 6 Mill Creek Center, P.O. Box 667, Brevard, NC 28712, Phone: 828-884-3490, Fax: 828-877-3100, www.smith-systems-inc.com



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

III

Displacement

23 **Resistive Displacement Sensors** *Keith Antonelli, James Ko, and Shyan Ku* **23**-1
Precision Potentiometers • Measurement Techniques • Implementation • Costs and Sources • Evaluation • References

24 **Inductive Displacement Sensors** *Halit Eren* **24**-1
Theory • Linear and Rotary Variable-Reluctance Transducer • Microsyn • Synchros • Moving-Coil Inductive Displacement Sensors • Linear-Variable-Differential Transformer • Rotary-Variable-Differential Transformer • Eddy Current Proximity Sensors • Gradiometer-Inductive Displacement Sensors • Miniature-Inductive Displacement Sensors • Shielding and Sensitivity of Inductive Sensors to Electromagnetic Interference • Bibliography • Partial List of Manufacturers and Suppliers

25 **Capacitive Sensors: Displacement, Humidity, Force** *Halit Eren* **25**-1
Introduction • Variable Distance Displacement Sensors • Variable Area Displacement Sensors • Variable Dielectric Displacement Sensors • Differential Capacitive Displacement Sensors • Integrated Circuit Smart Capacitive Displacement Sensors • Capacitive Pressure Sensors • Capacitive Accelerometers and Force Transducers • Capacitive Liquid-Level Measurement • Capacitive Humidity and Moisture Sensors • Signal Processing • Bibliography • Partial List of Manufacturers and Suppliers

26 **Piezoelectric Sensors and Transducers** *Ahmad Safari, Victor F. Janas, Amit Bandyopadhyay, and Andrei Kholkine* **26**-1
Governing Equations and Coefficients • Piezoelectric Materials • Measurements of Piezoelectric Effect • Applications • Defining Terms • References • Further Readings

27 **Laser Interferometer Displacement Sensor** *Bernhard Günther Zagar* **27**-1
Introduction • Helium–Neon Laser • Refractive Index of Air • Michelson Interferometer • Two-Frequency Heterodyne Interferometer • Single-Mode Homodyne Interferometer • Displacement Sensor Based on a Fabry–Pérot Interferometer • Interferometer Signal Processing • Typical Interferometer Applications • Conclusions • Appendix • Defining Terms • References

28 **Bore-Gaging Displacement Sensors** *Viktor P. Astakhov* **28**-1
Introduction • Bore Tolerancing • Components of Gage Accuracy • Bore-Gage Classification and Specification • Bore-Gage Types • Form-Measuring Instruments • Gage Repeatability and Reproducibility • References • Further Information

29 **Ultrasonic Displacement Sensors** *Nils Karlsson and Ole Pedersen* **29**-1
Physical Characteristics of Sound Waves • Ultrasound Transducer • Principles of Time-of-Flight Systems • References • Partial List of Vendors and Suppliers

30 Optical Encoder Displacement Sensors *J.R. René Mayer* **30**-1
Encoder Signals and Processing Circuitry • Incremental Encoder Quadrature Signals • Encoding Principles • Components and Technology • References • Further Readings

31 Magnetic Displacement Sensors *David S. Nyce* **31**-1
Introduction • Magnetic Field Terminology: Defining Terms • Magnetostrictive Sensors • Hall Effect Sensors • Magnetic Encoders • Conclusion • References • Further Information

32 Synchro/Resolver Displacement Sensors *Robert M. Hyatt, Jr. and David Dayton* **32**-1
Induction Potentiometers • Resolvers • Operating Parameters and Specifications for Resolvers • Equipment Needed for Testing Resolvers • Resolver Benefits • Synchros • Modular Solution • Sensible Design Alternative for Shaft Angle Encoding • Type II Servo Loop • Applications • Further Readings

33 Optical Fiber Displacement Sensors *Richard O. Claus, Vikram Bhatia, and Anbo Wang* **33**-1
Extrinsic Fabry–Perot Interferometric Sensor • Intrinsic Fabry–Perot Interferometric Sensor • Fiber Bragg-Grating Sensor • Long-Period Grating Sensor • Comparison of Sensing Schemes • Conclusion • References • Further Readings

34 Optical Beam Deflection Sensors *Grover C. Wetsel* **34**-1
Theory • Characterization of PSDs • Summary • References • Further Readings

35 Velocity Measurement *Charles P. Pinney and William E. Baker* **35**-1
Introduction • Measurement of Linear Velocity • Reference-Based Linear Velocity Measurement • Light Interference Methods for Linear Velocity Measurement • Seismic Devices for Linear Velocity • Angular Velocity Measurement • Relative Angular Sensors • Absolute Angular Velocity Sensors • Conclusion • References • Further Readings • Partial List of Vendors and Suppliers

23

Resistive Displacement Sensors

Keith Antonelli
Kinetic Sciences, Inc.

James Ko
Kinetic Sciences, Inc.

Shyan Ku
Kinetic Sciences, Inc.

23.1 Precision Potentiometers **23**-1
Types of Precision Potentiometers • Resistive Element • Electrical Characteristics • Mechanical Characteristics • Mechanical Mounting Methods
23.2 Measurement Techniques **23**-9
23.3 Implementation **23**-11
Coupling to the Pot
23.4 Costs and Sources **23**-12
23.5 Evaluation **23**-12
References **23**-14

Resistive displacement sensors are commonly termed potentiometers or "pots." A pot is an electromechanical device containing an electrically conductive *wiper* that slides against a fixed *resistive element* according to the position or angle of an external shaft (see Figure 23.1). Electrically, the resistive element is "divided" at the point of wiper contact. To measure displacement, a pot is typically wired in a "voltage divider" configuration, as shown in Figure 23.2. The circuit's output, a function of the position of the wiper, is an analog voltage available for direct use or digitization. Calibration maps the output voltage to units of displacement.

Table 23.1 lists some attributes inherent to pots. This chapter describes the different types of pots available, their electrical and mechanical characteristics, and practical approaches to using them for precision measurement. Sources and typical prices are also discussed. Versatile, inexpensive, and easy-to-use pots are a popular choice for precision measurement.

23.1 Precision Potentiometers

Pots are available in great variety, with specific kinds optimized for specific applications. Position measurement requires a high-quality pot designed for extended operation. Avoid pots classified as trimmers, rheostats, attenuators, volume controls, panel controls, etc. Instead, look for *precision potentiometers*.

23.1.1 Types of Precision Potentiometers

Precision pots are available in *rotary, linear-motion*, and *string pot* forms. String pots—also called *cable pots*, *yo-yo pots*, *cable extension transducers*, and *draw wire transducers*—measure the extended length of a spring-loaded cable. Rotary pots are available with single- or multiturn abilities: commonly 3, 5, or 10 turns. Linear-motion pots are available with maximum strokes ranging from roughly 5 mm to over 4 m [1,2]. String pots are available with maximum extensions exceeding 50 m [3]. Pot manufacturers

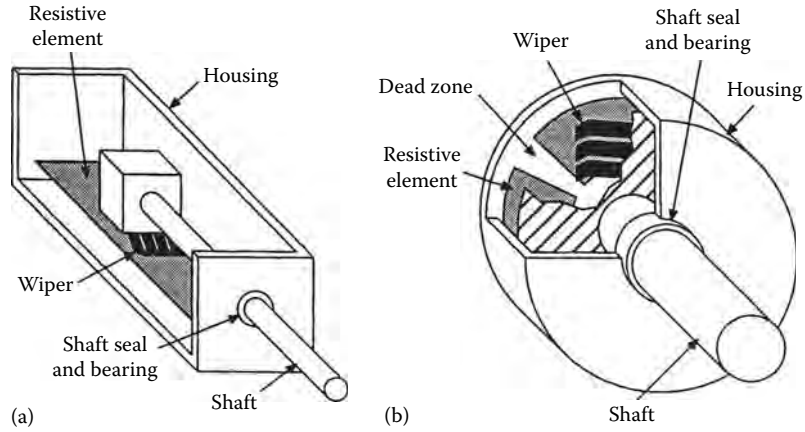


FIGURE 23.1 Representative cutaways of (a) linear-motion and (b) rotary potentiometers.

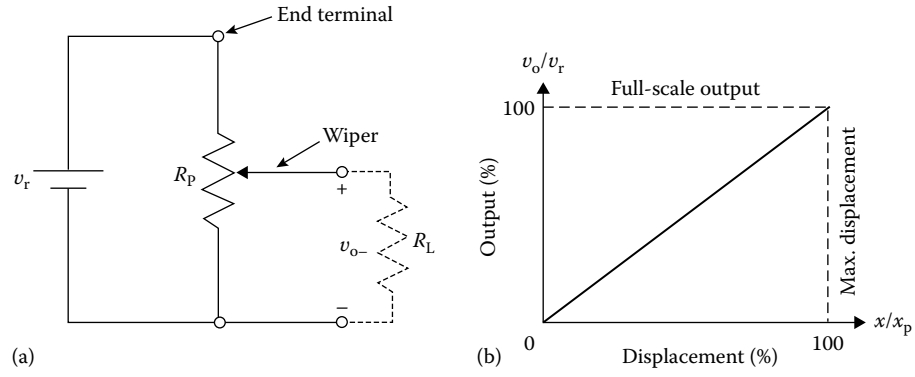


FIGURE 23.2 (a) Schematic diagrams depict a potentiometer as a resistor with an arrow representing the wiper. This schematic shows a pot used as a variable voltage divider—the preferred configuration for precision measurement. R_P is the total resistance of the pot, R_L is the load resistance, v_r is the reference or supply voltage, and v_o is the output voltage. (b) Shows an ideal linear output function where x represents the wiper position and x_p is its maximum position.

TABLE 23.1 Fundamental Potentiometer Characteristics

Advantages	Disadvantages
Easy to use	Limited bandwidth
Low cost	Frictional loading
Nonelectronic	Inertial loading
High-amplitude output signal	Wear
Proven technology	

usually specify a pot's type, dimensions, resistive element composition, electrical and mechanical parameters, and mounting method.

23.1.2 Resistive Element

Broadly, a pot's resistive element can be classified as either *wirewound* or *nonwirewound*. Wirewound elements contain tight coils of resistive wire that quantize measurement in steplike increments.

TABLE 23.2 Characteristics of Conductive Plastic, Wirewound, and Hybrid Resistive Elements

	Conductive Plastic	Wirewound	Hybrid
Resolution	Infinitesimal	Quantized	Infinitesimal
Power rating	Low	High	Low
Temperature stability	Poor	Excellent	Very good
Noise	Very low	Low, but degrades with time	Low
Life	10^6–10^8 cycles	10^5–10^6 cycles	10^6–10^7 cycles

In contrast, nonwirewound elements present a continuous sheet of resistive material capable of essentially unlimited measurement resolution.

Wirewound elements offer excellent temperature stability and high power dissipation abilities. The coils quantize measurement according to wire size and spacing. Providing the resolution limits are acceptable, wirewound elements can be a satisfactory choice for precision measurement; however, conductive plastic or hybrid elements will usually perform better and for considerably more cycles. These and other popular nonwirewound elements are described in more detail in the following.

Conductive plastic elements feature a smooth film with unlimited resolution, low friction, low noise, and long operational life. They are sensitive to temperature and other environmental factors and their power dissipation abilities are low; however, they are an excellent choice for most precision measurement applications.

Hybrid elements feature a wirewound core with a conductive plastic coating, combining wirewound and conductive plastic technologies to realize some of the more desirable attributes of both. The plastic limits power dissipation abilities in exchange for low noise, long life, and unlimited resolution. Like wirewounds, hybrids offer excellent temperature stability. They make an excellent choice for precision measurement.

Cermet elements made from a ceramic–metal alloy offer unlimited resolution and reasonable noise levels. Their advantages include high power dissipation abilities and excellent stability in adverse conditions. Cermet elements are rarely applied to precision measurement because conductive plastic elements offer lower noise, lower friction, and longer life.

Carbon composition elements, molded under pressure from a carbon–plastic mixture, are inexpensive and very popular for general use, but not for precision measurement. They offer unlimited resolution and low noise but are sensitive to environmental stresses (e.g., temperature, humidity) and are subject to wear.

Table 23.2 summarizes the distinguishing characteristics of the preferred resistive elements for precision measurement.

23.1.3 Electrical Characteristics

Before selecting a pot and integrating it into a measurement system, the following electrical characteristics should be considered.

23.1.3.1 Terminals and Taps

Table 23.3 shows the conventional markings found on the pot housing [4,5]; CW and CCW indicate clockwise and counterclockwise rotation as seen from the front end. Soldering studs and eyelets, integral connectors, and flying leads are common means for electrical connection. In addition to the wiper and end terminals, a pot may possess one or more terminals for *taps*. A tap enables an electrical connection to be made with a particular point along the resistive element. Sometimes, a *shunt resistor* is connected to a tap in order to modify the output function. End terminations and taps can exhibit different electrical characteristics depending on how they are manufactured (see Ref. [2] for more details).

TABLE 23.3 Potentiometer Terminal Markings

Terminal	Possible	Color	Codings	Rotary Pot	Linear-Motion Pot
1	Yellow	Red	Black	CCW limit	Fully retracted limit
2	Red	Green	White	Wiper	Wiper
3	Green	Black	Red	CW limit	Fully extended limit

23.1.3.2 Taper

Pots are available in a variety of different tapers that determine the shape of the output function. With a linear-taper pot, the output varies linearly with wiper motion, as shown in Figure 23.2. (Note that a pot with a linear taper should not be confused with a linear-motion pot, which is sometimes called a "linear pot.") Linear-taper pots are the most commonly available and are widely used in sensing and control applications. Pots with nonlinear tapers (e.g., logarithmic, sine, cosine, tangent, square, and cube) can also be useful, especially where computer control is not involved. Nonstandard tapers can be custom manufactured or alternatively, certain types of output functions can be produced using shunt resistors, by combining outputs from ganged pots or by other means (refer to Refs. [6,7] for more details). Of course, if a computer is involved, the output function can always be altered through a software lookup table or mapping function.

23.1.3.3 Electrical Travel

Figure 23.2 shows how the ideal output of a pot changes with wiper position. In practice, there is a small region at both ends where output remains constant until the wiper hits a mechanical stop. *Mechanical travel* is the total motion range of the wiper, and *electrical travel* is the slightly smaller motion range over which the electrical output is "valid." Thus, when using a pot as a sensor, it is important to ensure that the wiper motion falls within the electrical travel limits.

23.1.3.4 Linearity

Linearity is the maximum deviation of the output function from an ideal straight line. *Independent linearity* is commonly specified, where the straight line is defined as the line that minimizes the linearity error over a series of sampled points, not necessarily measured over the full range of the pot (see Figure 23.3). Other linearity metrics, such as *terminal-based linearity*, *absolute linearity*, and *zero-based linearity*, are also sometimes used (refer to Ref. [8] for more details). Pots are commonly available with independent linearities ranging from under 0.1% to 1%. When dealing with nonlinear output functions, *conformity* is specified since it is the more general term used to describe deviation from any ideal function. Conformity and linearity are usually expressed as a percentage of full-scale output (FSO).

23.1.3.5 Electrical Loading

Loading can significantly affect the linearity of measurements, regardless of a pot's quality and construction. Consider an ideal linear pot connected to an infinite load impedance (i.e., as in Figure 23.2). Since no current flows through the load, the output changes perfectly linear as the wiper travels along the length of the pot. However, if the load impedance is finite, the load draws some current, thereby affecting the output as illustrated in Figure 23.4. Circuit analysis shows that

$$\frac{v_o}{v_r} = \frac{(x/x_P)(R_L/R_P)}{(R_L/R_P)+(x/x_P)-(x/x_P)^2} \tag{23.1}$$

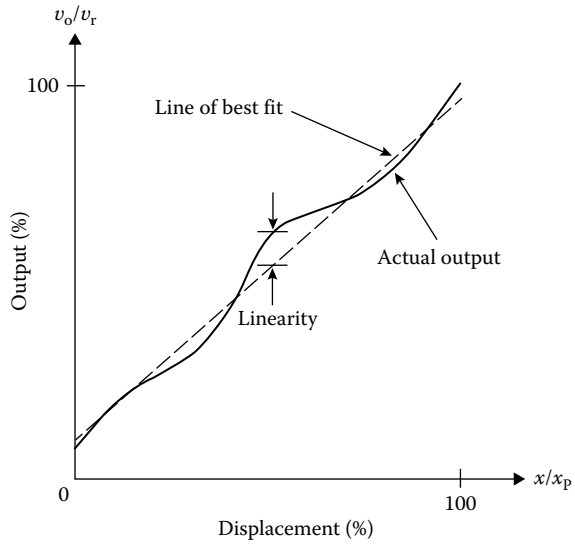


FIGURE 23.3 Independent linearity is the maximum amount by which the actual output function deviates from a line of best fit.

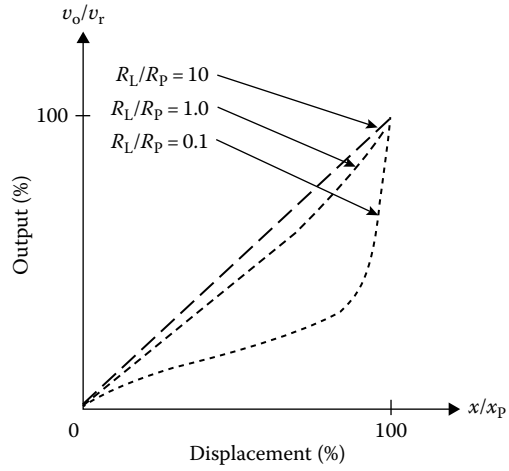


FIGURE 23.4 Linearity can be greatly influenced by the ratio of load resistance, R_L, to potentiometer resistance, R_P.

Therefore, R_L/R_P should be maximized to reduce loading effects (this also involves other trade-offs, to be discussed). A minimum R_L/R_P value of 10 is sometimes used as a guideline since loading error is then limited to less than 1% of FSO. Also, some manufacturers recommend a minimum load impedance or maximum wiper current in order to minimize loading effects and prevent damage to the wiper contacts. The following are some additional strategies that can be taken:

- Use a regulated voltage source whose output is stable with load variations.
- Use high-input-impedance signal-conditioning or data acquisition circuitry.
- Use only a portion of the pot's full travel.

23.1.3.6 Resolution

Resolution defines the smallest possible change in output that can be produced and detected. In wire-wound pots, the motion of the wiper over the coil generates a quantized response. Therefore, the best

attainable resolution is $r = (1/N) \times 100\%$, where N is the number of turns in the coil. Nonwirewound pots produce a smooth response with essentially unlimited resolution. Hybrid pots also fall into this category. In practice, resolution is always limited by factors such as

- Electrical noise, usually specified as *noise* for wirewound pots and *smoothness* for nonwirewound pots, both expressed as a percentage of FSO [10]
- Stability of the voltage supply, which can introduce additional noise into the measurement signal
- Analog-to-digital converter (ADC) resolution, usually expressed in "bits" (e.g., 10 mm travel digitized using a 12 bit ADC results in 10 mm/4096 = 0.0024 mm resolution at best)
- Mechanical effects such as stiction

23.1.3.7 Power Rating

The power dissipated by a pot is $P = v_r^2/R_P$; therefore, power rating determines the maximum voltage that can be applied to the pot at a given temperature. With greater voltage supplied to the pot, greater output (and noise) is produced but more power is dissipated, leading to greater thermal effects. In general, wirewound and cermet pots are better able to dissipate heat and thus have the highest power ratings.

23.1.3.8 Temperature Coefficient

As temperature increases, pot resistance also increases. However, a pot connected as shown in Figure 23.2 will divide the voltage equally well, regardless of its total resistance. Thus, temperature effects are not usually a major concern as long as the changes in resistance are uniform and the pot operates within its ratings. However, an increase in pot resistance also increases loading nonlinearities. Therefore, temperature coefficients can become an important consideration. The temperature coefficient, typically specified in ppm°C^{-1}, can be expressed as $\alpha = (\Delta R_P/R_P)/\Delta t$, where Δt is the change in temperature and ΔR_P is the corresponding change in total resistance. In general, wirewound pots possess the lowest temperature coefficients. Temperature-compensating signal-conditioning circuitry can also be used.

23.1.3.9 Resistance

Since a pot divides voltage equally well regardless of its total resistance, resistance tolerance is not usually a major concern. However, total resistance can have a great impact on loading effects. If resistance is large, less current flows through the pot, thus reducing temperature effects but also increasing loading.

23.1.3.10 AC Excitation

Pots can operate using either a DC or an AC voltage source. However, wirewound pots are susceptible to capacitive and inductive effects that can be substantial at moderate to high frequencies.

23.1.4 Mechanical Characteristics

The following mechanical characteristics influence measurement quality and system reliability and thus should be considered when selecting a pot.

23.1.4.1 Mechanical Loading

A pot adds inertia and friction to the moving parts of the system that it is measuring. As a result, it increases the force required to move these parts. This effect is referred to as *mechanical loading*.

To quantify mechanical loading, rotary pot manufacturers commonly list three values: the equivalent *mass moment of inertia* of the pot's rotating parts, the *dynamic* (or *running*) *torque* required to maintain rotation in a pot shaft, and the *starting torque* required to initiate shaft rotation. For linear-motion pots, the three analogous loading terms are *mass*, *starting force*, and *dynamic* (or *running*) *force*.

In extreme cases, mechanical loading can adversely affect the operating characteristics of a system. When including a pot in a design, ensure that the inertia added to the system is insignificant or that the inertia is considered when analyzing the data from the pot. The starting and running force or torque values might also be considered, although they are generally small due to the use of bearings and low-friction resistive elements.

23.1.4.2 Mechanical Travel

Distinguished from electrical travel, *mechanical travel* is the wiper's total motion range. A mechanical stop delimits mechanical travel at each end of the wiper's range of motion. Stops can withstand small loads only and therefore should not be used as mechanical limits for the system. Manufacturers list maximum loads as the *static stopping strength* (for static loads) and the *dynamic stopping strength* (for moving loads).

Rotary pots are also available without mechanical stops. The shaft of such an "unlimited travel" pot can be rotated continuously in either direction; however, electrical travel is always less than 360° due to the discontinuity or "dead zone" where the resistive element begins and ends (see Figure 23.1). Multiple revolutions can be measured with an unlimited travel pot in conjunction with a counter: the counter maintains the number of full revolutions, while the pot measures subrevolution angular displacement.

23.1.4.3 Operating Temperature

When operated within its specified temperature range, a pot maintains good electrical linearity and mechanical integrity. Depending on construction, pots can operate at temperatures from as low as −65 °C to as high as 150 °C. Operating outside specified limits can cause material failure, either directly from temperature or from thermally induced misalignment.

23.1.4.4 Vibration, Shock, and Acceleration

Vibration, shock, and acceleration are all potential sources of contact discontinuities between the wiper and the resistive element. In general, a contact failure is considered to be a discontinuity equal to or greater than 0.1 ms [2]. The values quoted in specification sheets are in *g*s and depend greatly on the particular laboratory test. Some characterization tests use sinusoidal vibration, random vibration, sinusoidal shock, sawtooth shock, or acceleration to excite the pot. Manufacturers use mechanical design strategies to eliminate weaknesses in a pot's dynamic response. For example, one technique minimizes vibration-induced contact discontinuities using multiple wipers of differing resonant frequencies.

23.1.4.5 Speed

Exceeding a pot's specified maximum speed can cause premature wear or discontinuous values through effects such as wiper bounce. As a general rule, the slower the shaft motion, the longer the unit will last (in total number of cycles). Speed limitations depend on the materials involved. For rotary pots, wirewound models have preferred maximum speeds on the order of 100 rpm, while conductive plastic models have allowable speeds as high as 2000 rpm. Linear-motion pots have preferred maximum velocities up to 10 m s^{-1}.

23.1.4.6 Life

Despite constant mechanical wear, a pot's expected lifetime is on the order of a million cycles when used under proper conditions. A quality film pot can last hundreds of millions of cycles. Of wirewound, hybrid, and conductive plastic pots, the uneven surface of a wirewound resistive element inherently experiences the most wear and thus has the shortest expected operating life. Hybrids improve on this by using a wirewound construction in combination with a smooth conductive film coating. Conductive plastic pots generally have the longest life expectancy due to the smooth surface of their resistive element.

23.1.4.7 Contamination and Seals

Foreign material contaminating pots can promote wear and increase friction between the wiper and the resistive element. Consequences range from increased mechanical loading to outright failure (e.g., seizing, contact discontinuity). Fortunately, sealed pots are available from most manufacturers for industrial applications where dirt and liquids are often unavoidable. To aid selection, specifications often include the type of *case sealing* (i.e., mechanisms and materials) and the *seal resistance* to cleaning solvents and other commonly encountered fluids.

23.1.4.8 Misalignment

Shaft misalignment in a pot can prematurely wear its bearing surfaces and increase its mechanical loading effects. A good design minimizes misalignment (see *Implementation* in the following). Manufacturers list a number of alignment tolerances. In linear-motion pots, *shaft misalignment* is the maximum amount a shaft can deviate from its axis. The degree to which a shaft can rotate around its axis is listed under *shaft rotation*. In rotary pots, *shaft end play* and *shaft radial play* both describe the amount of shaft deflection due to a radial load. *Shaft runout* denotes the shaft diameter eccentricity when a shaft is rotated under a radial load.

23.1.5 Mechanical Mounting Methods

Hardware features on a pot's housing determine the mounting method. Options vary with manufacturer and among rotary, linear-motion, and string pots. Offerings include custom bases, holes, tabs, flanges, and brackets—all of which secure with machine screws—and threaded studs, which secure with nuts. Linear-motion pots are available with rod or slider actuation, some with internal or external return springs. Mounting is typically accomplished by movable clamps, often supplied by the pot manufacturer. Other linear-motion pots mount via a threaded housing. For rotary pots, the two most popular mounting methods are the *bushing mount* and the *servo mount* (see Figure 23.5).

23.1.5.1 Bushing Mount

The pot provides a shaft-concentric, threaded sleeve that invades a hole in a mounting fixture and secures with a nut and lock washer. An off-axis tab or pin prevents housing rotation. Implementing a bushing mount requires little more than drilling a hole; however, limited rotational freedom and considerable play before tightening complicate precise setup.

23.1.5.2 Servo Mount

The pot provides a flanged, shaft-concentric, precision-machined rim that slips into a precision-bored hole in a mounting fixture. The flange secures with symmetrically arranged, quick-releasing *servo mount clamps*, available from Timber-Top, Inc. [9] and also from the sources listed in Table 23.4. (These

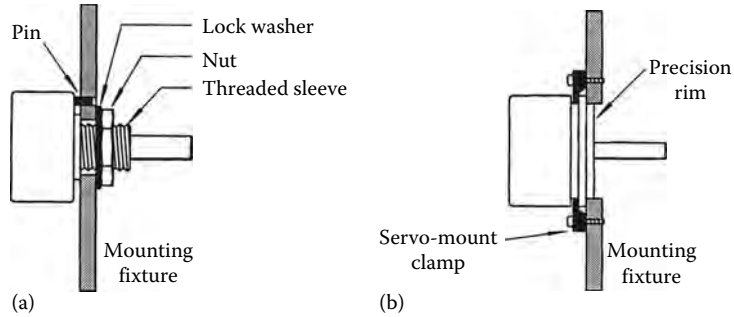


FIGURE 23.5 The two most common rotary pot mounts are (a) the bushing mount and (b) the servo mount.

TABLE 23.4 Sources of Small Mechanical Components

PIC Design, 86 Benson Road, P.O. Box 1004, Middlebury, CT 06762-1004, Tel: (800) 243-6125, (203) 758-8272; Fax: (203) 758-8271, http://www.pic-design.com/
Stock Drive Products/Sterling Instrument, 2101 Jericho Turnpike, Box 5416
New Hyde Park, NY 11042-5416, Tel: (516) 328-3300; Fax: (800) 737-7436, (516) 326-8827, http://www.sdp-si.com/
W.M. Berg, Inc., 499 Ocean Ave., East Rockaway, NY 11518, Tel: (800) 232-2374, (516) 599-5010; Fax: (800) 455-2374, (516) 599-3274, http://www.wmberg.com/

clamps are also called *synchro mount clamps* and *motor mount cleats*, since servo-mounting synchros and stepper motors are also available.) Servo mounts are precise and easy to adjust but entail the expense of precision machining.

23.2 Measurement Techniques

To measure displacement, a pot must attach to mechanical fixtures and components. The housing typically mounts to a stationary reference frame, while the shaft couples to a moving element. The *input motion* (i.e., the motion of interest) can couple directly or indirectly to the pot's shaft. A direct connection, although straightforward, carries certain limitations:

- The input motion maps 1:1 to the shaft motion.
- The input motion cannot exceed the pot's mechanical travel limits.
- Angle measurement requires a rotary pot; position measurement requires a linear-motion pot.
- The pot must mount close to the motion source.
- The input motion must be near-perfectly collinear or coaxial with the shaft axis.

Figure 23.6 shows ways to overcome these limitations. Mechanisms with a mechanical advantage scale motion and adjust travel limits. Mechanisms that convert between linear and rotary motion enable any type of pot to measure any kind of motion. Transmission mechanisms distance a pot from the measured motion. Compliant mechanisms compensate for misalignment. Examples and more details follow. Most of the described mechanisms can be realized with components available from the sources in Table 23.4.

Gears scale the mapping between input and pot shaft motions according to gear ratio. They also displace rotation axes to a parallel or perpendicular plane according to type of gear (e.g., spur vs. bevel). Gears introduce backlash. Friction rollers are a variation on the gear theme, immune to backlash but prone to slippage. The ratio of roller diameters scales the mapping between input and pot shaft motions.

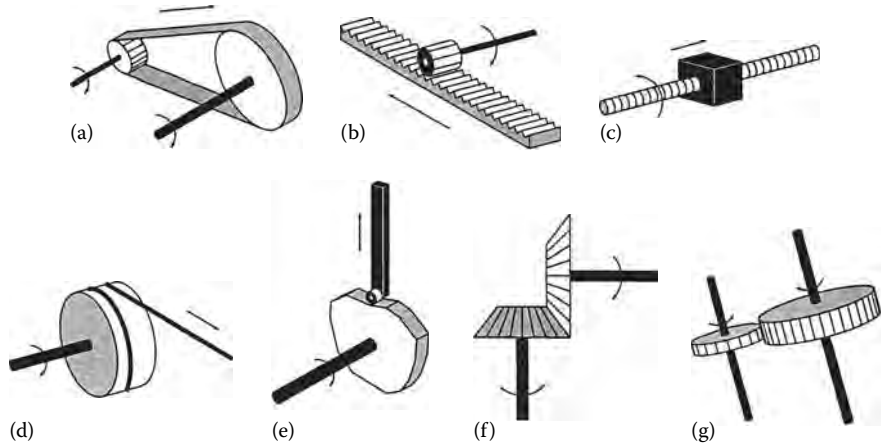


FIGURE 23.6 Mechanisms that extend a precision potentiometer's capabilities include (a) belts and pulleys, (b) rack and pinions, (c) lead screws, (d) cabled drums, (e) cams, (f) bevel gears, and (g) spur gears.

Rack-and-pinion mechanisms convert between linear and rotary motion. Mapping is determined by the rack's *linear pitch* (i.e., tooth-to-tooth spacing) compared to the number of teeth on the pinion. Backlash is inevitable.

Lead screws convert rotary motion to linear motion via the screw principle. Certain low-friction types (e.g., ball screws) are also capable of the reverse transformation (i.e., linear to rotary). Either way, mapping is controlled by the screw's *lead*—the distance the nut travels in one revolution. Lead screws are subject to backlash.

Cabled drums convert between linear and rotary motion according to the drum circumference, since one turn of diameter D wraps or unwraps a length πD of cable. An external force (e.g., supplied by a spring or a weight) might be necessary to maintain cable tension.

Pulleys can direct a string pot's cable over a complex path to a motion source. Mapping is 1:1 unless the routing provides a mechanical advantage.

Pulleys and *belts* transmit rotary motion scaled according to relative pulley diameters. The belt converts between linear and rotary motion (see Figure 23.6a). The empty area between pulleys provides a convenient passageway for other components. Sprocket wheels and chain have similar characteristics. Matched pulley–belt systems are available that operate with negligible slip, backlash, and stretch.

Cams map rotary motion into linear motion according to the function "programmed" into the cam profile (see Ref. [10] for more information).

Linkages can be designed to convert, scale, and transmit motion. Design and analysis can be quite complex and mapping characteristics tend to be highly nonlinear (see Ref. [11] for details).

Flexible shafts transmit rotary motion between two nonparallel axes with a 1:1 mapping, subject to torsional windup and hysteresis if the motion reverses.

Conduit, like a bicycle brake cable or *Bowden cable*, can route a cable over an arbitrary path to connect a pot to a remote motion source. The conduit should be incompressible and fixed at both ends. Mapping is 1:1 with some mechanical slop. Lubrication helps mitigate friction.

A mechanism's mapping characteristics impact measurement resolution and accuracy. Consider a stepper motor turning a lead screw to translate a nut. A linear-motion pot could measure the nut's position directly to some resolution. Alternatively, a rotary pot linked to the lead screw could measure the position with increased resolution if the mechanism mapped the same amount of nut travel to considerably more wiper travel. Weighing the resolution increase against the uncertainty due to backlash would determine which approach was more accurate.

23.3 Implementation

Integrating a pot into a measurement system requires consideration of various design issues, including the impact of the pot's physical characteristics, error sources, space restrictions, and wire routing. The pot's shaft type and bearings must be taken into consideration and protected against excessive loading. A good design will

- Give the pot mount the ability to accommodate minor misalignment
- Protect the shaft from thrust, side, and bending loads (i.e., not use the pot as a bearing)
- Provide hard limit stops within the pot's travel range (i.e., not use the pot's limit stops)
- Protect the pot from contaminants
- Strain relieve the pot's electrical connections

A thorough treatment of precision design issues appears in Ref. [12].

23.3.1 Coupling to the Pot

Successful implementation also requires practical techniques for mechanical attachment. A string pot's cable terminator usually fastens to other components with a screw. For other types of pots, coupling technique is partly influenced by the nature of the shaft. Rotary shafts come with various endings, including plain, single-flatted, double-flatted, slotted, and knurled. Linear-motion shafts usually terminate in threads but are also available with roller ends (to follow surfaces) or with spherical bearing ends (to accommodate misalignment). With care, a shaft can be cut, drilled, filed, threaded, etc.

In a typical measurement application, the pot shaft couples to a mechanical component (e.g., a gear, a pulley) or to another shaft of the same or different diameter. Successful couplings provide a positive link to the shaft without stressing the pot's mechanics. Satisfying these objectives with rotary and linear-motion pots requires a balance between careful alignment and compliant couplings. Alignment is not as critical with a string pot. Useful coupling methods include the following.

Compliant couplings: It is generally wise to put a compliant coupling between a pot's shaft and any other shafting. A compliant coupling joins two misaligned shafts of the same or different diameter. Offerings from the companies in Table 23.4 include bellow couplings, flex couplings, spring couplings, spider couplings, Oldham couplings, wafer spring couplings, flexible shafts, and universal joints. Each type has idiosyncrasies that impact measurement error; manufacturer catalogs provide details.

Sleeve couplings: Less expensive than a compliant coupling, a rigid sleeve coupling joins two shafts of the same or different diameter with the requirement that the shafts be perfectly aligned. Perfect alignment is difficult to achieve initially and impossible to maintain as the system ages. Imperfect alignment accelerates wear and risks damaging the pot. Sleeve couplings are available from the companies listed in Table 23.4.

Press fits: A press fit is particularly convenient when the bore of a small plastic part is nominally the same as the shaft diameter. Carefully force the part onto the shaft. Friction holds the part in place, but repeated reassembly will compromise the fit.

Shrink fits: Components with a bore slightly under the shaft diameter can be heated to expand sufficiently to slip over the shaft. A firm grip results as the part cools and the bore contracts.

Pinning: Small hubbed components can be pinned to a shaft. The pin should extend through the hub partway into the shaft, and the component should fit on the shaft without play. Use roll pins or spiral pins combined with a thread-locking compound (e.g., Loctite 242).

Setscrews: Small components are available with hubs that secure with setscrews. The component should fit on the shaft without play. For best results, use two setscrews against a shaft with perpendicular flats.

Dimple a plain shaft using the component's screw hole(s) as a drill guide. Apply a thread-locking compound (e.g., Loctite 242) to prevent the setscrews from working loose.

Clamping: Small components are also available with split hubs that grip a shaft when squeezed by a matching hub clamp. Clamping results in a secure fit without marring the shaft.

Adhesives: Retaining compounds (e.g., Loctite 609) can secure small components to a shaft. Follow manufacturer's instructions for best results.
Spring-loaded contact: A spring-loaded shaft will maintain positive contact against a surface that moves at reasonable speeds and without sudden acceleration.

23.4 Costs and Sources

Precision pots are inexpensive compared to other displacement measurement technologies. Table 23.5 lists approximate costs for off-the-shelf units in single quantity. Higher quality generally commands a higher price; however, excellent pots are often available at bargain prices due to volume production or surplus conditions. Electronic supply houses offer low-cost pots (i.e., under $20) that can suffice for short-term projects. Regardless of price, always check the manufacturer's specifications to confirm a pot's suitability for a given application.

Table 23.6 lists several sources of precision pots. Most manufacturers publish catalogs, and many have websites. In addition to a standard product line, most manufacturers will custom-build pots for high-volume applications.

23.5 Evaluation

Precision pots are a mature technology, effectively static except for occasional developments in materials, packaging, manufacturing, etc. Recent potentiometric innovations—including momentary-contact membrane pots [13] and solid-state digital pots [14]—are unavailing to precision measurement.

The variable voltage divider is the traditional configuration for precision measurement. The circuit's output, a high-amplitude DC voltage, is independent of variations in the pot's total resistance and is highly compatible with other circuits and systems. Other forms of output are possible with a precision pot configured as a variable resistor. For example, paired with a capacitor, a pot could supply a position-dependent *RC* time constant to modulate an oscillator's output frequency or duty cycle. In this setup, the pot's stability and AC characteristics would be important.

An alternative resistive displacement sensing technology is the *magnetoresistive potentiometer*, available in rotary and linear-motion forms. Magnetoresistive pots incorporate a noncontacting, permanent magnet "wiper" that rides above a pair of magnetoresistive elements. The elements, configured as a voltage divider, change their resistances according to the strength of the applied magnetic field and thus divide the voltage as a function of the magnet's position. The output is approximately linear over a limited range of motion (e.g., 90° in rotary models). Magnetoresistive pots offer unlimited resolution and exceptionally long life but may require temperature compensation circuitry (see Ref. [15] for more information).

TABLE 23.5 Typical Single-Quantity Prices ($US) for Commercially Available Pots

Potentiometer Type	Approximate Price Range
Rotary	$10–$350
Linear motion	$20–$2000
String	$250–$1000

TABLE 23.6 Sources of Precision Pots

Company	Potentiometer Types
Betatronix, Inc. 110 Nicon Court Hauppauge, NY 11788, Tel: (516) 582-6740; Fax: (516) 582-6038, http://www.betatronix.com/	Exotic linear motion, rotary
BI Technologies Corp. 4200 Bonita Place, Fullerton, CA 92635, Tel: (714) 447-2345; Fax: (714) 447-2500, http://www.bitechnologies.com/index.htm	Rotary
Bourns, Inc. Sensors & Controls Division, 2533 N. 1500 W., Ogden, UT 84404, Tel: (801) 786-6200; Fax: (801) 786-6203, http://www.bourns.com/	Mostly rotary, some linear motion
Celesco Transducer Products, Inc. 7800 Deering Avenue, Canoga Park, CA 91309, Tel: (800) 423-5483, (818) 884-6860, Fax: (818)340-1175, http://www.celesco.com/	String
Honeywell Data Instruments, Inc. 100 Discovery Way, Acton, MA 01720-3648, Tel: (800) 333-3282, (978) 264-9550, Fax: (978) 263-0630, http://sensing.honeywell.com/	Linear motion, rotary
BEI Sensors, Duncan Electronics Division BEI Sensors & Systems Company, 15771 Red Hill Avenue, Tustin, CA 92680, Tel: (714) 258-7500; Fax: (714) 258-8120, http://www.beiduncan.com/index.php	Linear motion, rotary
Dynamation Transducers Corp. 348 Marshall Street, Holliston, MA 01746-1441, Tel: (508) 429-8440; Fax: (508) 429-1317, http://www.dynamationtransducers.com/	Linear motion, rotary
JDK Controls, Inc. 424 Crown Pt. Circle, Grass Valley, CA 95945, Tel: (530) 273-4608; Fax: (530) 273-0769, http://www.jdkcontrols.com	Rotary, "do-it-yourself" rotor/wiper assemblies
Midori America Corp. 2555 E. Chapman Ave, Suite 400, Fullerton, CA 92631, Tel: (714) 449-0997; Fax: (714) 449-0139, http://www.midoriamerica.com/	Linear motion, rotary, string; also magnetoresistive
New England Instrument 245 Railroad Street, Woonsocket, RI 02895-1129, Tel: (401) 769-0703; Fax: (401) 769-0037, http://www.newenglandinstrument.com/	Linear motion, rotary, resistive elements
Novotechnik U.S., Inc. 237 Cedar Hill Street, Marlborough, MA 01752, Tel: (800) 667-7492, (508) 485-2244, Fax: (508) 485-2430, http://www.novotechnik.com/	Linear motion, rotary
Servo Systems, Co. 115 Main Road, PO Box 97, Montville, NJ 07045-0097, Tel: (800) 922-1103, (973) 335-1007, Fax: (973) 335-1661, http://www.servosystems.com/	Linear motion, rotary (surplus)
SpaceAge Control, Inc. 38850 20th Street East, Palmdale, CA 93550, Tel: (805) 273-3000; Fax: (805) 273-4240, http://spaceagecontrol.com/	String
UniMeasure, Inc. 501 S.W. 2nd Street, Corvallis, OR 97333, Tel: (541) 757-3158, Fax: (541) 757-0858, http://www.unimeasure.com/	String
Axsys Technologies, Inc. Vernitron Sensor Systems Division, Precision Potentiometer Division, 2800 Anvil Street North, St. Petersburg, FL 33710, Tel: (813) 347-2181; Fax: (813) 347-7520, http://www.axsys.com	Linear motion, rotary

References

1. Bourns, Inc., *Electronic Components RC4 Solutions Guide*, New York, 1995, p. 304.
2. Vernitron Motion Control Group, *Precision Potentiometers*, Catalog #752, New York, 1993.
3. Products, *Linear Position Transducers*, Unimeasure, Inc., http://unimeasure.com/ (accessed on March 24, 2013).
4. Instrument Society of America, *ISA-S37.12–1977* (*R1982*) *Specifications and Tests for Potentiometric Displacement Transducers*, Research Triangle Park, NC, 1982.
5. E. C. Jordan (Ed.), *Reference Data for Engineers: Radio, Electronics, Computer, and Communications*, 7th edn., Indianapolis, IN: H.W. Sams, 1985, pp. 5–16.
6. D. C. Greenwood, *Manual of Electromechanical Devices*, New York: McGraw-Hill, 1965, pp. 297–299.
7. E. S. Charkey, *Electromechanical System Components*, New York: Wiley-Interscience, 1972, pp. 302–303.
8. Variable Resistive Components Institute, *VRCI-P-100A Standard for Wirewound and Nonwirewound Precision Potentiometers*, Vista, CA, 1988.
9. Timber-Top, Inc., P.O. Box 517, Watertown, CT 06795, Tel: (860)-274–6706; Fax (860)-274–8041.
10. J. Angeles and C. S. López-Cajún, *Optimization of Cam Mechanisms*, Dordrecht, the Netherlands: Kluwer Academic, 1991.
11. P. W. Jensen, *Classical and Modern Mechanisms for Engineers and Inventors*, New York: Marcel Dekker, 1991.
12. A. H. Slocum, *Precision Machine Design*, Englewood Cliffs, NJ: Prentice Hall, 1992.
13. Spectra Symbol Inc., data sheet: *SoftPot®* (*Membrane Potentiometer*), Salt Lake City, UT, 1996.
14. Dallas Semiconductor Corp., *Digital Potentiometer Overview*, web page: www.dalsemi.com/Prod_info/Dig_Pots/, December 1997.
15. Midori America Corp., *Midori Position Sensors 1995 Catalog*, Fullerton, CA, 1995.

24

Inductive Displacement Sensors

24.1 Theory **24**-1
24.2 Linear and Rotary Variable-Reluctance Transducer **24**-3
Single-Coil Linear-Variable-Reluctance Sensor • Variable-Differential Reluctance Sensor • Variable-Reluctance Tachogenerators
24.3 Microsyn **24**-7
24.4 Synchros **24**-8
24.5 Moving-Coil Inductive Displacement Sensors **24**-9
Linear-Variable Inductor • Variable-Coupling Transducers • Induction Potentiometer
24.6 Linear-Variable-Differential Transformer **24**-12
24.7 Rotary-Variable-Differential Transformer **24**-17
24.8 Eddy Current Proximity Sensors **24**-17
24.9 Gradiometer-Inductive Displacement Sensors **24**-19
24.10 Miniature-Inductive Displacement Sensors **24**-19
24.11 Shielding and Sensitivity of Inductive Sensors to Electromagnetic Interference **24**-20
Bibliography **24**-21
Partial List of Manufacturers and Suppliers **24**-21

Halit Eren
Curtin University

24.1 Theory

Inductive sensors are widely used in industry in many diverse applications. They are robust and compact and are less affected by environmental factors (e.g., humidity, dust) in comparison to their capacitive counterparts.

Inductive sensors are based on the principles of magnetic circuits. They can be classified as self-generating sensors or passive sensors. The self-generating types utilize an electrical generator principle; that is, when there is a relative motion between a conductor and a magnetic field, a voltage is induced in the conductor. Or a varying magnetic field linking a stationary conductor produces voltage in the conductor. In instrumentation applications, the magnetic field may be varying with some frequency, and the conductor may also be moving at the same time. In inductive sensors, the relative motion between field and conductor is supplied by changes in the measurand, usually by means of some mechanical motion. On the other hand, the passive transducer requires an external source of power. In this case, the action of the transducer is simply the modulation of the excitation signal.

Inductive displacement sensors can be divided into two major types:

1. Air core sensors
2. Magnetic core sensors

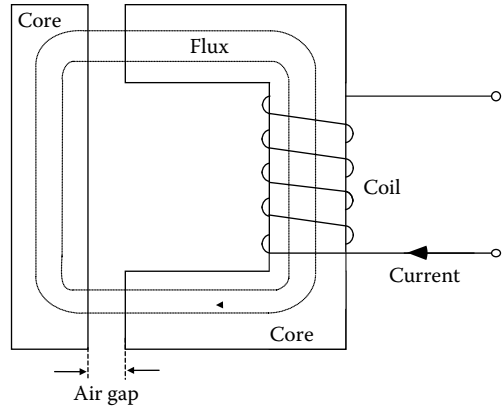


FIGURE 24.1 A basic inductive sensor consists of a magnetic circuit made from a ferromagnetic core with a coil wound on it. The coil acts as a source of mmf that drives the flux through the magnetic circuit through the core and the air gap. The presence of the air gap causes a large increase in circuit reluctance and a corresponding decrease in the flux. Hence, small variations in the air gap result in measurable changes in inductance.

Although air core sensors are used in some applications, they tend to have a lower sensitivity when compared to magnetic core sensors. Therefore, in this article, more emphasis will be given on magnetic core sensors. The basic principles are the same in both types.

A simple magnetic circuit is shown in Figure 24.1 to explain the basic principles of inductive sensors. In this figure, the magnetic circuit consists of a core, made from a ferromagnetic material together with a coil having n number of turns wound on it. The coil acts as a source of magnetomotive force (mmf), which drives the flux Φ through the magnetic circuit. If one assumes that the air gap is zero, the equation for the magnetic circuit can be expressed as

$$\text{mmf} = \text{Flux} \times \text{Reluctance} = \Phi \times \Re \text{ A-turns} \tag{24.1}$$

such that the reluctance $\Re$ limits the flux in a magnetic circuit similar to a resistance that limits the current in an electric circuit. By writing the mmf in terms of current, the magnetic flux may be expressed as

$$\Phi = \frac{ni}{\Re} \text{ weber} \tag{24.2}$$

In Figure 24.1, the flux linking a single turn is expressed by Equation 24.2, but for the total flux linking the entire n number of the turns of the coil, the equation becomes

$$\Psi = n\Phi = \frac{n^2 i}{\Re} \text{ weber} \tag{24.3}$$

Equation 24.3 leads to self-inductance L of the coil, which is described as the total flux (ψ weber) per unit current for that particular coil; that is

$$L = \frac{\Psi}{I} = \frac{n^2}{\Re} \tag{24.4}$$

This indicates that the self-inductance of an inductive element can be calculated by magnetic circuit properties. Expressing $\Re$ in terms of dimensions as

$$\Re = \frac{l}{\mu\mu_0 A} \tag{24.5}$$

where

l is the total length of the flux path
μ is the relative permeability of the magnetic circuit material
μ_0 is the permeability of free space ($=4\pi \times 10^{-7}$ H/m)
A is the cross-sectional area of the flux path

The arrangement illustrated in Figure 24.1 becomes a basic inductive sensor if the air gap is allowed to vary. In this case, the ferromagnetic core is separated into two parts by the air gap. The total reluctance of the circuit now is the addition of the reluctance of core and the reluctance of air gap. The relative permeability of air is close to unity, and the relative permeability of the ferromagnetic material can be of the order of a few thousand, indicating that the presence of the air gap causes a large increase in circuit reluctance and a corresponding decrease in the flux. Hence, as one would expect, a small variation in the air gap causes a measurable change in the inductance. Most of the inductive transducers are based on these principles and are discussed in the following in greater detail.

24.2 Linear and Rotary Variable-Reluctance Transducer

The variable-reluctance transducers are based on changes in the reluctance of a magnetic flux path. This type of transducer finds application particularly in acceleration measurements. However, they can be constructed in a suitable manner for sensing displacements as well as velocities. They come in many different forms, as described in the following.

24.2.1 Single-Coil Linear-Variable-Reluctance Sensor

A typical single-coil variable-reluctance displacement sensor is illustrated in Figure 24.2. The sensor consists of three elements: a ferromagnetic core in the shape of a semicircular ring, a variable air gap, and a ferromagnetic plate. The total reluctance of the magnetic circuit is the sum of the individual reluctances:

$$\Re_T = \Re_C + \Re_G + \Re_A \tag{24.6}$$

where $\Re_C$, $\Re_G$, and $\Re_A$ are the reluctances of the core, air gap, and armature, respectively.

Each one of these reluctances can be determined by using the properties of materials involved, as in Equation 24.5. In this particular case, the reluctance $\Re_T$ can be approximated as

$$\Re_T = \frac{R}{\mu_C \mu_0 r^2} + \frac{2d}{\mu_0 \pi r^2} + \frac{R}{\mu_A \mu_0 rt} \tag{24.7}$$

In obtaining Equation 24.7, the length of flux path in the core is taken as πR. The cross-sectional area is assumed to be uniform, with a value of πr^2. The total length of the flux path in air is $2d$, and it is assumed that there is no fringing or bending of the flux through the air gap, such that the cross-sectional area of the flux path in air will be close to that of the cross-sectional area of the core. The length of an average

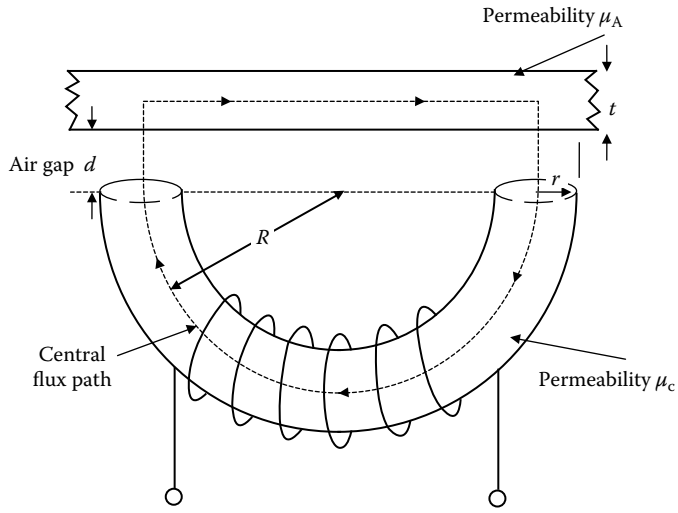


FIGURE 24.2 A typical single-coil, variable-reluctance displacement sensor. The sensor consists of three elements: a ferromagnetic core in the shape of a semicircular ring, a variable air gap, and a ferromagnetic plate. The reluctance of the coil is dependent on the single variable. The nonlinearly reluctance increases with the increasing air gap.

central flux path in the armature is $2R$. The calculation of the appropriate cross-sectional area of the armature is difficult, but it may be approximated to $2rt$, where t is the thickness of the armature.

In Equation 24.7, all of the parameters are fixed except for the one independent variable—the air gap. Hence, it can be simplified as

$$\Re_T = \Re_0 + kd \tag{24.8}$$

where

$$\Re_0 = \frac{R}{\mu_0 r}[1/\mu_C r + 1/\mu_A t]$$

$$k = \frac{2}{\mu_0 \pi r^2}$$

Using Equations 24.4 and 24.8, the inductance can be written as

$$L = \frac{n^2}{\Re_0 + kd} = \frac{L_0}{1 + \alpha d} \tag{24.9}$$

where

L_0 is the inductance at zero air gap

$\alpha = k/\Re_0$

The values of L_0 and α can be determined mathematically: they both depend on the core geometry and permeability, as explained earlier. From Equation 24.9, it can be seen that the relationship between L and α is nonlinear. Despite this nonlinearity, these types of single-coil sensors find applications in a number

of areas, such as force measurements and telemetry. In force measurements, the resultant change in inductance can be made to be a measure of the magnitude of the applied force. The coil usually forms one of the components of an *LC* oscillator, for which the output frequency varies with the applied force. Hence, the coil modulates the frequency of the local oscillator.

24.2.2 Variable-Differential Reluctance Sensor

The problem of the nonlinearity explained earlier can be eliminated by modifying the single-coil system into a variable-differential reluctance sensor (also known as push–pull sensor), as shown in Figure 24.3. This sensor consists of an armature moving between two identical cores and separated by a fixed distance of 2*d*. Now, Equation 24.9 can be written for both coils as

$$L_1 = \frac{L_{01}}{1+\alpha(d-x)}, \quad L_2 = \frac{L_{02}}{1+\alpha(d+x)} \tag{24.10}$$

Although the relationship between L_1 and L_2 is still nonlinear, the sensor can be incorporated into an ac deflection bridge to give a linear output for small movements. The hysteresis errors of these transducers are almost entirely limited to the mechanical components. These sensors respond to both static and dynamic measurements. They have continuous resolution and high outputs, but they may give erratic performance in response to presence of external magnetic fields. A typical sensor of this type has an input span of 1 cm, a coil inductance of 25 mH, and a coil resistance of 75 Ω. The resistance of the coil must carefully be considered when designing oscillator circuits. In many cases, the maximum nonlinearity is about 0.5%.

A typical commercially available variable-differential sensor is shown in Figure 24.4. The iron core is located halfway between the two E-shaped frames. The flux generated by primary coils depends on

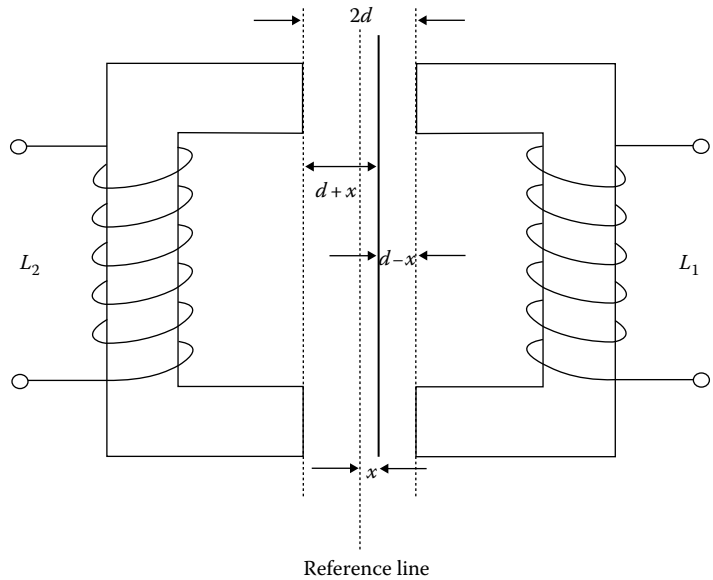


FIGURE 24.3 A variable-differential reluctance sensor consists of an armature moving between two identical cores separated by a fixed distance. The armature moves in the air gap in response to a mechanical input. This movement alters the reluctance of coils 1 and 2, thus altering their inductive properties. This arrangement overcomes the problem of nonlinearity inherent in single-coil sensors.

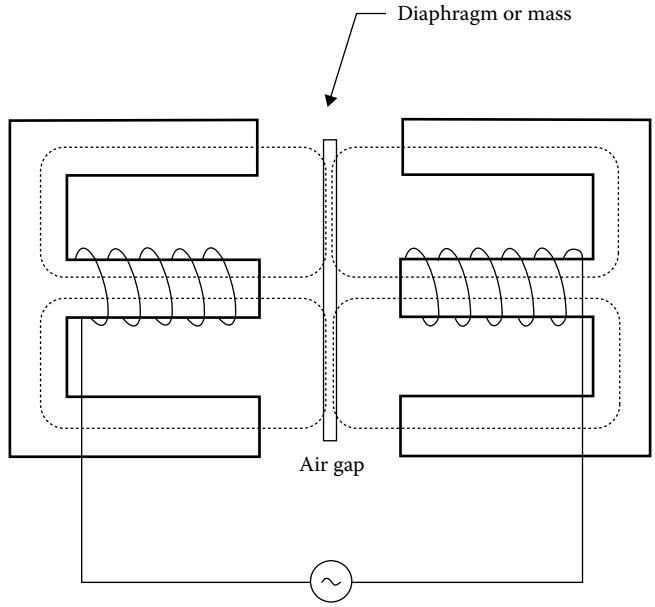


FIGURE 24.4 A typical commercial variable-differential sensor. The iron core is located halfway between the two E frames. Motion of the core increases the air gap for one of the E frames while decreasing the other side. This causes reluctances to change, thus inducing more voltage on one side than on the other. Motion in the other direction reverses the action, with a 180° phase shift occurring at null. The output voltage can be processed, depending on the requirements, by means of rectification, demodulation, or filtering. The full-scale motion may be extremely small, on the order of few thousandths of a centimeter.

the reluctance of the magnetic path, the main reluctance being the air gap. Any motion of the core increases the air gap on one side and decreases it on the other side, thus causing reluctance to change, in accordance with the principles explained earlier, and thereby inducing more voltage on one of the coils than on the other. Motion in the other direction reverses the action with a 180° phase shift occurring at null. The output voltage can be modified, depending on the requirements in signal processing, by means of rectification, demodulation, or filtering. In these instruments, full-scale motion may be extremely small—on the order of few thousandths of a centimeter.

In general, variable-reluctance transducers have small ranges and are used in specialized applications such as pressure transducers. Magnetic forces imposed on the armature can be quite large, and this severely limits their application. Nevertheless, the armature can be constructed as a diaphragm; hence, they become suitable for pressure measurements.

24.2.3 Variable-Reluctance Tachogenerators

Another example of a variable-reluctance sensor is shown in Figure 24.5. These sensors are based on Faraday's law of electromagnetic induction; therefore, they may also be referred to as electromagnetic sensors. Basically, the induced emf in the sensor depends on the linear or angular velocity of the motion.

The variable-reluctance tachogenerator consists of a ferromagnetic, toothed wheel attached to a rotating shaft and a coil wound onto a permanent magnet, extended by a soft iron pole piece. The wheel moves in close proximity to the pole piece, causing the flux linked by the coil to change, thus inducing an emf in the coil. The reluctance of the circuit depends on the width of the air gap between the rotating wheel and the pole piece. When the tooth is close to the pole piece, the reluctance is at

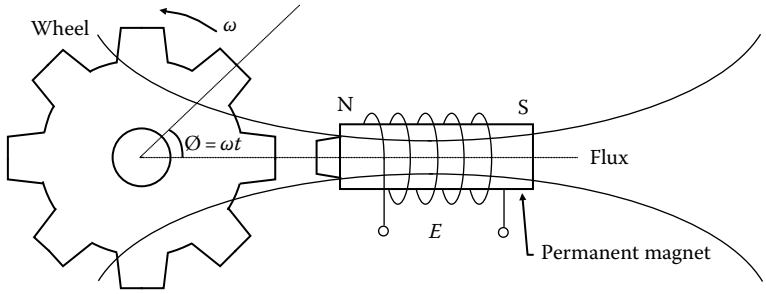


FIGURE 24.5 A variable-reluctance tachogenerator is a sensor that is based on Faraday's law of electromagnetic induction. It consists of a ferromagnetic toothed wheel attached to the rotating shaft and a coil wound onto a permanent magnet extended by a soft iron pole piece. The wheel rotates in close proximity to the pole piece, thus causing the flux linked by the coil to change. The change in flux causes an output in the coil similar to a square waveform whose frequency depends on the speed of the rotation of the wheel and the number of teeth.

its minimum, and it increases as the tooth moves away from the pole. When the wheel rotates with a velocity ω, the flux mathematically may be expressed as

$$\Psi(\theta) = A + B\cos m\theta \tag{24.11}$$

where
A is the mean flux
B is the amplitude of the flux variation
m is the number of teeth

The induced emf is given by

$$E = \frac{-d\Psi(\theta)}{dt} = -\left(\frac{d\Psi(\theta)}{d\theta}\right) \times \frac{d\theta}{dt} \tag{24.12}$$

or

$$E = bm\omega \sin m\omega t \tag{24.13}$$

Both amplitude and frequency of the generated voltage at the coil are proportional to the angular velocity of the wheel. In principle, the angular velocity ω can be found from either the amplitude or the frequency of the signal. In practice, the amplitude measured may be influenced by loading effects and electrical interference. In signal processing, the frequency is the preferred option because it can be converted into digital signals easily.

The variable-reluctance tachogenerators are most suitable for measuring angular velocities. They are also used in the volume flow rate measurements and the total volume flow determination of fluids.

24.3 Microsyn

Another commonly used example of variable-reluctance transducer is the Microsyn, as illustrated in Figure 24.6. In this arrangement, the coils are connected in such a manner that at the null position of the rotary element, the voltages induced in coils 1 and 3 are balanced by the voltages induced in coils 2 and 4.

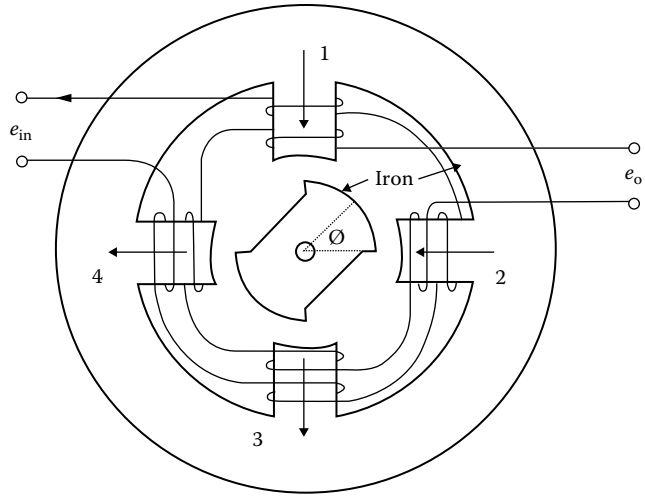


FIGURE 24.6 A Microsyn is a variable-reluctance transducer that consists of a ferromagnetic rotor and a stator carrying four coils. The stator coils are connected such that at the null position, the voltages induced in coils 1 and 3 are balanced by voltages induced in coils 2 and 4. The motion of the rotor in one direction increases the reluctance of two opposite coils while decreasing the reluctance in others, resulting in a net output voltage e_o. The movement in the opposite direction reverses this effect with a 180° phase shift.

The motion of the rotor in the clockwise direction increases the reluctance of coils 1 and 3 while decreasing the reluctance of coils 2 and 4, thus giving a net output voltage e_o. The movement in the counterclockwise direction causes a similar effect in coils 2 and 4 with a 180° phase shift. A direction-sensitive output can be obtained by using phase-sensitive demodulators, as explained in Section 24.6.

Microsyn transducers are used extensively in applications involving gyroscopes. By the use of microsyns, very small motions can be detected, giving output signals as low as 0.01° of changes in angles. The sensitivity of the device can be made as high as 5 V/degree of rotation. The nonlinearity may vary from 0.5% to 1.0% at full scale. The main advantage of these transducers is that the rotor does not have windings and slip-rings. Another advantage is that the magnetic reaction torque is also negligible.

24.4 Synchros

The term *synchro* is associated with another family of electromechanical devices that can be discussed under different headings. They are used primarily in angle measurements and are commonly applied in control engineering as parts of servomechanisms, machine tools, and antennas.

The construction of synchros is similar to that of wound-rotor induction motors, as shown in Figure 24.7. The rotation of the motor changes the mutual inductance between the rotor coil and the three stator coils. The three voltage signals from these coils define the angular position of the rotor. Synchros are used in connection with a variety of devices, including control transformers, Scott T transformers, resolvers, phase-sensitive demodulators, and analog-to-digital converters.

In some cases, a control transformer is attached to the outputs of the stator coils such that the output of the transformer produces a resultant mmf aligned in the same direction as that of the rotor of the synchro. In other words, the synchro rotor acts as a search coil in detecting the direction of the stator field of the control transformer. When the axis of this coil is aligned with the field, the maximum voltage is supplied to the transformer.

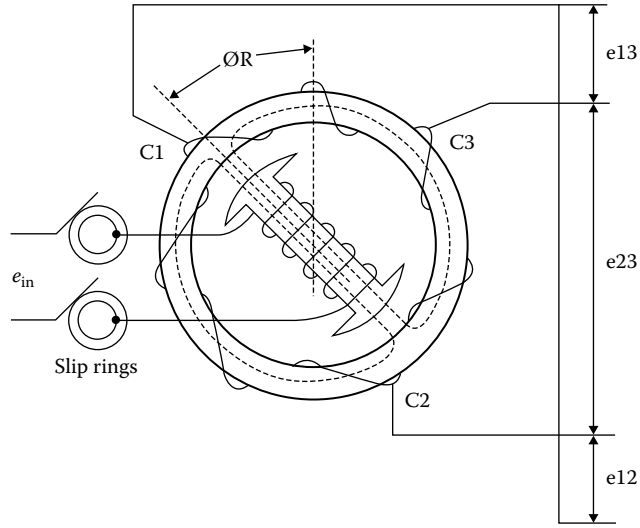


FIGURE 24.7 A synchro is similar to a wound-rotor induction motor. The rotation of the rotor changes the mutual inductance between the rotor coil and the stator coils. The voltages from these coils define the angular position of the rotor. They are primarily used in angle measurements and are commonly applied in control engineering as parts of servomechanisms, machine tools, and antennas.

In other cases, ac signals from the synchros are first applied to a Scott T transformer, which produces ac voltages with amplitudes proportional to the sine and cosine of the synchro's shaft angle. It is also possible to use phase-sensitive demodulations to convert the output signals to make suitable for the purpose of digital signal processing.

24.5 Moving-Coil Inductive Displacement Sensors

When the coil of the sensor rotates with a stable speed, it can pick up non-varying magnetic fields, but still comply with Faraday's Law of Induction. In some cases, the coil is made vibrate rather than rotate. Due to possibility of high-frequency vibrations, it is possible to make very small sensors with good geometric resolutions. Moving-coil inductive displacement sensors find limited applications in digital fluxmeters along with other sensors such as Hall and fluxgate sensors.

24.5.1 Linear-Variable Inductor

Mathematically, the principles of linear-variable inductors are very similar to the variable-reluctance type of transducer as explained at the beginning of the chapter. Also, in physical construction, there is a little distinction between variable-reluctance and variable-inductance transducers. The main difference is the way that the signals are picked up rather than the principles of operations. A typical linear-variable inductor consists of a movable iron core that provides the mechanical inputs and two coils forming two legs of an electrical bridge network. A typical example of such a transducer is the variable-coupling transducer, which is discussed next.

24.5.2 Variable-Coupling Transducers

These transducers consist of a former holding, a center-tapped coil, and a ferromagnetic plunger, as shown in Figure 24.8.

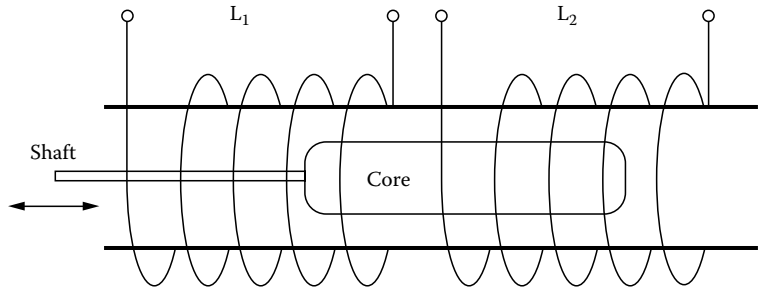


FIGURE 24.8 A typical linear-variable inductor consists of a movable iron core inside a former holding a center-tapped coil. The core and both of the coils have the same length L. When the core is in the reference position, each coil will have equal inductances of value L. As the core moves by δl, changes in inductances $+\delta L$ and $-\delta L$ create voltage outputs from the coils.

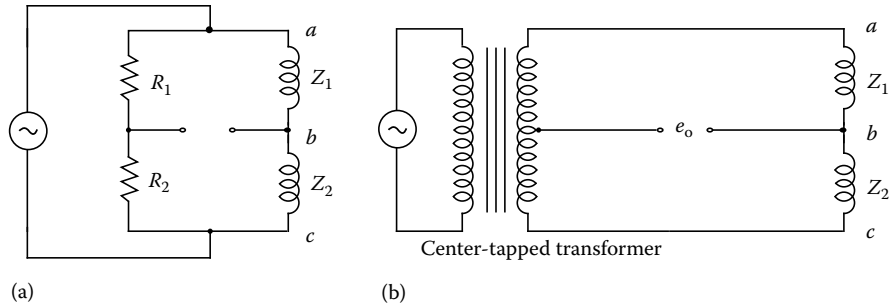


FIGURE 24.9 The two coils of a linear-variable inductor are usually placed to form two arms of a bridge circuit, also having two equal balancing resistors as in circuit (a). The bridge is excited with ac of 5–25 V with a frequency of 50 Hz–5 kHz. At a selected excitation frequency, the total transducer impedance at null conditions is set in the 100–1000 Ω range. By careful construction of the bridge, the output voltage can be made a linear function displacement of the core within a limited range. In some cases, in order to reduce power losses due to heating of resistors, center-tapped transformers may be used as a part of the bridge network (b).

The plunger and the two coils have the same length L. As the plunger moves, the inductances of the coils change. The two inductances are usually placed to form two arms of a bridge circuit with two equal balancing resistors, as shown in Figure 24.9a. The bridge is excited with ac of 5–25 V with a frequency of 50 Hz–5 kHz. At the selected excitation frequency, the total transducer impedance at null conditions is set in the 100–1000 Ω range. The resistors are set to have about the same value as transducer impedances. The load for the bridge output must be at least 10 times the resistance, R, value. When the plunger is in the reference position, each coil will have equal inductances of value L. As the plunger moves by δl, changes in inductances $+\delta L$ and $-\delta L$ creates a voltage output from the bridge. By constructing the bridge carefully, the output voltage can be made as a linear function displacement of the moving plunger within a rated range.

In some transducers, in order to reduce power losses due to heating of resistors, center-tapped transformers can be used as a part of the bridge network, as shown in Figure 24.9b. In this case, the circuit becomes more inductive and extra care must be taken to avoid the mutual coupling between the transformer and the transducer.

It is particularly easy to construct transducers of this type, by simply winding a center-tapped coil on a suitable former. The variable-inductance transducers are commercially available in strokes from

about 2 mm to 500 cm. The sensitivity ranges between 1.0% full scale to 0.02% in long stroke special constructions. These devices are also known as linear displacement transducers or LDTs, and they are available in various shape and sizes.

Apart from linear-variable inductors, there are rotary types available too. Their cores are specially shaped for rotational applications. Their nonlinearity can vary from 0.5% to 1.0% full scale over a range of 90° rotation. Their sensitivity can be up to 100 mV/degree of rotation.

24.5.3 Induction Potentiometer

One version of a rotary type linear inductor sensor is the induction potentiometer, shown in Figure 24.10. Two concentrated windings are wound on the stator and the rotor. The rotor winding is excited with an ac, thus inducing voltage in the stator windings. The amplitude of the output voltage is dependent on the mutual inductance between the two coils, where mutual inductance itself is dependent on the angle of rotation. For concentrated coil type induction potentiometers, the variation of the amplitude is sinusoidal, but linearity is restricted in the region of the null position. A linear distribution over an angle of 180° can be obtained by the construction of carefully designed distributed coils.

Standard commercial induction pots operate in a 50–400 Hz frequency range. They are small in size, from 1 to 6 cm, and their sensitivity can be in the order of 1 V/degree of rotation. Although the ranges of induction potentiometers are limited to less than 60° of rotation, it is possible to measure displacements in angles from 0° to full rotation by suitable arrangement of a number of induction pots. As in the case of many inductive sensors, the output of an induction pot may need phase-sensitive demodulators and suitable filters. In most cases, additional dummy coils are used to improve linearity and accuracy.

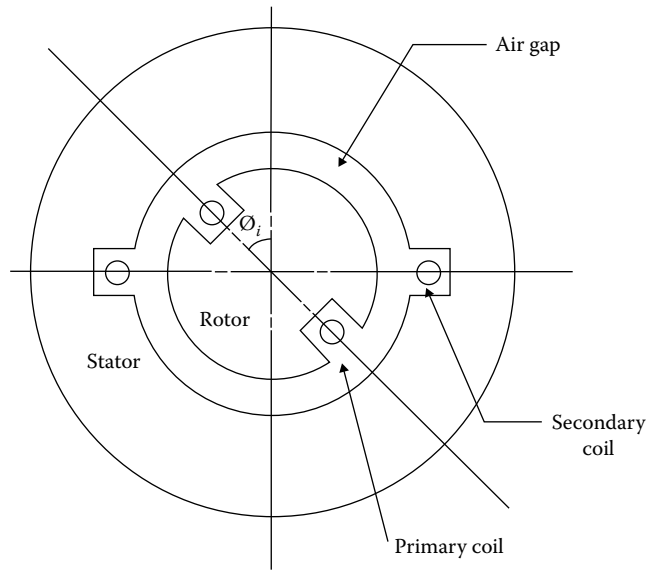


FIGURE 24.10 An induction potentiometer is a linear-variable inductor with two concentrated windings wound on the stator and on the rotor. The rotor winding is excited with ac, inducing voltage in the stator windings. The amplitude of the output voltage is dependent on the relative positions of the coils, as determined by the angle of rotation. For concentrated coils, the variation of the amplitude is sinusoidal, but linearity is restricted in the region of the null position. Different variations of induction potentiometers are available with distributed coils that give linear voltages over an angle of 180° of rotation.

24.6 Linear-Variable-Differential Transformer

The linear-variable-differential transformer (LVDT) is a passive inductive transducer. They find diverse range of applications. An LVDT consists of a single primary winding positioned between two identical secondary windings wound on a tubular ferromagnetic former, as shown in Figure 24.11. The primary winding is energized by a high frequency from 50 Hz to 20 kHz ac voltage. The two secondary windings are made identical by having an equal number of turns and similar geometry. They are connected in series opposition so that the induced output voltages oppose each other.

In many applications, the outputs are connected in opposing form, as shown in Figure 24.12a. The output voltages of individual secondaries v_1 and v_2 at null position are illustrated in Figure 24.12b. However, in opposing connection, any displacement in the core position x from the null point causes amplitude of the voltage output v_o and the phase difference α to change. The output waveform v_o in relation to core position is shown in Figure 24.12c. When the core is positioned in the middle, there is an equal coupling between the primary and secondary windings, thus giving a null point or reference point of the sensor. As long as the core remains near the center of the coil arrangement, output is linear. The linear ranges of commercial differential transformers are clearly specified, and the devices are seldom used outside this linear range.

The ferromagnetic core or plunger moves freely inside the former, thus altering the mutual inductance between the primary and secondaries. With the core in the center, or at the reference position, the induced emfs in the secondaries are equal, and since they oppose each other, the output voltage is zero. When the core moves, say to the left, from the center, more magnetic flux links with the left-hand coil than with the right-hand coil. The voltage induced in the left-hand coil is therefore larger than the induced voltage on the right-hand coil. The magnitude of the output voltage is then larger than at the null position and is equal to the difference between the two secondary voltages. The net output voltage is in phase with the voltage of the left-hand coil. The output of the device is then an indication of the displacement of the core. Similarly, movement in the opposite direction to the right from the center reverses this effect, and the output voltage is now in phase with the emf of the right-hand coil.

For mathematical analysis of the operation of LVDTs, Figure 24.12a can be used. The voltages induced in the secondary coils are dependent on the mutual inductance between the primary and

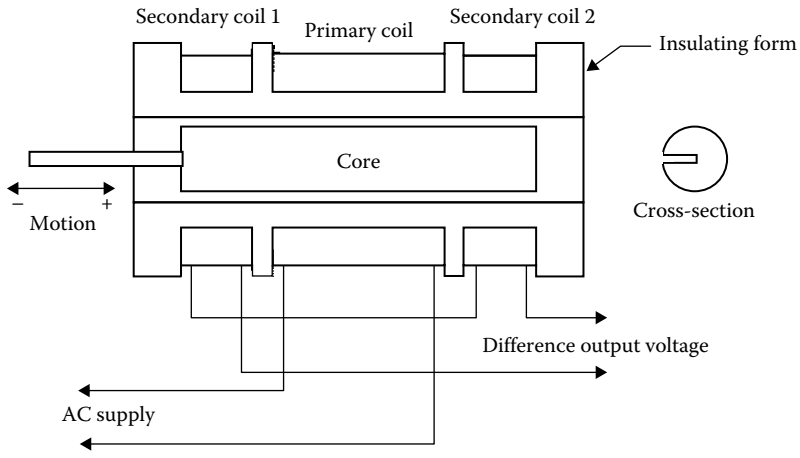


FIGURE 24.11 A LVDT is a passive inductive transducer consisting of a single primary winding positioned between two identical secondary windings wound on a tubular ferromagnetic former. As the core inside the former moves, the magnetic paths between primary and secondaries change, thus giving secondary outputs proportional to the movement. The two secondaries are made as identical as possible by ensuring equal sizes, shapes, and number of turns.

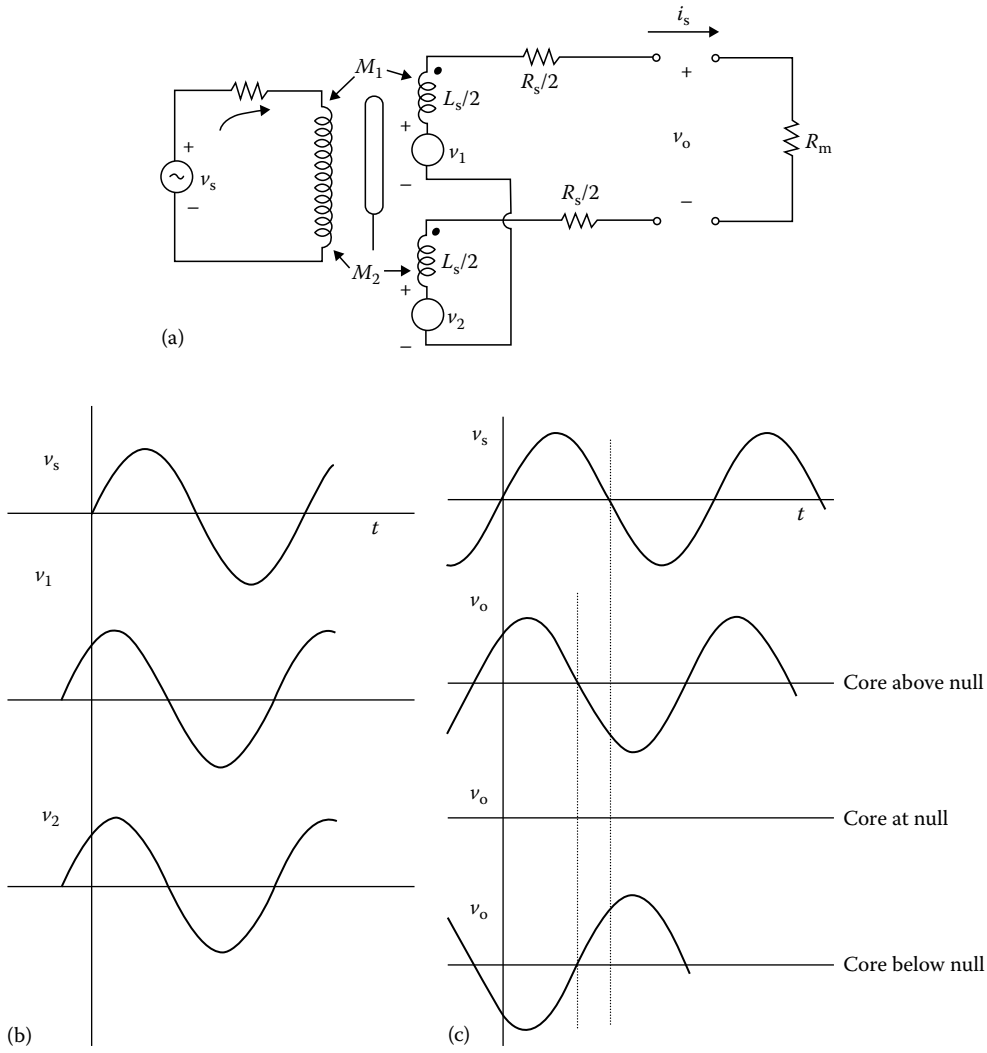


FIGURE 24.12 The voltages induced in the secondaries of an LVDT (a) may be processed in a number of ways. The output voltages of individual secondaries v_1 and v_2 at null position are illustrated in (b). In this case, the voltages of individual coils are equal and in phase with each other. Sometimes, the outputs are connected opposing each other, and the output waveform v_o becomes a function of core position x and phase angle α, as in (c). Note the phase shift of 180° as the core position changes above and below the null position.

individual secondary coils. Assuming that there is no cross coupling between the secondaries, the induced voltages may be written as

$$v_1 = M_1 s i_p \quad \text{and} \quad v_2 = M_2 s i_p \tag{24.14}$$

where

M_1 and M_2 are the mutual inductances between primary and secondary coils for a fixed core position
s is the Laplace operator
i_p is the primary current

In the case of opposing connection, no load output voltage v_o without any secondary current may be expressed as

$$v_o = v_1 - v_2 = (M_1 - M_2)si_p \tag{24.15}$$

writing

$$v_s = i_p(R + sL_p) \tag{24.16}$$

Substituting i_p in Equation 24.15 gives the transfer function of the transducer as

$$\frac{v_o}{v_s} = \frac{(M_1 - M_2)s}{R + sL_p} \tag{24.17}$$

However, if there is a current due to output signal processing, then describing equations may be modified as

$$v_o = R_m i_s \tag{24.18}$$

where

$$i_s = \frac{(M_1 - M_2)si_p}{R_s + R_m + sL_s}$$

and

$$v_s = i_p(R + sL_p) - (M_1 - M_2)si_s \tag{24.19}$$

Eliminating i_p and i_s from Equations 24.19 and 24.20 results in a transfer function:

$$\frac{v_o}{v_s} = \frac{R_m(M_1 - M_2)s}{[(M_1 - M_2)^2 + L_sL_p]s^2 + [L_p(R + R_m) + RL_s]s + (R_s + R_m) + R} \tag{24.20}$$

This is a second-order system, which indicates that due to the effect of the numerator of Equation 24.20, the phase angle of the system changes from +90° at low frequencies to −90° at high frequencies. In practical applications, the supply frequency is selected such that at the null position of the core, the phase angle of the system is 0°.

The amplitudes of the output voltages of secondary coils are dependent on the position of the core. These outputs may directly be processed from each individual secondary coil for slow movements of the core and when the direction of the movement of the core does not bear any importance. However, for fast movements of the core, the signals can be converted to dc and the direction of the movement from the null position can be detected. There are many options to do this; however, a *phase-sensitive demodulator* and filter arrangement are commonly used, as shown in Figure 24.13a. A typical output of the phase-sensitive demodulator is illustrated in Figure 24.13b, in relation to output voltage v_o, displacement x, and phase angle α.

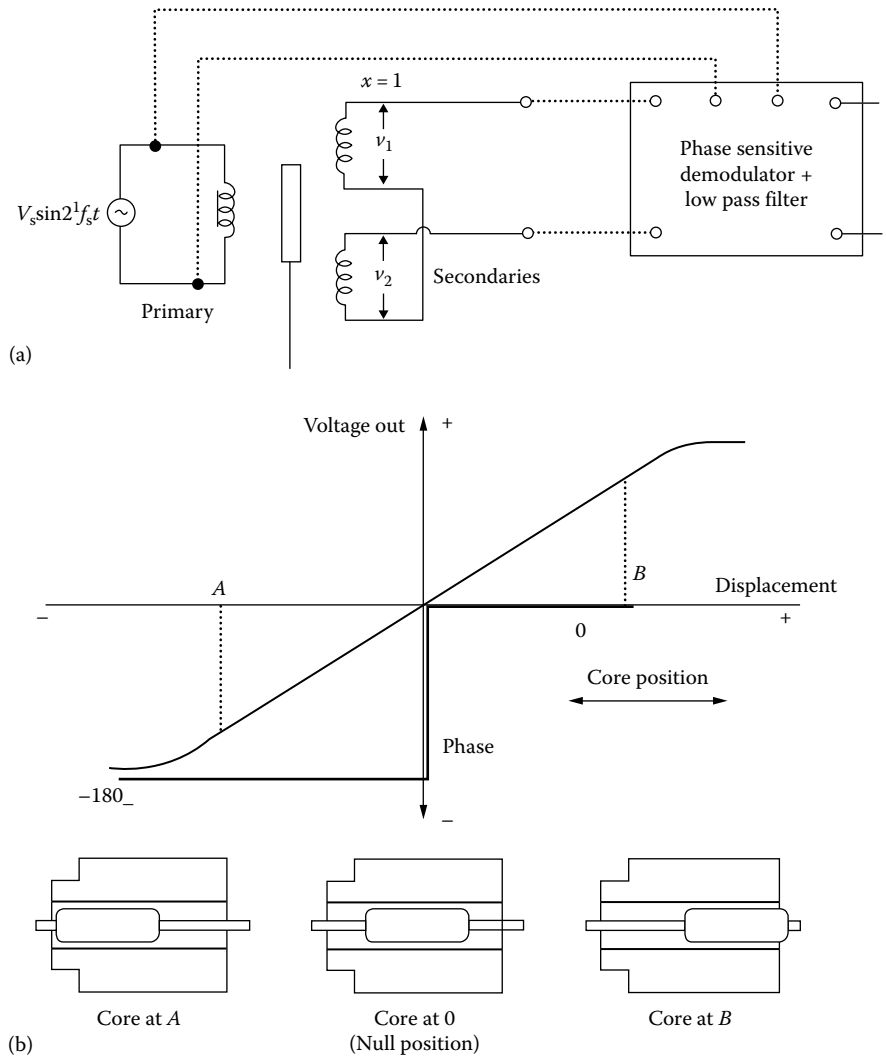


FIGURE 24.13 Phase-sensitive demodulator and (a) are commonly used to obtain displacement proportional signals from LVDTs and other differential-type inductive sensors. They convert the ac outputs from the sensors into dc values and also indicate the direction of movement of the core from the null position. A typical output of the phase-sensitive demodulator is shown in (b). The relationship between output voltage v_o and phase angle α is also shown against core position x.

The phase-sensitive demodulators are used extensively in differential-type inductive sensors. They basically convert the ac outputs to dc values and also indicate the direction of movement of the core from the null position. A typical phase-sensitive demodulation circuit may be constructed, based on diodes shown in Figure 24.14a. This arrangement is useful for very slow displacements, usually less than 1 or 2 Hz. In this figure, bridge 1 acts as a rectification circuit for secondary 1, and bridge 2 acts as a rectifier for secondary 2. The net output voltage is the difference between the outputs of two bridges, as in Figure 24.14b. The position of the core can be determined from the amplitude of the dc output, and the direction of the movement of the core can be determined from the polarity of the dc voltage. For rapid movements of the core, the outputs of the diode bridges need to be filtered, wherein only the frequencies of the movement of the core pass through and all the other frequencies produced by the modulation process are filtered. For this purpose, a suitably designed simple *RC* filter may be sufficient.

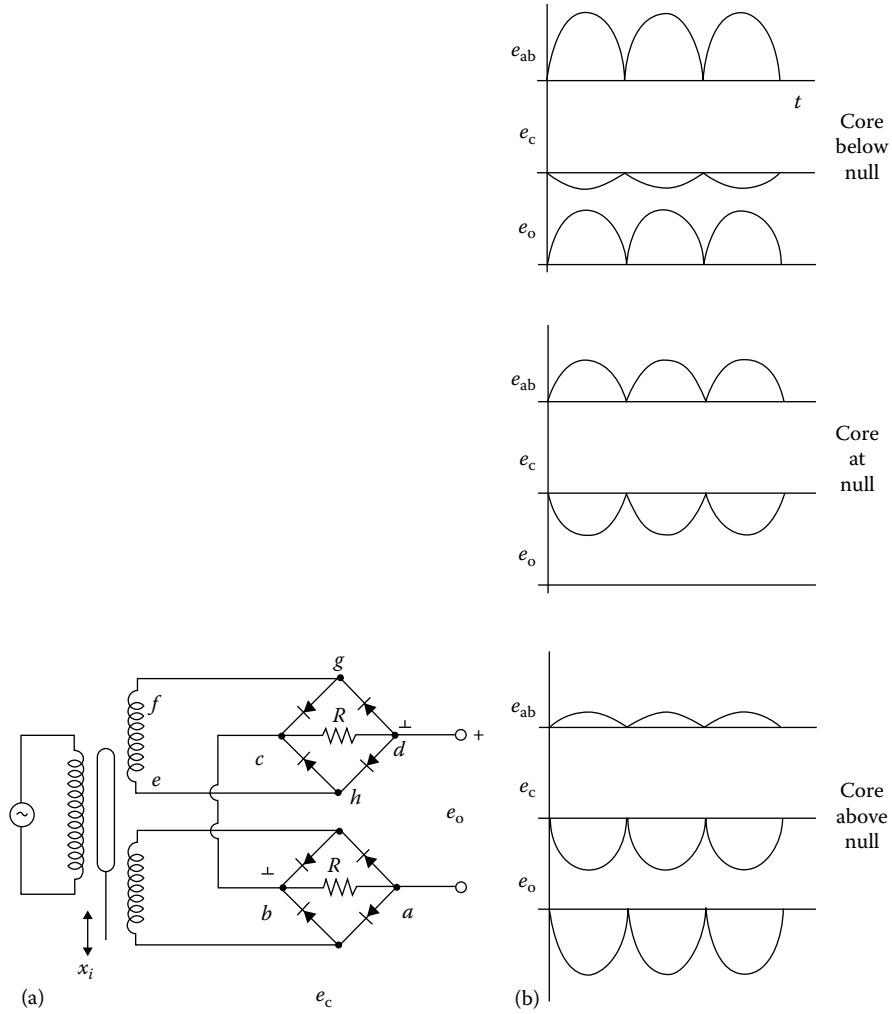


FIGURE 24.14 A typical phase-sensitive demodulation circuit based on diode bridges as in (a). Bridge 1 acts as a rectification circuit for secondary 1, and bridge 2 acts as a rectifier for secondary 2 where the net output voltage is the difference between the two bridges, as in (b). The position of the core can be determined from the amplitude of the dc output, and the direction of the movement of the core can be determined from the polarity of the voltage. For rapid movements of the core, the output of the diode bridges must be filtered; for this, a suitably designed simple *RC* filter may be sufficient.

There are phase-sensitive demodulator chips available in the marketplace, such as AD630 offered by Analog Devices Inc. These chips are highly versatile and flexible to use to suit particular application requirements. These chips offer many advantages over conventional phase-sensitive demodulation devices; for example, frequency off excitation may be adjusted to any value between 20 Hz and 20 kHz by connecting an external capacitor between two pins. The amplitude of the excitation voltage can be set up to 24 V. The internal filters may be set to required values by external capacitors. Connections to analog-to-digital converters are easily made by converting the bipolar output to unipolar scale.

The frequency response of LVDTs is primarily limited by the inertia characteristics of the device. In general, the frequency of the applied voltage should be 10 times the desired frequency response. Commercial LVDTs are available in a broad range of sizes, and they are widely used for displacement measurements in a variety of applications. These displacement sensors are available to cover ranges

from ±0.25 mm to ±7.5 cm. They are sensitive enough to be used to respond to displacements well below 0.0005 mm. They have operational temperature ranges from –265 °C to 600 °C. They are also available in radiation-resistant designs for operation in nuclear reactors. For a typical sensor of ±25 mm range, the recommended supply voltage is 4–6 V, with a nominal frequency of 5 kHz and a maximum nonlinearity of 1% full scale. Several commercial models are available that can produce a voltage output of 300 mV for 1 mm displacement of the core.

One important advantage of the LVDT is that there is no physical contact between the core and the coil form and hence no friction or wear. Nevertheless, there are radial and longitudinal magnetic forces on the core at all times. These magnetic forces may be regarded as magnetic springs that try to displace the core to its null position. This may be a critical factor in some applications.

A problem with LVDTs is that it may not be easy to make the two halves of the secondary identical; their inductance, resistance, and capacitance may be different, causing a large unwanted quadrature output in the balance position. Precision coil winding equipment may be required to reduce this problem to an acceptable value.

Another problem is associated with null position adjustments. The harmonics in the supply voltage and stray capacitances result in small null voltages. The null voltage may be reduced by proper grounding, which reduces the capacitive effects and center-tapped voltage source arrangements. In center-tapped supplies, a potentiometer may be used to obtain a minimum null reading.

The LVDTs have a variety of applications, including control for jet engines in close proximity to exhaust gases and measuring roll positions in the thickness of materials in hot-slab steel mills. Force and pressure measurements may also be made by LVDTs after some mechanical modifications.

24.7 Rotary-Variable-Differential Transformer

A variation from the LVDT is the rotary core differential transformer shown in Figures 24.15a and b. Here, the primary winding is wound on the center leg of an E core, and the secondary windings are wound on the outer legs of the core. The armature is rotated by an externally applied force about a pivot point above the center leg of the core. When the armature is displaced from its reference, the reluctance of the magnetic circuit through one of the secondary coils decreases and, simultaneously, increases the reluctance through the other coil. The induced emfs in the secondary windings, which are equal in the reference position of the armature, are now different in magnitude and phase as a result of the applied displacement. The induced emfs in the secondary coils are made to oppose each other, and the transformer operates in the same manner as an LVDT. The rotating variable transformers may be sensitive to vibrations. If a dc output is required, a demodulator network can be used, as in the case of LVDTs.

In most rotary LVDTs, the rotor mass is very small, usually less than 5 g. The nonlinearity in the output ranges between ±1% and ±3%, depending on the angle of rotation. The motion in the radial direction produces a small output signal that can affect the overall sensitivity. However, this transverse sensitivity is usually kept below 1% of the longitudinal sensitivity.

24.8 Eddy Current Proximity Sensors

Inductive transducers based on eddy currents are mainly probe types, containing two coils as shown in Figure 24.16. One of the coils, known as the active coil, is influenced by the presence of the conducting target. The second coil, known as the balance coil, serves to complete the bridge circuit and provides temperature compensation. The magnetic flux from the active coil passes into the conductive target by means of a probe. When the probe is brought close to the target, the flux from the probe links with the target produces eddy currents within the target.

The eddy current density is greatest at the target surface and becomes negligibly small at about three skin depths below the surface. The skin depth depends on the type of material used and the excitation frequency. While thinner targets can be used, a minimum of three skin depths is often necessary to

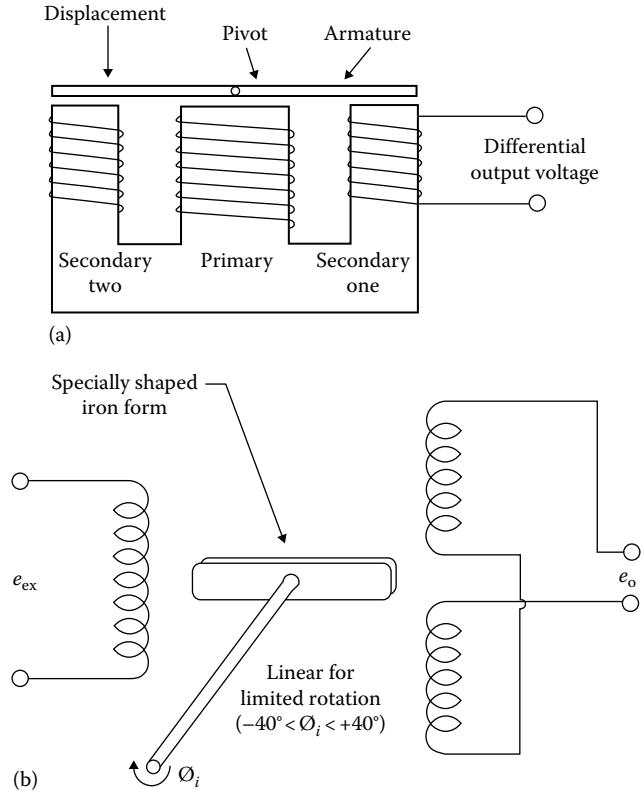


FIGURE 24.15 A rotary core differential transformer has an E-shaped core, carrying the primary winding on the center leg and the two secondaries on the outer legs, as in (a). The armature is rotated by an externally applied force about a pivot point above the center leg of the core (b). When the armature is displaced from its reference or balance position, the reluctance of the magnetic circuit through one secondary coil is decreased, increasing the reluctance through the other coil. The induced emfs in the secondary windings are different in magnitude and phase as a result of the applied displacement.

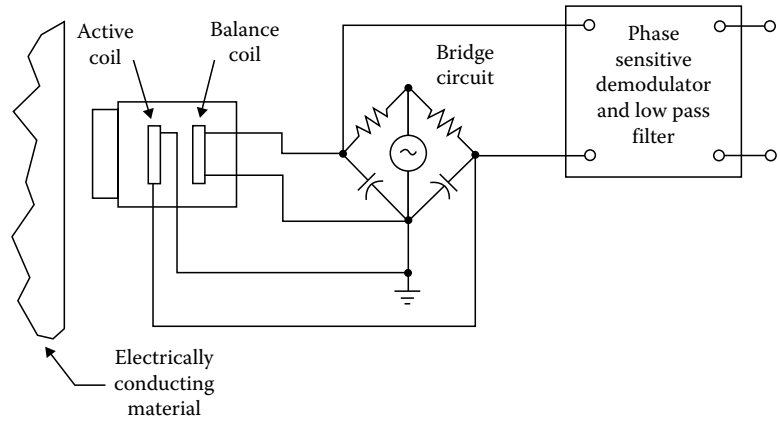


FIGURE 24.16 Eddy current transducers are inductive transducers using probes. The probes contain one active and one balance coil. The active coil responds to the presence of a conducting target, while the balance coil completes a bridge circuit and provides temperature compensation. When the probe is brought close the target, the flux from the probe links with the target, producing eddy currents within the target that alter the inductance of the active coil. This change in inductance is detected by a bridge circuit.

minimize the temperature effects. As the target comes closer to the probe, the eddy currents become stronger, causing the impedance of the active coil to change and thus altering the balance of the bridge in relation to the target position. This unbalance voltage of the bridge may be demodulated, filtered, and linearized to produce a dc output proportional to target displacement. The bridge oscillation may be as high as 1 MHz. High frequencies allow the use of thin targets and provide a good system frequency response.

Probes are commercially available with full-scale diameter ranging from 0.25 mm to 30 mm with a nonlinearity of 0.5% and a maximum resolution of 0.0001 mm. Targets are usually supplied by the clients, involving noncontact measurements of machine parts. For nonconductive targets, conductive materials of sufficient thickness must be attached to the surface by means of commercially available adhesives. Since the target material, shape, etc., influence the output, it is necessary to calibrate the system statistically for a specific target. The recommended measuring range of a given probe begins at a standoff distance equal to about 20% of the stated range of the probe. In some cases, a standoff distance of 10% is recommended for which the system is calibrated as standard. A distance greater than 10% of the measuring range can be used as long as the calibrated range is reduced by the same amount.

Flat targets must be of the same diameter as the probe or larger. If the target diameter is smaller than the probe diameter, the output drops considerably, thus becoming unreliable. Curved-surface targets can behave similar to flat surfaces if the diameter exceeds about three or four times the diameter of the probe. In this case, the target essentially becomes an infinite plane. This also allows some cross-axis movement without affecting the system output. Target diameters comparable to the sensor could result in detrimental effects in measurements due to cross-axis movements.

For curved or irregular-shaped targets, the system must be calibrated using the exact target that is seen in the operation. This tends to eliminate any errors caused by the curved surfaces during application. However, special multiprobe systems are available for orbital motions of rotating shafts. If the curved (shaft) target is about 10 times larger than the sensor diameter, it acts as an infinite plane and does not need special calibration. Care must be exercised to deal with electrical runout due to factors such as inhomogeneities in hardness, particularly valid for ferrous targets. However, nonferrous targets are free from electrical runout concerns.

24.9 Gradiometer-Inductive Displacement Sensors

Gradiometer sensors are used when the external magnetic field is generated by a large distant source. If, for example, two identical coils are placed in a uniform magnetic field, both will sense the same magnetic field. When both coils are connected differentially, the influence of the external (ambient) field will be eliminated. At the same time, if there is a smaller magnetic field due to a second source, the magnetic coil placed nearer the second source will be subjected larger field; hence, the differential output will change. This differential change can be processed suitably to detect the local source relative to the larger one. Gradiometer sensors can be made to detect variations in magnetic field of few femtotesla in the presence of few millitesla uniform ambient fields.

24.10 Miniature-Inductive Displacement Sensors

Recent developments in microchip technology allow the formation of inductive displacement sensors at the chip level. A typical configuration consists of multiple-coil system as illustrated in Figure 24.17. A structure with three excitation and three coils allows the corrections for the angular misalignments and lateral displacements. A typical application is in the implantable sensors to monitor activities in humans. They can achieve a sensitivity 5 μV/10 mm displacements. They can be as small as 1 mm in diameter 4 mm in length.

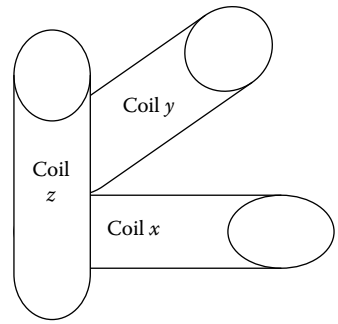


FIGURE 24.17 An implantable three-coil inductive sensor. A typical system consists of transmitter coil, which generates magnetic field and sensing coil that measures the field amplitude that depends on the distance and the mutual position between the coils.

24.11 Shielding and Sensitivity of Inductive Sensors to Electromagnetic Interference

Magnetic fields are produced by currents in wires and more strongly by the coils. The fields due to coils are important due to magnetic coupling, particularly when there are two or more coils in the circuit. The magnetic coupling between coils may be controlled by large spacing between coils, orientation of coils, and the shape of the coils and by shielding.

Inductive sensors come in different shapes and sizes. While some sensors have closed cores such as toroidal shapes, others have open cores and air gaps between cores and coils. Closed cores can have practically zero external fields, except for small leakage fluxes. Even if the sensors do not have closed cores, most variable inductor sensors have rather limited external fields, due to two neighboring sets of coils connected in opposite directions that minimize the external fields.

Inductive sensors are made from closed conductors. This implies that if the conductor moves in a magnetic field, a current will flow. Alternatively, a magnetic change produces current in a stationary closed conductor. Unless adequate measures are taken, there may be external magnetic fields linking (interference) with the sensor coils, thus producing currents and unwanted responses.

Due to inherent operations, inductive sensors are designed to have high sensitivity to magnetic flux changes. External electromagnetic interference and external fields can seriously affect the performance of the sensors. It is known that moderate magnetic fields are found near power transformers, electrical motors, and power lines. These small fields produce current in the inductive sensor elements. One way of eliminating external effects is accomplished by magnetic shielding of the sensors and by grounding appropriately. In magnetic shielding, one or more shells of high-permeability magnetic materials surround the part to be shielded. Multiple shells may be used to obtain very complete shielding. The ends of each individual shell are separated by insulation so that the shell does not act as a single shorted turn, thus accommodating high current flows. Similarly, in the case of multiple shielding, shells are isolated from each other by proper insulation.

Alternating magnetic fields are also screened by interposing highly conductive metal sheets such as copper or aluminum on the path of the magnetic flux. The eddy currents induced in the shield give a counter mmf that tends to cancel the interfering magnetic field. This type of shielding is particularly effective at high frequencies. Nevertheless, appropriate grounding must still be observed.

In many inductive sensors, stray capacitances can be a problem, especially at the null position of the moving core. If the capacitive effect is greater than a certain value, say 1% of the full-scale output, this effect may be reduced by the use of center-tapped supply and appropriate grounding.

Bibliography

Bentley, J. P., *Principles of Measurement Systems*, 2nd edn., Longman Scientific and Technical, Essex, U.K., 1988.

Chunhong, H. and Zhenghong, Q., Application of magnetic sensors in automation control, *J. Phys.: Conf. Ser*., 263, 012003, 2011.

Gatzen, H. H., Anreeva, E., and Iswahjudi, H., Eddy-current microsensors based on thin film technology, *IEEE Trans. Magn*., 38, 3368–3370, 2002.

Doebelin, E. O., *Measurement Systems: Application and Design*, 4th edn., McGraw-Hill, Singapore, 1990.

Duc, N. H. and Giang, D. T. H., Magnetic sensors based on piezoelectric-magnetostrictive composites, *J. Alloys Compd*., 449(1–2), 214–218, 2008.

Holman, J. P., *Experimental Methods for Engineers*, 5th edn., McGraw-Hill, Singapore, 1989.

Magnes, W. and Diaz-Michelana, M., Future directions for magnetic sensors for space applications, *IEEE Trans. Magn*., 45(10), 4493–4498, 2009.

Nippon Steel Technical Report, A survey method using magnetic sensors to detect the contact locations of other metal structures on pipelines, Nippon Steel Corp., No. 92, 2005, pp. 95–96.

Ripka, P., *Induction Sensors Magnetic Sensors and Magnetometers*, Artech House, Boston, MA, 2001, pp, 47–74.

Ripka, P. and Janosek, M., Advances in magnetic sensors, *Proceedings of the IEEE Sensors 2008*, Piscataway, NJ, 2008, pp. 1–4.

Partial List of Manufacturers and Suppliers

American Aerospace Controls, Inc. (AAC), 570A Smith St., Farmingdale, NY 11735, Phone: 631-694-5100, 888-873-8559 (toll free), Fax: 631-694-6739, www.a-a-c.com

AutomationDirect.com (AutomationDirect), 3505 Hutchinson Rd., Cumming, GA 30040-5860, Phone: 770-889-2858, 800-633-0405 (toll free), Fax: 770-889-7876, www.automationdirect.com

Automation Products Group, Inc. (APGSENSORS), 1025 W. 1700 N., Logan, UT 84321, Phone: 435-753-7300, 888-525-7300 (toll free), Fax: 435-753-7490, www.apgsensors.com

Autonics USA, 1353 Armour Blvd., Mundelein, IL 60060, Phone: 847-680-8160, 866-862-2602 (toll free), Fax: 847-680-8155, www.autonicsonline.com

Banner Engineering Corp., 9714 Tenth Ave. N., Minneapolis, MN 55440, Phone: 763-544-3164, 877-581-1749 (toll free), Fax: 763-544-3213, www.bannerengineering.com/en-US/

Baumer Ltd., 122 Spring St. Suite C-6, Southington, CT 06489-1534, Phone: 860-621-2121, 800-937-9336 (toll free), Fax: 860-628-6280, www.baumerelectric.com

Fargo Controls, Inc., P.O. Box 539, Eatontown, NJ 07724, Phone: 732-389-3376, 800-241-3755 (toll free), Fax: 732-542-3553, www.fargocontrols.com

Honeywell Sensing and Control, 1985 Douglas Dr., North, MN 10-192B Golden Valley, MN 55422, Phone: 763-954-4818, 800-446-6555 (toll free), www.honeywell.com/

IFM Efector, Inc., 782 Springdale Dr., Exton, PA 19341-2843, Phone: 610-524-2000, 800-441-8246 (toll free), Fax: 800-329-0436, www.ifm.com/ifmus/web/home.htm

ISSC-Kanson Electronics, Inc., 245 Forrest Ave., Hohenwald, TN 38462, Phone: 931-796-3050, 800-233-9354 (toll free), Fax: 931-796-3956, www.issc-kanson.com/

Leuze Electronic, Inc., 55395 Lyon Industrial Dr., New Hudson, MI 48165, Phone: 248-486-4466 Fax: 248-486-6699, www.leuzeusa.com

Macro Sensors, Division of Howard A. Schaevitz Technologies, 7300 US Rte. 130 N. Bldg. 22, Pennsauken, NJ 08110, Phone: 856-662-8000, Fax: 856-317-1005, www.macrosensors.com

MTI Instruments, Inc., 325 Washington Ave., Extension Albany, NY 12205-5505, Phone: 518-218-2550, Fax: 518-218-2506, www.mtiinstruments.com

OTP Industrial Solutions, 1900 Jetway Blvd., Columbus, OH 43219, Phone: 614-342-6123, Fax: 614-342-6490, www.otpnet.com/

Pilz Automation Safety L.P., 7150 Commerce Blvd., Canton, MI 48187, Phone: 866-889-1726 (toll free), Fax: 734-354-3355, www.pilz.us/index.jsp

Schneider Electric-Hyde Park, 1875 Founders Dr., Dayton, OH 45420, Phone: 937-514-7207, Fax: 937-258-5830, www.sesensors.com

SICK, Inc., 6900 W. 110th St., Bloomington, MN 55438, Phone: 952-941-6780, 800-325-7425 (toll free), Fax: 952-941-9287, www.sick.com/us/en-us/home/Pages/Homepag…

Smith Systems, Inc., 6 Mill Creek Center, P.O. Box 667, Brevard, NC 28712, Phone: 828-884-3490, Fax: 828-877-3100, www.smith-systems-inc.com

SoftNoze USA, Inc., 2216 Broad St., Frankfort, NY 13340-5100, Phone: 315-732-2726, Fax: 315-732-2963, www.softnoze.com

Spectec, 9 Polaris Way, P.O. Box 360, Emigrant, MT 59027, Phone: 406-333-4967, 800-735-9373 (toll free), Fax: 406-333-4259, www.spectecsensors.com

Turck, Inc., 3000 Campus Dr., Plymouth, MN 55441-2619, Phone: 763-553-7300, 800-554-4989 (toll free), Fax: 763-553-0708, www.turck.com

25

Capacitive Sensors: Displacement, Humidity, Force

25.1 Introduction **25**-1
Basic Principles of Capacitive Displacement Sensors
25.2 Variable Distance Displacement Sensors **25**-2
25.3 Variable Area Displacement Sensors **25**-3
25.4 Variable Dielectric Displacement Sensors **25**-3
25.5 Differential Capacitive Displacement Sensors **25**-4
25.6 Integrated Circuit Smart Capacitive Displacement Sensors **25**-5
25.7 Capacitive Pressure Sensors **25**-7
25.8 Capacitive Accelerometers and Force Transducers **25**-8
25.9 Capacitive Liquid-Level Measurement **25**-9
25.10 Capacitive Humidity and Moisture Sensors **25**-10
Aluminum Capacitive Humidity Sensors • Tantalum Capacitive Humidity Sensors • Silicon Capacitive Humidity Sensors • Polymer Capacitive Humidity Sensors • Capacitive Moisture Sensors
25.11 Signal Processing **25**-14
Operational Amplifier and Charge Amplifier-Based Circuits • Pulse Width Modulation • Square-Wave Linearization • Feedback Linearization • Oscillator Circuits
Bibliography **25**-17
Partial List of Manufacturers and Suppliers **25**-18

Halit Eren
Curtin University

25.1 Introduction

Capacitive sensors are based on changes in capacitance in response to physical variations. Capacitive displacement sensors are noncontact devices yielding high-resolution measurements of the position and/or change of position. They are used in a wide variety of applications such as assembly of equipment, disk drives, thickness measurements, machine tools, semiconductor devices, machining, and in some cases in density measurements. In some cases, the basic operational and sensing principles are common in dissimilar applications, and in other cases, different principles can be used for the same applications. For example, capacitive microphones are based on variations of spacing between plates in response to acoustical pressure, thus turning audio signals to variations in capacitance. On the other hand, a capacitive level indicator makes use of the changes in the relative permittivity between the plates.

However, capacitive sensors are best known to be applicable with displacement measurements for rotational or translational motions, as will be described in this chapter. Other applications of capacitance sensors such as humidity and moisture will also be discussed.

25.1.1 Basic Principles of Capacitive Displacement Sensors

The measurement of distances or displacements is an important aspect of many industrial, scientific, and engineering systems. The displacement is basically the vector representing a change in the position of a body or point with respect to a reference point. Capacitive displacement sensors satisfy the requirements of many applications where high linearity and wide ranges (from a few centimeters to a couple of nanometers) are needed.

The basic sensing element of a typical displacement sensor consists of two parallel electrodes with capacitance C. The capacitance is a function of the distance d (cm) between the electrodes, the surface area A (cm^2) of the electrodes, and the permittivity ε (8.85×10^{-12} F m^{-1} for air) of the dielectric between the electrodes; therefore,

$$C = f(d, A, \varepsilon) \tag{25.1}$$

There are three basic methods for realizing a capacitive displacement sensor, that is, by varying d, or A, or ε, as discussed in the following.

25.2 Variable Distance Displacement Sensors

A capacitor displacement sensor, made from two flat coplanar plates with a variable distance x apart, is illustrated in Figure 25.1. Ignoring fringe effects, the capacitance of this arrangement can be expressed by

$$C(x) = \frac{\varepsilon A}{x} = \frac{\varepsilon_r \varepsilon_0 A}{x} \tag{25.2}$$

where

ε is the dielectric constant or permittivity
ε_r is the relative dielectric constant (in air and vacuum $\varepsilon_r \approx 1$)
ε_0 is 8.854188×10^{-12} F m^{-1}, the dielectric constant of vacuum
x is the distance of the plates in m
A is the effective area of the plates in m^2

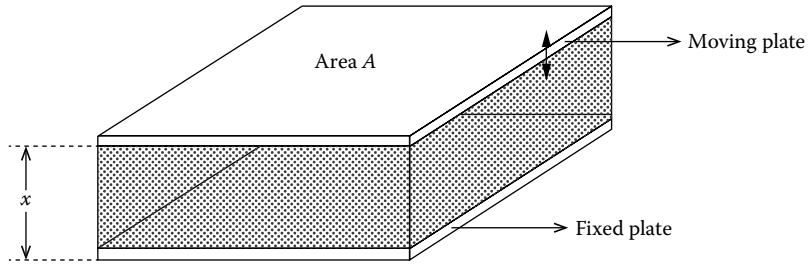


FIGURE 25.1 A variable distance capacitive displacement sensor. One of the plates of the capacitor moves to vary the distance between plates in response to changes in a physical variable. Generally, the outputs of these sensors are nonlinear with respect to displacement of an x having a hyperbolic transfer function characteristic. Appropriate signal processing must be employed for linearization.

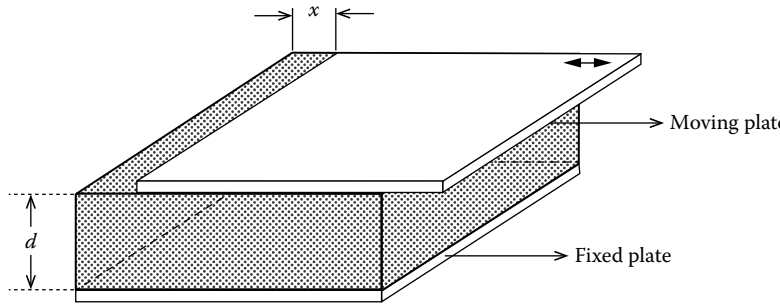


FIGURE 25.2 A variable area capacitive displacement sensor. The sensor operates on the variation in the effective area between plates of a flat-plate capacitor. The transducer output is linear with respect to displacement x. This type of sensor is normally implemented as a rotating capacitor for measuring angular displacements.

The capacitance of this transducer is nonlinear with respect to distance x, having a hyperbolic transfer function characteristic. The sensitivity of capacitance to changes in plate separation is

$$\frac{dC}{dx} = \frac{-\varepsilon_r \varepsilon_0 A}{x^2} \tag{25.3}$$

Equation 25.3 indicates that the sensitivity increases as x decreases. Nevertheless, from Equations 25.2 and 25.3, it follows that the percent change in C is proportional to the percent change in x. This can be expressed as

$$\frac{dC}{C} = \frac{-dx}{x} \tag{25.4}$$

This type of sensor is often used for measuring small incremental displacements without making contact with the object.

25.3 Variable Area Displacement Sensors

Alternatively, the displacements may be sensed by varying the surface area of the electrodes of a flat-plate capacitor, as illustrated in Figure 25.2. In this case, the capacitance would be

$$C = \frac{\varepsilon_r \varepsilon_0 (A - wx)}{d} \tag{25.5}$$

where

w is the width

wx is the reduction in the area due to movement of the plate

Then, the transducer output is linear with displacement x. This type of sensor is normally implemented as a rotating capacitor for measuring angular displacement. The rotating capacitor structures are also used as an output transducer for measuring electric voltages as capacitive voltmeters.

25.4 Variable Dielectric Displacement Sensors

In some cases, the displacement may be sensed by the relative movement of the dielectric material between the plates, as shown in Figure 25.3. The corresponding equations would be

$$C = \varepsilon_0 w[\varepsilon_2 l - (\varepsilon_2 - \varepsilon_1)x] \tag{25.6}$$

where

ε_1 is the relative permittivity of the dielectric material

ε_2 is the permittivity of the displacing material (e.g., liquid)

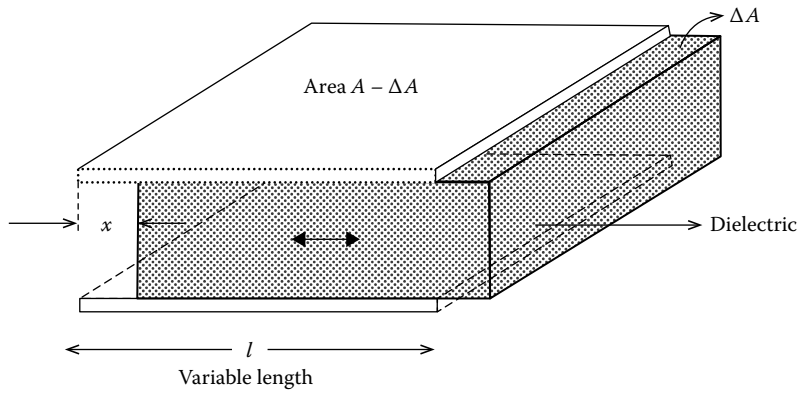


FIGURE 25.3 A variable dielectric capacitive displacement sensor. The dielectric material between the two parallel plate capacitors moves, varying the effective dielectric constant. The output of the sensor is linear.

In this case, the output of the transducer is also linear. This type of transducer is predominantly used in the form of two concentric cylinders for measuring the level of fluids in tanks. A nonconducting fluid forms the dielectric material. Further discussion will be included in the level measurements section.

25.5 Differential Capacitive Displacement Sensors

Some of the nonlinearity in capacitive sensors can be eliminated using differential capacitive arrangements. These sensors are three-terminal capacitors, as shown in Figure 25.4. Slight variations in the construction of these sensors find many different applications, including differential pressure measurements. In some versions, the central plate moves in response to physical variables with respect to the fixed plates. In others, the central plate is fixed and outer plates are allowed to move. The output from the center plate is zero at the central position and increases as it moves left or right. The range is equal to twice the separation d. For a displacement d, one obtains

$$2\delta C = C_1 - C_2 = \frac{\varepsilon_r \varepsilon_0 lw}{d - \delta d} - \frac{\varepsilon_r \varepsilon_0 lw}{d + \delta d} = \frac{2\varepsilon_r \varepsilon_0 lw\delta d}{d^2 + \delta d^2} \tag{25.7}$$

and

$$C_1 + C_2 = 2C = \frac{\varepsilon_r \varepsilon_0 lw}{d - \delta d} + \frac{\varepsilon_r \varepsilon_0 lw}{d + \delta d} = \frac{2\varepsilon_r \varepsilon_0 lwd}{d^2 + \delta d^2} \tag{25.8}$$

giving approximately

$$\frac{\delta C}{C} = \frac{\delta d}{d} \tag{25.9}$$

This indicates that the response of the device is linear compared to the response of the two-plate counterparts. However, in practice, some nonlinearity is still observed due to structural defects. Therefore, the outputs of these types of sensors still need to be treated carefully, as explained in the Section 25.11.

In some differential capacitive sensors, the two spherical depressions are ground into glass disks; then, these are gold plated to form the fixed plates of a differential capacitor. A thin, stainless-steel diaphragm is clamped between the disks and serves as a movable plate. With equal pressure applied to both ports,

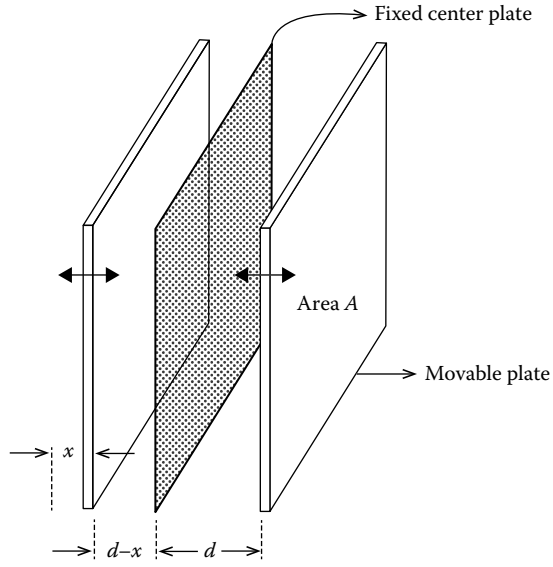


FIGURE 25.4 A differential capacitive sensor. They are essentially three-terminal capacitors with one fixed center plate and two outer plates. The response to physical variables is linear. In some versions, the central plate moves in response to physical variable with respect to two outer plates, and in others, the central plate is fixed and the outer plates are allowed to move.

the diaphragm is then in neutral position and the output is balanced at an electrical bridge. If the pressure on one side is greater than on the other side, the diaphragm deflects proportionally, thus giving an output due to the pressure difference. For the reverse pressure difference, there is a phase change of 180°. A direction-sensitive DC output voltage can be obtained by conventional phase-sensitive demodulation and appropriate filtering. Details of signal processing are given at the end of this chapter. In general, the differential capacitors exhibit better linearity than the single-capacitor types as explained earlier.

25.6 Integrated Circuit Smart Capacitive Displacement Sensors

Figure 25.5 shows a typical microstructure capacitive displacement sensor. The sensor consists of two electrodes with capacitance, C_x. Since the system is simple, the determination of the capacitance between the two electrodes is straightforward. The smaller electrode is surrounded by a guard electrode to make C_x

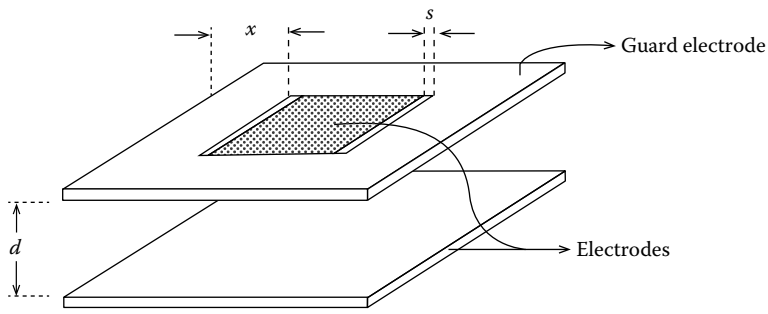


FIGURE 25.5 A typical smart capacitive position sensor. This type of microstructure position sensor contains three electrodes, two of which are fixed and the third electrode moves infinitesimally relative to the others. Although the response is highly nonlinear, the integrated chip contains linearization circuits. They feature a 0–1 mm measuring range with 1 μm accuracy.

independent of lateral and rotational movements of the system parallel to the electrode surface. However, the use of a guard electrode introduces relative deviations in the capacitance C_x between the two electrodes. This is partially true if the size of the guard electrode is smaller than

$$\delta < \exp\left(\frac{-\pi x}{d}\right) \tag{25.10}$$

where
x is the width of the guard
d is the distance between the electrodes

Since this deviation introduces nonlinearity, δ is required to be less than 100 part per million (ppm). Another form of deviation also exists between the small electrode and the surrounding guard, particularly for gaps:

$$\delta < \exp\left(\frac{-\pi d}{s}\right) \tag{25.11}$$

where s is the width of the gap. When the gap width, s, is less than 1/3 of the distance between electrodes, this deviation is negligible.

For signal processing, the system uses the three-signal concept. The capacitor C_x is connected to an inverting operational amplifier and oscillator. If the external movements are linear, by taking into account the parasitic capacitors and offsetting effects, the following equation can be written as

$$M_x = mC_x + M_{\text{off}} \tag{25.12}$$

where

m is the unknown gain
M_{off} is the unknown offset

By performing the measurement of a reference C_{ref}, by measuring the offset, M_{off}, and by making $m = 0$, the parameters m and M_{off} can be eliminated. The final measurement result for the position, P_{os}, can be defined as

$$P_{\text{os}} = \frac{M_{\text{ref}} - M_{\text{off}}}{M_x - M_{\text{off}}} \tag{25.13}$$

In this case, the sensor capacitance C_x can be simplified to

$$C_x = \frac{\varepsilon A_x}{d_0 + \Delta d} \tag{25.14}$$

where
A_x is the area of the electrode
d_0 is the initial distance between them
ε is the dielectric constant
Δd is the displacement to be measured

For the reference electrodes, the reference capacitance may be determined by

$$C_{\text{ref}} = \frac{\varepsilon A_{\text{ref}}}{d_{\text{ref}}} \tag{25.15}$$

with A_{ref} the area and d_{ref} the distance. Substitution of Equations 25.14 and 25.15 into Equations 25.12 and 25.13 yields

$$P_{\text{os}} = \frac{A_{\text{ref}}(d_0 + \Delta d)}{A_x d_{\text{ref}}} = a_1 \frac{\Delta d}{d_{\text{ref}}} + a_0 \tag{25.16}$$

P_{os} is a value representing the position if the stable constants a_1 and a_0 are unknown. The constant $a_1 = A_{\text{ref}}/A_x$ becomes a stable constant so long as there is good mechanical matching between the electrode areas. The constant $a_0 = (A_{\text{ref}}d_0)/(A_x d_{\text{ref}})$ is also a stable constant for fixed d_0 and d_{ref}. These constants are usually determined by calibrating repeatedly over a certain time span. In many applications, these calibrations are omitted if the displacement sensor is part of a larger system where an overall calibration is necessary. This overall calibration usually eliminates the requirement for a separate determination of a_1 and a_0.

The accuracy of this type of system could be as small as 1 μm over a 1 mm range. The total measuring time is better than 0.1 s. The capacitance range is from 1 pF to 50 fF.

25.7 Capacitive Pressure Sensors

A commonly used two-plate capacitive pressure sensor is made from one fixed metal plate and one flexible diaphragm, as shown in Figure 25.6. The flat circular diaphragm is clamped around its circumference and bent into a curved shape by an applied pressure *P*. The vertical displacement *y* of this system at any radius *r* is given by

$$y = \frac{3(1-\nu^2)(a^2 - r^2)\,P}{16\,Et^3} \tag{25.17}$$

where

a is the radius of diaphragm
t is the thickness of diaphragm
E is Young's modulus
ν is Poisson's ratio

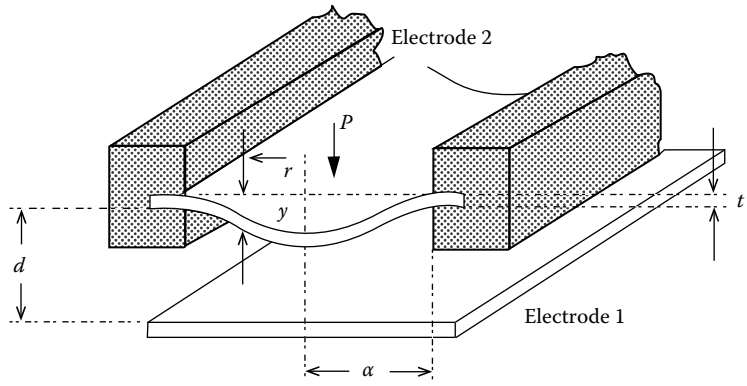


FIGURE 25.6 A capacitive pressure sensor. These pressure sensors are made from a fixed metal plate and a flexible diaphragm. The flat flexible diaphragm is clamped around its circumference. The bending of the flexible plate is proportional to the applied pressure *P*. The deformation of the diaphragm results in changes in capacitance.

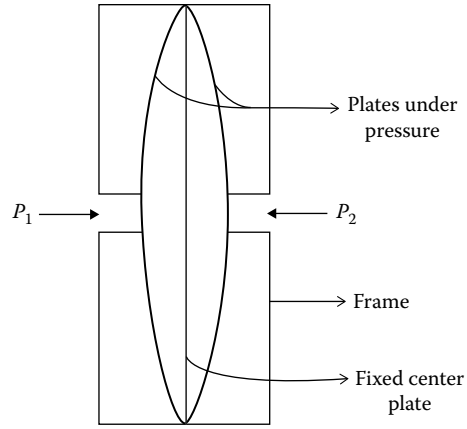


FIGURE 25.7 A differential capacitive pressure sensor. The capacitances C_1 and C_2 of the sensors change due to deformation of the outer plates, with respect to the fixed central plate, in response to the applied pressures P_1 and P_2.

Deformation of the diaphragm means that the average separation of the plates is reduced. Hence, the resulting increase in the capacitance ΔC can be calculated by

$$\frac{\Delta C}{C} = \frac{(1-\nu^2)a^4 P}{16Et^3} \tag{25.18}$$

where
d is the initial separation of the plates
C is the capacitance at zero pressure

Another type of sensor is the differential capacitance pressure sensor shown in Figure 25.7. The capacitances C_1 and C_2 of the sensor change with respect to the fixed central plate in response to the applied pressures P_1 and P_2. Hence, the output of the sensor is proportional to $(P_1 - P_2)$. The signals are processed using one of the techniques described in Section 25.11.

25.8 Capacitive Accelerometers and Force Transducers

Acceleration is the rate of change in velocity and the velocity is the rate or change in distance. As explained earlier, capacitive devices are commonly used in displacement measurements, and hence, they can be configured for acceleration measurement too. In recent years, capacitive-type micromachined accelerometers, as illustrated in Figure 25.8, are commonly used. These accelerometers are arranged such that a proof mass forms one of the plates of the capacitor and the other plate fixed as the base. When the sensor is accelerated, the proof mass moves altering the distance between itself and the base plate; thus, the capacitance changes. This change in capacitance corresponds to the applied acceleration.

In Figure 25.8, let $F(x)$ be the positive force in the direction in which x increases. Neglecting all losses (due to friction, resistance, etc.), the energy balance of the system can be written for an infinitesimally small displacement dx, electrical energy dE_e, and field energy dE_f of the electrical field between the electrodes as

$$dE_m + dE_c = dE_f \tag{25.19}$$

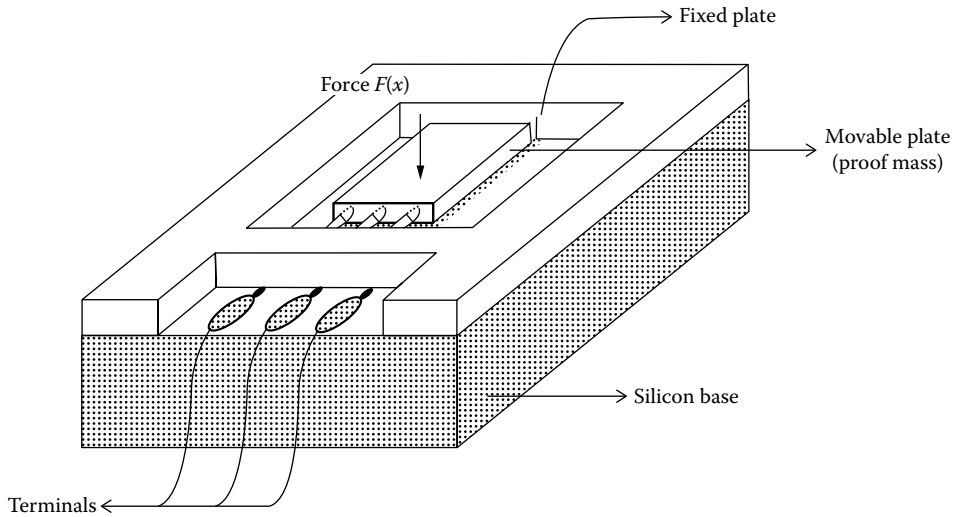


FIGURE 25.8 A capacitive force transducer. A typical capacitive micromachined accelerometer has one of the plates as the proof mass. The other plate is fixed, thus forming the base. When the sensor is accelerated, the proof mass tends to move, thus varying the distance between the two plates and altering the voltage across the capacitor. This change in voltage can be made directly proportional to the applied acceleration.

in which

$$dE_m = F(x)\ dx \tag{25.20}$$

Also,

$$dE_m = d(QV) = Q\ dV + V\ dQ \tag{25.21}$$

If the supply voltage V across the capacitor is kept constant, it follows that $dV = 0$. Since $Q = VC(x)$, the Coulomb force is given by

$$F(x) = -V^2 \frac{dC(x)}{dx} \tag{25.22}$$

Thus, if the movable electrode has complete freedom of motion, it will have assumed a position in which the capacitance is maximal; also, if C is a linear function of x, the force $F(x)$ becomes independent of x.

Capacitive silicon accelerometers are available in a wide range of specifications. A typical lightweight sensor will have a frequency range of 0–1000 Hz and a dynamic range of acceleration of ±2 g to ±500 g.

25.9 Capacitive Liquid-Level Measurement

The level of a nonconducting liquid can be determined by capacitive techniques. The method is generally based on the difference between the dielectric constant of the liquid and that of the gas or air above it. This is similar to variable dielectric displacement sensors as explained earlier. In capacitive liquid-level measurements, two concentric metal cylinders are used, as shown in Figure 25.9. The height of the liquid, h, is measured relative to the total height, l. An appropriate provision is made to ensure that the space between the cylindrical electrodes is filled by the liquid to the same height as the rest of the container.

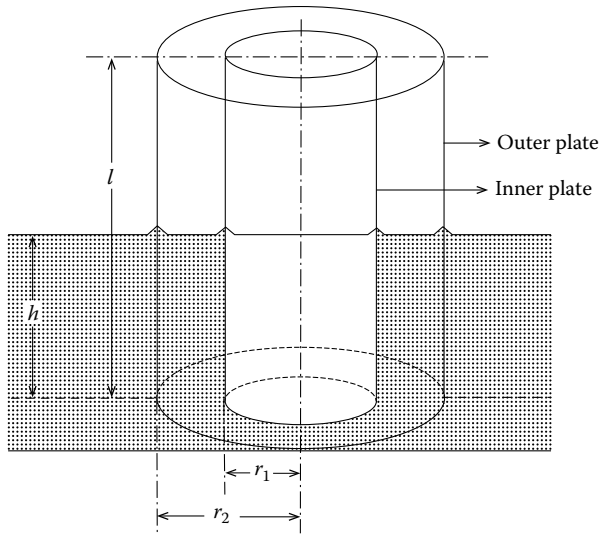


FIGURE 25.9 A capacitive liquid-level sensor. Two concentric metal cylinders are used as electrodes of a capacitor. The value of the capacitance depends on the permittivity of the liquid and that of the gas or air above it. The total permittivity changes depending on the liquid level. These devices are usually applied in nonconducting liquid applications.

The usual operational conditions dictate that the spacing between the electrodes, $s = r_2 - r_1$, should be much less than the radius of the inner electrode, r_1. Furthermore, the tank height should be much greater than r_2. When these conditions apply, the capacitance can be approximated by

$$C = \frac{\varepsilon_1(1) + \varepsilon_g(h-1)}{4.6\log\left[1-(s/r)\right]} \tag{25.23}$$

where ε_1 and ε_g are the dielectric constants of the liquid and gas (or air), respectively. The denominator of the earlier equation contains only terms that relate to the fixed system. Therefore, they become a single constant. A typical application is the measurement of the amount of gasoline in a tank in airplanes. The dielectric constant for most compounds commonly found in gasoline is approximately equal to 2, while that of air is approximately unity. A linear change in capacitance with gasoline level is expected for this situation. Quite high accuracy can be achieved if the denominator is kept quite small, thus accentuating the level differences. These sensors often incorporate an AC deflection bridge for signal processing.

25.10 Capacitive Humidity and Moisture Sensors

Capacitive humidity and moisture sensors are mostly based on the changes in the permittivity of the dielectric material, in principle, similar to variable dielectric displacement sensors explained previously. The permittivity of atmospheric air, some gases, and many solid materials are functions of their moisture contents and their temperature. The main disadvantage of this type of sensor is that a relatively small change in humidity results in a large change in capacitance sufficient enough for a sensitive detection.

Capacitive humidity sensors enjoy wide dynamic ranges, from 0.1 ppm to saturation points. They can function in saturated environments for long periods of time, a characteristic that would adversely affect many other humidity sensors. Their ability to function accurately and reliably extends over a wide range

of temperatures and pressures. Capacitive humidity sensors also exhibit low hysteresis and high stability with minimal maintenance requirements. These features make them viable for many specific operating conditions and ideally suitable for a system where uncertainty of unaccounted conditions exists during operations.

There are many types of capacitive humidity sensors. In here, aluminum, tantalum, silicon, and polymer types are introduced.

25.10.1 Aluminum Capacitive Humidity Sensors

The majority of capacitive humidity sensors are based on aluminum oxide-type dielectric material. In these types of sensors, high-purity aluminum is chemically oxidized to produce a prefilled insulating layer of partially hydrated aluminum oxide, which acts as the dielectric. A water-permeable but conductive gold film is deposited onto the oxide layer, usually by vacuum deposition, which forms the second electrode of the capacitor.

In another type, the aluminum–aluminum oxide sensor has a pore structure as illustrated in Figure 25.10. The oxide, with its pore structure, forms the active sensing material. Moisture in the air reaching the pores reduces the resistance and increases the capacitance. The decreased resistance can be thought of as being due to an increase in the conduction through the oxide. An increase in capacitance can be viewed due to an increase in the dielectric constant. The quantity measured can be either resistance or capacitance or impedance. High humidity is best measured by capacitance because resistance changes are vanishingly small at high levels.

In addition to the kind of transducer design illustrated in Figure 25.10, there are many others available with a number of substantial modifications due to their particular properties, such as the increased

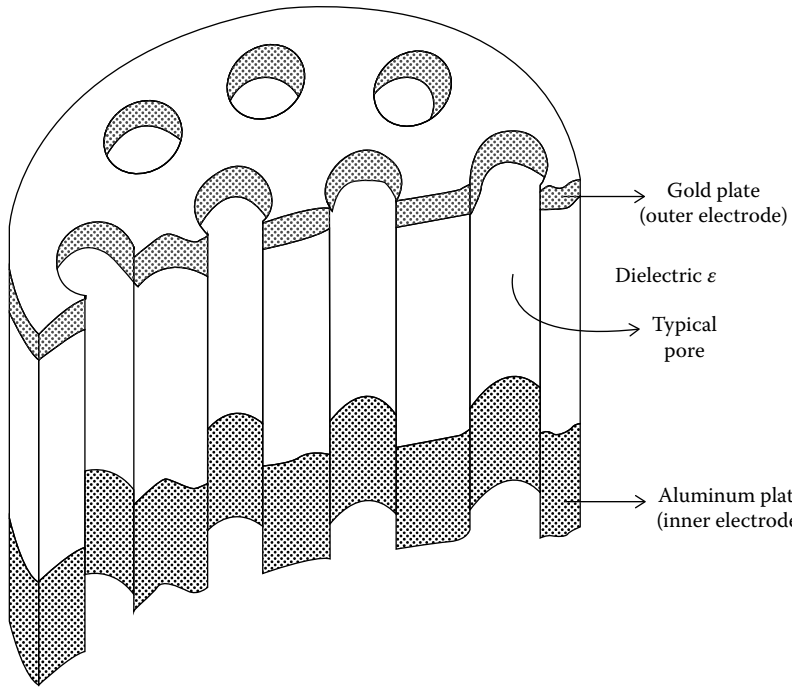


FIGURE 25.10 A typical capacitive humidity sensor. The sensors have pore or cracked mosaic structure for the moisture in air or gas to reach the dielectric material. The characteristics of the dielectric material change with the amount of water absorbed, thus reducing the resistance and increasing the capacitance. The quantity measured can be either the resistance or the capacitance or the impedance.

sensitivity and/or the faster response. Although most of these modifications can result in a change in physical dimensions or appearance, the basic sensing material—the aluminum oxide—remains the same.

In some versions, the oxide layer is formed by parallel tubular pores that are hexagonally packed and perpendicular to the plane of the base layer. These pores stop just before the aluminum layer, forming a very thin pore base. Water absorbed in these tubules is directly related to the moisture content of the gas in contact with it. The porous nature of the oxide layer produces a large area for the absorption of water vapor. At low humidity, the capacitance is due entirely to the mixed dielectric formed between the oxide, water vapor, and air. However, at higher humidity, parallel conductance paths through the absorbed water are formed down the pore surfaces. Near saturation, this pore surface resistance becomes negligible, implying that the measured capacitance is virtually in between the very thin pore base and the aluminum core.

25.10.2 Tantalum Capacitive Humidity Sensors

In some versions of capacitive humidity sensors, one of the capacitor plates consists of a layer of tantalum deposited on a glass substrate. A layer of polymer dielectric is then added, followed by a second plate made from a thin layer of chromium. The chromium layer is under high tensile stress such that it cracks into a fine mosaic structure that allows water molecules to penetrate inside the dielectric material. The stress in the chromium also causes the polymer to crack into a mosaic structure. A sensor of this type has an input range of 0%–100% relative humidity (RH). The capacitance is 375 pF at 0% RH and a linear sensitivity of 1.7 pF per% RH. The error is usually less than 2% due to nonlinearity and 1% due to hysteresis.

25.10.3 Silicon Capacitive Humidity Sensors

In some capacitive humidity sensors, silicon is used as the dielectric. The structure and operation of silicon humidity sensors are very similar to the aluminum oxide types. Some silicon-type humidity sensors also use the aluminum base and a thin-film gold layer as the two electrodes. The silicon dielectric has a very large surface area, which means that the sensitivity is still relatively large even if the sensing area is very small. This is an important feature with the increasing chances of miniaturization. Both sensor types are now typically found as extremely small wafer-shaped elements, placed on a mechanical mount with connecting lead wires. The formation of porous silicon is a very simple anodization process, and since no elaborate equipment is needed, devices can be made relatively at low costs. Also, by controlling the formation conditions, the structure of the porous silicon can easily be modified so devices can be tailored to suit particular applications.

In both the silicon and the aluminum oxide capacitive humidity sensors, the radii of the pores in the dielectric are such that they are specifically suited for water molecules. Most possible contaminants are too large in size to pollute the dielectric. However, contaminants can block the flow of water vapor into the sensor material, thus affecting the accuracy of the sensor. For example, in dust-contaminated streams, it may be possible to provide a simple physical barrier such as a sintered metal or plastic hood for the sensor heads. Many sensors come with specially designed casings to provide protection.

25.10.4 Polymer Capacitive Humidity Sensors

In some sensors, the dielectric consists of a polymer material that can absorb water molecules. The absorption of water vapor of the material results in changes in the dielectric constant of the capacitor. By careful design, the capacitance can be made directly proportional to percentage RH of the surrounding gas or atmosphere.

In general, an important key feature of capacitive humidity sensors is the chemical stability. Often, humidity sensing is required in an air sample that contains vapor contaminants (e.g., carbon monoxide) or the measurements are performed on a gas sample other than air (e.g., vaporized benzene). The performance of these sensors, and in particular the silicon types, is not affected by many of these gases. Hydrocarbons, carbon dioxide, carbon monoxide, and CFCs do not cause interference. However, the ionic nature of the aluminum oxide dielectric makes it susceptible to certain highly polar, corrosive gases such as ammonia, sulfur trioxide, and chlorine. Silicon is inert; its stable nature means that these polar gases affect the sensor element to a far lesser degree.

25.10.5 Capacitive Moisture Sensors

Capacitive moisture measurements are based on the changes in the permittivity of granular- or powder-type dielectric materials such as wheat and other grains containing water. Usually, the sensor consists of a large cylindrical chamber (e.g., 150 mm deep and 100 mm in diameter), as shown in Figure 25.11. The chamber is filled with samples under test. The variations in capacitance with respect to water content are processed. The capacitor is incorporated into an oscillatory circuit operating at a suitable frequency.

Capacitive moisture sensors must be calibrated if the samples are made from different materials, as the materials themselves demonstrate different permittivity and other properties. Accurate temperature is necessary as the dielectric constant may be highly dependent on temperature. Most of these devices are built to operate at temperature ranges of 0 °C–50 °C, supported by tight temperature compensation circuits. Once calibrated for a specific application, they are suitable for measuring moisture in the range of 0%–40%.

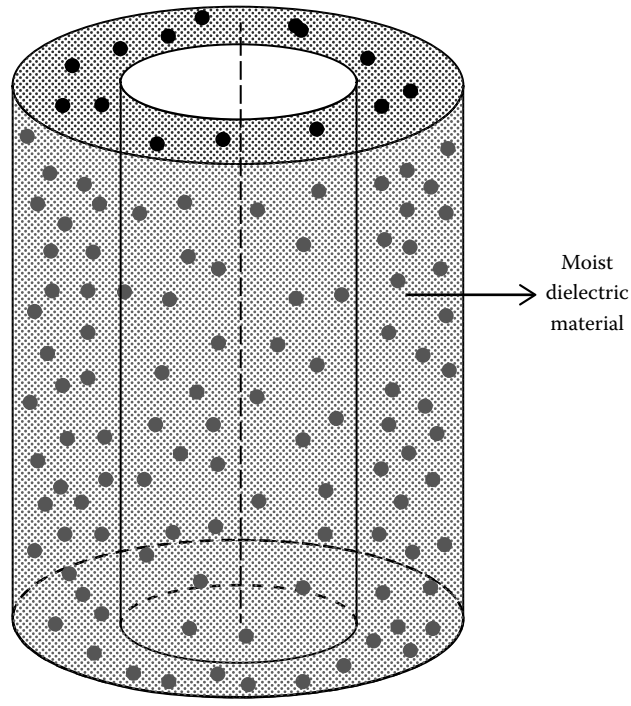


FIGURE 25.11 A capacitive moisture sensor. The permittivity of material between two cylindrical or parallel plates with fixed dimensions changes, depending on the moisture level of the material in the chamber. The variations in capacitance values with respect to water content are processed. The capacitor is incorporated as a part of an oscillatory circuit operating at a suitable frequency, usually at radio frequencies.

25.11 Signal Processing

Signal processing for capacitive sensors requires relatively complex circuitry in comparison to many other sensor types, but they have the advantage of mechanical simplicity. They have good sensitivity and range, exhibiting minimum mechanical loading effects. Capacitive sensors are usually incorporated either in some form of AC deflection bridges or oscillator circuits. In practice, capacitive sensors are not pure capacitances but have associated resistances representing losses in the dielectric. This can have an important influence in the design of the associated circuits, particularly in oscillator circuits. A number of the signal processing circuits are discussed in the following.

25.11.1 Operational Amplifier and Charge Amplifier-Based Circuits

An effective method of eliminating the nonlinearity of the relationship between the physical variable (e.g., two-plate displacement sensors) and capacitance, C, is by the use of operational amplifiers. An example of such circuits is illustrated in Figure 25.12. In this circuit, if the input impedance of the operational amplifier is high, the output is not saturated, and the input voltage is small, it is possible to write

$$\frac{1}{C_f} = \int i_f \, dt = e_{ex} - e_{ai} = e_{ex} \tag{25.24}$$

$$\frac{1}{C_x} = \int i_x \, dt = e_0 - e_{ai} = e_0 \tag{25.25}$$

$$i_f + i_x - i_{ai} = 0 = i_f + i_x \tag{25.26}$$

Manipulation of these equations yields to

$$e_0 = \frac{-C_f e_{ex}}{C_x} \tag{25.27}$$

Substituting the value of C_x gives

$$e_0 = \frac{-C_f x e_{ex}}{\varepsilon A} \tag{25.28}$$

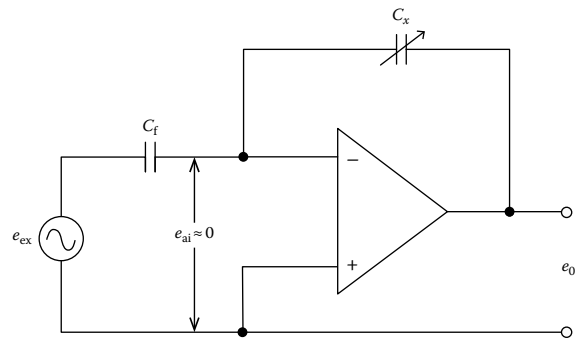


FIGURE 25.12 An operational amplifier signal processor. This method is useful to eliminate the nonlinearity in the signals generated by capacitive sensors. By this type of arrangement, the output voltage can be made directly proportional to variations in the signal representing the nonlinear operation of the device.

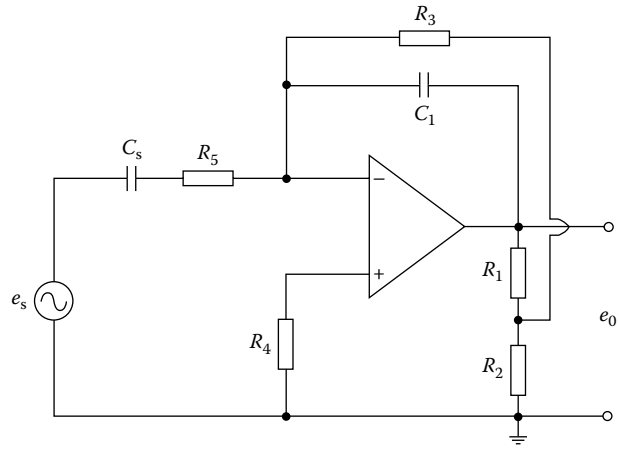


FIGURE 25.13 A practical charge amplifier. The effective feedback resistance is a function of other resistances. It is possible to reduce the output drift substantially by selecting the resistors suitably. The accuracy of this circuit can be improved further by cascading two or more amplifiers, thereby substantially improving the signal-to-noise ratio.

Equation 25.28 shows that the output voltage is directly proportional to the plate separation x, thus resulting in the linearity for all variations in the motion under measurement.

However, a practical circuit requires a resistance across C_f to limit drift in the output. The value of this resistance must be greater than the impedance of C_f at the lowest frequency of interest. Also, because the sensor impedance is assumed to be purely capacitive, the effective gain becomes independent of the frequency.

A practical charge amplifier circuit is depicted in Figure 25.13. In this case, the effective feedback resistance R_{ef} is given by

$$R_{ef} = \frac{R_3(R_1 + R_2)}{R_2} \tag{25.29}$$

It is possible to reduce the output drift substantially by the suitable selection of the resistors. The accuracy of this circuit can be improved substantially by cascading two or more amplifiers. In this way, a good improvement in the signal-to-noise ratio can be achieved. In the inverting input, the use of resistor R_4 is necessary because of the bias currents.

25.11.2 Pulse Width Modulation

As in the case of some capacitive vibration and displacement sensors, the output voltage may be in amplitude-modulated waveform, as shown in Figure 25.14. When rectified, the average value of this wave gives the mean separation between the plates. The vibration amplitude around this mean may be extracted by the use of a demodulator and a low-pass filter circuit. Then the output of the low-pass filter becomes a direct indication of the vibration, and the waveform can also be viewed on an oscilloscope.

25.11.3 Square-Wave Linearization

Another linearization technique applied in capacitive pressure sensors and accelerometers is the pulse width modulation. The sensor consists of two differential capacitors as shown in Figure 25.15. The voltages of these capacitors, e_1 and e_2, switch back and forth with a high excitation frequency (e.g., 400 kHz) between excitation voltage and ground. The system is arranged in such a way that the output voltage is

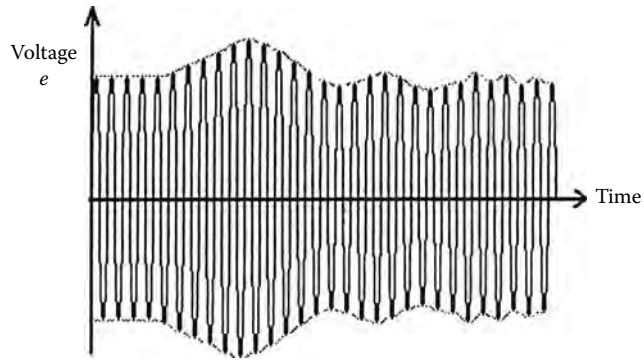


FIGURE 25.14 An amplitude-modulated signal. It is possible to configure some sensors to give amplitude-modulated signals, as in the case of capacitive vibration displacement sensors. When rectified, the average value of this wave gives the mean separation of the plates. The vibration amplitude around this mean position can be extracted by a demodulator and low-pass filter circuit. The output of the low-pass filter is a direct indication of the vibrations.

the average voltage difference between e_1 and e_2. At null position, $e_1 = e_2$, the output is a symmetrical square wave with zero average value. As the relative positions of the plates change, due to vibration, the average value of the output voltage shifts from the zero average value and becomes positive or negative depending on the direction of the displacement. The output voltage can be expressed by

$$e_o = \frac{e_{ex}(C_1 - C_2)}{C_1 + C_2} \tag{25.30}$$

Substituting

$$C_1 = \frac{C_0 x_0}{x_0 - x_i} \quad \text{and} \quad C_2 = \frac{C_0 x_0}{x_0 + x_i}$$

yields

$$e_o = \frac{e_{ex} x_i}{x_0} \tag{25.31}$$

Thus, the output is directly proportional to the variable x_i.

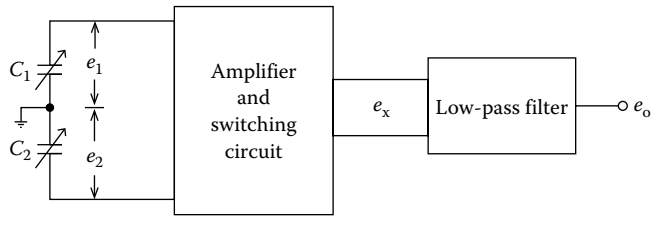


FIGURE 25.15 Block diagram of a square-wave linearization circuit. This is particularly useful for differential capacitance-type sensors. The voltages of these two capacitors are made to switch back and forth with a high excitation frequency between excitation voltage and ground. As the relative positions of the plates change due to vibration, the average value of the output voltage becomes positive or negative, depending on the direction of the displacement.

25.11.4 Feedback Linearization

Linearization of a capacitive sensor can also be obtained using a feedback system that adjusts capacitor current amplitude so that it stays constant at a reference value for all displacements. This is accomplished by obtaining a DC signal proportional to capacitor current from a demodulator, comparing this current with the reference current, and adjusting the voltage amplitude of the system excitation oscillator until the two currents agree. If the capacitor current is kept constant irrespective of the motion, then the voltage amplitude is linearly related to x as

$$e = Kx_i \tag{25.32}$$

where

$$K = \frac{|i_c|}{\omega C_0 x_0} \tag{25.33}$$

25.11.5 Oscillator Circuits

In many applications, the resultant changes in the capacitance of capacitive transducers can be measured with a suitable AC bridge such as Wein bridge or Schering bridge. However, in majority of cases, improvised versions of bridges are used as oscillator circuits for signal processing. The transducer is configured as a part of the oscillatory circuit that causes changes in the frequency of the oscillations. This change in frequency is scaled to be a measure of the magnitude of the physical variable.

As part of the oscillator circuits, the capacitive transducer has excellent frequency response and can measure both static and dynamic phenomena. Its disadvantages include sensitivity to temperature variations and the possibility of erratic or distorted signals due to long length of connection leads. Also, the receiving instrumentation may be large and complex, and it often includes a second fixed-frequency oscillator for heterodyning purposes. The difference frequency thus produced can be read by an appropriate output device such as an electronic counter.

Bibliography

Capacitive sensors—An overview, LION PRECISION, http://www.lionprecision.com/capacitive-sensors/index.html (accessed on September 9, 2011).

Capacitive sensors for noncontact position and displacement measurements and nano-positioning, ChenYang Technologies GmbH & Co. KG, http://www.chenyang-ism.com/CapaSensorPosi.htm (accessed on September 6, 2011).

Doebelin, E. O., *Measurement Systems: Application and Design*, 4th edn., New York, McGraw-Hill, 1990.

Eren, H., Displacement and angle sensors performance and selection, In *Handbook of Measuring and System Design*, P. Sydenham and R. Thorn (eds.), Wiley Online Library, 2005, http://onlinelibrary.wiley.com/book/10.1002/0471497398/homepage/EditorsContributors.html (accessed on September 4, 2011).

Hongshen, Ma., Capacitive displacement sensing for the nanogate, http://www.media.mit.edu/resenv/pubs/theses/HongMa-MSThesis.pdf (accessed on September 9, 2011).

Rhead, P. M., Patel, J., Atherton, P. D., Hicks, T. R., and McConnell, M., Capacitive displacement sensor—Design characterization of a long range, low noise non-contact, QUEENSGATE INTRUMENTS, http://www.azonano.com/article.aspx?ArticleID=1465 (accessed on August 30, 2011).

Xiaohui, H. and Wuqiang, Y., Planar capacitive sensors—Designs and applications, *Sensor Review*, 30(3), 24–39, 2010.

Yusof, N. B., Soin, N., and Dawal, S. Z. M., Capacitive interfacing for MEMS humidity and accelerometer sensors, *Proceedings of the IEEE International Conference for Technical Postgraduates*, Kuala Lumpur, Malaysia, Article 5412067, 2009.

Partial List of Manufacturers and Suppliers

AutomationDirect.com (AutomationDirect), 3505 Hutchinson Rd., Cumming, GA 30040-5860, Phone: 770-889-2858, 800-633-0405 (toll free), Fax: 770-889-7876, www.automationdirect.com

Automation Products Group, Inc. (APGSENSORS), 1025 W. 1700 N., Logan, UT 84321, Phone: 435-753-7300, 888-525-7300 (toll free), Fax: 435-753-7490, www.apgsensors.com

Autonics USA, 1353 Armour Blvd., Mundelein, IL 60060, Phone: 847-680-8160, 866-862-2602 (toll free), Fax: 847-680-8155, www.autonicsonline.com

Balluff, Inc., 8125 Holton Dr., Florence, KY 41042-0937, Phone: 859-727-2200, 800-543-8390 (toll free), Fax: 859-727-4823, www.balluff.com

Baumer Ltd., 122 Spring St., Suite C-6 Southington, CT 06489-1534, Phone: 860-621-2121, 800-937-9336 (toll free), Fax: 860-628-6280, www.baumerelectric.com

Capacitec, Inc., 87 Fitchburg Rd., P.O. Box 819 Ayer, MA 01432, Phone: 978-772-6033 Fax: 978-772-6036, www.capacitec.com

Custom Sensors & Technologies (CST), 14401 Princeton Ave., Moorpark, CA 93021-9707, Phone: 805-552-3599, Fax: 805-552-3577, www.cstsensors.com

Davis/Inotek, 1236 Mark St., Bensenville, IL 60106, Phone: 410-358-3900, 800-548-9409 (toll free), Fax: 410-358-0252, www.davis.com

EMRA Controls, Inc., 1502 Pike St., N.W. Suite 4B Auburn, WA 98001, Phone: 253-351-0292, 800-351-0293 (toll free), Fax: 253-804-2678, www.emracontrols.com

Fargo Controls, Inc., P.O. Box 539, Eatontown, NJ 07724, Phone: 732-389-3376, 800-241-3755 (toll free), Fax: 732-542-3553, www.fargocontrols.com

Honeywell Sensing and Control, 1985 Douglas Dr. North, MN10-192B Golden Valley, MN 55422, Phone: 763-954-4818, 800-446-6555 (toll free), www.honeywell.com/

IFM Efector, Inc., 782 Springdale Dr., Exton, PA 19341-2843, Phone: 610-524-2000, 800-441-8246 (toll free), Fax: 800-329-0436, www.ifm.com/ifmus/web/home.htm

Horizon Systems, Inc., 1101 Horizon Dr., Lawrence, KS 66046, Phone: 785-842-1299, Fax: 785-842-0181, www.horizonsystemsinc.com

ISSC-Kanson Electronics, Inc., 245 Forrest Ave., Hohenwald, TN 38462, Phone: 931-796-3050, 800-233-9354 (toll free), Fax: 931-796-3956, www.issc-kanson.com/

Leuze Electronic, Inc., 55395 Lyon Industrial Dr., New Hudson, MI 48165, Phone: 248-486-4466, Fax: 248-486-6699, www.leuzeusa.com

Lion Precision, Inc., 563 Shoreview Park Rd., Saint Paul, MN 55126-7014, Phone: 651-484-6544, 800-250-9297 (toll free), Fax: 651-484-6824, www.lionprecision.com

MicroSense, LLC, 70 Industrial Ave. E., Lowell, MA 01852, Phone: 978-843-7673, www.microsense.net

Maxcess, 2305 S.E. 8th Ave., Camas, WA 98607, Phone: 360-834-2345, Fax: 360-834-5865, www. maxcessintl.com

Measurement Specialties, Inc., 1000 Lucas Way, Hampton, VA 23666, Phone: 757-766-1500, 800-745-8008 (toll free), Fax: 757-766-4297, www.meas-spec.com

MicroEpsilon, 3200-110 Glen Royal Rd., Raleigh, NC 27617-7419, Phone: 919-787-9707, Fax: 919-787-9708, www.micro-epsilon.com

MKS Instruments, Inc., 2 Tech Dr., Suite 201 Andover, MA 01810, Phone: 978-645-5500, 800-227-8766 (toll free), Fax: 978-557-5100, www.mksinst.com

MTI Instruments, Inc., 325 Washington Ave., Extension Albany, NY 12205-5505, Phone: 518-218-2550, Fax: 518-218-2506, www.mtiinstruments.com

Pilz Automation Safety L.P., 7150 Commerce Blvd., Canton, MI 48187, Phone: 866-889-1726 (toll free), Fax: 734-354-3355, www.pilz.us/index.jsp

RDP Electrosense, Inc., 2216 Pottstown Pike, Pottstown, PA 19465, Phone: 610-469-0850, 800-334-5838 (toll free), Fax: 610-469-0852, www.rdpe.com

Schneider Electric-Hyde Park, 1875 Founders Dr., Dayton, OH 45420, Phone: 937-514-7207, Fax: 937-258-5830, www.sesensors.com

SICK, Inc., 6900 W. 110th St., Bloomington, MN 55438, Phone: 952-941-6780, 800-325-7425 (toll free), Fax: 952-941-9287, www.sick.com/us/en-us/home/Pages/Homepag…

TR Electronic, P.O. Box 2543, Station B, London, ON N6A4G9, Phone: 800-265-9483 (toll free), Fax: 519-452-1177, www.trelectronic.com

Turck, Inc., 3000 Campus Dr., Plymouth, MN 55441-2619, Phone: 763-553-7300, 800-554-4989 (toll free), Fax: 763-553-0708, www.turck.com

Vaisala, Inc., 10-D Gill St., Woburn, MA 01801-1068, Phone: 781-933-4500, 888-824-7252 (toll free), Fax: 781-933-8029, www.vaisala.com/en/lifesciencehightech/p…



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

26

Piezoelectric Sensors and Transducers

Ahmad Safari
Rutgers University

Victor F. Janas
Rutgers University

Amit Bandyopadhyay
Rutgers University

Andrei Kholkine
Rutgers University

26.1 Governing Equations and Coefficients **26**-1
26.2 Piezoelectric Materials **26**-3
Single Crystals • Piezoelectric Ceramics • Perovskites • Processing of Piezoelectric Ceramics • Piezoelectric Polymers • Piezoelectric Ceramic/Polymer Composites
26.3 Measurements of Piezoelectric Effect **26**-7
26.4 Applications **26**-10
Suppliers of Piezoelectric Materials
Defining Terms **26**-12
References **26**-13
Further Readings **26**-13

Piezoelectricity, discovered in Rochelle salt in 1880 by Jacques and Pierre Curie, is the term used to describe the ability of certain materials to develop an electric charge that is proportional to a direct applied mechanical stress. These materials also show the converse effect; that is, they will deform (strain) proportionally to an applied electric field. Some crystalline materials show piezoelectric behavior due to their unique crystal structure. The lattice structure of a crystal is described by the Bravais unit cell [1]. There are 230 microscopic symmetry types (space groups) in nature, based on the several symmetry elements such as translation, inversion center, mirror plane, or rotation axes. The combinations of these symmetry elements yield the macroscopic symmetry known as point groups. All natural crystals can be grouped into 32 different classes (point groups) based on their symmetry elements. The 32 point groups can be further classified into two subgroups: (1) crystals with a center of symmetry and (2) crystals with no center of symmetry. The 11 centrosymmetric subgroups do not show piezoelectricity. Of the 21 non-centrosymmetric groups, 20 show the piezoelectric effect along unique directional axes. An important class of piezoelectric materials includes ferroelectrics, in which the piezoelectric effect is closely related to the ferroelectric polarization that can be reversed by the application of sufficiently high electric field [2,3]. To induce piezoelectric properties in ferroelectric materials, a poling procedure is often required, which consists of the temporary application of a strong electric field. Poling is analogous to the magnetizing of a permanent magnet.

26.1 Governing Equations and Coefficients

The phenomenological master equation describing the deformations of an insulating crystal subject to both elastic and electric stresses is given by

$$x_{ij} = si_{ijkl}X_{kl} + d_{mij}E_m + M_{mnij}E_mE_n \tag{26.1}$$

where

x_{ij} are the components of elastic strain
s_{ijkl} is the elastic compliance tensor
X_{kl} are the stress components
d_{mij} are the piezoelectric tensor components
M_{mnij} is the electrostrictive tensor
E_m and E_n are the components of the electric field

Neglecting the second-order effects (electrostriction) and assuming that the material is under no stress ($X_{kl} = 0$), the elastic strain is given by

$$x_{ij} = d_{mij}E_m \tag{26.2}$$

Equation 26.2 is the mathematical definition of the converse piezoelectric effect, where induced strain is directly proportional to the first power of the field. The thermodynamically equivalent direct piezoelectric effect is given by

$$P_m = d_{mij}X_{ij} \tag{26.3}$$

where P_m is the component of electric polarization. The difference between the direct and the converse piezoelectric effect is shown schematically in Figure 26.1. The converse effect describes the actuating function of a piezoelectric, where a controlled electric field accurately changes the shape of a piezoelectric material. The sensing function of a piezoelectric is described by the direct effect, where a controlled stress on a piezoelectric material yields a charge proportional to the stress.

The third-rank tensor d_{mij} may be simplified using matrix notation [1] to d_{ij}, a second-rank tensor. In this form, d_{ij} is simply known as the piezoelectric charge coefficient, with units of coulombs per newton (C N^{-1}) or meters per volt (m V^{-1}). Another set of moduli that may be used to characterize piezoelectric materials are the piezoelectric voltage coefficients, g_{ij}, defined in matrix notation as

$$E_i = g_{ij}X_j \tag{26.4}$$

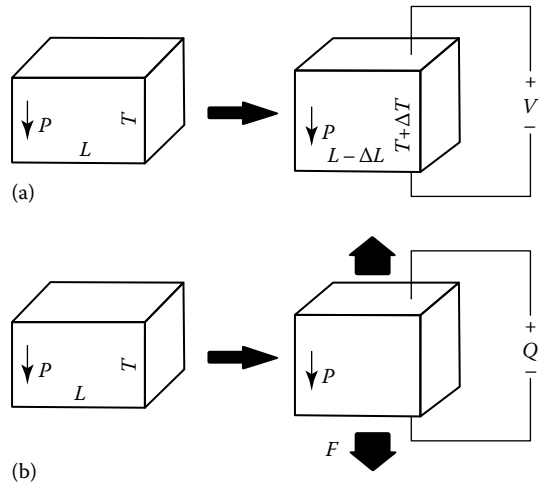


FIGURE 26.1 Schematic representations of the direct and converse piezoelectric effect: (a) An electric field applied to the material changes its shape, (b) a stress on the material yields a surface charge.

where E_i is the component of electric field arising due to the stress X_j. The d_{ij} and g_{ij} coefficients are related by

$$g_{ij} = \frac{d_{ij}}{\varepsilon_0 \varepsilon_{ii}} \tag{26.5}$$

where
ε_{ii} is the dielectric constant (relative permittivity)
ε_0 is the permittivity of free space (8.854×10^{-12} F m^{-1})

Another key property of piezoelectric materials is their electromechanical coupling coefficient k, defined as

$$k^2 = \frac{\text{Resulting mechanical energy}}{\text{Input electrical energy}} \tag{26.6}$$

or

$$k^2 = \frac{\text{Resulting electrical energy}}{\text{Input mechanical energy}} \tag{26.7}$$

The value of k represents the efficiency of the piezoelectric in converting one form of energy to another. Since energy conversion is never complete, the value of k^2 is always less than unity, so k is always less than 1.

Two final important parameters for piezoelectric (ferroelectric) materials are their Curie point (T_0) and Curie temperature (T_c). The Curie point is the temperature above which the material loses its ferroelectric and piezoelectric behavior. The Curie temperature is defined by the Curie–Weiss law:

$$\varepsilon = \varepsilon_0 + \frac{C}{T - T_c} \quad \text{for } T > T_c \tag{26.8}$$

where
ε is the dielectric constant of the material
C is the Curie–Weiss constant
T is the temperature

It represents the temperature where the material tends to have its highest dielectric constant. The Curie temperature is always lower (often within 10 °C) than the Curie point.

26.2 Piezoelectric Materials

26.2.1 Single Crystals

A number of single-crystal materials have demonstrated piezoelectricity. These materials dominate certain applications, such as frequency-stabilized oscillators in watches and radars and surface acoustic wave devices in television filters and analog signal correlators. A list of single-crystal piezoelectric materials includes quartz, lithium niobate and lithium tantalate, ammonium dihydrogen sulfate,

lithium sulfate monohydrate, and Rochelle salt. Recently, it was discovered that some relaxor-based ferroelectric single crystals of lead zinc niobate and lead magnesium niobate (PMN) and their solid solutions with lead titanate possess superior piezoelectric properties when compared to other piezoelectric materials.

Quartz, which is crystalline SiO_2, has a low value of d_{11} (12.3×10^{-12} C N^{-l}) [1]. The right-handed quartz develops a positive charge when put under compression and a negative charge when put under tension. The coupling coefficient k for quartz is also very low, typically around 0.1 [4]. In addition, the dielectric constant ε for quartz is small (~4) [5]. The Curie point, however, is relatively high ($T_0 \sim 573$ °C) [5], so quartz is stable for high-temperature applications. Despite the low piezoelectric properties, quartz is very abundant and inexpensive and has found a strong position in low-cost or high-temperature applications.

Piezoelectric behavior in lithium niobate ($LiNbO_3$) and lithium tantalate ($LiTaO_3$) was first studied in the mid-1960s [2]. Under shear, the d_{15} of $LiNbO_3$ and $LiTaO_3$ is 73 and 26 ($\times 10^{-12}$ C N^{-1}), respectively. Both have ε values of approximately 40. If cut correctly, they have coupling coefficient (k) values of 0.65 and 0.4, respectively. In addition, the Curie points for both are extremely high ($T_0 \sim 1210$ °C for $LiNbO_3$ and 620 °C for $LiTaO_3$). Both $LiNbO_3$ and $LiTaO_3$ are commonly used in infrared detectors.

Rochelle salt ($KNaC_4H_4O_6\ H_2O$) was first found to be piezoelectric in 1880 [2]. The d_{31} and k_{31} are 275×10^{-12} C N^{-1} and 0.65, respectively. The relative dielectric constant is approximately 350. Rochelle salt has two Curie points (lower $T_0 \sim 18$ °C and upper $T_0 \sim 24$ °C). It is highly soluble in water and is still extensively used in electroacoustic transducers.

Lead zinc niobate, $Pb(Zn_{1/3}Nb_{2/3})O_3$, and PMN, $Pb(Mg_{1/3}Nb_{2/3})O_3$, are typical relaxor materials characterized by the broad frequency-dependent maximum of dielectric constant versus temperature. The solid solutions of these materials with lead titanate, $PbTiO_3$, were shown to possess excellent piezoelectric properties when oriented along the [001] direction [6]. The piezoelectric charge coefficient d_{33} of 25×10^{-10} C N^{-1}, coupling coefficient k of more than 0.9, and the ultrahigh strain of 1.7% were achieved in $Pb(Zn_{1/3}Nb_{2/3})O_3$–$PbTiO_3$ solid solution. These single-crystal relaxor materials are now being intensively investigated and show great promise for future generations of piezoelectric transducers and sensors.

Other types of piezoelectric materials dominate the market for transducers. These materials include piezoelectric ceramics, piezoelectric polymers, and composites of piezoelectric ceramic with inactive polymers. The focus of the remainder of this chapter will be on non-single-crystal piezoelectric materials.

26.2.2 Piezoelectric Ceramics

In polycrystalline ceramics with polar grains, the randomness of the grains, as shown schematically in Figure 26.2a, yields a nonpiezoelectric material. The piezoelectric behavior is induced by "poling" the ceramic. By applying a strong dc electric field at a temperature just below the Curie temperature, the spontaneous

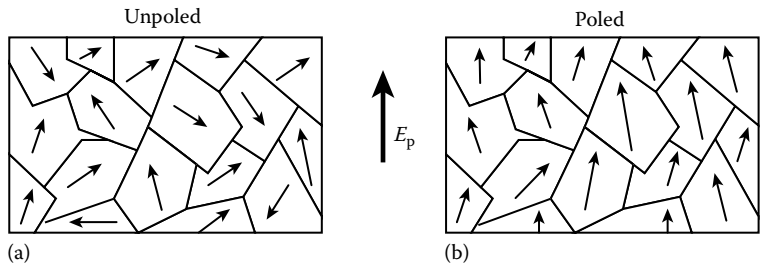


FIGURE 26.2 Schematic of the poling process in piezoelectric ceramics: (a) In the absence of an electric field, the domains have random orientation of polarization, (b) the polarization within the domains is aligned in the direction of the electric field.

polarization in each grain gets oriented toward the direction of the applied field. This is schematically shown in Figure 26.2b. Although all of the domains in a ceramic can never be fully aligned along the poling axis due to symmetry limitations, the ceramic ends up with a net polarization along the poling axis.

The largest class of piezoelectric ceramics is made up of mixed oxides containing corner-sharing octahedra of O^{2-} ions. The largest structure type, built with corner-shared oxygen octahedra, is the perovskite family, which is discussed in the following section.

26.2.3 Perovskites

Perovskite is the name given to a group of materials with general formula ABO_3 having the same structure as the mineral calcium titanate ($CaTiO_3$). Piezoelectric ceramics having this structure include barium titanate ($BaTiO_3$), lead titanate ($PbTiO_3$), lead zirconate titanate ($PbZr_xTi_{1-x}O_3$, or PZT), lead lanthanum zirconate titanate ($Pb_{1-x}La_x(Zr_yT_{1-y})_{1-x/4}O_3$, or PLZT), and PMN ($PbMg_{1/3}Nb_{2/3}O_3$). Several of these ceramics are discussed as follows.

The piezoelectric effect in $BaTiO_3$ was discovered in the 1940s [4], and it became the first piezoelectric ceramic developed. It replaced Rochelle salt because it is more stable, has a wider temperature range of operation, and is easily manufacturable. The Curie point, T_0, is about 130 °C. Above 130 °C, a nonpiezoelectric cubic phase is stable, where the center of positive charge (Ba^{2+} and Ti^{4+}) coincides with the center of the negative charge (O^{2-}) (Figure 26.3a). When cooled below the Curie point, a tetragonal structure (shown in Figure 26.3b) develops where the center of positive charge is displaced relative to the O^{2-} ions, leading to the formation of electric dipoles. Barium titanate has a relative dielectric constant ε_{33} of 1400 when unpoled and 1900 when poled [2,4]. The d_{15} and d_{33} coefficients of $BaTiO_3$ are 270 and 191 × 10^{-12} C N^{-1}, respectively. The k for $BaTiO_3$ is approximately 0.5. The large room temperature dielectric constant in barium titanate has led to its wide use in multilayer capacitor applications.

Lead titanate, $PbTiO_3$, first reported to be ferroelectric in 1950 [4], has a similar structure to $BaTiO_3$ but with a significantly higher Curie point (T_0 = 490 °C). Pure lead titanate is difficult to fabricate in bulk form. When cooled through the Curie point, the grains go through a cubic to tetragonal phase change, leading to large strain and ceramic fracturing. This spontaneous strain can be decreased by the addition of dopants such as Ca, Sr, Ba, Sn, and W. The calcium-doped $PbTiO_3$ [7] has a relative dielectric constant ε_{33} of 200, a d_{33} of 65 × 10^{-12} C N^{-1}, and a k of approximately 0.5. The addition of calcium results in a lowering of the Curie point to 225 °C. The main applications of lead titanate are hydrophones and sonobuoys.

PZT is a binary solid solution of $PbZrO_3$ (an antiferroelectric orthorhombic structure) and $PbTiO_3$ (a ferroelectric tetragonal perovskite structure) [2–4]. It has a perovskite structure, with the Zr^{4+} and Ti^{4+} ions occupying the B site of the ABO_3 structure at random. At the morphotropic phase

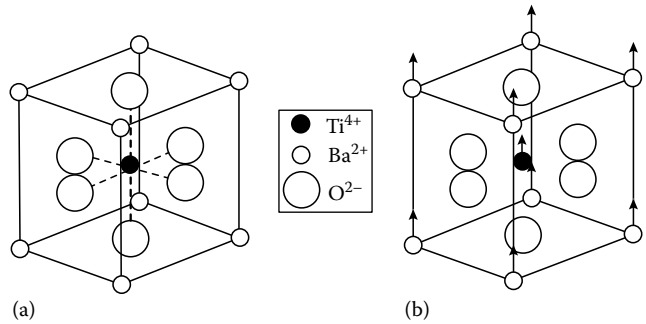


FIGURE 26.3 Crystal structure of $BaTiO_3$: (a) Above the Curie point, the cell is cubic; (b) below the Curie point, the cell is tetragonal with Ba^{2+} and Ti^{4+} ions displaced relative to O^{2-} ions.

boundary (MPB) separating the tetragonal and orthorhombic phases, PZT shows excellent piezoelectric properties. At room temperature, the MPB is at a Zr/Ti ratio of 52/48, resulting in a piezoelectric ceramic that is extremely easy to pole. Piezoelectric PZT at the MPB is usually doped by a variety of ions to form what are known as "hard" and "soft" PZTs. Hard PZT is doped with acceptor ions, such as K^+ or Na^+ at the A site or Fe^{3+}, Al^{3+}, or Mn^{3+} at the B site. This doping lowers the piezoelectric properties and makes the PZT more difficult to pole or de-pole. The typical piezoelectric properties of hard PZT include [5,7] a Curie point, T_0, of 365 °C; a ε_{33} of 1700–1750 (poled); a piezoelectric charge coefficient d_{33} of 360 to 370 × 10^{-12} C N^{-1}; and a coupling coefficient of about 0.7. Soft PZT is doped with donor ions such as La^{3+} at the A site or Nb^{5+} or Sb^{5+} at the B site. It has very high piezoelectric properties and is easy to pole or de-pole. The typical piezoelectric properties of soft PZT include [5,7] a Curie point, T_0, of 210 °C; a relative dielectric constant ε_{33} of 3200–3400 (poled); a d_{33} of 580 to 600 × 10^{-12} C N^{-1}; and a coupling coefficient k_{33} of 0.7.

26.2.4 Processing of Piezoelectric Ceramics

The electromechanical properties of piezoelectric ceramics are largely influenced by their processing conditions. Each step of the process must be carefully controlled to yield the best product. Figure 26.4 is a flowchart of a typical oxide manufacturing process for piezoelectric ceramics. The high-purity raw materials are accurately weighed according to their desired ratio and mechanically or chemically mixed. During the calcination step, the solid phases react to yield the piezoelectric phase. After calcining, the solid mixture is ground into fine particles by milling. Shaping is accomplished by a variety of ceramic processing techniques, including powder compaction, tape casting, slip casting, or extrusion. During the shaping operation, organic materials are typically added to the ceramic powder to improve its flow and binding characteristics. These organics are removed in a low-temperature (500 °C –600 °C) binder burn-off step.

After burnout, the ceramic structure is sintered to an optimum density at an elevated temperature. For the lead-containing piezoelectric ceramics ($PbTiO_3$, PZT, PLZT), sintering is performed in sealed crucibles with an optimized PbO atmosphere. This is because lead loss occurs in these ceramics above 800 °C. As mentioned earlier (Figure 26.2), the randomness of the ceramic grains yields a nonpiezoelectric material. By electroding the ceramic and applying a strong dc electric field at high temperature, the ceramic is poled. At this point, the piezoelectric ceramic is ready for final finishing and characterization.

26.2.5 Piezoelectric Polymers

The piezoelectric behavior of polymers was first reported in 1969 [8]. The behavior results from the crystalline regions formed in these polymers during solidification from the melt. When the polymer is drawn, or stretched, the regions become polar and can be poled by applying a high electric field. The most widely known piezoelectric polymers are polyvinylidene fluoride [9,10], also known as PVDF, PVDF—trifluoroethylene copolymer or P(VDF–TrFE) [9,10], and odd-number nylons, such as Nylon-11 [11].

The electromechanical properties of piezoelectric polymers are significantly lower than those of piezoelectric ceramics. The d_{33} values for PVDF and P(VDF–TrFE) are approximately 33 (× 10^{-12} C N^{-1}), and

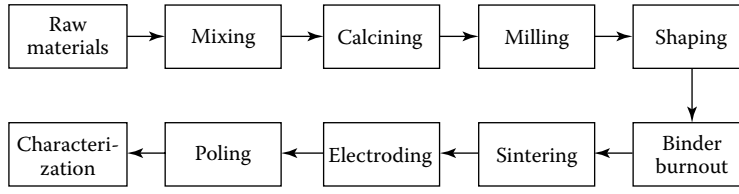


FIGURE 26.4 Flowchart for the processing of piezoelectric ceramics.

TABLE 26.1 Advantages (+) and Disadvantages (–) of Piezoelectric Ceramics, Polymers, and Composites

Parameter	Ceramic	Polymer	Ceramic/Polymer Composite
Acoustic impedance	High (–)	Low (+)	Low (+)
Coupling factor	High (+)	Low (–)	High (+)
Spurious modes	Many (–)	Few (+)	Few (+)
Dielectric constant	High (+)	Low (–)	Medium (+)
Flexibility	Stiff (–)	Flexible (+)	Flexible (+)
Cost	Cheap (+)	Expensive (–)	Medium (+)

Source: Adapted from Gururaja, T. R., *Am. Ceram. Soc. Bull.*, 73, 50, 1994.

the dielectric constant ε is in the range 6–12 [12,13]. They both have a coupling coefficient (k) of 0.20 and a Curie point (T_0) of approximately 100 °C. For Nylon-11, ε is around 2 [11], while k is approximately 0.11.

26.2.6 Piezoelectric Ceramic/Polymer Composites

As mentioned earlier, a number of single-crystal, ceramic, and polymer materials exhibit piezoelectric behavior. In addition to the monolithic materials, the composites of piezoelectric ceramics with polymers have also been formed. Table 26.1 [14] summarizes the advantages and disadvantages of each type of material. Ceramics are less expensive and easier to fabricate than polymers or composites. They also have relatively high dielectric constants and good electromechanical coupling. However, they have high acoustic impedance and are therefore a poor acoustic match to water, the media through which it is typically transmitting or receiving a signal. Also, since they are stiff and brittle, monolithic ceramics cannot be formed onto curved surfaces, limiting design flexibility in the transducer. Finally, they have a high degree of noise associated with their resonant modes. Piezoelectric polymers are acoustically well matched to water, are very flexible, and have few spurious modes. However, the applications for these polymers are limited by their low electromechanical coupling, low dielectric constant, and high cost of fabrication. Piezoelectric ceramic/polymer composites have shown superior properties when compared to single-phase materials. As shown in Table 26.1, they combine high coupling, low impedance, few spurious modes, and an intermediate dielectric constant. In addition, they are flexible and moderately priced.

26.3 Measurements of Piezoelectric Effect

Different means have been proposed to characterize the piezoelectric properties of materials. The resonance technique involves the measurement of the characteristic frequencies when the suitably shaped specimen (usually ceramic) is driven by a sinusoidally varying electric field. To a first approximation, the behavior of a piezoelectric sample close to its fundamental resonance frequency can be represented by an equivalent circuit as shown in Figure 26.5a. The schematic behavior of the reactance of the sample as a function of frequency is represented in Figure 26.5b. By measuring the characteristic frequencies of the sample, the material constants including piezoelectric coefficients can be calculated. The equations used for the calculations of the electromechanical properties are described in the IEEE Standard on piezoelectricity [15]. The simplest example of piezoelectric measurements by resonance technique relates to a piezoelectric ceramic rod (typically 6 mm in diameter and 15 mm long) poled along its length. It can be shown that the coupling coefficient k_{33} is expressed as a function of the series and parallel resonance frequencies, f_s and f_p, respectively:

$$k_{33}^2 = \frac{\pi}{2}\frac{f_s}{f_p}\tan\left(\frac{\pi}{2}\frac{f_p - f_s}{f_p}\right) \tag{26.9}$$

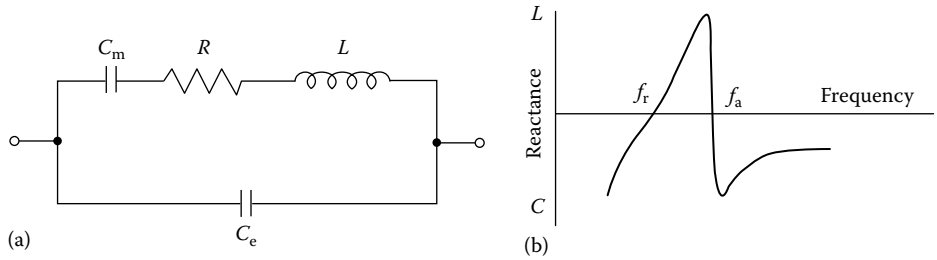


FIGURE 26.5 (a) Equivalent circuit of the piezoelectric sample near its fundamental electromechanical resonance (top branch represents the mechanical part and bottom branch represents the electric part of the circuit) and (b) electric reactance of the sample as a function of frequency.

The longitudinal piezoelectric coefficient d_{33} is calculated using k_{33}, elastic compliance s_{33}^E, and low-frequency dielectric constant ε_{33}^X:

$$d_{33} = k_{33}\left(\varepsilon_{33}^X s_{33}^E\right)^{1/2} \tag{26.10a}$$

Similarly, other electromechanical coupling coefficients and piezoelectric moduli can be derived using different vibration modes of the sample. The disadvantage of the resonance technique is that measurements are limited to the specific frequencies determined by the electromechanical resonance. It is used mostly for the rapid evaluation of the piezoelectric properties of ceramic samples whose dimensions can be easily adjusted for specific resonance conditions.

Subresonance techniques are frequently used to evaluate piezoelectric properties of materials at frequencies much lower than the fundamental resonance frequency of the sample. They include both the measurement of piezoelectric charge under the action of external mechanical force (direct effect) and the measurement of electric field–induced displacements (converse effect). In the latter case, the displacements are much smaller than in resonance; however, they still can be measured by using strain gages, capacitive sensors, linear variable differential transformer (LVDT) sensors, or optical interferometry [16,17].

A direct method is widely used to evaluate the sensor capabilities of piezoelectric materials at sufficiently low frequency. The mechanical deformations can be applied in different modes such as thickness expansion, transverse expansion, thickness shear, and face shear to obtain different components of the piezoelectric tensor. In the simplest case, the metal electrodes are placed onto the major surfaces of the piezoelectric transducer normal to its poling direction (direction of ferroelectric polarization), and the mechanical force is applied along this direction (Figure 26.1b). Thus, the charge is produced on the electrode plates under mechanical loading, which is proportional to the longitudinal piezoelectric coefficient d_{33} of the material. To relate the output voltage of the transducer to the piezoelectric charge, it is necessary to consider the equivalent circuit (Figure 26.6a). A circuit includes the charge generator, $Q = d_{33}F$; the leakage resistor of the transducer, R_s; the transducer capacitance, C_s; the capacitance of the connecting cables, C_c; and the input resistance and capacitance of the amplifier, R_a and C_a, respectively. Here, F denotes the force applied to the transducer (tensile or compressive). All the resistances and capacitances shown in Figure 26.6a can be combined, as shown in Figure 26.6b. A charge generator can be converted to a current generator, I, according to

$$I = \frac{dQ}{dt} = d_{33}\frac{dF}{dt} \tag{26.10b}$$

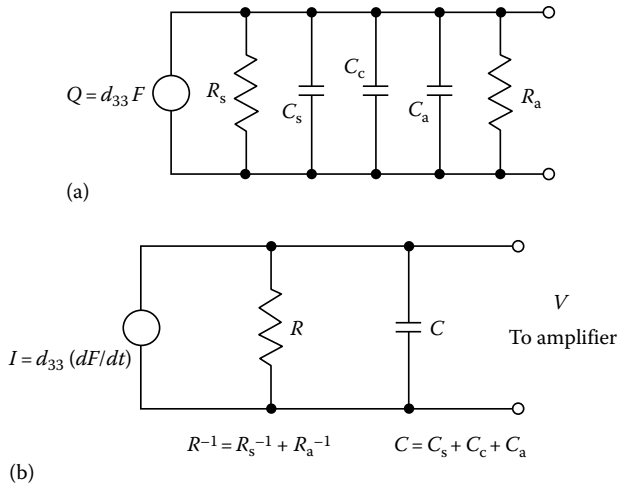


FIGURE 26.6 (a) Full and (b) simplified equivalent electric circuits of the piezoelectric sensor connected to the voltage amplifier.

Assuming that the amplifier does not draw any current, the output voltage V at a given frequency ω can be calculated:

$$V = \frac{d_{33}F}{C}\frac{j\omega\tau}{1+j\omega\tau} \tag{26.11}$$

where $\tau = RC$ is the time constant that depends on all resistances and capacitances of the circuit. For sufficiently high frequency, the measured response is frequency independent, and d_{33} can be easily evaluated from Equation 26.11 if the equivalent capacitance C is known. Since C is determined by the parallel capacitances of the sample, connecting cables, and amplifier (typically not exactly known), the standard capacitance is often added to the circuit, which is much greater than all the capacitances involved. However, according to Equation 26.11, the sensitivity of the circuit is greatly reduced with decreasing C. If τ is not large enough, the low-frequency cutoff does not allow piezoelectric measurements in quasistatic or low-frequency conditions.

To overcome the difficulties of using voltage amplifiers for piezoelectric measurements, a so-called charge amplifier was proposed. The idealized circuit of a charge amplifier connected to the piezoelectric transducer is shown in Figure 26.7. Note that an FET-input operational amplifier is used with a capacitor

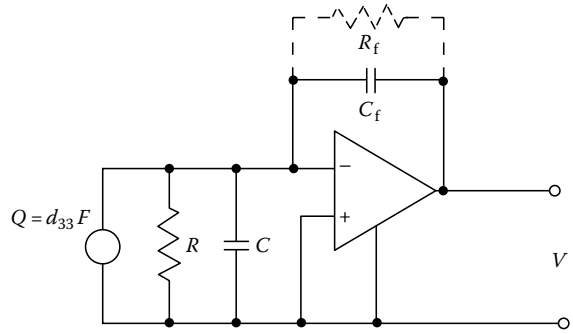


FIGURE 26.7 Equivalent electric circuit of the piezoelectric sensor connected to the charge amplifier.

C_f in the feedback loop. Assuming that the input current and the voltage of the operational amplifier are negligible, one can relate the charge on the transducer with the output voltage:

$$V = \frac{-Q}{C_f} = -d_{33}\frac{F}{C_f} \tag{26.12}$$

Equation 26.12 gives frequency-independent response where the output voltage is determined only by the piezoelectric coefficient d_{33} and the known capacitance C_f. Unfortunately, this advantage is difficult to realize since even a small input current of the amplifier will charge the feedback capacitor, leading to saturation of the amplifier. Therefore, a shunt resistor R_f is added to the circuit (dotted line in Figure 26.7), which prevents such charging. If one takes into account the *RC* circuit of the feedback loop, the output voltage will have the form of Equation 26.11, that is, becomes frequency dependent. In this case, the time constant τ is determined by the parameters of the feedback loop and does not depend on the capacitance of the transducer, connecting cables, or the input capacitance of the amplifier. This gives an important advantage to the charge amplifier when it is compared to the ordinary voltage amplifier.

26.4 Applications

The direct and converse piezoelectric effects in a number of materials have led to their use in electromechanical transducers. Electromechanical transducers convert electric energy to mechanical energy, and vice versa. These transducers have found applications where they are used in either passive or active modes. In the passive (sensor) mode, the transducer only receives signals. Here, the direct piezoelectric properties of the material are being exploited to obtain a voltage from an external stress. The applications in the passive mode include hydrophones, or underwater listening devices; microphones; phonograph pickups; gas igniters; dynamic strain gages; and vibrational sensors. In the active (actuator) mode, the transducer, using the converse piezoelectric properties of the material, changes its dimensions and sends an acoustic signal into a medium. The active-mode applications include nondestructive evaluation, fish/depth finders, ink-jet printers, micropositioners, micropumps, and medical ultrasonic imaging. Often, the same transducer is used for both sensor and actuator functions.

Two examples of piezoelectric sensors are given as follows. The first example is the ceramic transducer, which relates the deformation of the piezoelectric sensor to the output voltage via direct piezoelectric effect. Piezoceramics have high Young's moduli; therefore, large forces are required to generate strains in the transducer to produce measurable electric response. The compliance of the piezoelectric sensor can be greatly enhanced by making long strips or thin plates of the material and mounting them as cantilevers or diaphragms. The displacement of the cantilever end will result in a beam bending, leading to the mechanical stress in the piezoelectric material and the electric charge on the electrodes. A common configuration of the piezoelectric bender is shown in Figure 26.8a. Two beams poled in opposite directions are cemented together with one common electrode in the middle and two electrodes on the outer surfaces. The bending of such a bimorph will cause the upper beam to stretch and the lower beam to compress, resulting in a piezoelectric charge of the same polarity for two beams connected in series. To the first approximation, the charge Q appearing on the electrodes is proportional to the displacement Δl of the end of the bimorph via Equation 26.13 [18]:

$$Q = \frac{3}{8}\frac{Hw}{L}e_{31}\Delta l \tag{26.13}$$

where

H, w, and L are the thickness, the width, and the length of the bimorph, respectively

e_{31} is the transverse piezoelectric coefficient relating electric polarization and strain in a deformed piezoelectric material

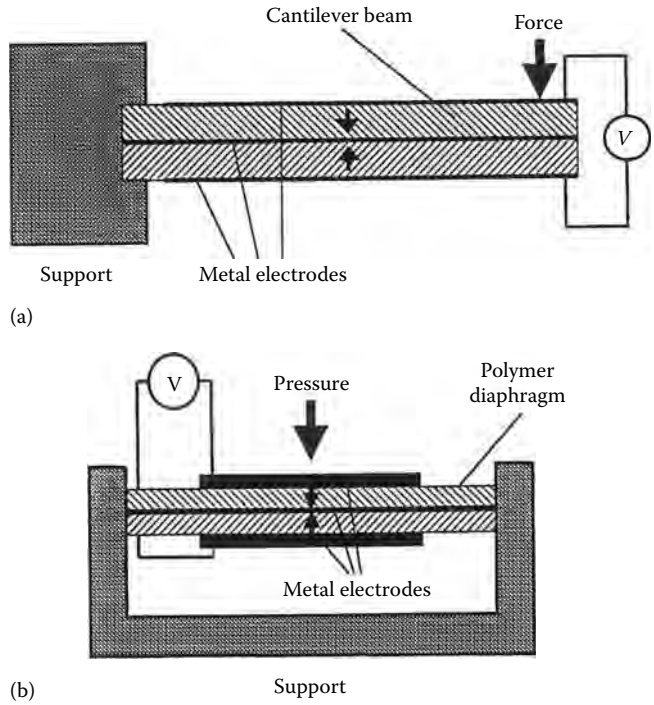


FIGURE 26.8 Schematic designs of the displacement sensor based on piezoelectric ceramic (a) and of the pressure sensor based on piezoelectric polymer film (b). The arrows indicate the directions of ferroelectric polarization in the piezoelectric material.

The charge can be measured either by the voltage amplifier (Figure 26.6) or by the charge amplifier (Figure 26.7).

In certain applications, the parameters of piezoelectric sensors can be improved by using ferroelectric polymers instead of single crystals and piezoceramics. Although the electromechanical properties of polymers are inferior to those of piezoelectric ceramics, their low dielectric constant offers the higher voltage response since they possess higher *g* piezoelectric coefficients. Also, the polymers are more mechanically robust and can be made in the form of thin layers (down to several micrometers). An example using the polymer bimorph as a pressure sensor is shown in Figure 26.8b. A circular diaphragm composed of two oppositely poled polymer films is clamped along its edges to a rigid surround, forming a microphone. The voltage appearing on the electrodes is proportional to the applied pressure *p* by Equation 26.14 [19]:

$$V = \frac{3}{16}\frac{d_{31}}{\varepsilon_{33}}\frac{D^2}{h}(1-\nu)p \tag{26.14}$$

where

D and h are the diameter and the thickness of the diaphragm, respectively

ν is the Poisson ratio

The high d_{31}/ε_{33} value for polymer sensors is advantageous to obtain higher voltage response. According to Equation 26.11, this advantage can be realized only if the high-input-impedance amplifier is used in close proximity to the transducer to reduce the influence of the connecting cables.

TABLE 26.2 Suppliers of Piezoelectric Materials and Sensors

Name	Address	Ceramic	Polymer	Composite
AMP Sensors	950 Fore Ave., Morristown, PA 19403, Phone: (610) 650-1500, Fax: (610) 650-1509		X	
Krautkramer Branson	50 Industrial Park Rd., Lewistown, PA 17044, Phone: (717) 242-0327, Fax: (717) 242-2606			X
Materials Systems, Inc.	531 Great Road, Littleton, MA 01460, Phone: (508) 486-0404, Fax: (508) 486-0706, http://www.materialsystems.com/			X
Morgan Matroc, Inc.	232 Forbes Rd., Bedford, OH 44146, Phone: (216) 232-8600, Fax: (216) 232-8731 http://www.morganelectroceramics.com/	X		
Sensor Technology Ltd.	20 Stewart Rd., P.O. Box 97, Collingwood, Ontario, Canada, Phone: +1 (705) 444-1440, Fax: +1 (705) 444-6787, http://www.sensortech.ca/site/index.cfm	X		
Staveley Sensors, Inc.	91 Prestige Park Circle, East Hartford, CT 06108, Phone: (860) 289-5428, Fax:(860) 289-3189, http://www.sensortech.ca/site/index.cfm			X
CTS Valpey Corporation	Hopkinton, MA 01748, Phone: (508) 435-6831, Fax: (508) 435-5289, http://www.ctsvalpey.com/		X	
Vermon USA	6288 SR 103 North Bldg. 37, Lewistown, PA 17044, Phone: (717) 248-6838, Fax:(717) 248-7066	X		
TRS Ceramics, Inc.	2820 E. College Ave., State College, PA 16801, Phone: (814) 238-7485, Fax: (814) 238-7539 http://www.trstechnologies.com/Products/piezoelectric_ceramics.php	X		

26.4.1 Suppliers of Piezoelectric Materials

Table 26.2 lists a number of the suppliers of piezoelectric materials, their addresses, and whether they supply piezoelectric ceramic, polymers, or composites. Most of them tailor the material to specific applications.

Defining Terms

Charge amplifier: An operational amplifier used to convert the input charge into output voltage by means of the capacitor in the feedback loop

Coupling coefficients: The materials constants that describe an ability of piezoelectric materials to convert electric energy into mechanical energy and vice versa

Ferroelectrics: A subgroup of piezoelectric materials possessing a net dipole moment (ferroelectric polarization) that can be reversed by the application of sufficiently high electric field

Piezoelectric coefficients: The materials constants that are used to describe the linear coupling between electric and mechanical parameters of the piezoelectric

Piezoelectric composites: The materials containing two or more components with different piezoelectric properties

Piezoelectric transducer: A device that converts the input electric energy into mechanical energy and vice versa via piezoelectric effect

Poling: The process of aligning the ferroelectric polarization along a unique (poling) direction

References

1. J. F. Nye, *Physical Properties of Crystals*, Oxford, U.K.: Oxford University Press, 1985.
2. Y. Xu, *Ferroelectric Materials and Their Applications*, Amsterdam, the Netherlands: North-Holland, 1991.
3. L. E. Cross, Ferroelectric ceramics: Tailoring properties for specific applications, In N. Setter and E. L. Colla (eds.), *Ferroelectric Ceramics: Tutorial Reviews, Theory, Processing, and Applications*, Basel, Germany: Birkhauser, 1993.
4. B. Jaffe, W. R. Cook, Jr., and H. Jaffe, *Piezoelectric Ceramics*, Marietta, OH: R.A.N., 1971.
5. *The User's Guide to Ultrasound & Optical Products*, Hopkinton, MA: Valpey-Fisher Corporation, 1996.
6. S.-E. Park and T. R. Shrout, Relaxor based ferroelectric single crystals with high piezoelectric performance, *Proceedings of the 8th US–Japan Seminar on Dielectric and Piezoelectric Ceramics*, October 15–18, Plymouth, MA, 1997, p. 235.
7. *Piezoelectric Products*, Collingwood, Ontario, Canada: Sensor Technology Limited, 1991.
8. H. Kawai, The piezoelectricity of poly(vinylidene fluoride), *Jpn. J. Appl. Phys*., 8, 975, 1969.
9. L. F. Brown, Ferroelectric polymers: Current and future ultrasonic applications, *Proceedings of the 1992 IEEE Ultrasonics Symposium*, New York, 1992, p. 539.
10. T. Furukawa, Recent advances in ferroelectric polymers, *Ferroelectrics*, 104, 229, 1990.
11. L. F. Brown, J. I. Scheinbeim, and B. A. Newman, High frequency dielectric and electromechanical properties of ferroelectric nylons, *Proceedings of the 1994 IEEE Ultrasonics Symposium*, New York, 1995, p. 337.
12. *Properties of Raytheon Polyvinylidene Fluoride (PVDF)*, Raytheon Research Division, Lexington, MA, 1990.
13. *Standard and Custom Piezo Film Components*, Valley Forge, PA: Atochem Sensors, Inc., 1991.
14. T. R. Gururaja, Piezoelectric transducers for medical ultrasonic imaging, *Am. Ceram. Soc. Bull*., 73, 50, 1994.
15. IEEE standards on piezoelectricity, *IEEE Std*. 176, 1978.
16. W. Y. Pan and L. E. Cross, A sensitive double beam laser interferometer for studying high-frequency piezoelectric and electrostrictive strains, *Rev. Sci. Instrum*., 60, 2701, 1989.
17. A. L. Kholkin, Ch. Wuethrich, D. V. Taylor, and N. Setter, Interferometric measurements of electric field-induced displacements in piezoelectric thin films, *Rev. Sci. Instrum*., 67, 1935, 1996.
18. A. J. Moulson and J. M. Herbert, *Electroceramics: Materials, Properties, Applications*, London, U.K.: Chapman and Hall, 1990.
19. J. M. Herbert, *Ferroelectric Transducers and Sensors*, New York: Gordon and Breach, 1982.

Further Readings

Achenbach, J. D., *Reciprocity for Piezoelectric System*, Cambridge, U.K.: Cambridge University Press, Online publication, December 2009, http://ebooks.cambridge.org/chapter.jsf?bid=CBO9780511550485&cid=CBO9780511550485A122 (accessed on June 25, 2012).

Arnau, A., *Piezoelectric Transducers and Applications*, Berlin, Germany: Springer, 2004.

Tichy, J., J. Erhart, E. Kittinger, and J. Privratska, *Fundamentals of Piezoelectric Sensorics*, Berlin, Germany: Springer, 2010.

Wolfbeis, O.S., C. Steinem, and A. Janshoff, *Piezoelectric Sensors*, Berlin, Germany: Springer, 2007.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

27

Laser Interferometer Displacement Sensor

27.1 Introduction **27**-1
27.2 Helium–Neon Laser **27**-2
Frequency Stabilization of He–Ne Lasers • Longitudinal Zeeman Effect
27.3 Refractive Index of Air **27**-5
Other Sources of Error in Interferometers
27.4 Michelson Interferometer **27**-7
27.5 Two-Frequency Heterodyne Interferometer **27**-10
27.6 Single-Mode Homodyne Interferometer **27**-10
27.7 Displacement Sensor Based on a Fabry–Pérot Interferometer **27**-12
27.8 Interferometer Signal Processing **27**-12
27.9 Typical Interferometer Applications **27**-14
27.10 Conclusions **27**-15
27.A Appendix **27**-15
Defining Terms **27**-16
References **27**-16

Bernhard Günther Zagar
Technical University of Graz

27.1 Introduction

In the past few years, very high precision, numerically controlled machine tools have been developed. To achieve the potential precision of these tools, length and displacement measurements whose resolution exceeds the least significant digit of the tool must be made. The measurement equipment typically would not rely on mechanical scales.

Laser interferometers compare the changes in optical path length to the wavelength of light, which can be chosen from atomic constants that can be determined with very little uncertainty.

In 1983, there was a redefinition of the meter [1] that was previously defined in 1960. The old definition was based on the wavelength of a certain radiation (the krypton-86 standard) that could not be realized to better than four parts in 10^9. The new definition, being based on frequency but not related to any particular radiation, opened the way to significant improvements in the precision with which the meter can be realized. As recommended in resolution 2 for the practical realization of the meter, the wavelength in vacuum λ_v of a plane electromagnetic wave of frequency f is $\lambda_v = c/f$, where c is the speed of light in vacuum, $c = 299,792,458$ m s^{-1} exactly. This way, the wavelength is related to *frequency* and *time*, which can be measured with the highest precision of all units within the Système International (SI).

In order to be independent of any environmental parameters, the meter is defined using the speed of light in a vacuum. However, interferometers usually must operate in ambient air.

Thus, environmental parameters that influence the speed of light in a particular medium (air) will affect and degrade the precision of the measurement.

Three major factors limit the absolute accuracy attainable with laser interferometers operating in ambient air: (1) the uncertainties of the vacuum wavelength, λ_v, of the laser source; (2) the uncertainty of the refractive index of the ambient air; and (3) the least-count resolution of the interferometer.

This chapter section is organized as follows. First, some basic laser principles are detailed, including ways to stabilize the vacuum wavelength of the laser. The effect most often used to stabilize lasers in commercial interferometers is the Zeeman effect, which yields relative uncertainties of 10^{-8}.

Second, the refractive index of air as another major factor limiting the attainable accuracy of laser interferometers operated in air is addressed. It is shown that it cannot be determined currently with uncertainty better than 5×10^{-8}.

And finally, the chapter section describes the most widely used Michelson interferometer and two of its variants for long-travel length measurement and gives their resolution.

27.2 Helium–Neon Laser

In order to attain the best possible accuracy, great care must be taken to ensure the highest wavelength stability of the light source. Almost all interferometric dimensional gages utilize a helium–neon laser because it has proven reliable, its emitted wavelength is in the visible range at about 633 nm, and it can be stabilized sufficiently well utilizing the Zeeman effect and to an even higher degree with the use of a very well-defined iodine absorption line also at ≈633 nm [1,2].

The helium–neon laser consists of a discharge tube as shown in Figure 27.1 [3] filled with the single-isotope gases helium (He^3) at a partial pressure of ≈105 Pa and neon (Ne^{20}) with a partial pressure of ≈13 Pa. It is pumped electrically using a voltage on the order of kilovolts with a current of a few milliamperes to excite both helium and neon atoms. Since the helium gas is the major component, it dominates the discharge properties of the laser tube. Neutral helium atoms collide with free electrons that are accelerated by the axial voltage and become excited and remain in two rather long-lived metastable states. These are close enough to the energy levels of certain excited states of neon atoms so that collisional energy transfer can take place between these two groups of atoms. Excited helium atoms may drop down to the ground state, while simultaneously neon atoms take up almost exactly the same amount of energy. Therefore, helium only serves to excite neon atoms; they do not contribute to the emission of light. The excited neon atoms remain in the excited state for a rather long period of time (on the order of 10^{-3} s). They return to lower energetic levels by stimulated emission of highly coherent light. This stimulated emission comes into effect when the light emitted by some neon atoms also prompts other atoms to emit. The mirrors of the laser cavity,

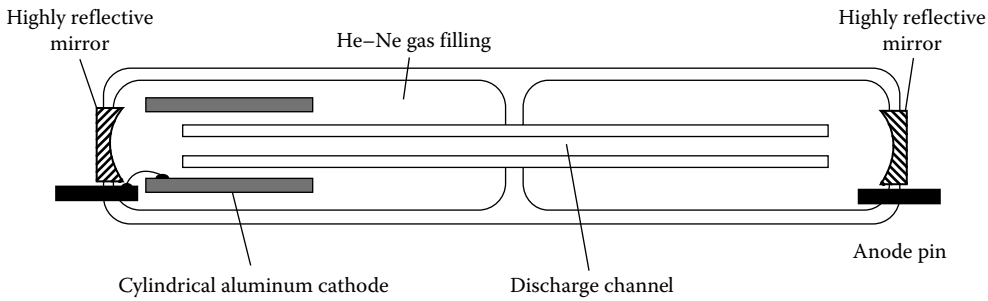


FIGURE 27.1 Schematics of the helium–neon laser. (Reprinted with permission from Siegman, A. E., *Lasers*, University Science Books, Mill Valley, CA.)

by reflecting most of the incident light, cause the light to traverse multiple paths through the active laser volume, thereby greatly amplifying the light if the cavity length L is an integer multiple m of half the wavelength λ:

$$L = m\frac{\lambda}{2} \tag{27.1}$$

The emitted light is fairly monochromatic but still has some finite spectral line width determined by the random emissions of Ne^{20}.

Brewster-angle [4] end windows of the discharge tube transmit light of the proper linear polarization if desired. The end mirrors have to be carefully polished. Their curvature radii have to satisfy the condition for stability. They have wavelength-selective dielectric coatings of very high reflectivity sometimes exceeding 99%.

Unless special precautions are taken, a He–Ne laser will emit several axial modes as shown schematically in Figure 27.1, resulting in a beat frequency that limits the temporal coherence and renders the laser unsuitable for interferometric purposes. Also, due to thermal expansion of the laser tube, the end mirrors will change their relative distance, thereby effectively tuning the wavelength within the line width of the gain curve.

Another important property of a high-quality He–Ne laser is its Gaussian cross-sectional profile, which is maintained along a propagating wave, that is, fundamental lateral mode. It is also a necessary condition for the wave fronts to remain quasiplanar.

27.2.1 Frequency Stabilization of He–Ne Lasers

The resonant frequency (and the wavelength λ) of the laser is determined in part by the distance between the two end mirrors and also by the refractive index n_M of the active medium (the He–Ne mixture). Since the line width of the gain profile of the active medium is usually in the gigahertz range, multiple axial modes can resonate in the cavity as is shown in Figure 27.2. The frequency difference Δf between two adjacent longitudinal modes is the *free spectral range* (FSR), which is given by Equation 27.2 and depends only on the cavity length, L; the refractive index of the active medium, n_M; and the speed of light in vacuum, c:

$$\Delta f = \frac{c}{2n_M L} \tag{27.2}$$

Due to thermal expansion of the laser cavity and/or thermally induced change in refractive index of the active medium, all resonating laser modes will move within the envelope of the gain profile. The effort undertaken in stabilizing the laser wavelength or equivalently stabilizing its frequency is aimed at locking the modes with respect to the gain profile and reducing the number of modes resonating simultaneously.

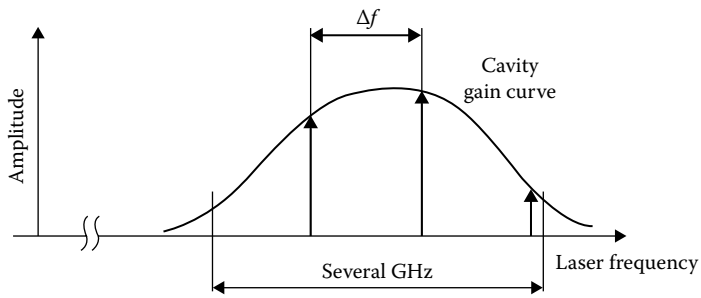


FIGURE 27.2 He–Ne laser can have multiple resonating modes (shown for ≈20 cm cavity length). (Reprinted with permission from Siegman, A. E., *Lasers*, University Science Books, Mill Valley, CA.).

27.2.2 Longitudinal Zeeman Effect

One of the most often used effects in stabilizing the frequency of a He–Ne laser for distance measurements is the Zeeman effect [3,5].

The tube of the laser whose frequency is to be stabilized is mounted inside an axial permanent magnet as shown in Figure 27.3. Because of a short cavity length chosen, the FSR Δf of the laser is large. Therefore, the laser emission appears as a single longitudinal mode when the cavity length is properly controlled. An externally applied magnetic field in longitudinal direction causes the transition frequencies of the neon atoms to split symmetrically to the nominal frequency, the separation in frequency being proportional to the magnetic field strength. This phenomenon is called the *longitudinal Zeeman effect*.

With the applied field strength, the single longitudinal laser mode of the He–Ne laser is split into right and left circularly polarized modes with a frequency difference of typically 1.5 MHz. The frequency difference depends on the laser frequency and exhibits an extreme value when the laser frequency matches exactly the atomic transition frequency. Frequency stabilization is accomplished with control of the cavity length so that the frequency is locked to the extreme value corresponding to the nominal frequency of the laser. In order to determine the actual laser frequency within the gain profile, the cavity length is periodically modulated with a piezoelectric actuator as shown in Figure 27.3. The frequency difference is constantly monitored at the minimum and maximum elongation of the modulation cycle. In the case when the extremal values are identical, the difference frequency will assume a maximum, and thus, the nominal value of the laser frequency is attained. This is achieved by controlling the length of the laser cavity. The cavity length is controlled both by using the thermal expansion of the laser tube, which is in close proximity to an electric heater whose operating current can be varied on a less dynamic scale, and on a very short time basis with a piezoelectric actuator attached to one of the end faces of the cavity [5].

This fast actuator allows for only limited amplitude, typically on the order of less than a wavelength. Therefore, the thermal expansion must be used to attain greater modulation amplitude. To obtain the highest possible stability of the laser frequency, the frequency differences are measured within a very short time interval, and the determined deviations from the nominal value are fed back in a closed control loop to the electromechanical and thermal positioning devices.

The short- and long-term frequency stability $\Delta f/f$ of such a stabilized laser can be as high as 2×10^{-8}, depending also on the parameters of the feedback loop.

Another phenomenon that exceeds the stabilization attainable with the Zeeman effect by far uses a particular absorption line of an iodine isotope ($^{127}I_2$, transition 11–5, R(127), component *i*) and locks the He–Ne laser [2,6] in that very well-defined frequency. Its wavelength $\lambda_1 = 632.9913981$ nm is very

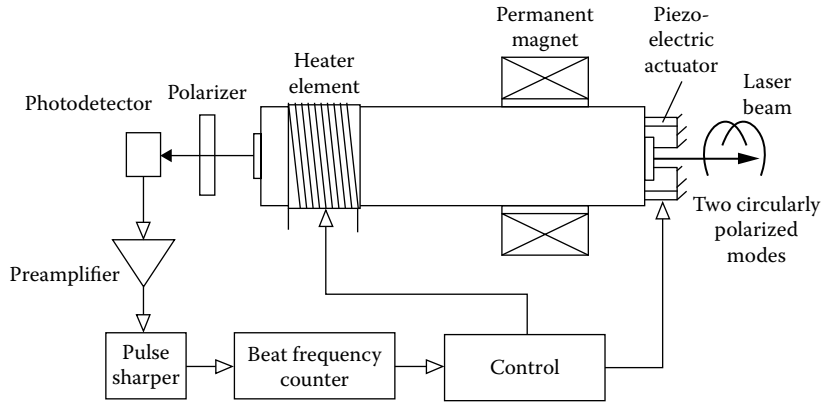


FIGURE 27.3 Laser wavelengths can be stabilized by utilizing the Zeeman effect. (Courtesy of Spindler & Hoyer, Inc., Milford, MA.)

close to one of the nominal wavelengths of the He–Ne laser, so this can easily be accomplished; however, the necessary equipment including the sophisticated electronics is rather involved [2]. For this reason, iodine-stabilized He–Ne lasers are only used where very high precision is required.

The overall estimated relative uncertainty of such a stabilized laser is $\pm 10^{-9}$ (which results from an estimated relative standard deviation of 3.4×10^{-10} [1]), which makes it suitable for the practical realization of the definition of primary and secondary standards.

27.3 Refractive Index of Air

The length scale of a displacement interferometer that is operated in ambient air is given by $\lambda_A = \lambda_V/n_A$, where λ_A is the wavelength of the source in air, λ_v is the wavelength in vacuum, and n_A is the refractive index of air. The vacuum wavelength λ_v is related to the speed of light by $\lambda_v = c/f_0$. The constant c, the speed of light in vacuum, was redefined in 1983 [1]. It is now defined to be exactly $c = 299{,}792{,}458$ m s^{-1}.

Thus, in order to measure distances interferometrically, it is essential to know the refractive index of air with an accuracy that is not less than the stability and the degree of certainty of the vacuum wavelength of the light source.

According to Edlén [7,8] and Owens [9], the refractive index is a function of the atmospheric pressure P, the temperature T, the relative humidity H, or alternatively the partial pressure of the water vapor in air e_s and the carbon dioxide concentration by volume D. The standard composition by volume of dry air is given by 78.03% N_2, 20.99% O_2, 0.933% Ar, 0.035% CO_2, and some other trace components, mainly noble gases [10]. The only component of the aforementioned list that might have variability is CO_2, which follows a long-term increasing behavior presumably associated with the combustion of fossil fuel [11]. The current value is ≈350 ppm by volume and is increasing by ≈1.4 ppm/year. The CO_2 concentration in air can also change in an industrial environment due to CO_2 emitters and can therefore show significant local variations.

More recently, the dependencies of the refractive index on nonnatural gaseous compounds like hydrocarbons in the air have been published [12] as well as corrections to Edlén's formulations [13–15]. Corrections to the refractive index due to those compounds may be necessary for very high-precision measurements in an industrial environment where chemical solvents or oils are in use. In Table 27.1, some nonnatural compounds, their maximum workplace concentrations (MWCs), and their effect on the refractive index if present at the location of the interferometer are listed.

TABLE 27.1 Refractive Indices n of Various Gaseous Compounds, Their MWC, and the Change in Refractive Index Caused by That Particular Concentration (for $T = 20$ °C, $P = 101.315$ hPa, $\lambda = 633$ nm)

Gas	Refractive Index $(n-1)10^4$	Concentration for $\Delta n/n = 10^{-7}$ in Air (ppm)	MWC (mg m^{-3})	$\Delta n/n \times 10^7$ due to MWC
Air	2.72			
Propane	10.3	130	1800	8
Butane	12.9	98	2350	10
Ethanol	8.1	190	1900	5
Ethyl acetate	13.0	97	1400	4
Dimethylketone	10.2	130	2400	9
Octane	23	50	2350	10
Chlorofluorocarbons (e.g., R12)	10.3	130	5000	7

Source: Reprinted with permission of VDI Verein Deutscher Ingenieure, G. Wilkening, Kompensation der Luftbrechzahl.

Jones [16] combined a precise determination of the density of moist air with Edlén's formulation to yield a somewhat simpler representation. For a typical iodine-stabilized He–Ne laser that has a vacuum wavelength of λ_v = 632.9913 nm, the Jones formulation is given by [17]

$$n(P,T,H,D) = 1 + A - B \tag{27.3}$$

where

$$\begin{aligned} A &= 78.603\left[1 + 0.540(D - 0.0003)\right]\frac{P}{TZ} \times 10^{-8} \\ B &= \left(0.00042066\, f_E e_s H\right) \times 10^{-8} \end{aligned} \tag{27.4}$$

where

P is the atmospheric pressure in pascals
T is the absolute temperature in kelvin
H is the relative humidity in %
D is the concentration of CO_2 in percent by volume

There are three additional factors in Jones' formulation that take the nonideal behavior of moist air as compared to an ideal gas into account. They are Z, a compressibility factor that reflects the nonideality of the air–water vapor mixture and that, for air containing reasonable amounts of CO_2 at a temperature between 15 °C and 28 °C and pressure of between 7×10^4 and 11×10^4 Pa, lies in the range between 0.99949 and 0.99979. f_E is an enhancement factor that expresses the fact that the effective saturation vapor pressure of water in air is greater than the saturation vapor pressure e_s. For the pressure and temperature ranges given earlier, f_E is bounded between 1.0030 and 1.0046 [16]. e_s is the saturation vapor pressure over a plane surface of pure liquid water and according to Jones is about 1705 Pa at a temperature of 15.0 °C and about 3,779 Pa for 28.0 °C. In the tables of Z, f_E and e_s are included in the Appendix of Jones' paper [16].

Table 27.2 gives an overview of the changes in environmental parameters that would cause a relative index change of 10^{-7}.

Edlén [8] and Jones [16] estimate that their empirical expressions for the dependency of the refractive index of air on the listed parameters have an absolute uncertainty of 5×10^{-8}.

Besides this fundamental limitation, there are some practical considerations that must be taken into account regarding the precision with which the environmental parameters can be measured. Estler [17] states that atmospheric pressure P can currently be determined with an uncertainty of ≈2.7 Pa, which can be assumed to be constant for the entire optical path of the interferometer if it is oriented horizontally.

TABLE 27.2 Parameters of Standard Air and Their Deviation to Cause a $\Delta n/n$ of 10^{-7}

Parameter	Standard Value	Variation for $\Delta n/n = +1 \times 10^{-7}$
Pressure P	101.3 kPa	+37.3 Pa
Temperature T	20.0°C	−0.1°C
Humidity H	40%	−10.0%
CO_2 concentration	350 ppm	+670 ppm

Source: Reprinted with permission of Estler, W. T., *J. Appl. Opt.*, 24(6), 808, 1985.

Please note that at sea level, the pressure gradient is ≈–13 Pa m^{-1}, resulting in a pressure-induced change in n_A of 3.4×10^{-8} m^{-1} if the measuring equipment is not kept level.

In an exceptionally well-controlled laboratory environment where special care is devoted to keep temperature gradients from affecting the refractive index along the optical path as much as possible, uncertainties of the temperature measurement can be as low as 0.01 °C according to [17]. Humidity measured with high-accuracy dew-point hygrometers can have uncertainties down to 0.5%. Changes in carbon dioxide concentrations have to be very significant (20% of the natural concentration) to cause a $\Delta n/n$ of 10^{-8}.

27.3.1 Other Sources of Error in Interferometers

Although the variation of the index of refraction of air is the most obvious source of error in typical interferometer applications, there are some often overlooked additional sources that are being discussed here briefly (from [18]):

- Cosine error results from the angular misalignment between the measurement axis and the axis of motion. Magnitude: $L[1 - \cos(\delta)]$ for $\delta = 0.1$ mrad and $L = 1$ m gives ≈ 4 nm in total.
- Optical alignment or better optical misalignment of the measurement beam with reference to the axis of motion can be visualized by observing the return beam of the measurement beam with respect to the reference beam. As the mechanical stage is moved, the reference signal will remain fixed, and any angular error (cosine error) of the measurement beam shows up as a runout in the beam overlap decreasing optical output level and thus increasing noise and uncertainty.
- *Target nonuniformity*: The target mirror must be flat to fractions of a wavelength λ in applications that require multiple axes of travel. A target mirror with a surface figure of $\lambda/10$ can contribute up to 63 (= $\lambda/10$) nm (for He–Ne lasers) of error as the stage travels along the axis parallel to the clear aperture of the mirror. In a measurement configuration where the beam reflects from the same location on the target mirror, this error source will be zero.
- Opposite axis error is often present in mechanical measuring systems. An opposite axis error is caused when perpendicular axes are not truly orthogonal to each other.
- Abbé error occurs when the axis of measurement is offset from the axis of interest. As first described by Dr. Ernst Abbé of Zeiss, "If errors of parallax are to be avoided, the measuring systems must be placed coaxially to the line in which displacement is to be measured on the workpiece."
- The mechanical stability or stiffness of the mechanical assembly is critical. If the physical relationship between the target optic and the point of interest changes during the measurement time, this is indistinguishable from actual motion. Vibration effects can be minimized by taking several measurements at one position and averaging them together.
- *Data age uncertainty*: Data age is defined as the difference in the time between when the object of interest is measured and when the user control system gets the position information. Data age uncertainty is defined as the maximum variation in the data age in a multi-axis system, due primarily to process variation in the electronics. Having minimum data age and data age uncertainty is critical for multi-axis high-velocity applications.

27.4 Michelson Interferometer

The basis for most interferometers used in interferometric dimensional gages is the classical Michelson interferometer [4] that is shown in Figure 27.4. The coherent monochromatic light of a wavelength-stabilized He–Ne laser is incident onto a beam splitter that splits the light into two equally intense beams (1) and (2).

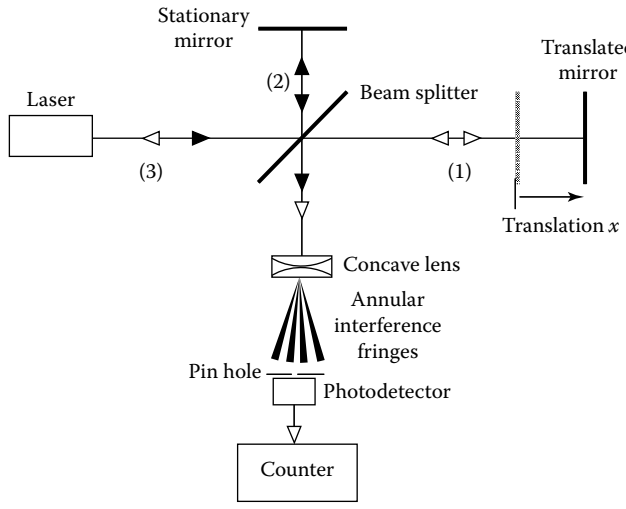


FIGURE 27.4 Schematics of the basic Michelson interferometer.

They are reflected off of both the stationary and the translatable mirror whose displacement x is to be measured and recombined at the splitter, where they are redirected toward a concave lens. Due to the coherence of the laser light, the wave fronts have a well-defined phase relation with respect to each other. This phase is determined by the difference between the optical path lengths of the two beams in arms 1 and 2. If this path difference is continuously changed by translating one of the mirrors, a sinusoidal

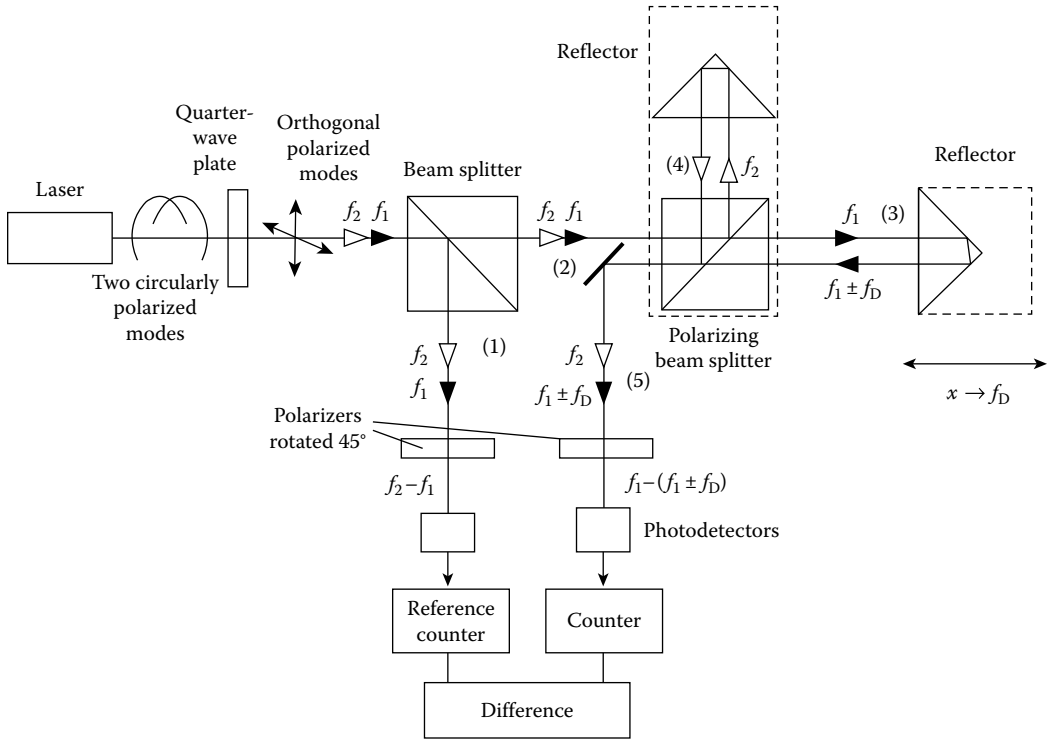


FIGURE 27.5 Two-frequency heterodyne interferometer. (Courtesy of Spindler & Hoyer, Inc., Milford, MA.)

intensity variation can be observed at a fixed location in space behind the lens used to introduce a beam divergence, resulting in an annular fringe pattern. The pinhole is used to define an exact location of observation, and the photodetector picks up the varying intensity for further processing. In the most basic signal processing setup, a number of bright and dark cycles are fed into a counter, which then counts changes in optical path length in integer multiples of $\lambda_A/2$. More sophisticated signal processing not only counts cycles but also determines relative phase changes in the sinusoidal varying intensity so that resolutions of $\lambda_A/512$ can ultimately be achieved.

When moving the mirror, one must guarantee a smooth motion without backward jitter of the mirror carriage to avoid double counts of interference fringes. Very high-quality linear bearings (such as air bearings) are necessary to accomplish just that.

As can be seen in Figure 27.4, the light reflected off both mirrors essentially retraces its own path and is at least partially incident onto the active volume of the laser source (3), thereby forming an external laser resonator that is able to detune the laser, effectively modulating its output power as well as its wavelength. To avoid this effect, commercial versions of Michelson interferometers employ corner-cube reflectors instead of plane mirrors as well as optical isolators, as shown in Figures 27.5 and 27.6. Some authors, however, report using optical arrangements that utilize this effect in conjunction with laser diodes to realize low-cost, short-travel displacement sensors [19,20]. These setups will not be discussed here, however.

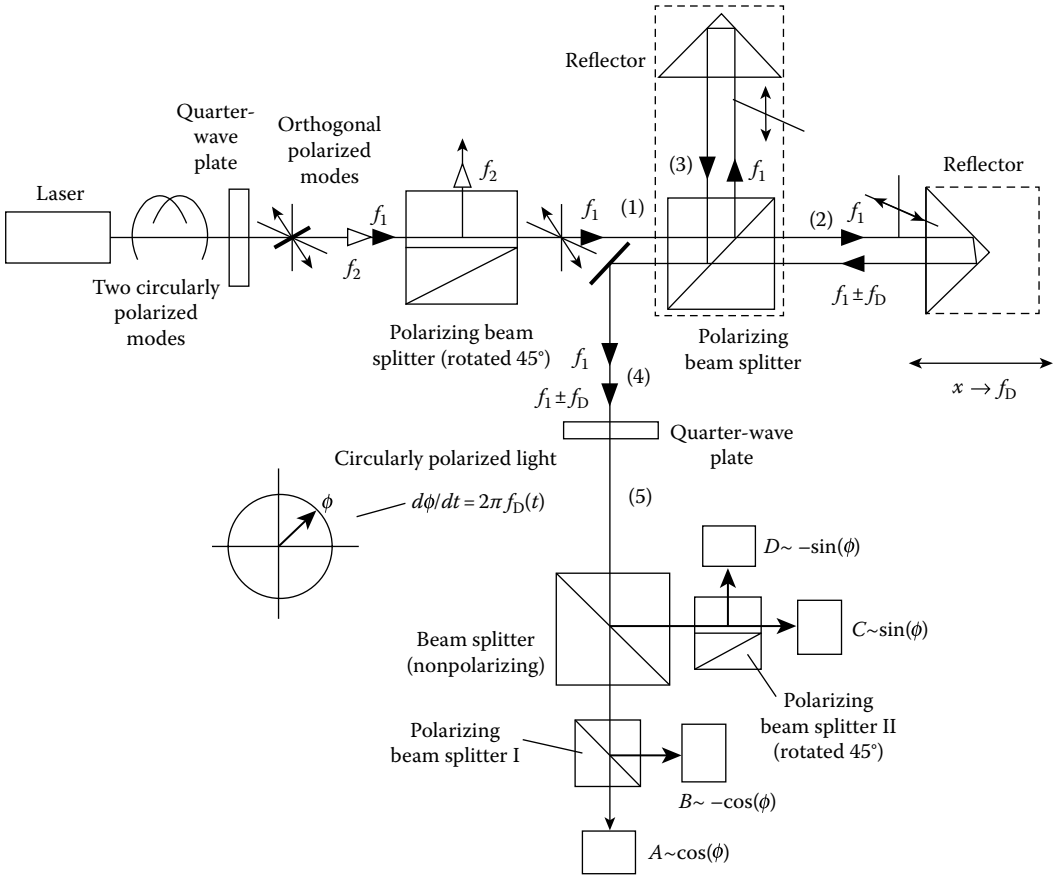


FIGURE 27.6 Single-mode homodyne interferometer. (Courtesy of Spindler & Hoyer, Inc., Milford, MA.)

27.5 Two-Frequency Heterodyne Interferometer

Figure 27.5 shows the commercially available two-frequency Michelson interferometer operating with a Zeeman-stabilized He–Ne laser source. This laser emits two longitudinal modes with frequencies f_1 and f_2 that are both circularly polarized in opposite directions. By passing the modes through a quarter-wave plate, two orthogonal linearly polarized waves are generated. Both are split by a nonpolarizing beam splitter. There is a polarizer located in arm 1 of that splitter, which is rotated by 45° with respect to both polarized waves impinging on it, thus effectively allowing them to interfere behind it yielding a difference frequency of $f_2 - f_1$ that is picked up by a photodetector and counted by a reference counter (frequency difference typically 1.5 MHz).

The orthogonal polarized waves in 2 are further split by a polarizing splitter. Spectral component $f_1 < f_2$ is transmitted into measuring arm 3, and frequency component f_2 is reflected into reference arm 4 of the interferometer. Due to the velocity v of the reflector in arm 3 resulting in a displacement x, the frequency f_1 is Doppler shifted by f_D (Equation 27.5). The movement of the reflector toward the interferometer results in a positive Doppler frequency $f_D > 0$. After recombining both waves from 3 and 4 in the beam splitter again, they are sent through a polarizer in arm 5 that also is rotated by 45° with respect to the direction of polarization of both recombined waves, thereby allowing them to interfere, yielding a difference frequency of $f_2 - f_1 - f_D$ at the location of the photodetector, which is counted by a second counter. By continuously forming the difference of both counts, the measurand (the displacement of x in multiples of $\lambda_A/2$) is calculated.

With this type of interferometer, the direction of motion is given by the sign of the resulting count. One disadvantage of the two-mode heterodyne interferometer is its limited dynamic range for the velocity v of the reflector moving toward the interferometer, since the Doppler frequency f_D, given by

$$f_D = \frac{2}{\lambda_A} v \tag{27.5}$$

is bound to be less than the initial frequency difference between f_2 and f_1 for stationary reflectors. Given a typical Zeeman effect–induced frequency difference f_z of 1.5 MHz, the velocity v is therefore bound to be less than

$$v < \frac{f_z}{2} \lambda_A = 0.474 \text{ m s}^{-1} \tag{27.6}$$

There is no such bound if the reflector is traveling away from the interferometer. By electronically interpolating the output signal of the photodetector, subwavelength resolution can be obtained [21].

27.6 Single-Mode Homodyne Interferometer

An interferometer setup that has no limitation on the maximum velocity in the aforementioned sense is the single-mode homodyne interferometer shown in Figure 27.6.

As in the two-frequency heterodyne interferometer, a Zeeman effect–stabilized laser source that emits two-frequency-displaced circularly polarized axial modes is usually used. After passing through a quarter-wave plate, two orthogonal polarized waves are generated, only one of which (f_1) is further used. The other (f_2) is reflected out of the optical path by an appropriately oriented polarizing beam splitter. The plane of polarization in arm 1 of the interferometer is tilted by 45° with respect to the plane defined by the two arms 2 and 3. The second polarizing beam splitter will transmit one horizontally oriented component into the measuring arm 2 and reflect a vertically oriented component into the reference arm 3 of the interferometer. The two arms of the interferometer maintain their orthogonal polarizations.

The frequency of the wave front in arm 2 is shifted by the Doppler effect (Equation 27.5) due to the motion of the reflector. The light reflected by the two triple mirrors is recombined in the polarizing beam splitter and redirected by a mirror. Since in this particular optical setup, the polarization states of the two beams in the two arms of the interferometer are orthogonal, there is no interference after the redirecting mirror in arm 4 as was the case with the basic Michelson interferometer setup (Figure 27.4). After passing a quarter-wave plate at 45°, two opposite circular polarized waves (one with frequency f_1, the other with frequency $f_1 \pm f_D$) are generated and can be described by a rotating phasor (characterized by $\Phi(t)$) with constant amplitude whose rate of rotation is dependent on the Doppler frequency (in arm 5). Amplitude fluctuations can be observed at the photodetectors A–D after this phasor has passed polarizers, which it does after being split by a nonpolarizing beam splitter.

The output of an interferometer has the general form

$$I(t) = I_o(t)\frac{1}{2}[1+\cos(\Phi(t))] \tag{27.7}$$

It is desired to infer $\Phi(t)$ from observation of $I(t)$. Note that $I_o(t)$, which is the intensity of the laser, can also fluctuate with time. The problems encountered with this are (1) the ambiguity in the sign of $\Phi(t)$ and (2) the dependence of the calculated phase on the intensity fluctuations due to the aging of the laser and optical components. The first problem stems from the fact that arccos(…) yields two solutions to Equation 27.7:

$$\Phi(t) = \pm\arccos\left[\frac{2I(t)}{I_o(t)} - 1\right] \tag{27.8}$$

The sign ambiguity can be resolved by also generating a $\sin(\Phi(t))$ yielding quadrature signals. In order to do so, a second output of the interferometer of the form

$$I_2(t) = I_o(t)\frac{1}{2}\left[1+\sin\left(\Phi(t)\right)\right] \tag{27.9}$$

is sought. Equations 27.7 and 27.9 will determine $\Phi(t)$ unambiguously only in the region $[0,2\pi)$, but there is still an ambiguity modulo 2π. The second problem can be dealt with by adding two more outputs of the form

$$I_3(t) = I_o(t)\frac{1}{2}\left[1-\cos\left(\Phi(t)\right)\right] \tag{27.10}$$

$$I_4(t) = I_o(t)\frac{1}{2}\left[1-\sin\left(\Phi(t)\right)\right] \tag{27.11}$$

Taking Equations 27.7 through 27.10 and 27.9 through 27.11, it is possible to obtain a zero crossing at the linear most sensitive point of inflection of the fringe where the effect of intensity fluctuations on the phase measurement is minimal. The setup of Figure 27.6 attempts to obtain these four outputs. Signal A represents the intensity variations as given in Equation 27.7, and signal B due to the nature of the splitting action is shifted with respect to A by 180° (Equation 27.10). There is another arm to the right of the nonpolarizing beam splitter incorporating the polarizing beam splitter II, which is rotated by 45° with respect to beam splitter I so that the attached detectors C and D are generating the signals defined by Equations 27.9 and 27.11.

Since the Doppler frequency is time dependent according to the velocity v of the measuring reflector, the distance, x, traveled by the reflector up to time T is given by Equation 27.12:

$$x = \int_0^T v(t)dt = \int_0^T f_D(t)\frac{\lambda_A}{2}dt = \int_0^T \frac{\partial \Phi(t)}{\partial t}\frac{\lambda_A}{4\pi}dt \tag{27.12}$$

It should be noted that the resolution of this interferometer is limited to $\lambda_A/4$ if no special hardware is used to interpolate between interference fringes.

This particular interferometer is not limited with respect to a maximum unambiguous Doppler frequency.

27.7 Displacement Sensor Based on a Fabry–Pérot Interferometer

The classical Michelson interferometer as detailed in the section earlier clearly is the system of choice for larger distances to be measured and if there is sufficient space to place the optical components in. In case there is a little available space to place a sensor but still there is the requirement for high precision measurements, a Fabry–Pérot-type sensor, as shown in Figure 27.7a, might be a good choice. A Fabry–Pérot interferometer is typically made of a resonator with two reflecting surfaces causing the light waves to multiply and bounce back and forth thus resulting in transmittance functions exhibiting steep transmission peaks dependent on the wavelength of the light source and the reflectivity of the surfaces as measured by the so-called finesse [20].

If a fixed laser wavelength λ is used and the finesse is kept rather small by using the nonpolished surface of the object as retroreflector (actually scatterer) and the end face of the monomode optical fiber as the second surface, one still obtains a Fabry–Pérot cavity if a collimating lens is properly placed between the fiber and the object. The optical output intensity is now varying as the distance x between the sensor (actually the fiber end face) and the retroreflecting surface is changed. This change in intensity $I_R(x)$ as a function of distance is shown in Figure 27.7b.

One can identify regions that allow for a large sensitivity (the steep slopes of the curve) but also regions (encircled) that do not show changes in output intensity if x is varied. To overcome this restricting drawback, the wavelength λ is periodically modulated (f in the kHz range) in the time domain by $\Delta\lambda$. In analyzing the systems response to this modulation, one can show [22] that the λ modulation results in an output signal whose modulation component is 90° phase shifted (yielding a quadrature component $\delta I_R(x)$) to the main signal, thus allowing to evaluate also in the flat and insensitive part of the Fabry–Pérot characteristic.

This extra feature not only overrules the usual Fabry Pérot blind spot limitations but also allows the sensing of displacement direction. Therefore, it paves the way toward ultrahigh resolution displacement sensing on the overall mirror displacement range. Moreover, by means of a special engineering technique, the fiber probe allows a tilt of ±0:4° at a 100 mm range, thus allowing a plug-and-measure application of the sensor without the need for prior careful alignment. The attainable measurement range is on the order of $x_{min} < x < 100$ mm with a repeatability of better than 10 nm combined with a maximum displacement velocity of 25 mm s^{-1}.

27.8 Interferometer Signal Processing

At the photodetectors ends the optical path of the interferometers. Since the photodetectors are sensitive to light intensity values, only the sinusoidal signal caused by the motion of the reflector will be superimposed onto a pedestal signal proportional to the mean intensity overtime or displacement, respectively. The electronics attached to the photodetectors is aimed to reliably detect zero crossings of the sinusoidal

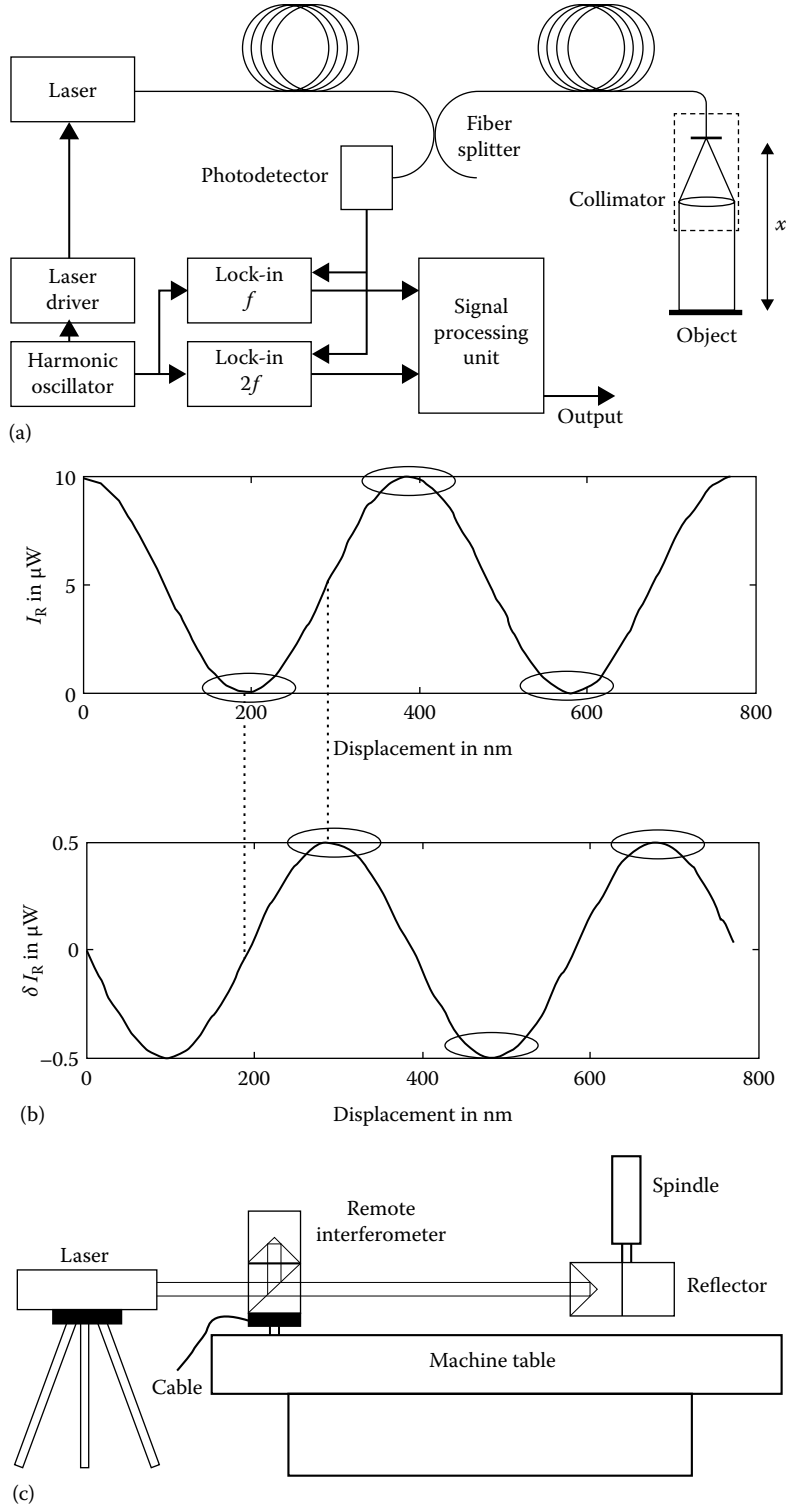


FIGURE 27.7 Displacement sensor based on a Fabry–Pérot interferometer: (a) a Fabry–Pérot-type sensor, (b) change in intensity as a function of distance, and (c) an experimental set up.

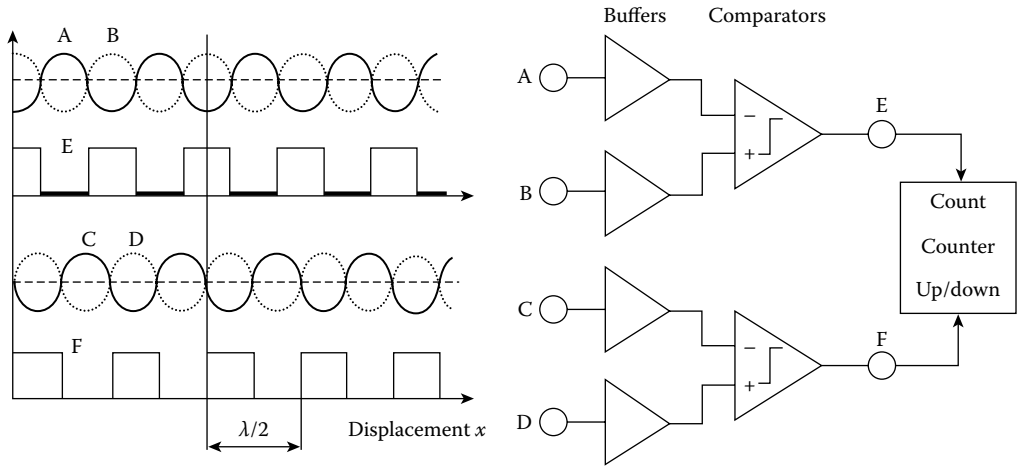


FIGURE 27.8 Simple signal conditioning circuit yielding quadrature signals E and F.

component of the signals (A through D) even under low-contrast conditions by subtracting out the pedestal signal component in the comparators. Low fringe contrast can be the result of small reflector tilt and/or vibration during periods of reflector movement, nonideal interferometer alignment, and imperfections in optical components like unequal splitting of the beam in beam splitters. Figure 27.8 shows a simple conditioning unit.

The main purpose of that setup is to produce digital quadrature signals (E and F in Figure 27.8) that can most easily be used to perform signal interpolation digitally [23,24] and also allow for the determination of the direction of movement by using, for example, signal E as count signal and F as an up/down indicator.

Both comparators compare sinusoidal components having a possibly slowly time-varying dc pedestal amplified by buffer amplifiers. All photodetector signals A through D will be affected by this dc pedestal so that forming their difference in the comparator will yield zero-crossing signals independent of the pedestal. The two comparator units are necessary to generate quadrature signals for direction detection. Furthermore, this electronics allows easily for a fringe interpolation by a factor of 4, bringing the least-count resolution of the interferometer down to $1 \approx 60$ nm. This interpolation can be done if instead of feeding the signal E into the count input, the exored signal E⊕F is used. In this case, the up/down terminal of the counter circuit needs to connect to the signal E&F. If an even higher resolution is sought, the digital interpolation of these quadrature signals can be performed [21,23,24].

27.9 Typical Interferometer Applications

Interferometers are typically used in calibration procedures for machine tools or coordinate-measuring machines (CMMs). A typical so-called remote interferometer optics arrangement is shown in Figure 27.7c. The advantage of the arrangement is twofold. Firstly, as compared to the interferometer optics directly attached to the laser, any vibrations between the laser head and the machine table do not influence (to a first degree) the measurement accuracy, and secondly, any heat dissipated by the laser is kept away from the interferometer optics, the optical paths, and clearly also from the machine under test. The last point is especially important since typical helium–neon lasers do dissipate on the order of 5–10 W; thus, they have the potential to cause significant warming of the high-accuracy machines resulting in a degradation of accuracy at the micron level.

Providing thermal drift has been eliminated this way; the accuracy of the linear laser measurement depends primarily on the performance of the weather station used to correct for air humidity, temperature, and pressure.

27.10 Conclusions

Using laser interferometers operating with highly stabilized laser sources, relative uncertainties in length measurements as low as 5×10^{-8} can be realized, which makes these kinds of equipment suitable to convey primary and secondary length standards into industrial measurement labs. Since, for practical purposes, these interferometers need to be operated in ambient air, the uncertainty in the refractive index of air gives the largest single factor limiting the overall precision. Decreasing the uncertainty stemming from that source necessitates the determination of environmental parameters such as relative humidity, temperature along the optical path, and atmospheric pressure with very little error. Laser refractometers able to determine the refractive index of air directly, on the other hand, allow for the direct measurement and compensation of the refractive index. Zeeman effect–stabilized laser sources can reach a relative uncertainty of less than 10^{-8} and thus contribute only a small portion of the overall error. The interferometer itself, depending on the signal processing involved, will have at least-count resolution of $\lambda_A/4$ without any optical or electronic interpolation and can have an order of magnitude less uncertainty if high-performance phase meters are used to subdivide the wavelength λ_A. The very high-quality air bearings and a very well-controlled measurement environment are necessary to reach these goals.

27.A Appendix

In Table 27.A.1, some companies are listed that manufacture either complete interferometer systems or major components thereof, such as beam splitters, retroreflectors, refractometers, and wavemeters. This list is by no means exhaustive. Furthermore, no price information is included because the system cost is too much dependent on the particular choice of system components.

TABLE 27.A.1 Companies that Produce Interferometers or Components

Manufacturer	Components	Complete Systems
Agilent Technologies, Inc., 5301 Stevens Creek Blvd, Santa Clara, CA 95051, (877) 424 4536	*	*
AttoCube Systems, Koeniginstr. 11, D-80539 Munich, Germany, +49 89 2877 809—19	*	*
Lasertex Sp., ul.Radzionkowska 17, 51-506 Warsaw, Poland, +48 509 495 023	*	—
CVI Melles Griot, Inc., 200 Dorado Place SE, Albuquerque, NM 87123, (505) 296 9541	*	—
Polytec PI, Inc., Auburn, MA 01501, (508) 832 3456	*	—
Renishaw plc New Mills, Wotton-under-Edge, Gloucestershire, GL12 8JR, United Kingdom, +44 1453 524524	—	*
SIOS Messtechnik GmbH, Am Vogelherd 46, D-98693 Ilmenau, Germany, +49 3677 64470	*	*
Zygo Corporation, Middlefield, CT 06455-0448, (860) 347 8506	*	*

Defining Terms

Heterodyne technique: The superposition of two harmonic signals with frequencies f_1 and f_2 in a nonlinear device that results in a signal containing both sum and differences frequencies. For interferometric purposes where each individual frequency is in the 10^{14} Hz range, the frequency difference might be low enough to be registered by photodetectors. They also serve as nonlinear devices because their response is to light intensity only, which is proportional to the light amplitude squared.

Interference: A phenomenon that strikingly illustrates the wave nature of light. It occurs when radiation follows more than one path from its source to the point of detection. It may be described as the local departures of the resultant intensity from the law of addition as the point of detection is moved, for the intensity oscillates about the sum of the separate intensities.

Laser: The acronym of *l*ight *a*mplification by *s*timulated *e*mission of *r*adiation. It was originally used to describe light amplification but has more recently come to mean an optical oscillator.

Refractive index: A number specifying the ratio of the propagation velocities of light in vacuum c to that in a medium. The refractive index for any medium is always somewhat larger than 1. Its value is dependent on the composition of the medium and the wavelength of the incident radiation.

Zeeman effect: An effect that is observed if excited atoms emit their radiation in the presence of a magnetic field. The longitudinal Zeeman effect causes a single emission line to split symmetrically into right and left circularly polarized lines.

References

1. Documents concerning the new definition of the metre, *Metrologia*, 19, 163–177, 1984.
2. E. Jaatinen and N. Brown, A simple external iodine stabilizer applied to 633 nm, 612 nm, and 543 nm He-Ne lasers, *Metrologia*, 32, 95–101, 1995.
3. A. E. Siegman, *Lasers*, Mill Valley, CA: University Science Books, 1986.
4. W. R. Steel, *Interferometry*, 2nd edn. (Cambridge studies in modern optics), Cambridge, U.K.: Cambridge University Press, 1985.
5. W. R. C. Rowley, The performance of a longitudinal Zeeman-stabilized He–Ne laser (633 nm) with thermal modulation and control, *Meas. Sci. Technol.*, 1, 348–351, 1990.
6. W. Demtröder, *Laser Spectroscopy: Basic Concepts and Instrumentation*, 2nd edn., Berlin, Germany: Springer, 1996.
7. B. Edlén, The dispersion of standard air, *J. Opt. Soc. Am.*, 43(5), 339–344, 1953.
8. B. Edlén, The refractive index of air, *Metrologia*, 2(2), 71–80, 1966.
9. J. C. Owens, Optical refractive index of air: Dependence on pressure, temperature and composition, *Appl. Opt.*, 6(1), 51–59, 1967.
10. H. D. Baehr, *Thermodynamik*, 6th edn., Berlin, Germany: Springer, 1988.
11. R. Revelle, Carbon dioxide and world climate, *Sci. Am.*, 247(2), 35, 1982.
12. K. P. Birch, F. Reinboth, R. W. Ward, and G. Wilkening, Evaluation of the effect of variations in the refractive index of air upon the uncertainty of industrial length measurement, *Metrologia*, 30(1), 7–14, 1993.
13. K. P. Birch and M. J. Downs, An updated Edlén equation for the refractive index of air, *Metrologia*, 30, 155–162, 1993.
14. K. P. Birch and M. J. Downs, Corrections to the updated Edlén equation for the refractive index of air, *Metrologia*, 31, 315–316, 1994.
15. P. E. Ciddor, Refractive index of air: new equations for the visible and near infrared, *Appl. Opt.*, 35(9), 1566–1573, 1996.
16. F. E. Jones, The refractivity of air, *J. Natl. Bur. Stand.*, 86(1), 27–32, 1981.

17. W. T. Estler, High-accuracy displacement interferometry in air, *J. Appl. Opt*., 24(6), 808–815, 1985.
18. Zygo Corporation, Error sources in displacement measuring interferometry, see http://www.zygo.com/library/papers/ZMI_Error_Sources.pdf (accessed on March 14, 2013).
19. J. A. Smith, U. W. Rathe, and C. P. Burger, Lasers with optical feedback as displacement sensors, *Opt. Eng*., 34(9), 2802–2810, 1995.
20. N. Takahashi, S. Kakuma, and R. Ohaba, Active heterodyne interferometric displacement measurement using optical feedback effects of laser diodes, *Opt. Eng*., 35, 802–907, 1996.
21. J. A. Smith and C. P. Burger, Digital phase demodulation in heterodyne sensors, *Opt. Eng*., 34, 2793–2801, 1995.
22. K. Karrai, Device for position detection, Patent No. U.S.2010/0259760 A1, October 14, 2010.
23. K. Oka, M. Tsukada, and Y. Ohtsuka, Real-time phase demodulator for optical heterodyne detection processes, *Meas. Sci. Technol*., 2, 106–110, 1991.
24. J. Waller, X. H. Shi, N. C. Altoveros, J. Howard, B. D. Blackwell, and G. B. Warr, Digital interface for quadrature demodulation of interferometer signals, *Rev. Sci. Instrum*., 66, 1171–1174, 1995.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

28

Bore-Gaging Displacement Sensors

28.1 Introduction **28**-1
28.2 Bore Tolerancing **28**-2
28.3 Components of Gage Accuracy **28**-2
28.4 Bore-Gage Classification and Specification **28**-2
28.5 Bore-Gage Types **28**-3
Fixed-Limit Gages • Variable-Indicating Gages: Two- and Three-Point Bore-Gage Options
28.6 Form-Measuring Instruments **28**-11
28.7 Gage Repeatability and Reproducibility **28**-13
Understanding the Terms • Application Notes
References **28**-16
Further Information **28**-16

Viktor P. Astakhov
Astakhov Tool Service

28.1 Introduction

Holes, also known as bores, in general are one of the most common features of mechanical component. In precision applications, eccentricity and diameter of the holes directly relate to the performance and even service life of the component. Deviation from the permissible tolerances can impede operation and result in premature failure of the system, It is thus important to inspect and validate the dimensional accuracy, such as the diameter, shape, and position (e.g., eccentricity) of the holes. To do these, many state-of-the-art hole inspection gages are available.

Bore gaging, already a significant factor in manufacturing, has become increasingly important in metrology because of the growing concern for total quality assurance (TQA). Compared with outside-diameter (OD) measurement, bore gaging creates more engineering challenges by the very nature of its special role in QA. This is especially true when measuring difficult-to-assess internal features such as splines, threads, and deep bores.

As machine capability increases, so does the demand for dimensional gaging to cope with tighter limits on tolerances and greater complexity of component parts. Manufacturers are concerned not only with making the component to a specific tolerance band but also with size variation from component to component because it gives them better control of their process. To meet this demand, the gage they use must be able to discriminate size variation better than before. Moreover, many manufacturers now require that any in-process or post-process gage record and send the measured size of a component to a statistical process control (SPC) or data-collection system. This chapter is to familiarize manufacturing professionals with the basic concepts of bore gaging.

Dimensions are a part of the total specification assigned to parts designed by engineering. However, the engineer in industry is constantly faced with the fact that no two objects in the material world can

ever be made exactly the same. The small variations that occur in repetitive production must be considered in the design. To inform the workman how much variation from exact size is permissible, the designer uses a tolerance or limit dimension technique. A *tolerance* is defined as the total permissible variation of size, or the difference between the limits of size. *Limit dimensions* are the maximum and minimum permissible dimensions. Proper tolerancing practice ensures that the finished product functions in its intended manner and operates for its expected life.

28.2 Bore Tolerancing

All bore dimensions applied to the drawing, except those specifically labeled as basic, gage, reference, maximum, or minimum, will have an exact tolerance, either applied directly to the dimension or indicated by means of general tolerance notes. For any directly tolerated decimal dimension, the tolerance has the same number of decimal places as the decimal portion of the dimension [1].

Engineering tolerances may broadly be divided into three groups: (1) *size tolerances* assigned to dimensions such as length, diameter, and angle; (2) *geometric tolerances* used to control a hole shape in the longitudinal and transverse directions; and (3) *positional tolerances* used to control the relative position of mating features. Interested readers may refer to [1,2].

The ISO system of limits and fits (ISO 286-2:2010 "Geometrical product specifications (GPS)—ISO code system for tolerances on linear sizes—Part 2: Tables of standard tolerance classes and limit deviations for holes and shafts") covers standard tolerances and deviations for sizes up to 3150 mm. The system is based on a series of tolerances graded to suit all classes of work from the finest to the coarsest, along with different types of fits that range from coarse clearance to heavy interference. Here, *fit* is the general term used to signify the range of tightness that may result from the application of a specific combination of tolerances in the design of mating parts.

There are 18 tolerance grades intended to meet the requirements of different classes of parts. These tolerance grades are referred to as ITs and range from IT 01, IT 02 (reserved for the future), and IT 1 to IT 16 (for today's use). In each grade, the tolerance values increase with size according to a formula that relates the value of a given constant to the mean diameter of a particular size range. The system provides 27 different fundamental deviations for sizes up to and including 500 mm and 14 for larger sizes to give different type of fits ranging from coarse clearance to heavy interference. Interested readers may refer to standard ISO/TS 14253-4:2010 Geometrical product specifications (GPS)—Inspection by measurement of workpieces and measuring equipment—Part 4: Background on functional limits and specification limits in decision rules.

28.3 Components of Gage Accuracy

The accuracy of a gage is determined by three factors:

- *Resolution*: For a gage to accurately determine whether or not a component is in tolerance, the general rule is that the gage resolution should be near 10% of the tolerance.
- Reproducibility and repeatability (R&R).
- *Linear accuracy*: This is the value of maximum deviation from the true size a gage will be capable of measuring across its entire working range. As this figure is intrinsically linked to the range of the gage, it means nothing unless considering both the value of deviation and the range together. Many manufacturers mistakenly use the 10% rule that establishes an appropriate resolution and apply it to linear accuracy. This is incorrect, because it does not take into account the range of the gage.

28.4 Bore-Gage Classification and Specification

To measure the aforementioned tolerances, modern manufacturing requires the use of gages. A *gage* is defined as a device for investigating the dimensional fitness of a part for specific function. *Gaging* is defined by ANSI as a process of measuring manufactured materials to assure the specified uniformity

of size and contour required by industries. Gaging thereby assures the proper functioning and interchangeability of parts; that is, one part will fit in the same place as any similar part and perform the same function, whether the part is for the original assembly or replacement in service [3].

Bore gages may be classified as follows:

1. Master gages
2. Inspection gages
3. Manufacturer's gages
4. Gages that control dimensions
5. Gages that control various parameters of bore geometry
6. Fixed-limit working gages
7. Variable-indicating gages
8. Post-process gages
9. In-process gages

Master gages are made to their basic dimensions as accurately as possible and are used for reference, such as for checking or setting inspection of manufacturer's gages. *Inspection gages* are used by inspectors to check the manufactured products. *Manufacturer's gages* are used for inspection of parts during production.

Post-process gages are used for inspecting parts after being manufactured. Basically, this kind of gage accomplishes two things: (1) it controls the dimensions of a product within the prescribed limitations, and (2) it segregates or rejects products that are outside these limits. Post-process gaging with feedback is a technique to improve part accuracy by using the results of part inspection to compensate for repeatable errors in the bore manufacturing operations. The process is normally applied to CNC (computer numerically controlled) machines using inspection data to modify the part program and on tracer machines using the same data to modify the part template.

In-process gages are used for inspecting parts during the machining cycle. In today's manufacturing strategy, in-process gages and data-collection software provide faster feedback on quality. Indeed, the data-collection and distribution aspect of 100% inspection has become as important as the gaging technology itself. Software specifically designed to capture information from multiple gages, measure dozens of products types and sizes, and make it available to both roving inspectors and supervising quality personnel as needed is quickly becoming part of quality control strategies as an integral part of the measurement systems analysis (MSA). In conjunction with CNC units, in-process gaging can automatically compensate for workpiece misalignment, tool length variations, and errors due to tool wear.

28.5 Bore-Gage Types

In choosing a bore gage, one should first eliminate those types of bore gages that are least appropriate for the application. For example, for a production environment, noncontact measurement techniques, including optical and laser methods, tend to be bulky, relatively expensive, and inflexible when measuring special internal features. Such gaging requires the part to be taken to the gage. Not all types of noncontact systems are suitable for a production environment, and many entail high maintenance costs.

Gages that control dimensions are used to control bore diameter [3]. These gages can be either post-process or in-process gages. Further, these gages can be either *fixed-limit gages* or *variable-indicating gages*. These cages are classified as (1) plug/pin gages, (2) needle gages, (3) chamfer gages, countersink gages, (4) dial caliper gages, (5) dial bore gages, (6) electronic bore gages, (7) hole micrometers/ingages, and (8) telescopic gages.

28.5.1 Fixed-Limit Gages

A *plug (pin) gage* is a fixed-limit working bore gage [3]. These inexpensive gages do not actually measure dimensions or geometry. They simply tell the operator whether the bore is oversized or undersized.



FIGURE 28.1 Typical plug gages: (a) double end (left go, right no-go) and (b) progressive.

The actual design of most plug gages is standard, being covered by American Gage Design (AGD) standards. A plug gage is usually made up of two members. One member is called the go end and the other the no-go or not-go end. The gage commonly has two parts: the gaging members, and a handle with the sign, go or no go, and the gage maker's tolerance marked on it. In today's manufacturing, three types of AGD standard plug/pin gages are used. The first is the double end (Figure 28.1a); the second is the progressive gage in which the go and no-go parts follow each other (Figure 28.1b); and the third is the pin gage (Figure 28.2) that comes with the set of pins of various bore sizes. Pin gages became the most popular in the automotive industry due to their versatility and fast interchangeability of pins. Plug gages are used not only to inspect round holes but also to inspect hole of any shape. For example, Figure 28.3 shows a plug gage for threaded holes. Interested readers may refer to [3].

Fixed-limit gage tolerance is generally determined from the amount of workpiece tolerance. A 10% rule is generally used for determining the amount of gage tolerance for fixed, limit-type gages. Four classes of gage makers' tolerances have been established by the American Gage Design Committee and are in general use [3]. These four classes establish maximum variation for any designed gage size. The degree of accuracy needed determines the class of gage to be used. Table 28.1 shows these four classes of gage makers' tolerances. Referring to Table 28.1, class XX gages are used primarily as master gages and for final close tolerance inspection. Class X gages are used for some types of master gage work and as close tolerance inspection and working gages. Class Y gages are used as inspection and working gages.



FIGURE 28.2 Typical pin gage and set of pins.



FIGURE 28.3 Plug thread gage.

TABLE 28.1 Standard Gage Makers' Tolerances (ANSI/ASME B89.1.5)

Diameter	XXX	XX	X	Y	Z	ZZ
Above-Including						
Inches						
0.010–0.825	0.000010	0.000020	0.000040	0.000070	0.0001	0.0002
0.825–1.1510	0.000015	0.000030	0.000060	0.000090	0.00012	0.00024
1.510–2.510	0.000020	0.000040	0.000080	0.00012	0.00016	0.00032
2.510–4.510	0.000025	0.000050	0.0001	0.00015	0.0002	0.0004
4.510–6.510	0.000033	0.000065	0.00013	0.00019	0.00025	0.0005
6.510–9.010	0.000040	0.000080	0.00016	0.00024	0.00032	0.00064
9.010–12.010	0.000050	0.0001	0.00002	0.0003	0.0004	0.0008
Millimeters						
0.254–20.96	0.00025	0.00051	0.00102	0.00178	0.00254	0.00508
20.96–38.35	0.00038	0.00076	0.00152	0.00229	0.00305	0.00610
38.35–63.75	0.00051	0.00102	0.00203	0.00305	0.00406	0.00813
63.75–114.55	0.00064	0.00127	0.00254	0.00381	0.00508	0.01016
114.55–165.35	0.00084	0.00165	0.00330	0.00483	0.00635	0.01217
165.35–228.85	0.00102	0.00203	0.00406	0.00610	0.00813	0.01626
228.85–305.05	0.00127	0.00254	0.00508	0.00762	0.01016	0.02032

TABLE 28.2 Class X MG-25 Series Pin Gage Sets by Meyer Gage Company, Inc.

	Nominal Size Range		
Gage Set No.	English (in.)	Metric (mm)	Set Price USD
MG1-25	0.0110–0.0609	0.28–1.55	240.79
MG2-25	0.0610–0.2509	1.55–6.37	119.79
MG3-25	0.2510–0.5009	6.38–12.72	138.10
MG4-25	0.5010–0.7499	12.73–19.05	190.58
MG5-25	0.7500–0.8129	19.05–20.65	266.82

Class Z gages are used as working gages where part tolerances are large. Table 28.2 shows the diameter ranges of pin gage sets manufactured by Meyer Gage Company, Inc.

28.5.2 Variable-Indicating Gages: Two- and Three-Point Bore-Gage Options

Two- and three-point contact measurement is a popular solution as a production gage and provides many options. The gage is typically calibrated to a reference master, such as a setting ring, which means that only the setting ring itself must be calibrated regularly. Depending on the type chosen, gages under this description can have a wide working range. Moreover, they offer the best flexibility for adaptation to special feature measurement. Both two- and three-point contact methods can measure deep into bores, have lower initial and maintenance costs, and are portable. Digital versions can interface with SPC systems.

Cylinder bore gage was originally developed to measure cylinder bores in the automotive industry. It is a two-point contact measuring system for use on components that may be subject to ovality problems. The design includes a mechanism to centralize the measuring head in the bore. It is a cost-effective solution for simple bore measurement and can be easily adapted to measure deep bores up to 6 ft (approx. 1.8 m) deep. Cylinder bore gages are not easily adapted to measure special features



FIGURE 28.4 Self-centering two-point bore gage.

and require multiple setting masters. Analog versions require the operator to decide when he has established the minimum value, which can introduce variability from operator to operator. Unless a definite tendency toward ovality exists in the manufacturing process, a three-point system is a better solution.

A mechanical self-centering two-point bore gage shown in Figure 28.4 is made for comparative measurements of bores along with their form errors. This self-centering tool can be used with dial indicators, precision comparator indicators, and electrical or digital probes. The bore gage will accept both 8 mm and 3/8 in. diameter stems, virtually any dial indicator, right off the shelf, even indicators that one may already have in the shop. No special housing is required. By using indicators reading in 0.0001 in. or 0.002 mm, one can have a bore gage that measures both metric and inch. Table 28.3 shows the diameter ranges and prices of self-centering two-point bore gages manufactured by TESA Co.

Micrometer bore gages come in a wide variety. The simplest micrometer bore gage is shown in Figure 28.5. The traditional internal micrometer has many advantages because of its flexibility, offering a choice of two- or three-point systems, a large working range on each measuring head, and adaptability to measuring special features. Depending on which design is chosen, it is possible to find systems that can measure from 0.040 in. (1 mm) to 12 in. (300 mm) as a standard range. Both analog (Figure 28.5) and digital (Figure 28.6) versions are available, with the digital available as either mechanical or electronic. Basic parameters of bore-gage micrometers by Bowers Metrology UK are presented in Table 28.4. Only the electronic versions have the ability to send data to an SPC system, printer, or data collector.

TABLE 28.3 TESA Veribor Bore Gages

Range	Veribor Number	Order Number	Sale Price (USD)
0.7–1.4 in. 18–35 mm	05790090	29457	437.00
1.4–2.5 in. 35–63 mm	05790091	29458	426.00
2–6 in. 50–152 mm	05790092	29459	491.00
0.7–6 in. 18–152 mm	05790093	29479	833.00



FIGURE 28.5 Simplest bore-gage micrometer.



FIGURE 28.6 Bore-gage micrometer with a digital readout by Bowers Metrology UK.

For example, Marposs Corp. developed a new wireless bore gage called iWave (Figure 28.7). It is composed of a standard MBG plug gage directly connected to a special handle housing the wireless (*Bluetooth*® wireless technology) transmitter and the power supply batteries. By pushing a button, real-time measurement values are displayed on the receiving electronic unit, with the possibility to store the measurement data. The robust and reliable StarLock system allows gage head changeover in just few seconds, without the need of any tool accommodating various mechanical gage heads such as (a) Marposs M1 Star MBG plug heads and any existing plug head on the market, (b) Marposs M3 Star snap heads and rings, (c) depth gages, (d) sweep gages, and (e) gages for chamfers, spot facings, and countersinks. A dedicated computer (Merlin™) that receives the signal is capable of statistical analysis with graphic display (histogram, control charts, value chart, etc.), numeric summary (e.g., spread and average), and data segregation (by machine, product batch, and analysis purposes), and measured part counters local and remote data storage and access.

TABLE 28.4 Parameters of Bore-Gage Micrometers by Bowers Metrology UK

Product Code (mm)	Range (mm)	Product Code (mm)	Range (mm)	Accuracy		Measuring Depth		Setting Ring	
				mm	in.	mm	in.	mm	in.
XTD6M	6–8	XTD6i	1/4–5/16	0.003	0.0001	58	2.28	8	0.312
XTD8M	8–10	XTD8i	5/16–3/8	0.003	0.0001	58	2.28	8	0.312
XTD10M	10–12.5	XTD10i	3/8–1/2	0.003	0.0001	58	2.28	12.5	0.500
XTD12M	12.5–16	XTD12i	1/2–5/8	0.003	0.0001	62	2.44	12.5	0.500
XTD16M	16–20	XTD16i	5/8–3/4	0.003	0.0001	62	2.44	20	0.750
XTD20M	20–25	XTD20i	3/4–1	0.003	0.00015	66	2.62	20	0.750
XTD25M	25–35	XTD25i	1–1 3/8	0.003	0.00015	66	2.62	35	1.375
XTD35M	35–50	XTD35i	1 3/8–2	0.003	0.00015	80	3.15	35	1.375
XTD50M	50–65	XTD50i	2–2 5/8	0.004	0.00015	80	3.15	65.0	2.625
XTD65M	65–80	XTD65i	2 5/8–3 1/4	0.004	0.00015	80	3.15	65.0	2.625
XTD80M	80–100	XTD80i	3 1/4–4	0.004	0.00015	100	3.94	80.0	3.250
XTD100M	100–125	XTD100i	4–5	0.005	0.0002	115	4.53	125	5.000
XTD125M	125–150	XTD125i	5–6	0.005	0.0002	115	4.53	125	5.000
XTD150M	150–175	XTD150i	6–7	0.005	0.0002	115	4.53	175	7.000
XTD175M	175–200	XTD175i	7–8	0.005	0.0002	115	4.53	175	7.000



FIGURE 28.7 iWave wireless bore gage by Marposs Corp.

Pistol-grip bore gages (Figure 28.8) offer all the advantages of the traditional internal micrometer, and also allow a fast, single-handed operation. The fact that there is a constant gaging force, independent of operator "feel," with the pistol-grip bore gage, facilitates good gage repeatability and reproducibility results. It is very adaptable to measuring special features such as spline pitch diameter, ball races, and ball nuts.



FIGURE 28.8 Pistol-grip bore gage and its use.

Multipoint contact bore gages shown in Figure 28.9 are still the most popular for high-precision inspections. This line of bore gages has been the industry standard for more than 60 years. Its features are as follows: (1) Comfortable and ergonomic well-balanced handle can be removed to gain clearance in restricted areas; (2) precision-machined aluminum dial unit housing, stainless steel bore-gage extension tubing in wide variety of standard and special lengths; (3) thermally insulated high-impact resistant shield protects extension tubing, replicable contacts available in steel, carbide, or diamond. Proven bell crank transfer is accurate and linear over its entire range; (4) precisely machined cast-phosphor-bronze gaging head for sensitivity and durability; and (5) all bore-gage sizes utilize the same dial unit housing for extreme versatility.

Air gages are used in the measurement of multiple features of precision bores. Air gage sensors measure the backpressure created when an air nozzle is brought in close proximity to the workpiece. Figure 28.10 shows the principle of air gaging: the air gage readout measures this pressure and displays the dimensional change in the air gap. The following features of air gages should be pointed out: (a) *Noncontact*—the air nozzle throttles the air at the point where the air exits the jet hole, thus the average height of a spot on the surface opposite the jet hole is sensed by the air readout. Since the nozzle does not contact the workpiece, wear does not directly affect the accuracy of the gage; moreover, the pressure of the air has a self-cleaning effect on the workpieces, making air gages perform exceptionally well in shop-floor environments. (b) *Sensor size*—air nozzles can be made with jet holes from 0.010 to 0.100 in. (0.25–2.54 mm),



FIGURE 28.9 Multipoint contact bore gages.

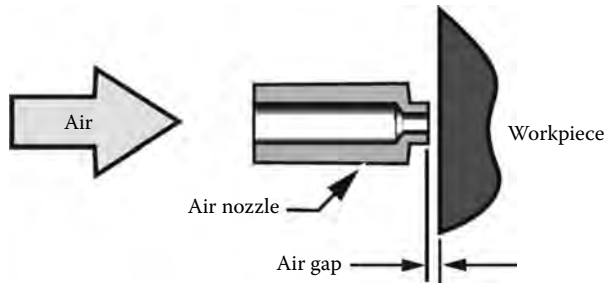


FIGURE 28.10 Principle of air gaging.

with 0.050 in. (1.25 mm) being preferred for most applications. (c) *Gaging ranges*—air gages have relatively limited ranges. For good linearity, the gaging range of air gage member must be less than 8% of the jet hole size. For most applications, practical limits on air flow limit jet sizes to 0.079 in. (2 mm) and gaging ranges to 0.006 in. (0.1 mm) or less. Furthermore, applications requiring exceptionally high accuracies, such as diametrical tolerance limits of 0.00008 in. (0.002 mm), are best measured by air gage members designed for ranges 0.0006 in. (0.015 mm) or less. (d) *Response time* depends on the volume of the air circuit and the size of the gaging nozzles. Response times can vary from less than 100 ms to several seconds; except for gages with very small jet sizes, typical response times will be less than 2 s. (e) *Cost*—air gage nozzles are less expensive than electromagnetic gaging cartridges. For applications that require the summing of inputs from two or more sensors, they are significantly less expensive than electromagnetic gaging.

Figure 28.11 shows the construction of a typical air probe. The probe comprises a hardened steel body in which air passages are drilled to two or more gaging nozzles. The body is precision ground to slip into

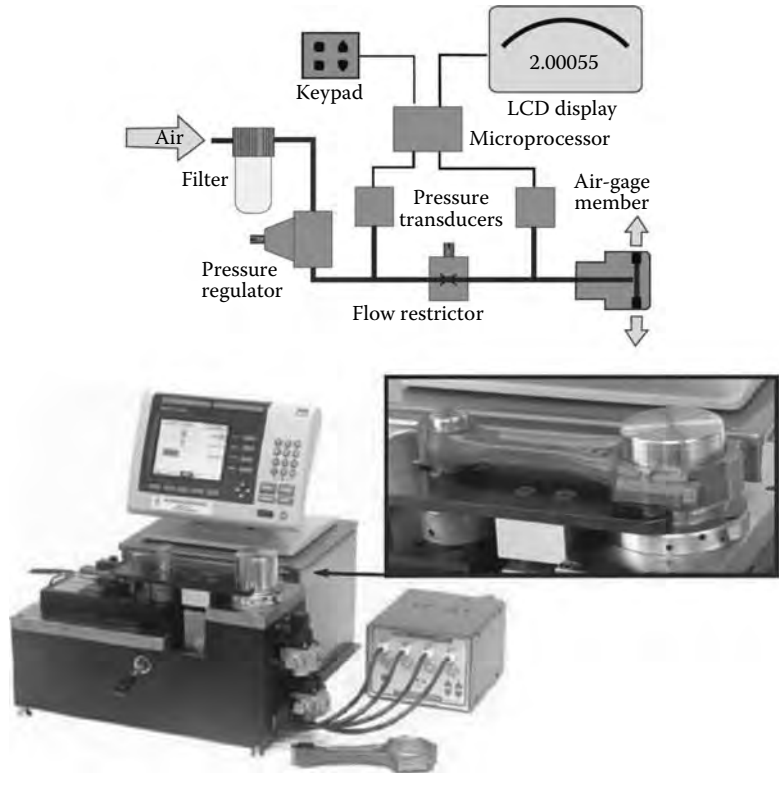


FIGURE 28.11 Electronic air gage by Western Gage Corp.

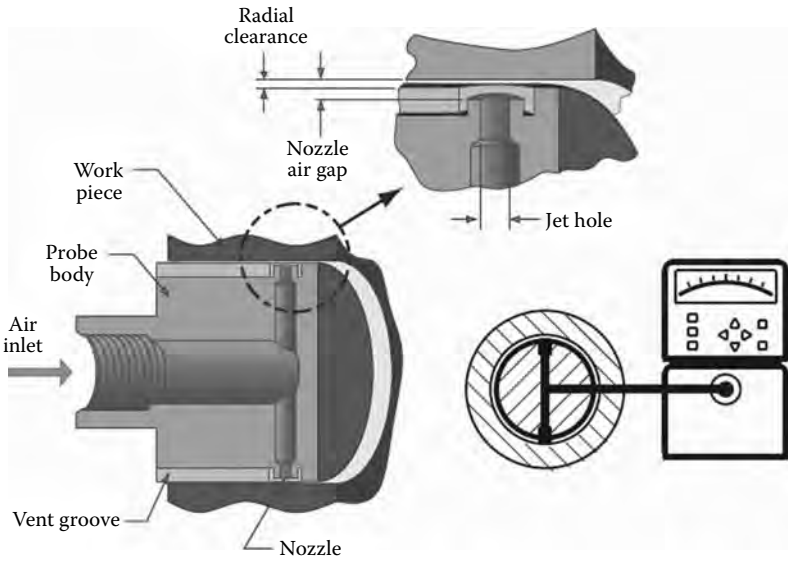


FIGURE 28.12 Air probe features.

the bore at the low limit of the product tolerance; Figure 28.12 shows the nozzle tips are recessed a small amount below the probe body as shown in the magnified view of the air gage nozzle. By recessing the air nozzles below the probe body, the measurement is made noncontact so that wear does not directly affect the accuracy of the gage. The air flow purges the gaging surface of contaminants, thus making air gage measurements highly reliable. The opposed nozzle design of the probe creates a "differential" type of measurement that is independent of how the probe is positioned radially within the test bore, that is, radial movement causing an increase in air flow in one nozzle is offset by a corresponding decrease in flow in the opposing nozzle. These features are key factors in attaining fast accurate gaging with unskilled operators.

28.6 Form-Measuring Instruments

Most modern form-measuring instruments stage the part on a turntable and provide a means to position a gage head against the part (Figure 28.13). As the turntable rotates, the gage head measures deviation from the true circle. Those gages where the gage heads are supported by a simple, rigid, manual, or motorized stand that does not provide precise control over positioning are capable of performing the following measurements: roundness, concentricity, circular runout, circular flatness, perpendicularity, plane runout, top and bottom face runout, circular parallelism, and coaxiality. Modern fully automatic machines are the most sophisticated measuring instruments. Normally, they are equipped with a Windows™-based, PC-compatible graphical user interface to perform real-time data acquisition and processing.

Mitutoyo/MTI Corp. produces a wide range of these machines. Basic characteristics of machines from Series 211—roundness/cylindricity measuring system—are shown in Table 28.5. The RA-2200 round tester (Figure 28.14) provides high accuracy, high speed, and high performance in roundness measurement. The fully automatic or a DAT (digital adjustment table) function aided manual workpiece centering and leveling turns what used to be a difficult and finicky task into one that is simple enough for even untrained users to measure with high accuracy. This facilitates substantial reductions in overall measurement time. The RA-2200 system comes complete with powerful data analysis software ROUNDPAK, which requires only simple manipulation using a mouse and icon, achieving enhanced functionality and ease of operation.

See Table 28.6 for a list of companies that make bore gages.

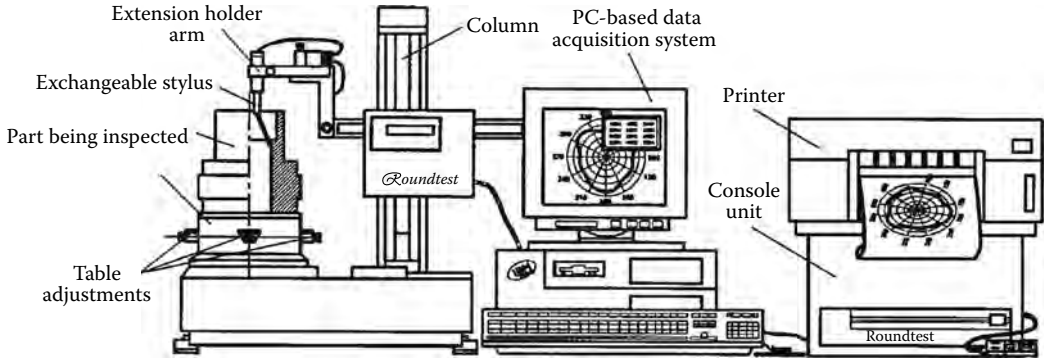


FIGURE 28.13 Form-measuring machine.

TABLE 28.5 Specification of Mitutoyo Round Tester Series 211

Model No.	RA-2200AS	RA-2200DS	RA-2200AH	RA-2200DH
Order No.	211–511A (mm/in.)	211–514A (in.)	211–512A (mm/in.)	211–515A (mm/in.)
Effective table diameter	9.25 in. (235 mm)	8 in. (200 mm)	9.25 in. (235 mm)	8 in. (200 mm)
Centering/leveling adjustment	Automatic centering/leveling turntable	Digital adjustment table	Automatic centering/ leveling turntable	Digital adjustment table
Centering range	±0.118 in. (±3 mm)	±0.197 in. (±5 mm)	±0.118 in. (±3 mm)	±0.197 in. (±5 mm)
Column travel	12 in. (300 mm) (standard column)		20 in. (500 mm) (high column)	
Basic unit mass	396 lb (180 kg)		440 lb (200 kg)	



FIGURE 28.14 Round test RA-2200AS/DS/AH/DH by Mitutoyo.

TABLE 28.6 Companies That Make Bore Gages

Company	Products
Bowers Metrology UK, Unit 15, Bordon Trading Estate, Bordon, Hampshire GU35 9HH, Tel.: 08708 50 90 50, Fax: 08708 50 90 60, http://www.bowers.co.uk/	Wide choice of cost-effective, quality measuring instruments from simple calipers to wireless data transmission gages
Western Gage Corporation, 3316 A Maya Linda Camarillo, CA 93012-8059, Tel: (800) 423-5062 (805) 445-1410 Fax: (805) 445-7530, http://www.westerngage.com/index.htm	Custom manufacturer of *gages*. Types of *gages* include air probe, air ring, plug, comparator, and master *gages*
Meyer Gage Company, Inc., 230 Burnham St., South Windsor, CT 06074, Tel.: 1-888-Pin-Gage (1-888-746-4243), Fax: 1-860-528-1428, http://www.meyergage.com/	Manufacturer of plug gages, pin gages, trilock, and taperlock gages
Marposs S.p.A., Company with a sole shareholder, Via Saliceto, 13-40010 Bentivoglio (BO), Italy; 3300 Cross Creek Pkwy., Auburn Hills, MI 48326, Tel.: 248-370-0404, 888-627-7677, Fax: 248-370-0621, http://www.marposs.com/	Leading global supplier of precision gaging equipment for shop-floor inspection, measurement, and process control solutions
The L. S. Starrett Company, 121 Crescent Street, Athol, MA 01 331, Tel.: 010 1 978-249-3551, Main Fax: 010 1 978-249-8495, http://www.starrett.com/	Manufactures more than 5000 variations of precision tools, gages, measuring instruments, and saw blades for industrial, professional, and consumer markets
Mitutoyo America Corporation, 965 Corporate Blvd., Aurora, IL 60502, Tel.: 630-820-9666, Fax: 630-978-3501, http://www.mitutoyo.com/	Offers the widest array of precision measuring products in the world, as well as the infrastructure and expertise to support them. It includes Mitutoyo hardware and software and Mitutoyo Engineered Systems expertise.
Brown & Sharpe, brand of, Hexagon Metrology, Inc., 250 Circuit Drive, North Kingstown, RI 02852, Tel.: 800-343-7933, 401-886-2000, http://brownandsharpe.com/	Offers a complete line of dimensional metrology products. Brown & Sharpe and TESA-branded precision hand tools and 1D to 3D measuring instruments, the TESAStar CMM probing system
Mahr GmbH, Carl Mahr Str.1, D-37073 Göttingen, Germany, Tel.: +49(0)551-70730, http://www.mahr.com	The Mahr Group is the world's third largest manufacturer of a complete range of production dimensional measuring technology. Mahr measuring instruments range from calipers to optical coordinate measuring systems.
Dyer Company, 1500 McGovernville Rd., P.O. Box 4966, Lancaster, PA 17604, Tel.: 1-800-631-3333, http://www.dyergage.com/	Production gages for precision bores, groves, minimum wall thickness, and special applications
Flexbar Machine Corporation, 250 Gibbs Road, Islandia, NY 11749, Tel.: 800.879.7575, Fax: 631.582.8487, http://www.flexbar.com/	Offers a wide variety of precision measuring instruments, machine tool accessories, machine safety guards, hardness testers, surface roughness gages, and optical and video inspection systems
Taylor Hobson Ltd, PO Box 36, 2 New Star Road, Leicester LE4 9JQ, United Kingdom, Tel.: +44 (0) 116 276 3771, Fax: +44 (0) 116 246 0579, http://www.taylor-hobson.com	Taylor Hobson is a high precision technology company, operating at the highest levels of accuracy within the field of surface, roundness, and form measurement. We provide contact and noncontact metrology solutions for the most demanding applications, with a worldwide infrastructure to support our clients on a global basis.

28.7 Gage Repeatability and Reproducibility

28.7.1 Understanding the Terms

Repeatability is the variability of the measurements obtained by one person while measuring the same item repeatedly. This is also known as the inherent precision of the measurement equipment (gage). Consider the probability density functions shown in Figure 28.15. The density functions were constructed from measurements of the bore diameter with gage A and gage B. The density functions demonstrate that gage B is more repeatable than gage A.

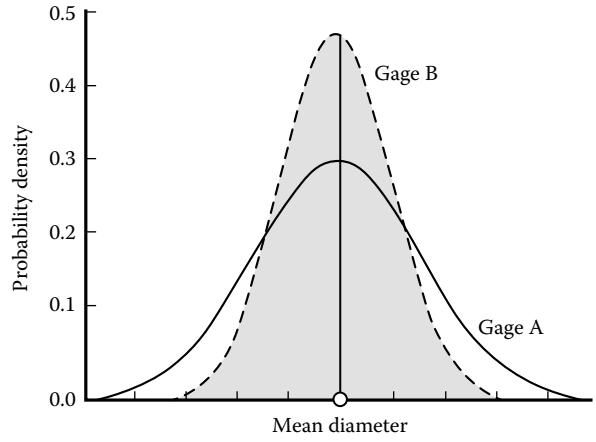


FIGURE 28.15 Probability density functions.

Reproducibility is the variability of the measurement system caused by differences in operator behavior. Mathematically, it is the variability of the average values obtained by several operators while measuring the same item. Figure 28.16 displays the probability density functions of the measurements for three operators. The variability of the individual operators is the same, but because each operator has a different bias, the total variability of the measurement system is higher when three operators are used than when one operator is used. Figure 28.16a also displays the probability density functions of the measurements for three operators using the same scale as Figure 28.16a. Notice that there is more difference in the means of the measurements shown in Figure 28.16b than those shown in Figure 28.16a. Therefore, the reproducibility of the system shown in Figure 28.16b is higher than the reproducibility of the system shown in Figure 28.16a.

Gage R&R methodology was developed in the 1960s to address the estimation of measurement system variation as applied to manufacturing. It was developed by ASTM for raw materials testing. Later on, the automotive industry developed the MSA as one of the core components of the Six Sigma approach to quality control. It was found that many problems encountered with SPC and design of experiments (DOE) is caused by problems with measurement systems. As a result, the MSA (gage R&R studies) became a core component of the QS9000/TS16949 Automotive Quality System Requirements. The standard is supported by manuals published by the Automotive Industry Action Group (AIAG).

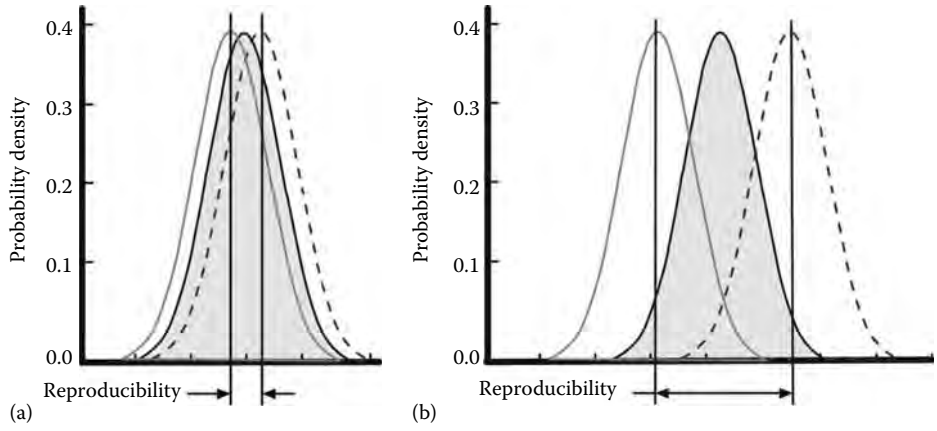


FIGURE 28.16 Reproducibility demonstration based on the results of measurements: (a) High reproducibility, and (b) low reproducibility.

The MSA standard is described in the third edition of the Measurement Systems Analysis Manual AIAG MSA-3 published in May 2003.

28.7.2 Application Notes

The role of reproducibility in gage R&R methodology is to define the variation among operator averages. Here, the term operator, or appraiser, refers to the person who uses the gage. Operator averages are thought to vary from person to person, and R&R is attempting to estimate the variance due to operator differences in practice. The statistical model in gage R&R methodology (without interaction) is as follows (Equation 28.1):

$$y_{ijk} = \chi_i + \alpha_j + \varepsilon_{ijk} \tag{28.1}$$

where

y_{ijk} is the kth repeat measurement of the ith part by the jth operator
χ_i component is the true value of the ith part dimension
α_j component is the reproducibility effect associated with operator j
ε_{ijk} is the random repeatability error that occurs with each measurement

Each measurement, y, is composed of these three components. The reproducibility term (α) may be thought of as a kind of personal bias associated with an operator, that is, each operator measures the various parts somewhat differently than the true value x, and this is the individual's α-effect. When several operators are involved in a gage R&R study, they effectively pick of a random sample (of operators) from a potentially infinite universe of all such possible operators. The α-terms are assumed to have a mean of 0 and an unknown variance θ^2.

The total variance of all measurements, y, has a variance equal to the sum of the individual variance components as in Equation 28.2:

$$\text{var}(y) = \nu^2 + \theta^2 + \sigma^2 \tag{28.2}$$

A gage R&R study is a designed experiment used to estimate the individual variance components. Typically, θ^2 and σ^2 are the main components of interest. The method, based on sample ranges, has enjoyed continued popularity, particularly for small to modest sample sizes, for many years. Today, many computer packages perform gage R&R using the analysis of variance (or ANOVA) technique [4]. The measurement study (classic) typically utilizes one to three appraisers for one measuring instrument that is measuring a single characteristic. Each appraiser measures five to ten units selected from a process two or three times (replications). Before proceeding with the analysis of the study, the ranges for the replications of the measurements made by each appraiser on each part are determined and used to calculate control limits for the range chart. Then each range is checked to determine if it falls inside the limits. Those measurements that result in a range outside the limits should be excluded from further analysis or should be redone. Operative assumptions include the following:

1. The measuring instrument stays in calibration (central location does not change)
2. Appraisers use the same method of measurement
3. Parts are measured in the same place (if the assumption that the parts are measured in the same place is incorrect, the possibility of within-part variation will need to be considered)

Among the many known methods, "average and range" and "ANOVA" methods are mostly used today. However, "ANOVA" method has a more widespread usage as it constitutes the basis of a number of modern commercial software packages to analyze gage R&R (e.g., GAGE). In the selection of a software package suitable for a given application, the first and foremost requirement to any gage R&R

software package is that it should directly follow the recommendations specified in AIAG's third edition of the MSA manual and help your organization conform to ISO/TS 16949-2002, ISO 9000 Section 4.11, FDA 21 CRF11.1, and other critical gage calibration standards.

References

1. Gooldy, G., *Dimensioning, Tolerancing, and Gaging Applied*: Ref. ASME Y14.5M-1994 ANSI B4.4 (B89.3.6) AMD ISO 1101, Prentice Hall, Upper Saddle River, NJ, 1998.
2. Smith, G. T., *Industrial Metrology: Surfaces and Roundness*, Springer, London, U.K., 2002.
3. Nee, J., *Fundamentals of Tool Design*, 6th edn., SME, Deborn, MI, 2010.
4. Burdick, R. K., Borror, C. M., and Montgomery, D. C., Design and analysis of gauge R&R studies: Making decisions with confidence intervals in random and mixed ANOVA models, ASA-SIAM Series on Statistics and Applied Probability, Alexandria, VA, 2003.

Further Information

Evans, J. R. and Lindsay, W. M., *The Management and Control of Quality*, Thomson South-Western, Mason, OH, 2008—provides a reader the strength and proven experience of well-known authors. Presents the fundamental principles and historical foundations of total quality.

Figliola, R. S. and Beasley, D. E., *Theory and Design for Mechanical Measurements*, 4th edn., Wiley, New York, 2005—provides a fundamental treatment for developing, operating, and analyzing measurements systems and for reporting results.

Meadows, G. D., *Geometric Dimensioning and Tolerancing Handbook: Applications, Analysis & Measurement*, James D. Meadows & Associates, Inc., Hendersonville, TN, 2009—explains the new ASME Y14.5-2009 standard. This is one of the most comprehensive GD&T volumes ever written and has the unprecedented ability to cover almost every facet of dimensioning and tolerancing. It includes step-by-step procedures for dimensioning and tolerancing parts and assemblies.

Pyzdek, T. and Keller, P. A., *The Six Sigma Handbook*, 3rd edn., McGraw Hill, New York, 2009—provides comprehensive reference to make Six Sigma strategy and operations, as well as successful project management of quality assurance.

29

Ultrasonic Displacement Sensors

Nils Karlsson
Swedish Defense Research Agency

Ole Pedersen
Linköping University

29.1 Physical Characteristics of Sound Waves **29**-2
29.2 Ultrasound Transducer **29**-4
29.3 Principles of Time-of-Flight Systems **29**-5
Suppliers and Vendors
References **29**-7
Partial List of Vendors and Suppliers **29**-8

Ultrasound is an acoustic wave with a frequency higher than the audible range of the human ear, which is 20 kHz. Ultrasound can be within the audible range for some animals, like dogs, bats, or dolphins. Around 1883, Sir Francis Galton performed the first known experiments with whistles generating ultrasound. Many decades later, people started to find ultrasound applications in engineering, medicine, and daily life. The basic principle for the use of ultrasound as a measurement tool is the *time-of-flight technique*. The pulse-echo method is one example. In the pulse-echo method, a pulse of ultrasound is transmitted in a medium. When the pulse reaches another medium, it is totally or partially reflected, and the elapsed time from emission to detection of the reflected pulse is measured. This time depends on the distance and the velocity of the sound. When sound travels with a known velocity c, the time t elapsed between the outgoing signal and its incoming echo is a measure of the distance d to the object causing the echo:

$$d = \frac{ct}{2} \tag{29.1}$$

Figure 29.1 shows a simple pulse-echo system. The transmitter and the receiver could be the same device, but they are separated for clarity in this figure.

The oscillator generates an electric signal with a typical frequency of 40 kHz. This electric signal is transformed into mechanical vibrations of the same frequency in the transmitter. These vibrations generate sound waves that are reflected by the object. The reflected sound echo causes an electric signal in the receiver. For precise measurements, the speed of sound is a crucial parameter. A typical value in air at 1 atm pressure and room temperature is 343 m s^{-1}, but the speed of sound is influenced by air pressure, air temperature, and the chemical composition of air (water, CO_2, etc.). For example, the speed of sound is proportional to the square root of absolute temperature. Measuring distances in an environment with large temperature gradients can result in erroneously calculated distances. As an advantage, ultrasound waves are robust against other disturbances such as light, smoke, and electromagnetic interference [1–9].

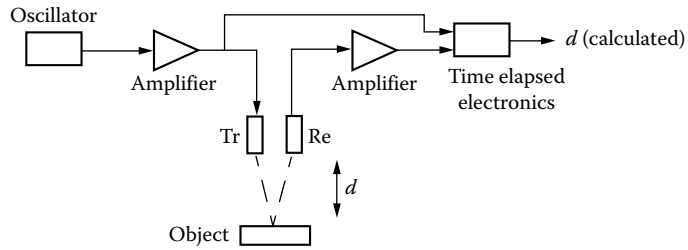


FIGURE 29.1 Principle of a pulse-echo ultrasound system for distance measurements (Tr, transmitter; Re, receiver).

29.1 Physical Characteristics of Sound Waves

Sound is a vibration in matter. It propagates as a longitudinal wave, i.e., the displacement in the material is in the direction of the sound propagation.

The velocity of sound depends on the medium in which it propagates. In a homogeneous and isotropic solid, the velocity depends on the density, ρ, and the modulus of elasticity, E, according to following equation:

$$c = \sqrt{\frac{E}{\rho}} \tag{29.2}$$

In a liquid, the velocity depends on the density and the adiabatic compressibility κ; Equation 29.2 can be expressed as

$$c = \sqrt{\frac{1}{\kappa \cdot \rho}} \tag{29.3}$$

In gases, the velocity of sound is described by equation. Here g represents the ratio of the specific heat constant at constant pressure (c_p) to the specific heat constant at constant volume (c_v), p is pressure, R is the universal gas constant, T is the absolute temperature, and M is the molecular weight:

$$c = \sqrt{\frac{gRT}{M}} = \sqrt{\frac{c_p}{c_v}\frac{p}{\rho}} \tag{29.4}$$

The energy of a sound wave passing a point in the medium consists of alternatively the kinetic energy of particles in motion and the potential energy due to pressure differences. The energy passing a unit surface area per a unit of time is the intensity, I. If this intensity propagates from a point source equally in all directions, the intensity at a distance R is spread over the area of a sphere of radius R. The intensity at a distance R is then proportional to $1/R^2$. The sound wave is also attenuated due to friction and absorption. The attenuation increases with higher frequencies. The signal also decreases due to scattering. The sound beam from a transmitting transducer concentrates the energy to a fraction of the area of the sphere, but the $1/R^2$ dependence still remains. The directivity of a transducer increases with its size and signal frequency. To get a narrow beam, the transmitter membrane diameter must be large with respect to the wavelength. High-frequency ultrasound cannot be the general solution as ultrasound of a high frequency is absorbed faster than ultrasound of a lower frequency.

An important quantity is the acoustic impedance. In the far field, it can be treated as the characteristic impedance R_a, which is the product of the density of the medium, ρ, and the sound of speed, c:

$$R_a = \rho c \tag{29.5}$$

When the acoustic wave reaches the boundary of two different mediums, they have different acoustic impedance. This will cause a reflection of the wave. With the increase in difference in impedance, a larger portion of the sound is reflected. The acoustic impedance of a target to detect shall differ significantly from the medium in which the sound propagates, while the medium for propagation should be as uniform as possible:

$$\text{Reflection coefficient} = \frac{I_{\text{reflected}}}{I_{\text{incident}}} = \frac{(R_1 - R_2)^2}{(R_1 + R_2)^2} \tag{29.6}$$

The reflection of a target also depends on its size and shape. This can be expressed as the effective scattering area or the "equivalent cross section." The target is when compared with a sphere and what size this sphere must have to give the same reflection. In the case when this reflection is received by a receiver at distance R at the same place as the transmitter, the intensity spreading due range is $1/R^2$ again. This gives total range dependence for the received echo of $1/R^4$ [1–9].

The practical importance of the acoustic impedance is realized when the ultrasonic pulse-echo system shown in Figure 29.1 is considered. First, the electric energy is converted into mechanical vibrations of a membrane in the transmitter. Second, the vibrations (the sound wave) have to pass through the boundary between the membrane (usually a solid material) and free air. Because the transmitter membrane and the free air have different characteristic impedances, much of the acoustic intensity is reflected (Figure 29.2).

The transmitted ultrasound in free air will first propagate in a parallel beam (near field of the transducer), but after a distance L, the beam diverges (the far field of the transducer). See Equation 29.7 and Figure 29.3:

$$L \approx \frac{D^2}{4\lambda} \tag{29.7}$$

where

D is the diameter of the circular transmitter

λ is the wavelength of the ultrasound

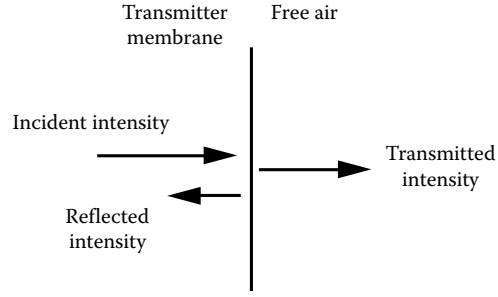


FIGURE 29.2 Reflection and transmission of a sound wave at the interface between media of different characteristic impedances.

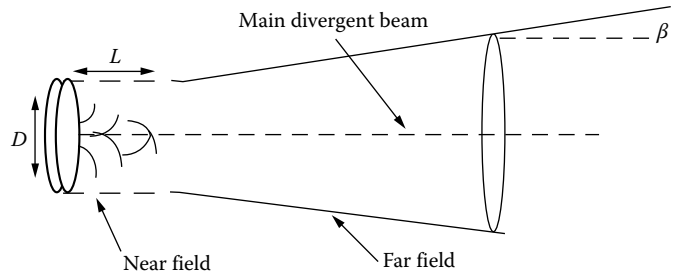


FIGURE 29.3 Illustration of the ultrasound beam in the near field and the far field of the transducer.

The sound intensity in the near field is complicated due to interference effects of sound originating from different parts of the transducer membrane. In the far field, the intensity is approximately uniform and the beam spread follows:

$$\sin\beta = 1.22\frac{\lambda}{D} \tag{29.8}$$

where β is the half lobe angle.

To get a narrow beam, the transmitter membrane diameter must be large with respect to the wavelength. High-frequency ultrasound cannot be the general solution as ultrasound of a high frequency is absorbed faster than ultrasound of a low frequency.

29.2 Ultrasound Transducer

Ultrasonic transducers can be based on different techniques. The most common types of in-air transducers are

- Mechanical
- Electromagnetic
- Piezoelectric
- Electrostatic
- Magnetostrictive

The simplest type, mechanical transducers such as whistles and sirens are used up to approximately 50 kHz. This type only works as a transmitter. Electromagnetic transducers such as loudspeakers and microphones can be used, but they are mainly suited for lower frequencies. The piezoelectric transducer is more suitable for ultrasonic frequencies and is quite common. Piezoelectric crystals change dimensions when they are exposed to an electric field. When an alternating voltage is applied over the piezoelectric material, it changes its dimensions with the frequency of the voltage. The transducer is mainly suited for use at frequencies near the mechanical resonance frequency of the crystal. The piezoelectric transducer works as both a transmitter and a receiver. When a piezoelectric material is forced to vibrate by a sound pulse, it generates a voltage. Some natural crystals, such as quartz, are piezoelectric. Ceramics and some polymers, like PVDF (polyvinylidene fluoride) can be polarized to become piezoelectric. Polymers are suitable as transducers in air since their acoustic impedance is low compared to other standard piezoelectric materials.

The electrostatic transducer (Figure 29.4) is a plate capacitor with one plate fixed and the other free to vibrate as a membrane. When a voltage is applied between the plates, the electrostatic forces tend to attract or repel the plates relative to each other depending on the polarity of the voltage. This transducer can be used as both a transmitter and a receiver. The magnetostrictive transducer is based on the

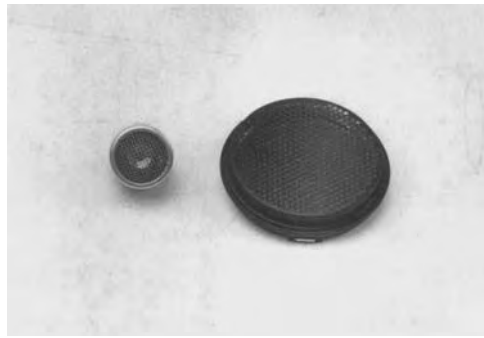


FIGURE 29.4 Ultrasonic transducers: piezoelectric (left) and electrostatic (right).

phenomenon of magnetostriction, which means that the dimensions of a ferromagnetic rod change due to the changes of an externally applied magnetic field. This transducer can also act as both a transmitter and a receiver [10–16].

Transducers are often integrated with electronics to a complete system for the measurement.

29.3 Principles of Time-of-Flight Systems

There are several techniques for ultrasonic range measurements. They are all based on the principle of measuring the time difference between transmitted and received ultrasound signal as shown in Figure 29.5.

The *pulse-echo method* is the simplest one. Usually, this method has a low signal-to-noise ratio (SNR) because of the low transmitted energy due to the short duration of the pulse. Multireflections are detectable.

In the *phase-angle method*, the phase angle is measured between the continuous transmitted signal and the continuous received signal and is used to measure the distance. The method is relatively insensitive to disturbances. Multireflections are not detected in a meaningful way. When the distance is longer than one wavelength, another method must be used to monitor the distance.

The *frequency-modulation method* uses transmitted signals that are linearly frequency modulated. Thus, detected signals are a delayed replica of the transmitted signal at an earlier frequency. The frequency shift is proportional to the time of flight. The method is robust against disturbing signals and multireflections are detectable.

The *correlation method* determines the cross-correlation function between transmitted and received signals. When the transmitted signal is a random sequence, the cross-correlation function estimates the

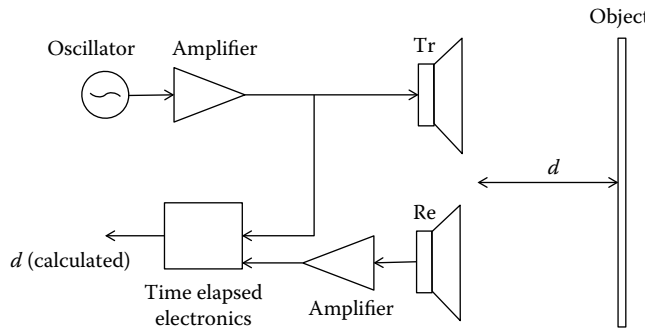


FIGURE 29.5 Principle of an ultrasonic distance measurement system.

impulse response of the system, which in turn is a good indicator of all possible time delays. The method is robust against disturbances, and multireflections are detectable [17–21].

Industrial acoustic noise can affect the received signals in an ultrasound time-of-flight system. The noise can be generated from leaking compressed air pipes, noisy machines, or other ultrasonic systems. This external noise is not correlated with the relevant echo signals of the transmitted noise and can therefore be eliminated by the use of correlation methods. Disturbances correlated with the relevant echo signal (e.g., unwanted reflections) will not be eliminated by the use of correlation methods.

The impulse response $h(t, t_0)$ is used as a sensitive indicator of time delay between transmitted signal at time t_0 and received signal at time t. The impulse response is given by

$$h(t, t_0) = F^{-1}\left[\frac{S_{xy}}{S_{xx}}\right] \tag{29.9}$$

where

F^{-1} is the inverse Fourier transform
x is the transmitted signal
y is the received signal
$S_{xy}(f)$ is the cross-spectral density function [the Fourier transform of the cross-correlation function of the transmitted signal $x(t)$ and the received signal $y(t)$]
S_{xx} is the power density function (the Fourier transform of the autocorrelation function of the transmitted signal)

To analyze the transfer channel by data acquisition requires a high sampling rate, theoretically at least two times the highest frequency component in the received signal (and in practice as high as 10 times the highest frequency). One way to reduce the sampling rate is to first convert the signal from its band-pass characteristics around a center frequency f_0 (approx. 50 kHz) to low-pass characteristics from dc to $B/2$, where B is the appropriate bandwidth. The accuracy of the range estimate, and hence the time interval $t - t_0$, can be improved by processing the estimate of the impulse response $h(t - t_0)$ with a curve-fitting (least square) method and digital filtering in a computer. A correlation-based time-of-flight system is described in Ref. [18].

Table 29.1 lists some advantages and drawbacks of the described time-of-flight methods.

29.3.1 Suppliers and Vendors

There are many manufacturers and vendors of ultrasonic sensors for displacement measurements. There are several types of sensors depending on physical principle of the sensor, mounting, and capsuling.

TABLE 29.1 Advantages and Disadvantages of Time-of-Flight Methods

Method	Main Advantage	Main Disadvantage
Pulse-echo method	Simple	Low SNR
Phase-angle method	Rather insensitive to disturbances	Cannot be used directly at distances longer than the wavelength of the ultrasound
Frequency-modulation method	Robust against disturbances; multireflections detectable	Can give ambiguous results in long distance measurements, but can give similar results as phase-angle method in short distance measurements
Correlation method	Very robust against disturbances	Make relatively high demands on hardware and/or computations

The sensor is often integrated with electronics for distance measurements including more or less advanced intelligence, handling unwanted phenomena, and also prepared with an interface to other systems. A partial list of vendors and suppliers is given at the end of this chapter.

References

1. J. Blitz, *Elements of Acoustics*, London, U.K.: Butterworth, 1964.
2. L. E. Kinsler, A. R. Frey, A. B. Coppens, and t. V. Sanders, *Fundamentals of Acoustics*, 3rd edn., New York: John Wiley & Sons, 1982.
3. G. Lindstedt, Borrowing the bat's ear for automation. Ultrasonic measurements in an industrial environment, Department of Industrial Electrical Engineering and Automation, Lund Institute of Technology, Lund, Sweden, 1996.
4. G. S. Kino, *Acoustic Waves: Devices, Imaging and Analog Signal Processing*, Englewood Cliffs, NJ: Prentice Hall, 1987.
5. J. Burczynski, Introduction to the use of sonar systems for estimating fish biomass, FAO Fisheries Technical Paper No. 191, Revision 1, Food and Agriculture Organization of the United Nations, Rome, Italy, 1982.
6. F. Widmann, Parking aid with a wheel sensor, United States Patent No. 5,602,542, February 1997.
7. V. Mágori, Ultrasonic sensors in air, pp 471–481, *Proceedings of the IEEE Ultrasonic Symposium*, New York, 1994.
8. D. P. Massa, Choosing an ultrasonic sensor for proximity or distance measurement part 1: Acoustic considerations, *Sensors*, 16/2 (3), 1999. Available online at Acoustic/Ultrasound, http://www.sensorsmag.com/sensors/acoustic-ultrasound/choosing-ultrasonic-sensor-proximity-or-distance-measurement-825 (accessed on March 14, 2013).
9. D. P. Massa, Choosing an ultrasonic sensor for proximity or distance measurement part 2: Optimizing sensor selection, 16/3 (10), 1999. Available online at Acoustic/Ultrasound, http://www.sensorsmag.com/sensors/acoustic-ultrasound/choosing-ultrasonic-sensor-proximity-or-distance-measurement-838 (accessed on March 14, 2013).
10. S. R. Ruocco, *Robot Sensors and Transducers*, New York: John Wiley & Sons 1987.
11. P. H. Sydenham and R. Thorn, *Handbook of Measurement Science*, Vol. 3, *Elements of Change*, New York: John Wiley & Sons, 1992.
12. J. Fraden, *Handbook of Modern Sensors, Physics, Design and Applications*, 4th edn., New York: Springer, 2010.
13. H. R. Gallantree, Review of transducer application of polyvinylidene fluoride, *IEEE Proceedings*, 130, 219–224, 1983.
14. T. T. Wang and J. M. Herbert, *The Applications of Ferroelectric Polymers*, London, U.K.: Chapman & Hall, 1988.
15. C. Z. Rosen, B. V. Hiremath, and R. Newham, *Piezoelectricity*, New York: American Institute of Physics, 1992.
16. P. Mattila, F. Tsuzuki, H. Väätäjä, and K. Sasaki, Electroacoustic model for electrostatic ultrasonic transducers with V-grooved backplates, *IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control*, 42(1), 1–7, 1995.
17. J. A. Kleppe, *Engineering Applications of Acoustics*, Boston, MA: Artech House Publishers, 1989.
18. P. Holmberg, Robust ultrasonic range finder—An FFT analysis, *Measurement Science and Technology*, 3, 1025–1037, 1992.
19. T. Kimura, S. Wadaka, K. Misu, T. Nagatsuka, T. Tajime, and M. Koike, A high resolution range measurement method using double frequencies and phase detection, pp. 737–741, *Proceedings of the IEEE Ultrasonic Symposium*, Seattle, WA, 1995.

20. M. M. Saad, C. J. Bleakley, and S. Dobson, Robust high-accuracy ultrasonic range measurement system, *IEEE Transactions on Instrumentation and Measurement*, 60, 3334–3341, 2011.
21. M. Parilla, J. J. Anaya, and C. Fritsch, Digital signal processing techniques for high accuracy ultrasonic range measurements, *IEEE Transactions on Instrumentation and Measurement*, 40(4), 759–763, 1991.

Partial List of Vendors and Suppliers

Advanced Sonic Processing Systems, 324 Christian St., Oxford, CT 06478-1023, Phone: 203-266-4440, Fax: 203-266-4441, www.advancedsonics.com

Banner Engineering Corp, 9714 Tenth Ave. N., Minneapolis, MN 55440, Phone: 763-544-3164, 877-581-1749 (toll free), Fax: 763-544-3213, www.bannerengineering.com/en-US/

Blue Wave Ultrasonics, Inc., P.O. Box 4347, Davenport, IA 52808-4347, Phone: 563-322-0144, 800-373-0144 (toll free), Fax: 563-322-7180, www.bluewaveinc.com

DeFelsko Corp., 802 Proctor Ave., Ogdensburg, NY 13669, Phone: 315-393-4450, 800-448-3835 (toll free), Fax: 315-393-8471, www.defelsko.com

Olympus NDT, 48 Woerd Ave., Waltham, MA 02453, Phone: 781-419-3900, 800-225-8330 (toll free), Fax: 781-419-3980, www.olympus-ims.com/?apd=1

Senix Corp, 52 Maple St., Bristol, VT 05443, Phone: 866-919-2671 (toll free), Fax: 802-453-2549, www.senix.com

Sonic and Materials, Inc., 53 Church Hill Rd., Newtown, CT 06470, Phone: 203-270-4600, 800-745-1105 (toll free), Fax: 203-270-4610, www.sonics.com

Ultrasonic Power Corp., 239 E. Stephenson St., Freeport, IL 61032, Phone: 815-235-6020, 800-575-0168 (toll free), Fax: 815-232-2150, www.upcorp.com

Transducers Direct, 12115 Ellington Court, Cincinnati, OH 45249, Phone: 513-583-9491, Fax: 513-583-9476, www.transducersdirect.com

30

Optical Encoder Displacement Sensors

30.1 Encoder Signals and Processing Circuitry **30**-2
Absolute Encoders
30.2 Incremental Encoder Quadrature Signals **30**-7
Digital Quadrature Signals • Analog Quadrature Signals
30.3 Encoding Principles **30**-13
Geometric Masking • Moiré Fringes • Diffraction-Based Encoders • Rotary Encoders • Linear Encoders
30.4 Components and Technology **30**-19
References **30**-21
Further Readings **30**-22

J.R. René Mayer
Ecole Polytechnique de Montreal

The detection of angular and linear motion is a key function in a multitude of systems such as machine tools, industrial robots, a variety of instruments, and computer mice. Although they are one of the many techniques capable of such measurements, the ease with which they are interfaced to digital systems has made them very popular.

Optical encoders are used to measure either angular or linear positions. Those used for angular detection are commonly called rotary or shaft encoders, since they usually detect the rotation of a shaft. Optical encoders encompass a variety of devices, all of which use light as the means to transform movement into electrical signals. All devices have two basic building blocks: a main grating and a detection system. It is the position of one with respect to the other that is detected. The main grating represents the measurement standard. For linear measurements, the main grating, commonly called the scale, is one or more sets of parallel lines of constant or specially coded pitch supported by a substrate. Similarly, a rotary encoder has a grating with radial lines on a disk.

Both linear and rotary encoders can, in principle, be absolute or incremental, although in practice, linear absolute encoders employing optical principles are quite uncommon and have drastically limited performance characteristics (accuracy, resolution, and/or maximum operating speed). Figure 30.1 shows a simplified classification of optical encoders. This classification refers to the nature of the information generated. The incremental encoder detects movement relative to a reference point. As a result, some form of reference signal is usually supplied by the encoder at a fixed position in order to define a reference position. The current position is then incremented (or decremented) as appropriate. Multiple reference marks can also be used, where the distance between successive marks is unique so that as soon as two successive marks have been detected, it becomes possible to establish absolute position from then on. The reference point can also be mechanical. Should power be lost or a signal transmission error occur, then the absolute position is lost and the encoder must return to one or more reference points in order to reset its counters. Unfortunately, a loss of count may not be detected until a reference point is reaccessed. Furthermore, reading errors may accumulate. On the other hand, absolute encoders

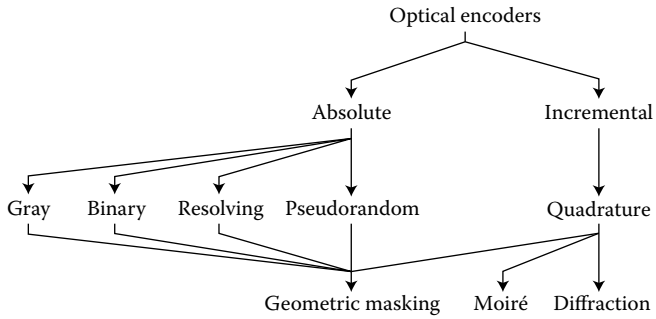


FIGURE 30.1 Classifications of optical encoders based on (1) the nature of the final information provided, (2) the type of signals generated, and (3) the technology used to generate the signals.

produce a set of binary signals from which the absolute position can be deduced without the knowledge of the previous motion history. The current position is known right from powering on. In the case of absolute rotary encoders, single and multiturn devices are available. Multiturn devices use an internal mechanical transmission system to drive a second grating that serves as a turn counter.

Most incremental encoders use quadrature signals as output to carry the motion information. Some encoders use one square-wave signal, which is used for position in one direction only. Also, this single square wave can be fed into either a PLC (programmable logic controller) or another electronic interface that converts this signal to a rate or revolution per minute (RPM) for speed indication. However, whenever bidirectional operation is required, quadrature signals are necessary. Quadrature signals come in analog or digital form. The analog form consists simply of a sine and a cosine signal. The number of sinusoidal cycles per unit change of the measured variable (a revolution or 360° for a rotary encoder) determines the basic resolution of the encoder prior to interpolation. The digital form consists of two square-wave trains, 90° (often called electrical degree) out of phase. The 90° phase lag is indispensable in detecting the motion direction and hence increment or decrement of the position counter accordingly. The main optical techniques to generate the quadrature signals are geometric masking, Moiré fringes, and diffraction based. For linear encoders, the basic resolution is related to the distance traveled by the grating in order for the encoder to produce one full quadrature cycle. For rotary encoders, the basic resolution is usually described as the number of quadrature cycles per turn. The resolution of an encoder system can be increased by electronic means. With analog quadrature signals, it is possible to interpolate within each quadrature cycle. The limit of the interpolation factor depends on the quality (mark space, quadrature separation, and jitter) of the basic signals. With square-wave signals, multiplication by a factor of two or four is easily achieved. Increasing the resolution in this manner does not, however, improve the trueness, often called accuracy (or linearity) of the measurement.

Absolute encoders are classified according to the type of code used. The main four codes are Gray, binary (usually read by vee-scan detection), optical resolving, and pseudorandom. All absolute encoders use geometric masking to generate the code.

30.1 Encoder Signals and Processing Circuitry

30.1.1 Absolute Encoders

30.1.1.1 Direct Binary

Figure 30.2a illustrates the concept of an absolute linear optical encoder using a direct binary-encoded scale. The fixed scale has n tracks (here n is 4), each providing one bit of a direct binary number. The lowest track (first track from the center of the disk for a rotary encoder) is the most significant digit and has a weight, 2^{n-1} (here 2^3), while the upper track is the least significant digit with a weight 2^0. The track providing the least significant digit has 2^{n-1} cycles of light and dark records, while the most significant

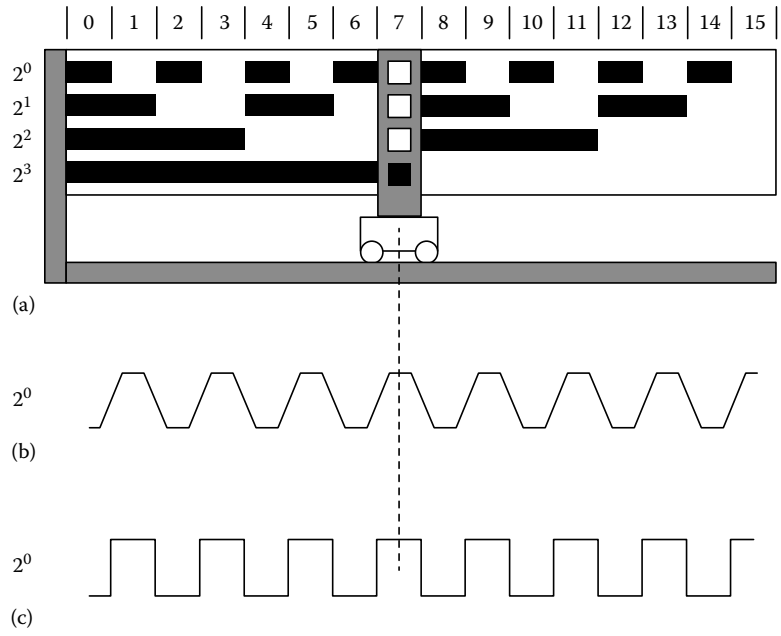


FIGURE 30.2 (a) Absolute encoders using a natural binary code of four digits. Four tracks are required. The moving read head has four apertures and is shown in position 7 along the scale. (b) The output of the read-head aperture corresponding to the least significant track. It represents the proportion of light area covering the aperture. (c) The binary digit obtained after squaring the raw output signal.

track has 2^0 or 1 such cycle. For each track, the moving read head has a readout unit consisting of a light source, a mask, and a photodetector. Figure 30.2b shows the output from the photodetector, which represents the total intensity of light reaching its surface. As the mask passes over a clear region of the grating, the photodetector output increases and then decreases. In theory, a truncated triangular wave is obtained, which can easily be converted to a square wave (Figure 30.2c) by a suitably chosen threshold level. The result is a high or 1 for a light record and a low or 0 for a dark one. The position, in base 10, corresponding to the reading head position in Figure 30.2 is

$$1\times 2^0 + 1\times 2^1 + 1\times 2^2 + 0\times 2^3 = 7 \tag{30.1}$$

The code configuration just described is not suitable for practical use because some transitions require that two or more bit values change simultaneously. For example, from positions 7 to 8, all bits change values. Unless the change is simultaneous, incorrect position readout results at some position. This would require that the scale is geometrically perfect, that the read head be perfectly aligned with the scale, and that the electronics are perfectly adjusted and stable over time. This problem is solved either by the use of a vee-scan detection method or by the use of a unit-distance code such as the Gray code.

30.1.1.2 Vee-Scan

The vee-scan method uses a V-shape pattern of readout units that removes the potential reading ambiguity of direct binary scales. Stephens and Davies [1] indicate the read points at which transitions are detected in Figure 30.3a. They also describe the conversion of the threshold output signals to binary code using combinational logic. The primary advantage is that the location tolerance of the transition point of each reading unit need only be ±1/8 of the cycle length for that particular track. For example, ±45° is for the most significant track of a rotary encoder disk. Figure 30.3b shows a direct binary word obtained through logic combinations of the vee-scan readings.

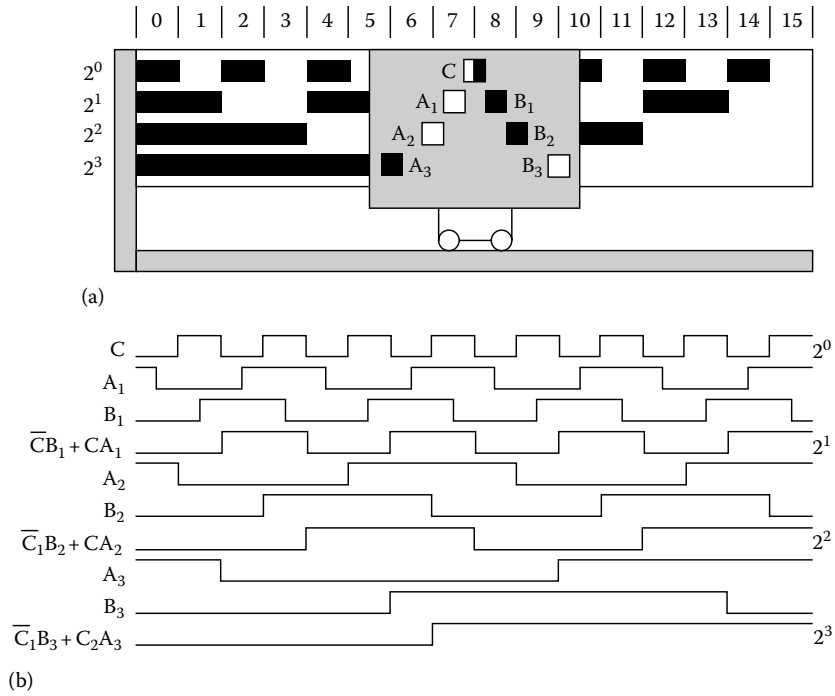


FIGURE 30.3 Vee-scan configuration of reading units in (a) removes the ambiguity associated with a natural binary scale. Simple combinational logic is then used to generate the natural binary readout in (b).

30.1.1.3 Gray Code

The use of vee-scan requires additional reading heads as well as processing electronics. The Gray code is a unit-distance code, and so only one bit of data changes between representations of two consecutive numbers or successive positions. This removes the possibility of ambiguous readout. It has the following advantages: (1) It is easily converted to direct binary code, and (2) the finest tracks are twice the width of equivalent direct binary code tracks. Figure 30.4a shows a Gray code linear scale. Figure 30.4b shows a scheme for the conversion from Gray code to binary code and proceeds as follows: (1) The most significant bit (msb) of the binary code equals the msb of the Gray-coded number, (2) add (modulo 2) the msb of the binary number to the next significant bit of the Gray-coded number to obtain the next binary bit, and (3) repeat step (2) until all bits of the Gray-coded number have been added modulo 2. The resultant number is the binary equivalent of the Gray-coded number. The modulo-2 addition is equivalent to the action of an exclusive OR. Figure 30.4c shows a simple circuit using combinational logic to perform Gray to binary conversion. Sente and Buyse [2] suggest the use of an external ROM to convert the code where the input-coded word is the address and the output-coded word is the data. Using a 16-bit ROM for a 12-bit code, Sente and Buyse [2] suggest using the remaining 4 bits to implement a direction signal. Stephens et al. [1] describe the use of vee-scan with a Gray code scale for even better robustness.

30.1.1.4 Pseudorandom Code

Pseudorandom encoding allows the use of only two tracks, and more recently only one, to produce an absolute encoder. One track contains the pattern used to identify the current position, while the other is used to synchronize the reading of the encoded track and remove ambiguity problems. A pseudorandom binary sequence (PRBS) is a series of binary records or numbers, generated in such a way that any consecutive series of n digits is unique. Such code is called chain code, and it has the property that the first $n - 1$ digits of an

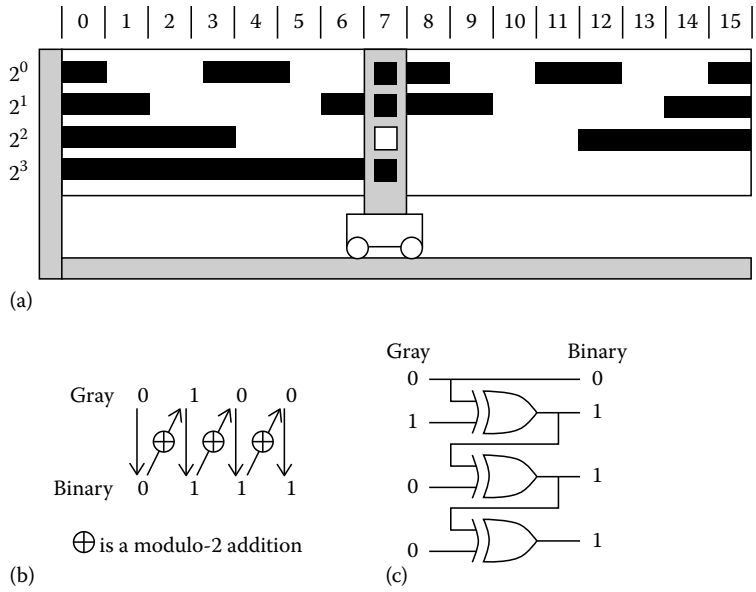


FIGURE 30.4 (a) Gray code allows a transition on only one track between each successive position so that no ambiguity arises. A scheme based on modulo-2 additions converts the Gray code to natural binary code in (b). Exclusive ORs implement the conversion in (c).

n-bit word are identical to the last $n-1$ digits of the previous code word. This allows their partial overlapping on a single track. A PRBS of length $2^n - 1$ is defined by

$$XN(j)\big|j = 0, 1, ..., 2^{n-1} \tag{30.2}$$

The code can be generated by reading the *n*th stage of a feedback shift register after *j* shifts. The register must be initialized so that at least one of the registers is nonzero and the feedback connection implements the formula

$$X(0) = X(n) \oplus c(n-1)X(n-1) \oplus \cdots \oplus c(1)X(1) \tag{30.3}$$

where the *c* coefficients are 0 or 1. The feedback registers for which *c* is 1 are listed in Table 30.1 for values of *n* from 4 to 14. The following is a PRBS for $n = 4$ with the pseudocode obtained using all registers set to 1 initially, 111101011001000. For a rotary encoder disk, the 15 sectors would have a 24° width.

Petriu [3,4] describes a possible configuration of a PRBS disk that uses a PRBS track and a synchronization track (Figure 30.5), together with the processing method to reconstitute the position in natural binary. Table 30.1 gives the reverse feedback configuration. The shift register is initially loaded with the current *n*-tuple. Then the reverse logic is applied recurrently until the initial sequence of the PRBS is reached. At this point, the *n*-bit counter represents the value of *j*. For a rotary encoder, $(j * 360)/(2^n - 1)$ is the current angular position, whereas for a linear encoder, the position is $j * P$ where *P* is the scale record length or pitch. Petriu [4] suggests that in order to allow nonambiguous bidirectional reading, $n + 1$ heads are used on the PRBS track. The synchronization track has a series of 0s and 1s in records of the same width as the PRBS track and in phase. There is a *P*/2 shift, where *P* is the record's length between the A head on the synchronization track and the $n + 1$ read heads on the PRBS track. The $n + 1$ records are updated on a trigger from the A signal. This ensures that the $n + 1$ heads are closely aligned with the PRBS record mid-position. A second read head called *B* on the synchronization track is shifted by *P/2* relative to *A*. *A* and *B* are in quadrature, which allows their simultaneous use to generate a motion direction signal. The correct *n*-tuple, i.e., the lower or upper subset, is selected on the basis of the moving direction and is then converted to natural binary by reverse feedback. Petriu [4] also suggests a simple means of increasing the

TABLE 30.1 Shift-Register Feedback Connections for Generating Pseudorandom Binary Sequences

N	Length	Direct Sequence	Reverse Sequence
4	15	1, 4	3, 4
5	31	2, 5	3, 5
6	63	1, 6	5, 6
7	127	3, 7	4, 7
8	255	2, 3, 4, 8	4, 5, 6, 8
9	511	4, 9	5, 9
10	1023	3, 10	7, 10
11	2047	2, 11	9, 11
12	4095	1, 4, 6, 12	6, 8, 11, 12
13	8191	1, 3, 4, 13	9, 10, 12, 13
14	16,383	1, 6, 10, 14	4, 8, 13, 14

resolution by a factor of 2 using some additional electronics. He also proposes a scheme to use an arbitrary (not $2^n - 1$) number of sectors, but this requires a third track and some additional correction electronics to handle the last $n - 1$ records of a disk, since these are no longer PRBS patterns. Tomlinson [5] proposes another method for truncation of the PRBS sequence that does not require a third track. Instead, particular codes were removed by applying additional logic in the direct and reverse feedback logic.

Ross and Taylor [6] and Arsic and Denic [7] suggest ways of reducing the number of reading heads by accumulating readings into a shift register so that a minimum of two heads are sufficient to read the PRBS track. However, on a start-up, the correct position is not known until the encoder has moved so that all registers have been updated. This type of encoder is therefore not completely absolute because it does not indicate its correct position on the start-up. Finally, Arazi [8] mentions the use of a ROM that stores the translation table, as an alternative to a logic circuit.

30.1.1.5 Optical Resolving

Renishaw Inc. now commercializes a range of single-track absolute linear and rotary encoders that use a track consisting of a long non-repeated binary pattern akin to a barcode. A camera takes a picture of the current visible track portion, which has embedded absolute position code and phase information in a single track of graduations that appear as a long, non-repeating bar code with a 30 μm pitch, and decodes it into an absolute position, claiming a resolution of 1 nm, even at speeds up to 100 m/s. The photo image includes redundancy, so only about a quarter of the code captured in the image needs to be correct, providing good immunity to dirt and contamination. A high-speed digital signal processor (DSP) is used to process the position determination and checking algorithms and answer the position request in less than 15 μs. The system is capable of processing and answering up to 25,000 position requests per second. The encoder is primarily linear but is also used for rotary applications by putting the scale on the outer perimeter of a disk.

This method has similarities with its electromagnetic counterpart and depends on the generation of a sine and a cosine signal pair per encoder shaft revolution. The resolution and accuracy of this encoder depend on its ability to generate signals that conform to their ideal waveforms and the resolving power of the electronic circuit responsible for performing the rectangular to polar conversion to produce angular output data. The methods for accomplishing this conversion are similar to those described in the sections on analog quadrature signal interpolation. The code disk of such a device does not incorporate a series of parallel lines of constant pitch, but rather depends on special code tracks having analog profile signal outputs. These can be affected by changing the code tracks' cross sections or by graduating their optical density around the revolution. In the latter case, a masking reticle need not be employed.

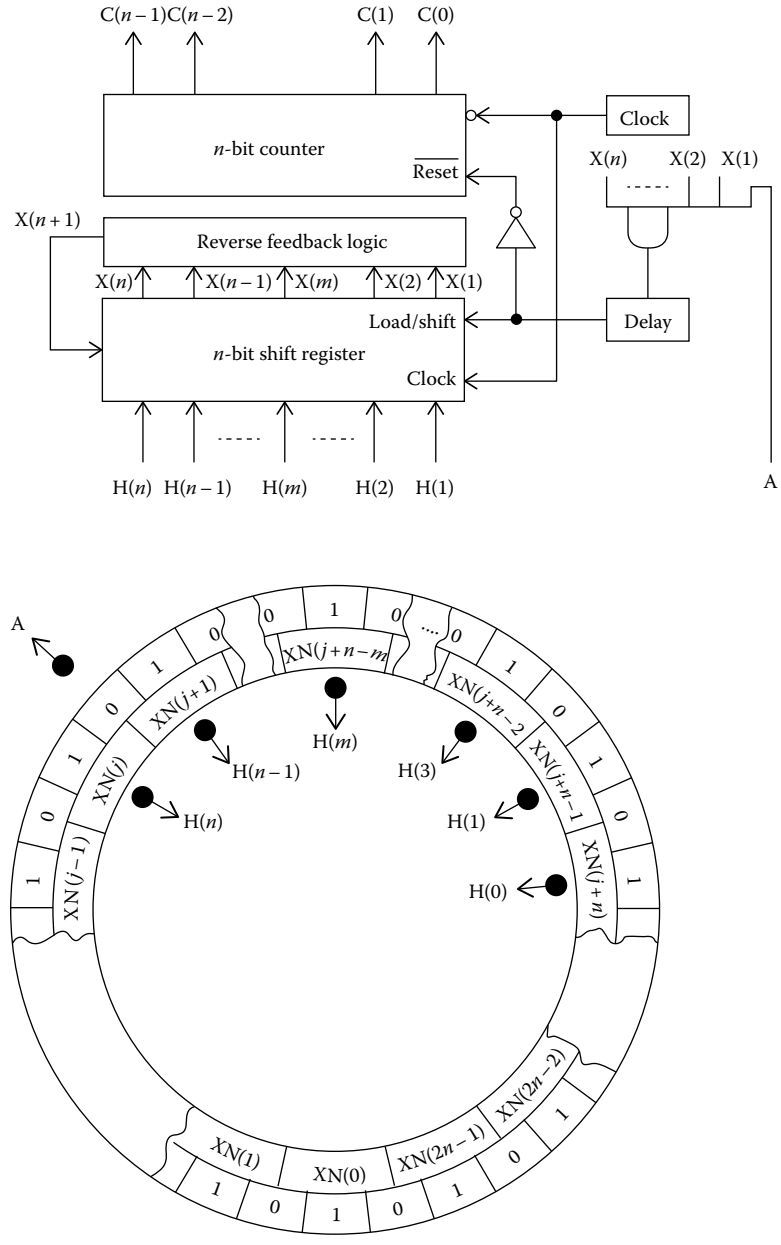


FIGURE 30.5 Pseudorandom shaft encoder with a simple synchronization track to validate the readout from the n-tuple read by the reading heads $H(i)$, $i = n, \ldots, 1$. The circuit is a simplified code conversion to binary based on the reverse feedback logic. The counter counts the number of steps required to return to the initial sequence.

30.2 Incremental Encoder Quadrature Signals

30.2.1 Digital Quadrature Signals

Figure 30.6 illustrates the concept of an incremental rotary optical encoder. The shaft-mounted disk has a series of alternating dark and light sectors of equal length and in equal numbers. The dark and light code is detected by a stationary mask with two apertures, *A* and *B*, displaced one quarter of a cycle from

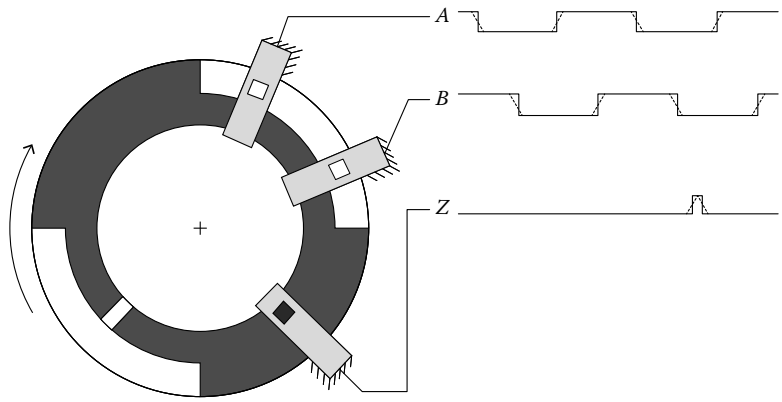
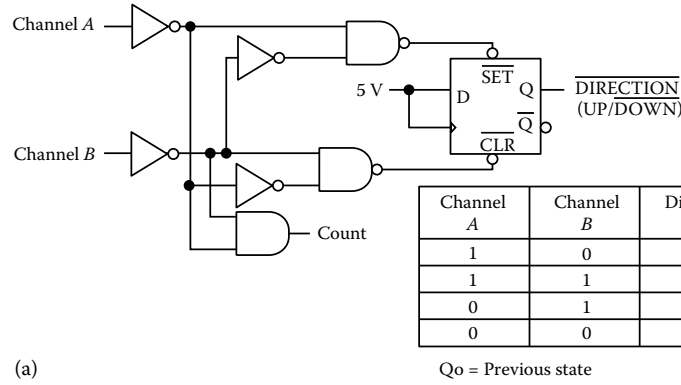


FIGURE 30.6 The lowest possible resolution for a rotary incremental encoder. The outer track provides $n = 2$ cycles of the alternating black and clear records. Two heads A and B displaced by 90° give, after squaring, two square waves in quadrature.

each other. When a light sector covers a window, a 1 signal is produced, and a 0 results from a dark sector. At a transition, a rising or falling signal occurs. These signals require some pretreatment to square the signals and avoid problems associated with slow movement at the transition positions, resulting in slow rise time and fall time, in the presence of low noise. The resulting cleaned A and B signals are two square waves 90° out of phase and are called quadrature signals. Since the A and B signals have only four possible states, they clearly do not provide a means of distinguishing more than four different locations. As a result, absolute position discrimination is only possible within one quadrature cycle. Instead, the quadrature signals are used to increment or decrement a counter that gives the actual position. The counter is initialized (usually reset) on a z-phase signal produced on a separate track, the innermost track in Figure 30.6. It is also possible to have a number of z-phase signals over the encoder range with a distance-coded variation between the z-phase markers. The code can be used to initialize the counter at nonzero positions.

Figure 30.7a shows a simple circuit from Conner [9] to provide one count per cycle of the quadrature signal. The direction information necessary to decide whether to increment or decrement the counters is also produced. Note that Schmitt triggers provide some amount of hysteresis and redress the falling and rising edges from the encoder. Further circuitry is required to increase the amount of hysteresis in very noisy environments. Additional treatment may be required before feeding these signals to counters. The reason is that if, following a count, there is a small movement in the opposite direction without a reverse count signal being issued followed by a forward motion, then the forward motion produces a second count at the same position. Kuzdrall [10] proposes a more reliable circuit in Figure 30.7b. The circuit drives a 74HC193 counter with a down clock and an up clock. Both flip-flops are normally in the high state. Whenever one of them switches to a low state, it is soon reset by the use of that output as a set signal. That low output is only possible when phase B is low and there is a transition on A. Depending on the direction of that transition, one of the flip-flops produces the brief low state, causing an appropriate up or down count. The problem associated with small oscillations around a position is reduced since the up and down counts occur at the same encoder position. Venugopal [11] proposes a circuit, in Figure 30.7c, which produces a similar effect. A count can only occur when B is low and there is a transition on A. Depending on the transition direction, one of two the monostables triggers and effectively allows the count at the appropriate input of the 74LS193 counter.

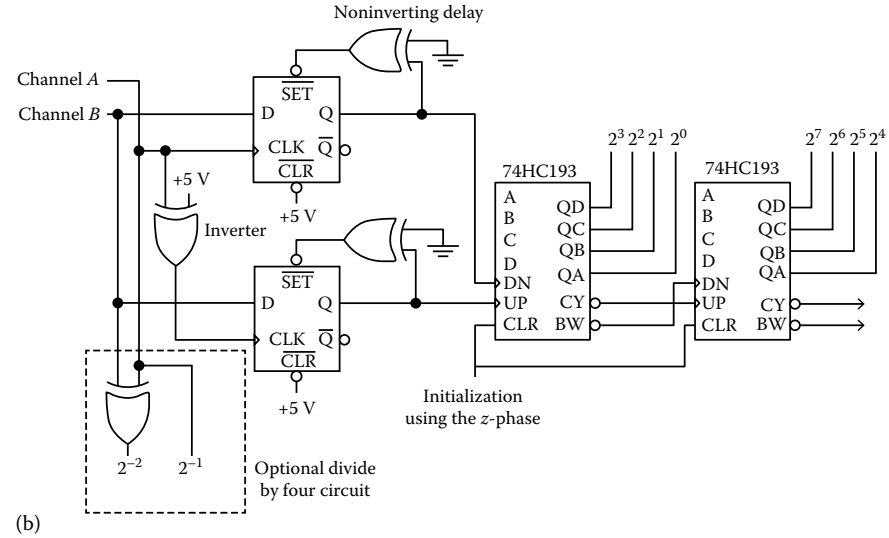
In cases where a noisy environment is present, Holle [12] describes a digital filter to clean the A and B signals further, a filter that uses a small number of logic gates to form a 4-bit delay filter. Wigmore [13] proposes other circuits for count and direction detection.



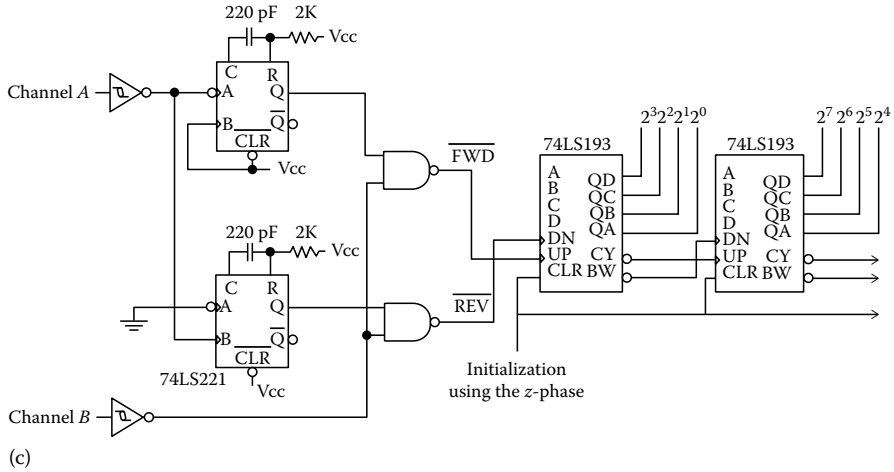
Channel *A*	Channel *B*	Direction	Count
1	0	0	0
1	1	Qo	0
0	1	1	0
0	0	Qo	1

Qo = Previous state

(a)



(b)



(c)

FIGURE 30.7 Circuit producing a direction and a count signal. (From Conner, D., *EDN*, 35(9), 57, 1990.) (a) Caution must be exercised when using this circuit for counting purposes since multiple counts are possible when the encoder oscillates around the $A = 0$ and $B = 0$ and the $A = 0$ and $B = 1$ states. (From Kuzdrall, J.A., *Electron. Des.*, 17, 81, 1992.) (b) Counts up and down at the same states transition but depending on the direction of movement.

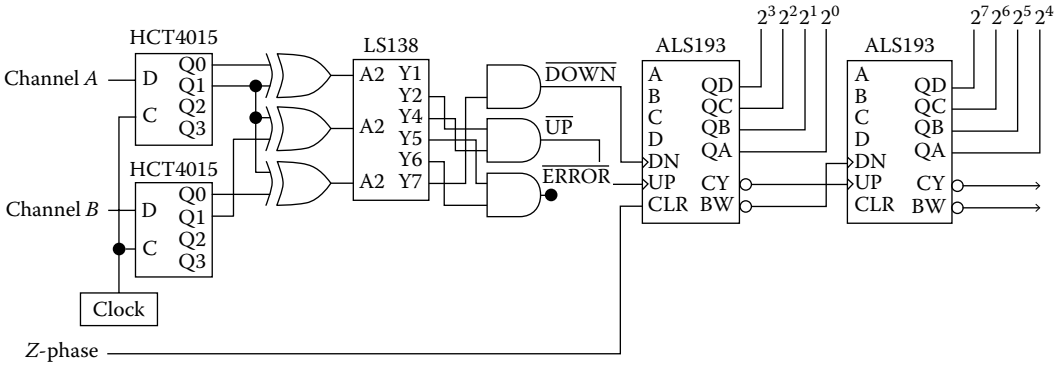


FIGURE 30.8 Divide-by-four circuit producing four counts per cycle of the quadrature signals. Based on except for the LS138 demultiplexer that replaces the suggested 4051 multiplexer because of the latter's unequal fall time and rise time. The clocked shift registers generate a 4-bit code to the demultiplexer for each clock cycle. The clock frequency should be eight times the maximum frequency of the quadrature signals. (From Butler, M.M., *Electronics*, 20, 128, 1980.)

Tables 30.6 and 30.7 list commercial chips and their suppliers' details. Some chips produce count and direction signals or up and down clocking signals. Others also include counters.

The aforementioned circuits do not fully exploit the information contained in the quadrature signals since only one of the four edges (or states) within one quadrature cycle is used to count. Figure 30.8 shows a slightly modified version of a divide-by-four counter circuit proposed by Butler [14]. Two 4-bit shift registers, three exclusive OR gates, and an 8-channel demultiplexer derive the up and down count signals. The clocked shift registers generate a 4-bit code to the demultiplexer for each clock cycle. To ensure that no encoder transitions are missed, the clock frequency should be at least 8 *NS*, where *N* is the number of cycles produced by the encoder for each shaft revolutions and *S* is the maximum speed in revolutions per second. The up and down signals can be fed to cascaded 74ALS193 counters. This circuit analyzes the current and previous states of the *A* and *B* channels to either count up, count down, or issue an error flag when an improper sequence occurs.

Kuzdrall [10] proposes to view a quadrature signal cycle as a Gray code so that the two least significant bits of the count are obtained by a Gray code to binary code conversion as in Figure 30.7b. Phase *A* generates bit 1 of the natural binary position, and phase *B* is exclusive ORed with phase A to produce the least significant bit. The counter provides the remaining bits of the natural binary position. Marty [15] proposes a state machine that stores both the actual *A* and *B* values and the previous values using *D*-type flip-flops (in a similar way to Butler) to form a hexadecimal number. In all, eight different hex digits are generated, four for each direction of motion. A 4-line to 16-line decoder then feeds into two 4-input NAND gates to produce an up or down count signal. As for Butler, this last circuit is not dependent on propagation delays as with Kuzdrall's circuit.

30.2.2 Analog Quadrature Signals

30.2.2.1 Interpolation by Resistor Network

Some optical encoders deliver analog quadrature signals in the form of a $\sin(\theta)$ signal, *A*, and a $\cos(\theta)$ signal, *B*, where θ is the phase (electrical degree) within one cycle of the quadrature signal. θ does not equal shaft angle but is related to it: 360 electrical degrees of θ corresponds to $360/N$ mechanical degrees, where *N* is the number of analog quadrature signal cycles per shaft revolution. Indeed, θ still exists for linear encoders of this type. Although these signals can be squared and fed to a divide-by-four counter, they can also be processed directly to generate a finer resolution. The main techniques are (1) multiple phase-shifted signals, (2) lookup table, and (3) arctangent processor.

The multiple phase-shifted signals method relies solely on electronics to increase the frequency of the final digital quadrature signals by an integer amount. Benzaid and Bird [16] propose the circuit of Figure 30.9a, which has been designed in this particular case for a fourfold frequency increase. This can then be followed by a digital divide-by-four circuit (not shown). The A and B signals are combined to produce an additional six phase-shifted signals, three of which are shifted by $\pi/8$, $\pi/4$, and $\pi/8$, respectively, from the A signal and three others that are shifted similarly from the B signal. The result is a total of eight available sinusoidal signals, phase shifted by $\alpha = i\,\pi/8$ with $i = 0, 1, \ldots, 7$. This can be generalized, saying that for an m-fold increase in resolution, $2m$ signals that are phase shifted by $\alpha i\,\pi/2m$ with $i = 0, 1, \ldots, 2m - 1$ are required. The phase-shifted sinusoidal signals are then squared using TTL

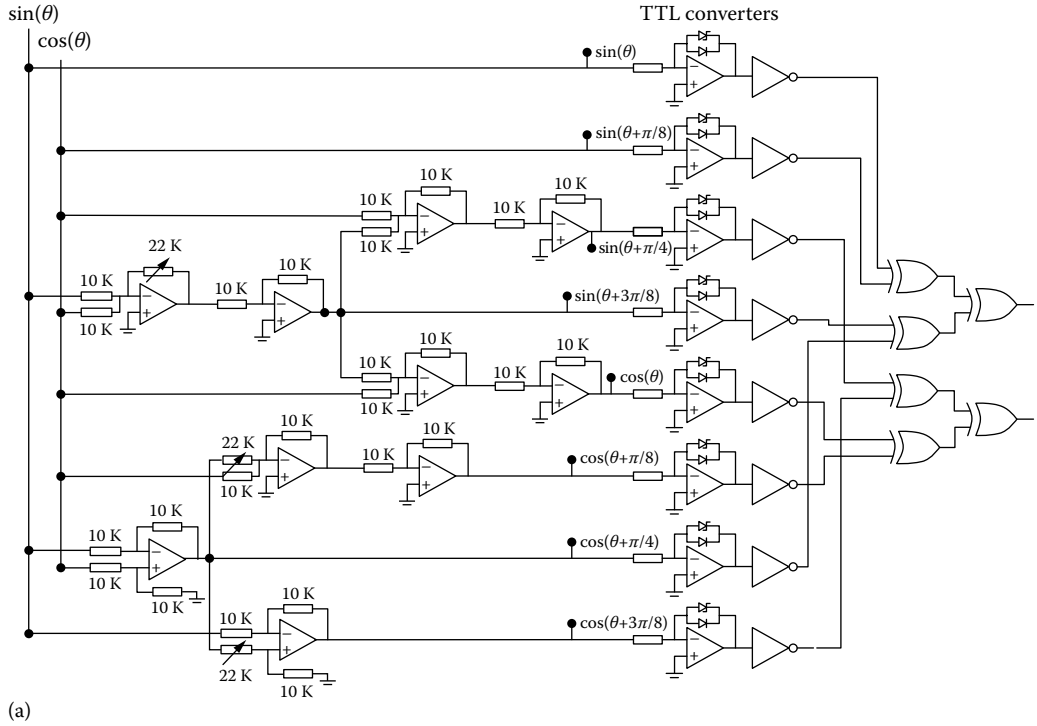


FIGURE 30.9 Interpolation circuit. The sine and cosine signals from the encoder are vector combined in (a) to produce various phase-shifted signals. Following TTL conversion, they are exclusive ORed. The squared phase-shifted signals and the final result are in (b). (From Benzaid, O. and Bird, B.M., Interpolation techniques for incremental encoders, *Proceedings of 23rd International Conference on Intelligent Motion*, June 22–24, pp. 165–172, 1993.)

converters. Figure 30.9b shows the resulting square waves. The addition, modulo 2, of the signals for even values of i gives A' and similarly for the signals with odd values of i gives B'. The modulo-2 sum is performed via exclusive ORs.

The vector additions of the initial A and B signals result in phase-shifted signals. The weights of A and B are calculated from the trigonometric relations:

$$\sin(\theta + \alpha) = \cos(\alpha)\sin(\theta) + \sin(\alpha)\cos(\theta) \tag{30.4}$$

and

$$\cos(\theta + \alpha) = -\sin(\alpha)\sin(\theta) + \cos(\alpha)\cos(\theta) \tag{30.5}$$

where $\sin(\theta) = A$ and $\cos(\theta) = B$.

Thus,

$$\sin(\theta + \alpha) = \cos(\alpha)\ A + \sin(\alpha)\ B \tag{30.6}$$

and

$$\cos(\theta + \alpha) = -\sin(\alpha)\ A + \cos(\alpha)\ B \tag{30.7}$$

Note that the amplitudes of the phase-shifted signals are not critical, since it is their zero-crossing points that produce the square-wave transition. Also, in the circuit of Figure 30.9a, the weights were slightly modified to simplify the implementation, which results in a small variation of the duty cycle of A' and B'.

The phase-shifted signals can alternatively be produced using voltage dividers and Schmitt triggers, where the divider resistors are in the ratio $\tan(\alpha)$ [17]. As m increases, the precision of the weights becomes more stringent, and the speed of the electronics processing the high-frequency square signals might limit the upper value of m possible with this method.

30.2.2.2 Interpolation by Sampling and Numerical Processing

A number of methods digitize the analog quadrature signals in order to perform a digital interpolation within one cycle of the quadrature signals. These techniques permit an even higher interpolation. The signals are periodically sampled in sample-and-hold circuitry and digitized by an analog-to-digital converter (ADC). One technique uses an interpolation table, while the other performs arctangent calculations.

Hagiwara [18] uses an m-bit ADC to digitize the analog A and B signals into the binary numbers Da and Db. Figure 30.10a shows how these binary numbers are used as the addresses of a grid, here built for $m = 3$ as a 2^m by 2^m grid. Hagiwara then associates a phase angle with the center of each cell of the matrix using simple arctangent calculations. The angle is calculated with respect to the center of the grid. This phase angle is then associated with one of $2^n - 1$ phase codes using

$$\text{Phase code} = \text{ integer part of } \left(2^n \frac{\theta}{2\pi} \right) \tag{30.8}$$

as shown in Figure 30.10b.

The proposed circuit includes the means of performing the accumulation of interpolated values to deliver the encoder absolute position. A modification to this circuit is also proposed that can compensate the phase code for known inaccuracies in the encoder signals with respect to the actual encoder position.

For the highest possible level of interpolation, Mayer [19] describes an arctangent calculator method. It uses a microprocessor to perform high-level trigonometric calculations on the digitized analog signals. As this process is performed, the analog signals are continuously squared and fed to a

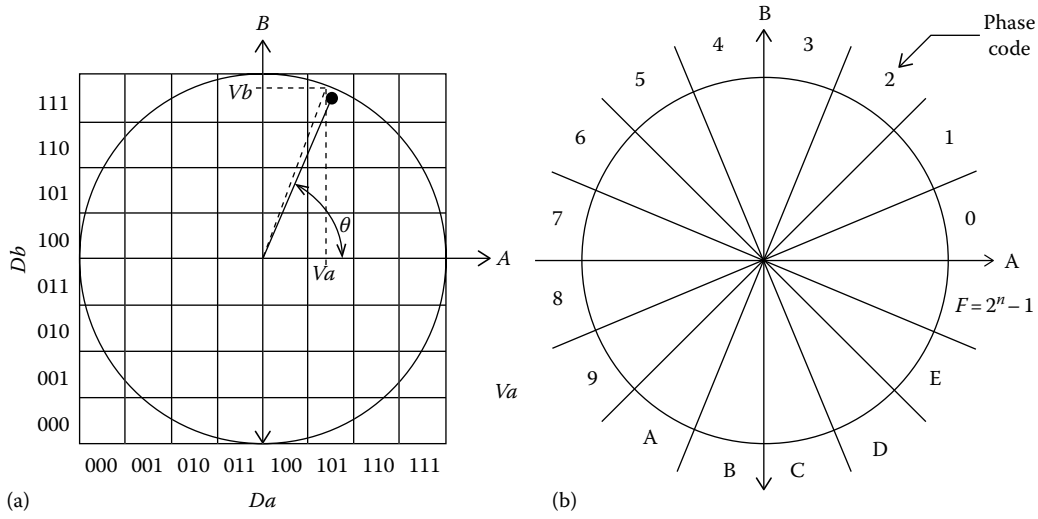


FIGURE 30.10 Hagiwara proposes a 2D lookup table that associates a quantized phase value to a set of digitized values of the quadrature signals. The phase code associated with a grid address may be adjusted to compensate for known errors in the quadrature signals: (a) the use of binary numbers for addressing a grid, and (b) calculation of angles with respect to the center of the grid. (From Hagiwara, N., *IEEE Trans. Instrum. Meas.*, 41(1), 98, 1992.)

divide-by-four counter. The two information sources are combined to produce the encoder's position with a very high level of resolution. Note, however, that although the resulting theoretical resolution is limited only by the ADC, in practice, it is the quality of the analog signals in relation to the physical position being measured that will limit the precision obtained.

30.3 Encoding Principles

Optical encoders use one of three techniques to generate the electrical signals from the relative movement of the grating and the reading heads. They are (1) geometric masking, (2) Moiré effects, and (3) laser interference. Note that absolute encoders mainly use geometric masking. Geometric masking relies on geometric optics theory and considers light as traveling in a straight line. However, as the grating period reduces and the resolution of the encoder increases, Moiré fringe effects are observed and used to produce the signals. Although diffraction effects then become nonnegligible, their influence can be controlled by careful design. Finally, for very high resolution, diffraction effects are directly exploited to perform the measurements.

30.3.1 Geometric Masking

Geometric masking is applied to absolute and incremental encoders. The electromagnetic field associated with the propagation of visible light is characterized by very rapid oscillations (frequencies of the order of 10^{14} s^{-1}). It may therefore be expected that a good first-order approximation to the propagation laws of light is obtained by neglecting the wavelength of light. In this case, diffraction phenomena may be ignored, and light may be thought to propagate in a straight line. Geometry can then be used to analyze the behavior of light. The approximation is valid whenever light rays propagate through an encoder grating with fairly coarse pitch (say, more than 10 μm). Figure 30.11 shows a portion of an encoder scale being scanned by a reading head with multiple slits in order to send more light onto the photosensitive area. The light source is collimated (rays are made parallel to each other after having been emitted in multiple directions from a small source), passes through the stationary grating or index grating, and propagates through the moving grating. In the case of an incremental encoder, a second head would read from the same moving grating but

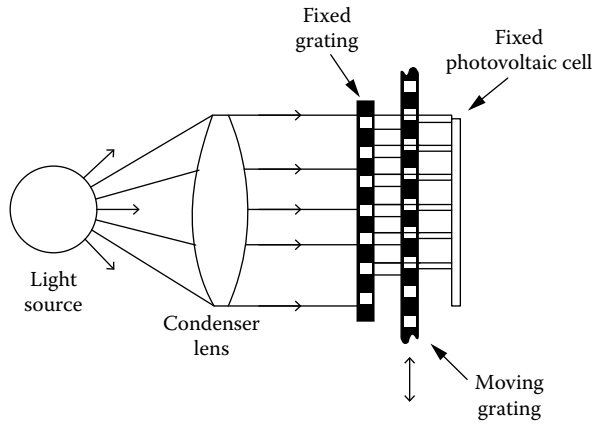


FIGURE 30.11 The reading head contains a light source that can be an incandescent bulb or a light-emitting diode. The light may require a condenser in order to collimate it. This redirects the light rays from the source so that they travel perpendicular to the gratings. The stationary index grating structures the light beam into bands that then cross the moving grating. Depending on the relative position of the fixed and the moving gratings, more or less light reaches the photodetector. The detector produces a signal proportional to the total amount of light on its sensitive surface.

be displaced by n + 1/4 pitch in order to produce a second signal with a phase of 90°. When the slits of the two gratings are aligned, a maximum amount of light reaches the detector. Similarly, a minimum amount of light is transmitted when the gratings are out of phase by 180°. Depending on the design of the grating, such as the duty cycle of the transmitting and opaque sectors, various cyclic signal patterns are obtained. Generally, it is noteworthy that optical encoders can also be used as a reflective as opposed to transmissive design. Also, in some absolute optical encoders, the encoding principle does not rely on a primary grating, per se, or the detection system does not necessarily incorporate any masking element.

30.3.2 Moiré Fringes

Moiré fringe methods are primarily associated with incremental encoders. Moiré fringes are observed when light passes through two similar periodic patterns brought close to each other and with their line pattern nearly parallel. Figure 30.12 shows linear and radial gratings producing Moiré fringes. Take a linear grating with sinusoidal amplitude transmittance as proposed by Gasvik [20]:

$$f(x,y) = a + a\cos\left(\frac{2\pi x}{P}\right) \tag{30.9}$$

where

P is the grating period
a is the amplitude
x is measured perpendicularly to the grating lines

It is also possible to represent a square-wave-type grating using a Fourier series in the case of a radial grating and a Fourier integral for a linear grating. When two linear gratings, a and b, are laid in contact, the resulting transmittance, f_c, is the product of their individual transmittances f_a and f_b:

$$f_c = f_a \times f_b \tag{30.10}$$

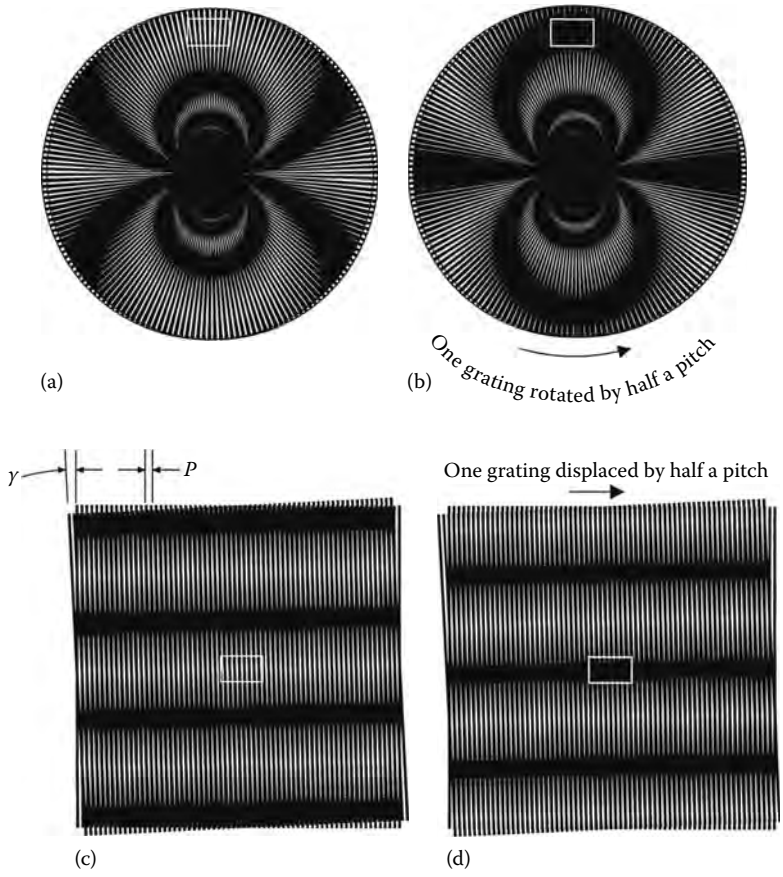


FIGURE 30.12 Moiré fringes produced when two gratings of equal pitch are overimposed. In (a), two radial gratings with 180 cycles have their centers displaced horizontally by approximately half of the grating pitch (as measured on the perimeter). This could be due to encoder-assembly errors. The fringes are always circular and pass by the two centers. In (b), the same displacement was followed by a rotation of half a pitch as a result of the normal movement of the encoder disk. This causes a phase change of 180° of the fringe pattern at any fixed observation point, as represented by the rectangular observation region. The width of the fringes depends on the offset displacement between the two grating centers. If they coincide, the whole pattern changes from a dark to a relatively bright fringe. In (c), two linear gratings are rotated by one-tenth (in radian) of their grating pitch. This could be due to encoder assembly or mounting errors. As a result, Moiré fringes occur at 90° to the grating and with a period 10 times larger. The normal displacement of one grating by half of the grating pitch causes a phase change of 180° of the Moiré fringe pattern at any fixed observation point, as represented by the rectangular observation region. (From Ditchburn, R.W., *Light*, Vol. 1, Academic Press, New York, 1976.)

In the case when a misalignment γ exists between the two gratings, the separation of the Moiré fringe is d with

$$d = \frac{P}{2\sin(\gamma/2)} \tag{30.11}$$

and for small γ,

$$d = \frac{P}{\gamma} \tag{30.12}$$

It follows that when the two gratings are moved relative to each other by a distance x' in the direction perpendicular to their lines, then the phase $\phi(x')$ of the Moiré fringes at a stationary location changes by

$$\phi(x') = \frac{2\pi x'}{P} \tag{30.13}$$

In effect, every time x' equals P, the entire Moiré fringe pattern undergoes one complete cycle of fluctuation. If the two gratings are adjusted to near parallelism, then the fringes are "fluffed out" (see Burch [21]) to an infinite width. Under such conditions, there is no fringe motion to be observed, and the entire region simply goes from brightness to darkness. These effects are mathematically predicted using a convolution integral:

$$f_c(x') = \frac{\int_{x_1}^{x_2} f_a(x) f_b(x - x')}{dx} \tag{30.14}$$

where x_1 to x_2 represents the region covered by the photodetector. If $g_a(k)$ and $g_b(k)$ are the exponential Fourier transforms of f_a *and* f_b, then the photodetector response f_c will have the Fourier transform $g_c(k)$:

$$g_c(k) = g_a(k) g_b(-k) \tag{30.15}$$

Moiré effects can be obtained by transmission or by reflection at the grating.

It must be kept in mind, however, which as the grating pitch reduces, diffraction effects become significant; careful design and adjustment of the gap between the stationary and the moving grating and also lighting and detection arrangements become critical. Since a large number of lines of the gratings are used to generate the photodetector response, small local imperfections of the grating pitch are averaged out, resulting in improved measurement accuracy.

30.3.3 Diffraction-Based Encoders

Diffraction-based encoders are used for the highest levels of precision. They successfully exploit diffraction effects and are referred to as physical optics encoders as opposed to geometrical optics encoders. Figure 30.13 shows diffraction effects observed when coherent light encounters a pattern of slits [20]. The slit is a few wavelengths wide, so strong diffraction effects are observed. Each slit may be thought of as acting as a coherent light source producing circular wave fronts. Along certain directions, the portions of these wave fronts are in phase with each other. These directions are given by

$$\sin\theta_m = \frac{m\lambda}{P} \tag{30.16}$$

where

m is the positive or negative integer and is the order of diffraction
P is the diffraction grating period

Suppose now that the slits are moving by a distance x as shown in Figure 30.13. Then the phase of the wave front at a stationary location will increase for m positive and decrease for m negative by $2\pi mx/P$.

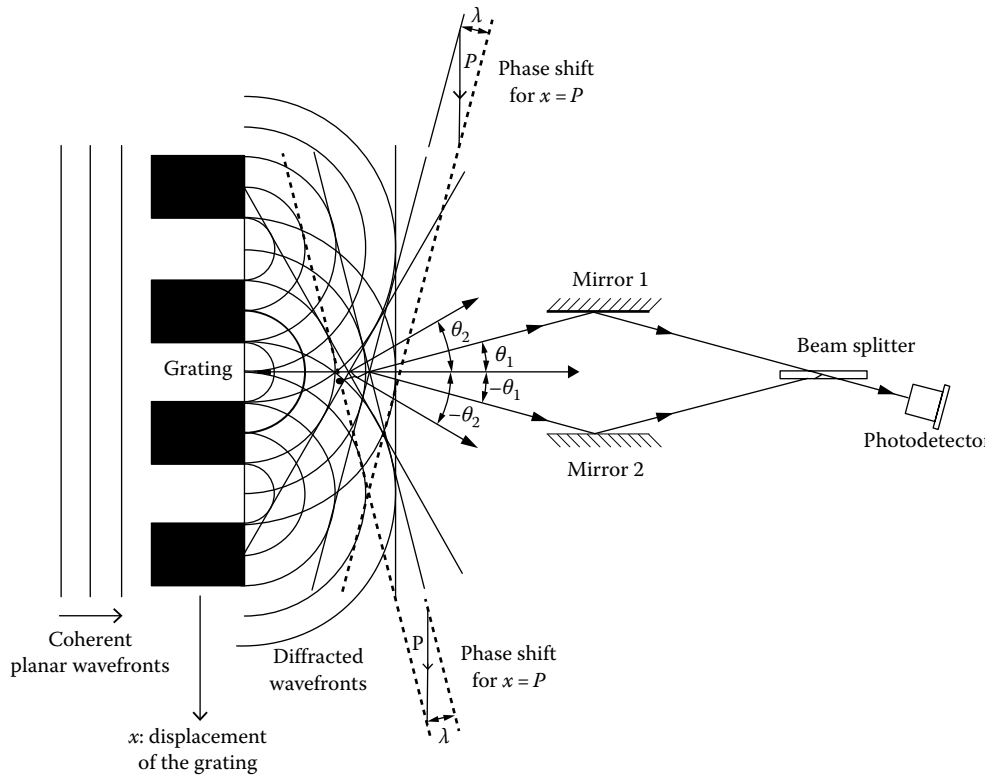


FIGURE 30.13 The grating diffracts the coherent planar wave front coming from the left into a series of circular wave fronts. The coherent circular wave fronts are in phase along certain directions resulting in diffracted planar wave fronts. These planar wave fronts correspond to the common tangents of the circles. For example, the third innermost circle of the uppermost slit is in phase with the fourth innermost circle of the middle slit and also with the fifth innermost circle of the lowest slit. This diffracted wave front corresponds to the first-order diffraction ($m = 1$) and makes an angle θ_1, with the initial planar wave-front direction. A similar order wave front ($m = -1$) has a direction $-\theta_1$. These two wave fronts are initially in phase. A displacement of the grating downward by x causes the $m = 1$ wave front to move by a distance $x\lambda/P$ and the $m = -1$ wave front by $-x\lambda/P$, thus causing a relative phase shift between the two wave fronts of $2x\lambda/P$. This relative phase shift produces a movement of interference fringes at the photodetector.

When the slit pattern has moved by one pattern cycle, then the two wave fronts will have developed a relative phase change of $2m$ cycles. The two phase-shifted wave fronts may be recombined to produce interference fringes.

30.3.4 Rotary Encoders

The Canon laser rotary encoder uses an optical configuration that generates four cycles of interference fringes per cycle of the diffraction grating. This is achieved by splitting the original wave front in two and interrogating the grating at two diametrically opposed locations. Reflecting both diffracted beams back through the grating results in a further doubling of the resolution. The two beams are finally recombined for interference fringe counting. Furthermore, as explained by Nishimura and Ishizuka [22], using diametrically opposed portions of the disk attenuates the effect of eccentricity. An encoder with an external diameter of 36 mm produces 81,000 analog quadrature cycles/revolution. The grating has a pitch of approximately 5 μm and the laser light has a wavelength of 780 nm.

A grating motion by P causes four fringe cycles.

30.3.5 Linear Encoders

Sony markets a laser linear encoder with a volume hologram grating of 0.55 μm pitch [23]. Figure 30.14 gives some insight into the principle of operation of this type of encoder. In theory, for maximum intensity of the reflected light beam through the hologram grating, the path length difference from successive plane mirrors must be equal to λ [20]. The Sony Laserscale encoder uses a semiconductor laser as the coherent light source. The beam splitter produces two beams to interrogate the grating at two separate locations. The beams are then diffracted by the hologram grating, followed by a reflection at the mirrors and pass a second time through the grating. They are finally recombined at the photodetector to produce interference fringes. Because the two beams are diffracted by the grating in opposite directions, and because they pass twice through the grating, four signal cycles are obtained when the grating moves by one pitch.

The Heidenhain Company produces both linear and rotary encoders of very high resolution, also using the principles of diffraction. Their approach, unlike those previously mentioned, uses an index grating and a reflecting scale grating [17,24].

The Renishaw Company uses a fine reflective grating of 20 μm pitch that diffuses the light from an infrared light-emitting diode. In order to avoid the problems caused by diffraction effects, the index grating of a conventional Moiré approach would require a very small gap between the index and

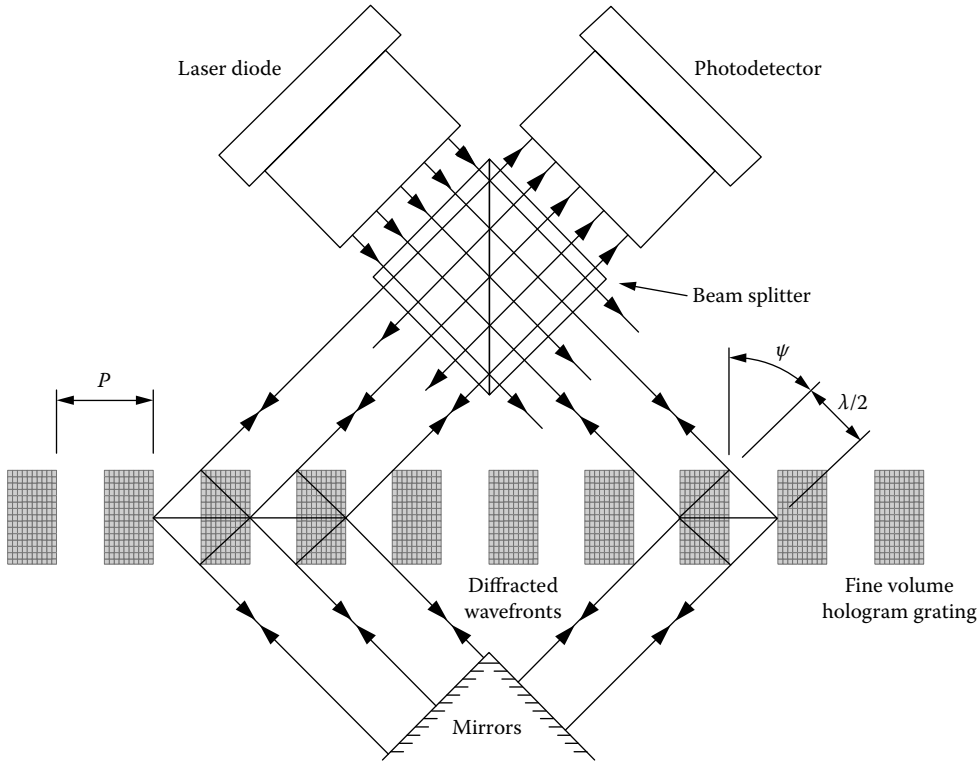


FIGURE 30.14 The coherent planar wave front of the laser diode is split into two at the beam splitter. Both beams can be regarded as being partially reflected at the grating by minute mirrors separated by a pitch, *P*. Maximum reflection is achieved when $2P \sin \psi = \lambda$. Each reflected portion of the incident beam is phase shifted by one λ with respect to that reflected by an adjacent mirror so that the reflected wave front remains coherent. The reflected wave fronts are truly reflected at the orthogonal mirrors for a second pass through the grating, after which they are recombined by the splitter to interfere at the photodetector. A displacement of the grating by x along its axis causes one beam path to be shortened by $2x\lambda/P$, while the other is lengthened by the same amount.

the main gratings. Instead, the index grating is located at a distance of 2.5 mm. The index grating is then able to diffract the diffused light from the main grating and to image it 2.5 mm further. There, a fringe pattern of the Moiré type is produced and photoelectrically analyzed [25].

30.4 Components and Technology

The choice of an encoder demands careful consideration of a number of factors, such as (1) the required resolution, repeatability, and accuracy (linearity); (2) the maximum and minimum operating speeds; (3) the environmental conditions (temperature range, relative humidity (condensing or not condensing), contaminants such as dust, water, and oil); (4) the minimum friction torque (or force) acceptable; (5) the maximum inertia acceptable; (6) the available space; (7) the position to be known immediately after a power loss; (8) the range in degrees (or mm); (9) the mounting and shaft loading; and (10) the price.

Rotary encoders also sometimes employ reflective tapes attached to, or markings directly etched into, the surface of a drum or spindle to be read from the side. In special cases, tapes or engravings of this type can be made to conform to the surface of an elliptical cross-sectional cam or other irregular contour. Cylindrical primary gratings employing transmissive readout have also been produced.

The main gratings (scale) come in a wide variety of materials, both for the substrate and for the marking. Flexible scales based on a metal tape substrate are also available from Renishaw, allowing very long (tens of meters) continuous reading of linear motion. Tables 30.2 through 30.4 give a partial list of encoders currently available on the market. The list is not exhaustive in terms of suppliers and does not cover all the models of the suppliers listed. Table 30.5 gives the address and telephone numbers of the suppliers. Tables 30.6 and 30.7 list some suppliers of quadrature-decoding circuits.

TABLE 30.2 Commercial Optical Rotary Incremental Encoders

Manufacturer	Model No.	Output Type	Counts[a]	Resolution[b]	Price[c]
BEI	H25	Square wave	2,540	50,800	$340
Canon	Kl	Sine wave	81,000	1,296,000	$2,700
Canon	X-1M	Sine wave	225,000	18,000,000	$14,000
Dynamics	25	Sine wave	3,000	60,000	$290
Dynamics	35	Sine wave	9,000	360,000	$1,150
Gurley	911	Square wave	1,800	144,000	$1,300
Gurley	920	Square wave	4,500	144,000	$500
Gurley	835	Square wave	11,250	360,000	$1,500
Heidenhain	ROD 426	Square wave	50–10,000	200–40,000	$370
Heidenhain	ROD 905	Analog	36,000	0.035 arcsec	$12,600
Lucas Ledex	LD20	Square wave	100	—	$180
Lucas Ledex	LD20	Square wave	1,000	—	$195
Lucas Ledex	DG60L	Square wave	5,000	—	$245
Renco	RM21	Square wave	2,048	—	$176
TR Electronic	IE58	Square wave	10,000	—	$428

[a] Number of quadrature cycles per revolution without electronic divide-by-four or interpolation.

[b] Unless otherwise specified, is the number of counts per revolution with electronic interpolation, either internal or external to the encoder, supplied by the manufacturer as an option.

[c] Based on orders of one unit. Must not be used to compare products since many other characteristics, not listed in this table, determine the price.

TABLE 30.3 Commercial Optical Rotary Absolute Encoders

Manufacturer	Model No.	Steps per Turn	No. of Turn	Price[a]
BEI	M25	65,536	1	$2130
BEI	MT40	512	16	$1240
BEI	MT40	65,530	512	$5000
Gurley	25/04S	131,072	1	$1900
Heidenhain	ROC 424	4,096	4096	
Lucas Ledex	AG60E	360 or 512	1	$486
Lucas Ledex	AG661	4,096	4096	$1260
TR Electronic	CE65[b]	8,192	4096	$1408

[a] Based on orders of one unit. Must not be used to compare products since many other characteristics, not listed in this table, determine the price.
[b] Programmable output.

TABLE 30.4 Commercial Optical Linear Incremental Encoders

Manufacturer	Model No.	Output Type	Pitch[a] (μm)	Resolution[b] (μm)	Length (mm)	Price[c] (Length)
Canon	ML-16+	Sine wave	1.6	0.4	To 300	$1525 (50 mm)
Canon	ML-08+	Sine wave	0.8	0.2	To 150	$3100
Gurley	LE18	Square wave	20	0.1	To 1500	$750 (1000 mm)
Gurley	LE25	Square wave	20	0.1	To 3000	$800 (1000 mm)
Heidenhain	LS603	Sine wave	20	5	To 3040	$932 (1020 mm)
Heidenhain	LIP401	Sine wave	2	0.005	To 420	$4000 (100 mm)
Renishaw	RG2	RS422A	20	0.5	To 60,000	$640 + $360/1000 mm
Sony	BS75A-30NS	Square wave	0.14	0.05	30	$2628

[a] Period of the quadrature cycle without electronic divide-by-four or interpolation.
[b] With electronic interpolation supplied by the manufacturer.
[c] Based on orders of one unit. Must not be used to compare products since many other characteristics, not listed in this table, determine the price.

TABLE 30.5 Companies That Make Optical Encoders

BEI Sensors and Motion Systems Company
Encoder Systems Division, 13100 Telfair Avenue, Sylmar, CA. Tel.: (848) 341-6161, http://www.beisensors.com/index.html

Canon USA Inc.
Components Division, New York Headquarters, One Canon Plaza, Lake Success, NY 11042, Tel.: (516) 488-6700, http://www.usa.canon.com/cusa/home

Dr. Johannes Heidenhain GmbH
DR.-Johannes-Heidenhain-Strasse 5, D83301 Traunreut, Deutschland, Tel.: (08669)31-0,

Gurley Precision Instruments Inc.
514 Fulton Street, Troy, NY 12181-0088, Tel.: (518) 272-6300, http://www.gurley.com/

Ledex Products
Lucas Control Systems Products, 801 Scholz Drive, P.O. Box 427, Vandalia, OH 45377-0427, Tel.: (513) 454-2345, http://www.ledex.com/

Renco Encoders Inc.
26 Coromar Drive, Goleta, CA 93117, Tel.: (805) 968-1525, http://www.renco.com/

Renishaw plc,
Transducer Systems Division, Old Town, Wotton-under-Edge, Gloucestershire GL12 7DH, U.K., Tel.: +44 1453 844302, http://www.renishaw.com/en/renishaw-enhancing-efficiency-in-manufacturing-and-healthcare—1030

TR Electronic GmbH
Eglishalde 6, Postfach 1552, D-7218 Trossingen, Germany, Tel.: 0 74 25/228-0, http://www.tr-electronic.de/trgroup/index.html

Sony Magnescale Inc.
Toyo Building, 9-17, Nishigotanda 3-chome, Shinagawa-ku, Tokyo, 141 Japan, Tel.: (03)-3490-9481, http://www.magnescale.com/mgs/index.html

TABLE 30.6 Commercial Digital Quadrature Signal Decoder Circuits

Manufacturer	Model No.	Output	Decoding Factor	Counter	Price
Hewlett-Packard	HCTL-2000	Count	12 bits	x4	$12.75
Hewlett-Packard	HCTL-2016	Count	16 bits	x4	$12.75
Hewlett-Packard	HCTL-2020	Count	16 bits and cascade o/p	x4	$14.55
U.S. Digital Corp.	LS7083	Up and down clock		xl or x4	$3.05
U.S. Digital Corp.	LS7084	Count and direction		xl or x4	$3.60

TABLE 30.7 Companies That Make Divide-by-Four Decoders

U.S. Digital Corporation
3800 N.E. 68th Street, Suite A3, Vancouver, WA 98661-1353, Tel.: (360) 696-2468

Hewlett-Packard Company
Direct Marketing Organization
5301 Stevens Creek Boulevard, P.O. Box 58059, MS 51LSJ, Santa Clara, CA 95052-8059, Tel.: (408) 246-4300

References

1. P. E. Stephens and G. G. Davies, New developments in optical shaft-angle encoder design, *Marconi Rev*., 46(228), 26–42, 1983.
2. P. Sente and H. Buyse, From smart sensors to smart actuators: Application of digital encoders for position and speed measurements in numerical control systems, *Measurement*, 15(1), 25–32, 1995.
3. E. M. Petriu, Absolute-type position transducers using a pseudorandom encoding, *IEEE Trans. Instrum. Meas*., IM-36, 950–955, 1987.
4. E. M. Petriu, Scanning method for absolute pseudorandom position encoders, *Electron. Lett*., 24, 1236–1237, 1988.
5. G. H. Tomlinson, Absolute-type shaft encoder using shift register sequences, *Electron. Lett*., 23, 398–400, 1987.
6. J. N. Ross and P. A. Taylor, Incremental digital position encoder with error detection and correction, *Electron. Lett*., 25, 1436–1437, 1989.
7. M. Arsic and D. Denic, New pseudorandom code reading method applied to position encoders, *Electron. Lett*., 29, 893–894, 1993.
8. B. Arazi, Position recovery using binary sequences, *Electron. Lett*., 20, 61–62, 1984.
9. D. Conner, Long-lived devices offer high resolution, *EDN*, 35(9), 57–64, 1990.
10. J. A. Kuzdrall, Build an error-free encoder interface, *Electron. Design*, September 17, 81–86, 1992.
11. P. Venugopal, Reflective optical SMT module reduces encoder size, *Power Conver. Intell. Motion*, 21(5), 60–62, 1995.
12. S. Holle, Incremental encoder basics, *Sensors*, 7(4), 22–30, 1990.
13. T. Wigmore, Optical shaft encoder from sharp, *Elektor Electron*., 15(169), 60–62, 1989.
14. M. M. Butler, Simplified multiplier improves standard shaft encoder, *Electronics*, November 20, 128–129, 1980.
15. B. Marty, Design a robust quadrature encoder, *Electron. Design*, June 24, 71–72, 74, 76, 1993.
16. O. Benzaid and B. M. Bird, Interpolation techniques for incremental encoders, *Proceedings of 23rd International Conference on Intelligent Motion*, June 22–24, pp. 165–172, 1993.
17. Heidenhain General Catalog, Dr. Johannes Heidenhain GmbH, DR.-Johannes-Heidenhain-Strasse 5, D83301 Traunreut, Deutschland, November 1993, 8.
18. N. Hagiwara, Y. Suzuki, and H. Murase, A method of improving the resolution and accuracy of rotary encoders using a code compensation technique, *IEEE Trans. Instrum. Meas*., 41(1), 98–101, 1992.

19. J. R. R. Mayer, High-resolution of rotary encoder analog quadrature signals, *IEEE Trans. Instrum. Meas.*, 43(3), 494–498, 1994.
20. K. J. Gasvik, *Optical Metrology*, New York: John Wiley & Sons, 1987.
21. J. M. Burch, The metrological applications of diffraction gratings, in E. Wolf (Ed.), *Progress in Optics,* Vol. II, Amsterdam, the Netherlands: North-Holland Publishing, 1963.
22. T. Nishimura and K. Ishizuka (From Canon, Inc., Tokyo), Laser Rotary Encoders, Motion: Official Journal of the Electronic Motion Control Association, September/October 1986, Reprint obtained from Canon.
23. Anonymous from Sony Magnescale America Inc., Hologram technology goes to work, *Mach. Des.*, January 12, 32, 1995.
24. Anonymous from Heidenhain, Encoding systems Vorsprung durch Heidenhain, *Eng. Mater. Des.*, September 1989, 53–54.
25. Jim Henshaw of Renishaw, Linear encoder offers superior flexibility, *Des. Eng.*, September 1995.
26. R. W. Ditchburn, *Light*, Vol. 1, New York: Academic Press, 1976.

Further Readings

Latest Encoder News, BEI Sensors. http://www.beiied.com/news/BEI_industrial-encoders_for-dummies.html, (accessed on June 10, 2012).

Linear and Rotary Encoders, National Instruments. http://www.magnescale.com/mgs/index.html (accessed on March 25, 2012).

O. Solgaard, *Photonic Microsystems: Micro and Nanotechnology Applied to Optical devices and Systems*, Berlin, Germany: Springer Science, 2009.

31

Magnetic Displacement Sensors

31.1 Introduction **31**-1
31.2 Magnetic Field Terminology: Defining Terms **31**-2
31.3 Magnetostrictive Sensors **31**-5
Distributed Impedance Sensors • Magnetoresistive Sensors
31.4 Hall Effect Sensors **31**-9
31.5 Magnetic Encoders **31**-10
31.6 Conclusion **31**-12
References **31**-12
Further Information **31**-12

David S. Nyce
Revolution Sensor Company

31.1 Introduction

Several types of linear and angular displacement measuring devices rely on electromagnetic fields, and the magnetic properties of materials, in the operation of their basic sensing elements. Some may not commonly be referred to as magnetic sensors but are instead named according to their specific sensing technique. The magnetic sensors presented here use either a permanent magnet, an AC- or DC-powered electromagnet, or a coiled conductor. Together with various components and materials used to sense the magnetic field, the combination is arranged to obtain a response indicating angular or linear displacement or position. Either the sensor is caused to operate by a magnetic field or the properties of the sensor are derived from the use of a magnetic field. Types of magnetic sensors presented in this section include magnetostrictive, distributed impedance, magnetoresistive, Hall effect, and magnetic encoders. Some versions of synchro/resolvers and related sensors loosely meet the characteristics listed earlier but are included in another section of this handbook and thus will not be included in this section. Linear variable differential transformers (LVDTs) and inductive proximity sensors are types of magnetic displacement sensors, but measure displacement over a limited range, and are also covered in another section of this handbook.

An important aspect of magnetic sensors is that they utilize a noncontact sensing element. There is no mechanical connection or linkage between the stationary members and the movable members of the sensor. In some devices that sense a position magnet or core, the sensor can even be designed to allow removal of the magnet or core from the sensitive element, when readings are not required. Noncontact implies that the lifetime of the sensing element is not limited to a finite number of cycles by friction-induced wear. This is important in some industrial machinery and in many applications requiring high reliability. All of the sensors presented in this chapter utilize noncontact sensing techniques.

Displacement refers to a change in position, or the amount by which an object has been displaced, rather than to a single, absolute position. In common industrial practice and by many sensor manufacturers, however, the terms displacement sensor and position sensor are commonly confused, and displacement sensors are typically labeled as either incremental or absolute. Nonetheless, a displacement

sensor indicates the amount of change between the present location and a previous location by providing a magnitude and direction of change with respect to the previously reported location. For example, with a rotation sensor, the indication may be a number of counts clockwise from the previously recorded location, where the number of counts is the magnitude and clockwise is the direction. If the information that describes the current location is lost, due to power loss or other disturbance, the system must be reset, re-zeroed, or "homed." During the reset, the sensor must be in a reference, or zero, location or can be moved to a "homing" position where the absolute location is known. Magnetic encoders can be designed as either incremental or absolute reading (i.e., displacement or position, respectively). Optical encoders, inductosyns, and synchro/resolvers are types of displacement and position sensors that can be designed as either incremental or absolute reading but are covered in other chapters of this handbook.

Most sensors described in this section are absolute reading. They supply a reading of distance or angle from a fixed datum, rather than from a previous position. Consecutive readings can be subtracted to give an incremental indication, thus emulating a displacement sensor. Such emulation can be useful when replacing an incremental encoder in the field with an absolute-reading type of sensor. An absolute sensor indicates the current position without the need for knowledge of the previous position. It never needs to be reset to a reference location or homed in order to derive the measured location. As mentioned earlier, such absolute-reading sensors are called *position sensors*.

Magnetic sensor types will be described here based on the technology employed, rather than the application. Relative usefulness for making linear or angular measurements will be indicated for each type of sensor.

Noncontact magnetic sensor technology described here for position or displacement measurement includes magnetostrictive, distributed impedance, magnetoresistive, Hall effect, and magnetic encoders.

31.2 Magnetic Field Terminology: Defining Terms

Magnetic field intensity (H) or magnetizing force: The force that drives the generation of magnetic flux in a material. H is measured in A m^{-1} (where A is for amperes and m is for meters).

Magnetic flux density (B): The amount of magnetic flux resulting from the applied magnetizing force. B is measured in N/(A · m) (where N is for Newton).

Magnetic permeability (μ): The ability of a material to support magnetic lines of flux. The μ of a material (pronounced myoo) is the product of the relative permeability of that material and the permeability of free space. The relative permeability of most nonferrous materials is near unity. In free space, magnetic flux density is related to magnetic field intensity by the formula:

$$B = \mu_0 H \tag{31.1}$$

where μ_0 is the permeability of free space, having the value $4\pi \times 10^{-7}$ H m^{-1} (where H is for henries). In other materials, the magnetic flux density at a point is related to the magnetic intensity at the same point by

$$B = \mu H \tag{31.2}$$

where

$$\mu = \mu_0 \mu_r \tag{31.3}$$

and μ_r is the relative permeability [1].

See Table 31.1 for a list of materials with approximate relative magnetic permeabilities.

Hysteresis: A phenomenon in which the state of a system does not reversibly follow changes in an external parameter [2]. In a displacement sensor, it is the difference in output readings obtained at a given point when approaching that point repeatedly from upscale and downscale readings. Figure 31.1 is a typical output versus input graph.

TABLE 31.1 Various Materials and Their Relative Permeabilities

Material	Approximate Relative Permeability
Vacuum	1
Air	1
Water	1
Aluminum	1
Copper	1
316 stainless steel (annealed)	1
Cold rolled steel	180
Nickel	600
Iron	5,000
Permalloy	8,000
Mu metal	20,000

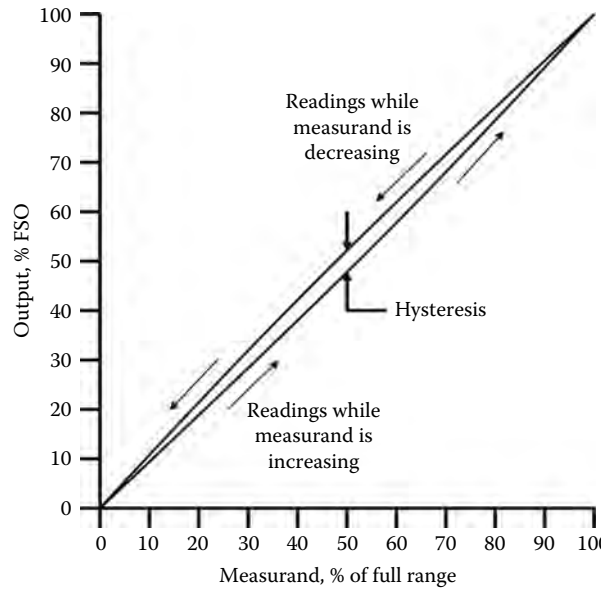


FIGURE 31.1 Hysteresis: output versus input. FSO, full scale output.

Magnetic hysteresis: Depicted in the hysteresis loop, Figure 31.2. When a ferromagnetic material is placed in an alternating magnetic field, the flux density (B) lags behind the magnetizing force (H) that causes it. The area under the hysteresis loop is the hysteresis loss per cycle and is high for permanent magnets and low for high permeability, low-loss magnetic materials [3].

Magnetic saturation: The upper limit to the ability of ferromagnetic materials to carry magnetic flux.

Magnetization curve: Shows the amount of magnetizing force needed for a ferromagnetic material to become saturated. It is a graph with B as the ordinate and H as the abscissa (also known as the B–H curve). A magnetization curve for a specific material would look the same as Figure 31.2, with the addition of calibration marks and the curve adjusted to describe the characteristic of that material.

Permittivity: Is a measure of the resistance encountered regarding the formation of an electric field in a medium or how much an electric field affects a dielectric medium (a dielectric material is a nonconductor of electricity). Permittivity quantifies the degree to which a material is able to "permit" an

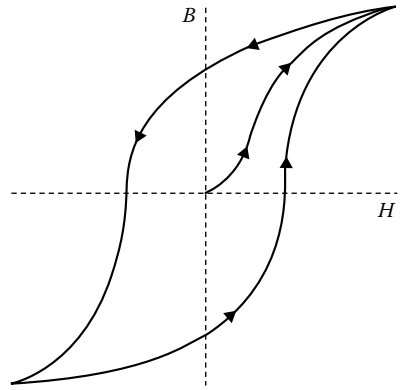


FIGURE 31.2 Magnetic hysteresis.

electric field to be formed, rather than polarizing the material (polarizing a material would reduce the magnitude of electric field to be formed within the material). Whereas permeability is measured in henries per meter, absolute permittivity, ε_0, is measured in farads per meter. In many calculations, relative permittivity, ε_r, is used, which assumes that the relative permittivity of vacuum is 1. Then the relative permittivity of a material is the ratio of its permittivity to that of vacuum. Air has a relative permittivity of approximately 1.

See Table 31.2 for a list of materials with approximate relative magnetic permeabilities.

Field propagation velocities: The higher the relative permeability of a material, the slower a magnetic field will travel, and the larger will be the inductance of a coil wound on this material, having a given size and turns count [4]. Likewise, the higher the relative permittivity of a material, the slower an electric field will travel, and the larger will be the capacitance of a given size and spacing of parallel plates [4]. The absolute speed, c_0, of propagation of an electromagnetic wave is given by

$$c_0 = \frac{1}{\left(\varepsilon_0 \mu_0\right)^{1/2}} \tag{31.4}$$

This relationship also describes the speed of light.

TABLE 31.2 Various Materials and Their Relative Permittivities

Material	Approximate Relative Permittivity
Vacuum	1
Air	1
Water	1
Gasoline	2
Methanol	30
Ethylene glycol	37
Paper	3.5
PTFE	2.1
Polyethylene	2.3
Polypropylene	2.3
Concrete	4.5

31.3 Magnetostrictive Sensors

A magnetostrictive sensor is commonly used to measure displacement, although fundamentally being an absolute position sensor. It uses a ferromagnetic element to detect the location of a position magnet that is displaced along its length. The position magnet is attached to a member whose position is to be sensed, and the sensor body remains stationary (see Figure 31.3). The position magnet moves along the measuring length without contacting the sensing element.

Ferromagnetic materials such as iron and nickel display the property called *magnetostriction*. Application of a magnetic field to these materials causes a strain in the crystal structure, resulting in a change in the size and shape of the material. A material exhibiting positive magnetostriction will expand when magnetized. Conversely, with negative magnetostriction, the material contracts when magnetized [5].

The ferromagnetic materials used in magnetostrictive displacement sensors are transition metals, such as iron, nickel, and cobalt. In these metals, the 3*d* electron shell is not completely filled, which allows the formation of a magnetic moment (i.e., the shells closer to the nucleus are complete, and they do not contribute to the magnetic moment). As electron spins are rotated by a magnetic field, coupling between the electron spin and the electron orbit causes electron energies to change. The crystal strains so that electrons at the surface can relax to states of lower energy [6].

This physical response of a ferromagnetic material is due to the presence of magnetic moments and can be understood by considering the material as a collection of tiny permanent magnets, called *domains*. Each domain consists of many atoms. When a material is not magnetized, the domains are randomly arranged. However, when the material is magnetized, the domains are oriented with their axes approximately parallel to each other. Interaction of an external magnetic field with the domains causes the magnetostrictive effect (see Figure 31.4). This effect can be optimized by controlling the ordering of domains through alloy selection, thermal annealing, cold working, and magnetic field strength.

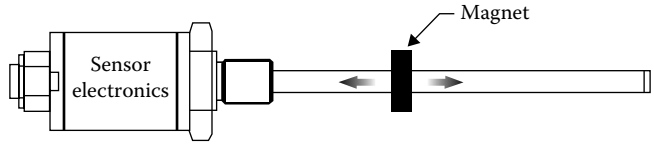


FIGURE 31.3 Magnetostrictive sensor with position magnet.

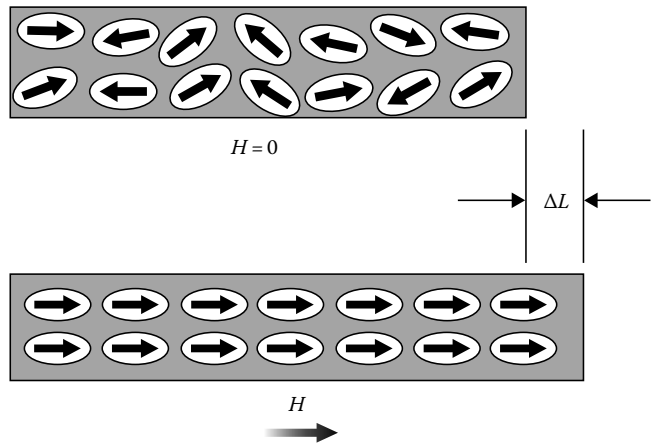


FIGURE 31.4 Magnetic domains: alignment with magnetic field, "*H*," causes dimensional changes.

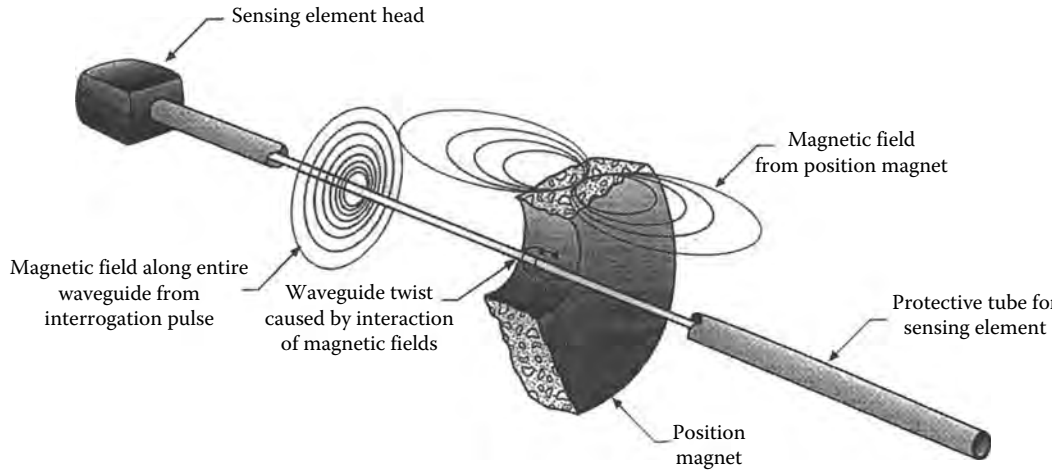


FIGURE 31.5 Operation of magnetostrictive position sensor.

While application of a magnetic field causes the physical strain, as described earlier, the reverse is also true: exerting stress causes the magnetic properties (i.e., permeability, susceptibility) to change. This is called the *Villari effect*.

In magnetostrictive sensors, uniform distortions of length, as shown in Figure 31.4, offer limited usefulness. Usually, the magnetization is rotated with a small field to induce a local distortion, described as the *Wiedemann effect*. This is a mechanical torsion that occurs at a point along a magnetostrictive wire or tube when an electric current is passed through the wire or tube while it is subjected to an axial magnetic field. The torsion occurs at the location of the axial magnetic field, which is usually provided by a small permanent magnet called the position magnet.

In a position (or displacement) sensor, a ferromagnetic wire or tube, called the waveguide, is used as the sensing element (see Figure 31.5). The sensor measures the distance between the position magnet and the pickup. To start a measurement, a current pulse I (called the interrogation pulse) is applied to the waveguide. This causes a magnetic field to instantly surround the waveguide along its full length.

In a magnetostrictive position sensor, the current is a pulse of approximately 1–2 μs duration. A torsional mechanical wave is formed at the location of the position magnet due to the Wiedemann effect. Portions of this wave travel from the position magnet both toward and away from the pickup. The wave traveling along the waveguide toward the pickup is detected when it arrives at the pickup. The time measurement between application of the current pulse (forming of the torsional wave at the position magnet) and its detection by the pickup is indicative of the distance between the position magnet and the pickup. The speed of the wave is typically in a range of about 2500–3000 m s^{-1}. The portion of the wave traveling away from the pickup could act as an interfering signal after it is reflected from the waveguide tip. So instead, it is damped by a damping element when it reaches the end of the waveguide opposite the pickup. Damping is usually accomplished by attaching one of various configurations of elastomeric materials to the end of the waveguide. The end of the waveguide within the damping element is unusable for position determination and therefore called the "dead zone." In cases where the waveguide is interrogated infrequently, the reflected wave is eventually attenuated without using a damping element, and therefore a damping element may not be needed.

The time measurement can be buffered and used directly as the sensor output, or it can be conditioned inside the sensor to provide various output types, including analog voltage or current, pulse width modulation, controller area network (CAN) bus, synchronous serial interface (SSI), highway addressable remote transducer (HART), and PROFIBUS. Magnetostrictive position sensors can be made as short as 1 cm long or up to more than 30 m long. Resolution of some magnetostrictive position sensors can be as fine as 1 μm. Temperature coefficients of 2–5 ppm $°C^{-1}$ can be achieved, depending on the material selected for the waveguide. The sensors are inherently stable, since the

measurement mainly relies on the physical properties of the waveguide material, rather than the stability of electronic circuits. Longer sensors become very cost effective because the same electronics package can drive sensors of varying length; only the waveguide and its packaging are increased in length to make the sensor longer, and sometimes the waveguide drive pulse may be modified.

The magnetostrictive wire can be straight for a linear sensor or shaped to provide curved or rotary measurements. Curved sensors are sometimes used to measure angular or nonlinear motion in industrial applications, although rotary magnetostrictive sensors are not yet very popular.

31.3.1 Distributed Impedance Sensors

Inductive sensors have long been used, in which a magnetically permeable target is allowed to move in the vicinity of a coil and in which the coil inductance varies with movement of the target. This type of sensor and the related LVDT have been included in another chapter of this handbook. A new type of magnetic sensor that responds to changes in an electromagnetic field is called the distributed impedance sensor. In this patented technology, a linear or rotary sensor is formed of a sensing element and driving electronics. The sensing element comprises inductance, capacitance, and resistance, all of which are distributed along the length of the sensing element, rather than treated as a lump inductance as would be done with an inductive sensor (in an inductive sensor, the capacitance is considered to be negligible) or a lump capacitance as would be done with a capacitive sensor (in which the inductance would be negligible). A configuration of the sensing element that is preferred for high performance and high manufacturability is that of a flat conductor wrapped in a helical pattern around a temperature stable rod or tube. Two such helices form an approximate double helix, with the helices wound in opposite phase from one another. This configuration can be seen in Figure 31.6, a photo of a typical position sensor.

The sensing element is driven in antiphase mode by an oscillator circuit operating at a frequency where the capacitive reactance is nonnegligible as compared with the inductive reactance. A surface wave travels along the sensing element, providing a time delay depending upon the position of a target tube. The target tube is formed of an electrically conductive material and covers various portions of the sensing element, depending on the position to be measured. The time delay can be directly measured or can be converted into a frequency that varies with position, thereby providing a signal that is simple to interface with digital circuits.

Separating the series inductances along a sensing axis by parallel capacitances provides distributed impedance whereby both the inductances and capacitances are distributed along the length of the sensing axis. This enables shaping of the electromagnetic field along the sensing element so that the magnetic field penetrates into a measured material (such as the target tube) but is concentrated in a volume close to the sensing element in order to avoid sensing unwanted conductive materials that may be nearby [7].

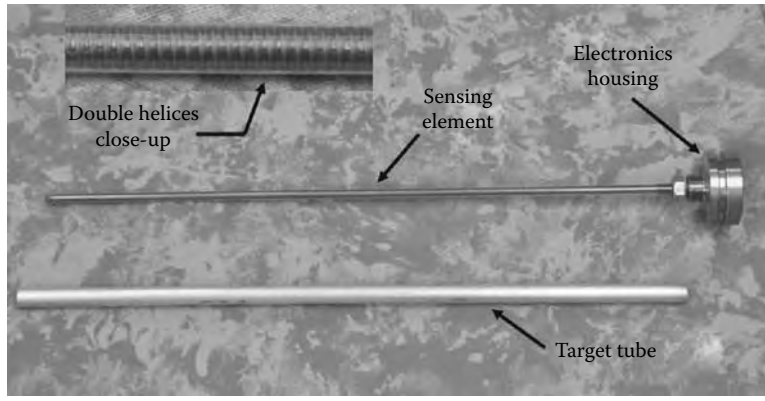


FIGURE 31.6 A distributed impedance sensor with target tube and a close-up view of the helices that are inside of the sensing element. (Photo courtesy of Nyce, D.S., Revolution Sensor Company, Apex, NC.)

The target tube can be formed of any electrically conductive material and does not need to be of high magnetic permeability. Because of this, a tube, which is already a part of the measured assembly, can serve as the target, such as part of the rod or cylinder in a suspension strut or hydraulic cylinder. This means that the target does not need to be changed to nonmagnetic stainless steel, as is needed with a magnetostrictive sensor and as sometimes needed with other magnetic or inductive sensors. It also does not require a position magnet as with a magnetostrictive sensor.

This technology has been trademarked as NyceWave™ by the Revolution Sensor Company. Typical sensors are available with analog and digital outputs including voltages (0–5 V, 0.5–4.5 V, 1–5 V, 0–10 V, etc.), currents (0–20 mA, 4–20 mA), SSI, and CAN bus.

The solid construction of a distributed impedance position sensor provides a very robust product that can operate over a wide temperature range. Typical industrial products can operate over a range of −40 °C to 150 °C, including the sensing element as well as the electronics head. This positions the distributed impedance sensor as a technology that is well suited for high-reliability applications, such as in aircraft, space vehicles, vehicle suspensions, mobile hydraulics, and others. Rotary sensors are also available.

31.3.2 Magnetoresistive Sensors

In most magnetic materials, electrical resistance decreases when a magnetic field is applied and the magnetization is *perpendicular* to the current flow (a current will be flowing any time electrical resistance is measured) (see Figure 31.7). The resistance decreases as the magnetic flux density increases, until the material reaches magnetic saturation.

The rate of resistance decrease is less as the magnetic material nears saturation. The amount of resistance change is on the order of about 1% at room temperature (0.3% in iron, 2% in nickel). When the magnetic field is *parallel* to the current, the resistance increases with increasing magnetic field strength. Sensitivity is greatest when the magnetic field is perpendicular to the current flow. These are properties of the phenomenon called *magnetoresistance* (MR). The MR effect is due to the combination of two component parts. These are a reduction in forward carrier velocity as a result of the carriers being forced to move sideways as well as forward and a reduction in the effective cross-sectional area of the conductor as a result of the carriers being crowded to one side [8].

When a position magnet is brought close to a single MR sensing element, the resistance change is at its maximum as the magnet passes over the approximate center of the element and then reduces until the magnet is past the element. The resistance changes according to

$$\text{Resistivity} = \frac{\text{Voltage}}{(\text{Carrier density} \times \text{Carrier velocity})} \tag{31.5}$$

By using multiple MR elements arranged along a line, a longer displacement measuring device can be fashioned. The signals from the string of sensors are decoded to find which elements are being affected by the magnet. Then the individual readings are used to determine the magnet position more precisely. Relatively high-performance sensors can be manufactured. Temperature sensitivity of the MR elements needs to be compensated, and longer sensors contain many individual sensing elements. Because of this, longer sensors become more difficult to manufacture and are expensive.

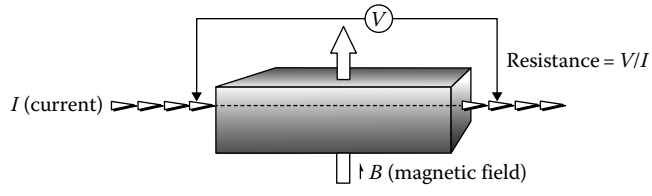


FIGURE 31.7 Magnetoresistance.

Anisotropic MR materials are capable of resistance changes in the range of 1%–2%. The MR of a conductor body can be increased by making it a composite of two or more layers of materials having different levels of MR. Multilayered structures of exotic materials (sometimes more than 10 layers) have enabled development of materials that exhibit much greater magnetoresistive effect and saturate at larger applied fields. This has been named giant magnetoresistance (GMR). Some commercial sensors based on GMR are currently available. The GMR elements can be arranged in a four-element bridge connection for greater sensitivity. In this arrangement, two of the elements are shielded from the applied magnetic field. The other two elements are sensitive to the applied field. Sensitivity can also be increased by incorporating flux concentrators on the sensitive elements. In a bridge connection, the output voltage can vary by more than 5% of the supply voltage [9]. Rotary sensors can be constructed by attaching a pole piece to a rotating shaft. One or more permanent magnets and the pole piece are arranged to cause the magnetic field around the MR element to change with angular displacement.

Further research is being conducted on MR materials to improve the sensitivity by lowering the strength of magnetic field needed and increasing the amount of resistance change. The next higher level of MR performance is called colossal magnetoresistance (CMR). CMR is not yet practical for industrial sensors because of severe limitations on the operating temperature range.

Although MR, GMR, and CMR are limited for use in displacement sensors at this time by cost, temperature, and fabrication constraints, much research is in progress. Maybe humongous magnetoresistance (HMR) is next?

31.4 Hall Effect Sensors

The Hall effect is a property exhibited in a conductor affected by a magnetic field. A voltage potential V_H, called the Hall voltage, appears across the conductor when a magnetic field is applied at right angles to the current flow. Its direction is perpendicular to both the magnetic field and the current. The magnitude of the Hall voltage is proportional to both the magnetic flux density and the current. The magnetic field causes a gradient of carrier concentration across the conductor. The larger number of carriers on one side of the conductor, compared to the other side, causes the voltage potential V_H. A pictorial representation is shown in Figure 31.8. The amplitude of the voltage varies with the current and magnetic field according to [10]

$$V_H = K_H\left(\frac{\beta I}{z}\right) \tag{31.6}$$

where

V_H is the Hall voltage
K_H is the Hall constant
β is the magnetic flux density
I is the current flowing through the conductor
z is the thickness of the conductor

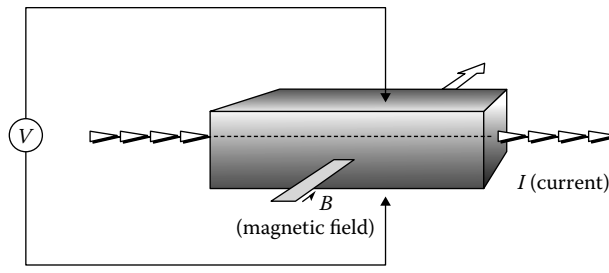


FIGURE 31.8 Hall effect.

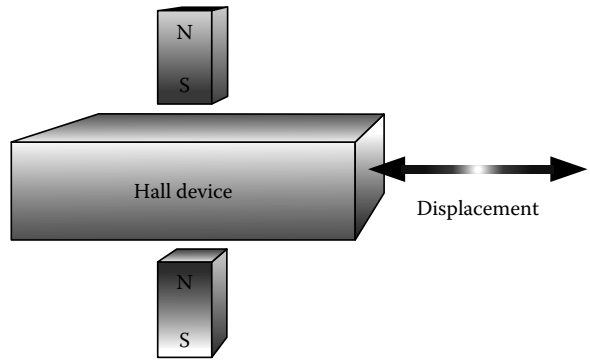


FIGURE 31.9 Two-magnet Hall effect sensors.

Sensors utilizing the Hall effect typically are constructed of semiconductor material, giving the advantage of allowing conditioning electronics to be deposited right onto the same material. Either *p*- or *n*-type semiconductor material can be used, with the associated polarity of current flow. The greatest output is achieved with a large Hall constant, which requires high carrier mobility. Low resistivity will limit thermal noise voltage, for a more useful signal-to-noise ratio (SNR). These conditions are optimized using an *n*-type semiconductor [7].

A displacement sensor can be made with a Hall sensing element and a movable magnet, with an output proportional to the distance between the two. Two magnets can be arranged with one Hall sensor as in Figure 31.9 to yield a near-zero field intensity when the sensor is equidistant between the magnets.

These single Hall effect device configurations have a very limited linear range. Longer range displacement sensors can be built by using multiple Hall sensors spaced along a carrier substrate. A magnet is moved along in close proximity to the carrier. As the magnet approaches and then moves away from each Hall element, the respective sensors will have increasing or decreasing outputs, respectively. The output from the battery of sensors is derived by reading the individual outputs of the sensors closest to the magnet and also decoding the locations of those particular sensors being read. This method can produce relatively high-performance displacement sensors of up to several meters long. Longer sensors become increasingly more difficult to produce and are expensive because of the large number of sensors being multiplexed. Rotary, as well as linear, sensors can be produced by mechanical arrangement of the sensing elements to cause magnetic field variation with the desired angular or linear input.

31.5 Magnetic Encoders

Magnetic encoders use a strip or disk of magnetic media onto which digital information is stored. This information is recorded at the location it describes and is in the form of a collection of magnetized and nonmagnetized areas. A magnetic encoder includes this sensing element, as well as one or more read heads, electronics, and a mechanical enclosure with input shaft and bushings. The input shaft moves in and out for a linear sensor. It has wipes to prevent ingestion of foreign material and bushings designed to accept sideloading. An angular sensor has a shaft that rotates and includes bushings to withstand thrust and sideloading. The encoded media is implemented as either a strip in a linear sensor or as a disk in an angular sensor.

As a read head passes above the encoded area, it picks up the magnetic variations and reads the position information. The information, digital ones and zeroes, will usually be encoded in several parallel tracks to represent the binary digits of the position information. A standard binary code presents a problem for encoders in that some numbers require the changing of several of the bits at one time to indicate a single increment of the number represented. If all the changing bits are not perfectly aligned

TABLE 31.3 Gray Code, as Used in Some Encoders

$Base_{10}$ Number	"Natural" Binary	Gray Code	Binary Coded Decimal (BCD) Tens Units	Hexadecimal (HEX)
0	0000	0000	0000 0000	0
1	0001	0001	0000 0001	1
2	0010	0011	0000 0010	2
3	0011	0010	0000 0011	3
4	0100	0110	0000 0100	4
5	0101	0111	0000 0101	5
6	0110	0101	0000 0110	6
7	0111	0100	0000 0111	7
8	1000	1100	0000 1000	8
9	1001	1101	0000 1001	9
10	1010	1111	0001 0000	A
11	1011	1110	0001 0001	B
12	1100	1010	0001 0010	C
13	1101	1011	0001 0011	D
14	1110	1001	0001 0100	E
15	1111	1000	0001 0101	F

with each other, instantaneous erroneous readings will result. To avoid this problem, a special adaptation of the binary code called "Gray code" is used (see Table 31.3). A single increment of the number represented causes a change of only 1 bit in the Gray code.

The read head incorporates a ferromagnetic core wound with input and output windings. A read pulse is applied to the input winding, and information is read on the output winding. If the core is above a magnetized area of the magnetic media, the core becomes saturated, no output pulse is generated, and a logic 0 results [11]. If the core is above a nonmagnetized area when the read pulse is applied, an output pulse occurs and produces a logic 1. Another arrangement that is practical for angular, but not linear, encoders uses a ring-shaped multipole permanent magnet. The magnet is rotated past a pair of sensors to yield an incremental reading with sine and cosine outputs (called "quadrature" output, see Figure 31.10).

The waveforms can be square, sinusoidal, or triangular. *A* and *B* outputs are used to indicate the displacement and the direction of the displacement. The number of transitions or "counts" is proportional to the displacement magnitude. The direction of displacement (i.e., + or –) can be found by comparing the two phases. For example, in Figure 31.10, at the time the *A* changes from a logic 0 to a logic 1, the status of the *B* will indicate the direction of travel. A logic 0 on the *B* could equal the positive direction; a logic 1 could equal the negative direction.

Magnetic encoders can be incremental or absolute (i.e., indicating displacement or position). In an incremental configuration, equally spaced pulses encoded on the magnetic media are read from one or more tracks. The pulses are collected by an up/down counter, and the counter output

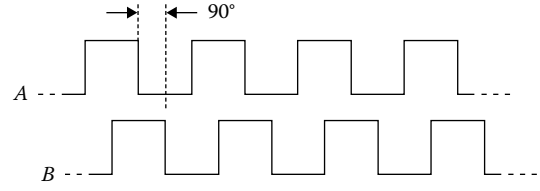


FIGURE 31.10 Quadrature output.

TABLE 31.4 Sensors and Manufacturers

Technology	Manufacturers	Description	Price
Magnetostrictive	MTS Systems Corp., Balluff Inc.	Lengths to 20 m, 2 μm resolution, analog/digital outputs	$100–$3,000
Magnetoresistive	Nonvolatile Electronics, Midori America	Linear GMR, rotary MR, linear MR up to 30 mm	<$10 up to $500
Distributed impedance	LRT Sensors, Alliance Sensors	Lengths to 2 m, infinite resolution, inaccuracy to 0.02%	$50–$350
Hall effect	Optec Technology, Spectec	Standard and custom sensors	Approximately $10–$100
Magnetic encoder	Heidenhain, Sony Precision Technology	Rotary and linear encoders	$100–$2,000

represents the position. Quadrature outputs can be coded to tell the direction of displacement, as described earlier. The zero position is set by resetting the counter.

Absolute magnetic encoders have the digital code representing the position encoded directly at that position. No counter is needed. The Gray code can be interpreted to yield the position in engineering units. Nonlinear coding, such as sine or cosine, is sometimes used.

31.6 Conclusion

Some technologies, such as magnetostrictive, can provide very low inaccuracies over a long measurement range. Other technologies, such as distributed impedance, can provide a very high level of reliability and resistance to shock and vibration. The cost of sensors varies widely by technology and application. Accordingly, several sensor technologies should be considered for a given application, before choosing the best overall match combining the requirements for performance, reliability, and cost. Table 31.4 provides a list of sensors and some manufacturers.

References

1. O. Esbach, *Handbook of Engineering Fundamentals*, New York: John Wiley & Sons, 1975, p. 957.
2. R. Lerner and G. Trigg, *Encyclopedia of Physics*, New York: VCH Publishers, 1990, p. 529.
3. P. Neelakanta, *Handbook of Electromagnetic Materials*, Boca Raton, FL: CRC Press, 1995, p. 333.
4. D. S. Nyce, *Linear Position Sensors, Theory & Application*, Hoboken, NJ: John Wiley & Sons, 2004, pp. 66, 84.
5. D. S. Nyce, Magnetostriction-based linear position sensors, *Sensors Magazine*, 11(4), 22, 1994.
6. R. Philippe, *Electrical and Magnetic Properties of Materials*, Norwood, MA: Artech House, 1988.
7. H. Burke, *Handbook of Magnetic Phenomena*, New York: Van Nostrand Reinhold, 1986.
8. D. S. Nyce, *Distributed Impedance Sensor*, U.S. Patent # 7,340,951, March 11, 2008.
9. Nonvolatile Electronics Inc. NVSB series datasheet. March 1996.
10. J. R. Carstens, *Electrical Sensors and Transducers*, Englewood Cliffs, NJ: Regents/Prentice-Hall, 1992, p. 125.
11. H. Norton, *Handbook of Transducers*, Englewood Cliffs, NJ: Prentice-Hall, 1989, pp. 106–112.

Further Information

Askeland, D., *The Science and Engineering of Materials*, Boston, MA: PWS-Kent Publishing, 1989.
Boll, R., *Soft Magnetic Materials*, London, U.K.: Heyden & Son, 1977.
Craik, D., *Magnetism Principles and Applications*, New York: John Wiley & Sons, 1995.
Cullity, B. D., *Introduction to Magnetic Materials*, Reading, MA: Addison-Wesley, 1972.
Herceg, E., *Handbook of Measurement and Control*, Pennsauken, NJ: Schaevitz Engineering, 1976.

Jiles, D., *Introduction to Magnetism and Magnetic Materials*, London, U.K.: Chapman and Hall, 1991.
Lorrain, P. and Corson, D., *Electromagnetic Fields and Waves*, San Francisco, CA: WH. Freeman, 1962.
Mazda, F., *Electronics Engineer's Reference Book*, 6th edn., London, U.K.: Butterworth, 1989.
Nyce, D., Magnetostrictive Linear Position Sensors, *Sensors Magazine*, November 1999, also available at http://archives.sensorsmag.com/articles/1199/38_1199/index.htm (accessed on March 4, 2013).
Olson, H., *Dynamical Analogies*, New York: D. Van Nostrand, 1943.
Rose, R., Shepard, L., and Wulff, J., *The Structure and Properties of Materials*, New York: John Wiley & Sons, 1966.
Shackelford, J., *Introduction to Materials Science for Engineers*, New York: Macmillan, 1985.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

32

Synchro/Resolver Displacement Sensors

32.1 Induction Potentiometers **32**-1
32.2 Resolvers **32**-2
32.3 Operating Parameters and Specifications for Resolvers **32**-5
32.4 Equipment Needed for Testing Resolvers **32**-5
32.5 Resolver Benefits **32**-6
Multispeed Units • Applications
32.6 Synchros **32**-7
32.7 Modular Solution **32**-7
32.8 Sensible Design Alternative for Shaft Angle Encoding **32**-8
Resolver-to-Digital Converters • Closed-Loop Feedback
32.9 Type II Servo Loop **32**-9
32.10 Applications **32**-10
Resolver-to-Digital Conversion • Bandwidth Optimization • Encoder Emulation • Determining Position Lag Error Due to Acceleration • Large Step Settling Time
Further Readings **32**-14

Robert M. Hyatt, Jr.
Howell Electric Motors

David Dayton
ILC Data Device Corp.

Most electromagnetic position transducers are based on transformer technology. Transformers work by exciting the primary winding with a continuously changing voltage and inducing a voltage in the secondary winding by subjecting it to the changing magnetic field set up by the primary. They are ac-only devices, which make all electromagnetically coupled position sensors ac transformer coupled. They are inductive by nature, consisting of wound coils. By varying the amount of coupling from the primary (excited) winding of a transformer to the secondary (coupled) winding with respect to either linear or rotary displacement, an analog signal can be generated that represents the displacement. This coupling variation is accomplished by moving either one of the windings or a core element that provides a flux path between the two windings.

One of the simplest forms of electromagnetic position transducers is the LVDT, which is described in the section on inductive sensors. If the displacement of the core in an LVDT-type unit is changed from linear to rotary, the device becomes an RVDT (rotary variable differential transformer).

32.1 Induction Potentiometers

The component designer can "boost" the output, increase accuracy, and achieve a slightly greater angular range if windings are placed on the rotor as shown in Figure 32.1. The disadvantages to this method are (1) additional windings, (2) more physical space required, (3) greater variations over temperature, and (4) greater phase shift due to the additional windings.

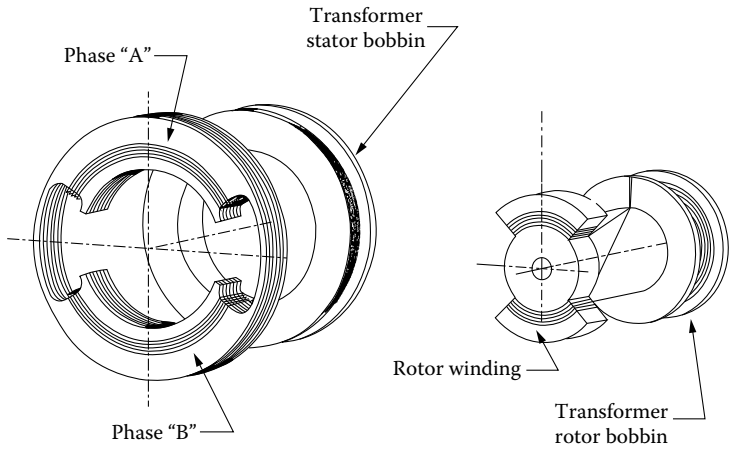


FIGURE 32.1 The induction potentiometer has windings on the rotor and the stator.

The advantage of the induction pot design is greater sensitivity (more volts per degree), resulting in better signal to noise and higher accuracy in most cases.

32.2 Resolvers

If the two-slot lamination in the stator stack shown in Figure 32.1 is replaced by a multislot lamination (see Figure 32.2), and two sets of windings are designed in concentric coil sets and distributed in each quadrant of the laminated stack, a close approximation to a sine wave can be generated on one of the secondary windings, and a close approximation to a cosine wave can be generated on the other set of windings. Rotary transformers of this design are called *resolvers*. Using a multislot rotor lamination and distributing the windings in the rotor, the sine–cosine waveforms can be improved even further.

A resolver effectively amplitude-modulates the ac excitation signal placed on the rotor windings in proportion to the sine and the cosine of the angle of mechanical rotation. This sine–cosine electrical output information measured across the stator windings may be used for position and velocity data. In this manner, the resolver is an analog trigonometric function generator. Most resolvers have two primary windings that are located at right angles to each other in the stator, and two secondary windings also at right angles to each other, located on the rotor (see Figure 32.3).

If the rotor winding (R1–R3) is excited with the rated input voltage (see Figure 32.4), the amplitude of the output winding of the stator (S2–S4) will be proportional to the sine of the rotor angle θ,

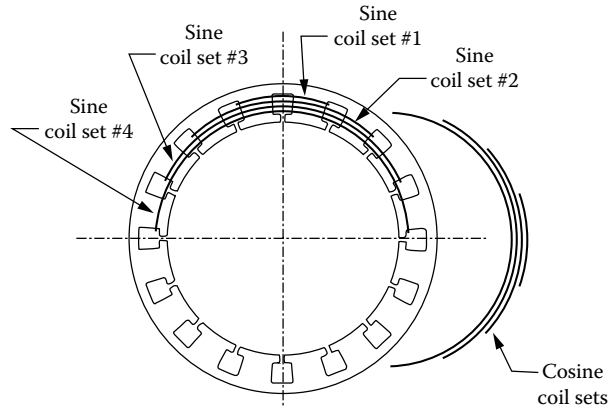


FIGURE 32.2 The resolver stator has distributed coil windings on a 16-slot lamination to generate a sine wave.

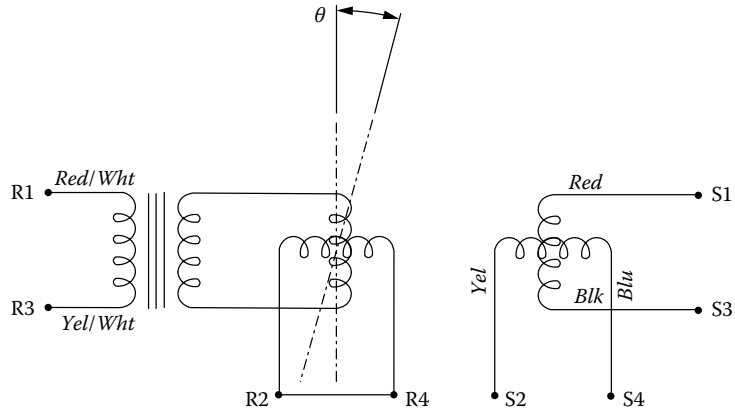


FIGURE 32.3 A brushless resolver modulates the ac excitation on the rotor by the rotation angle.

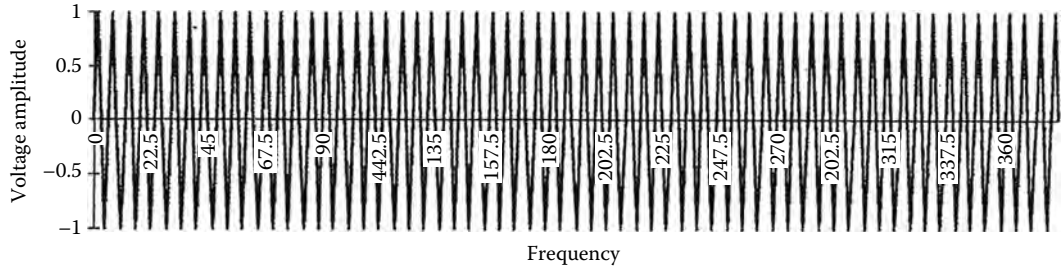


FIGURE 32.4 The resolver rotor winding is excited with the rated input voltage.

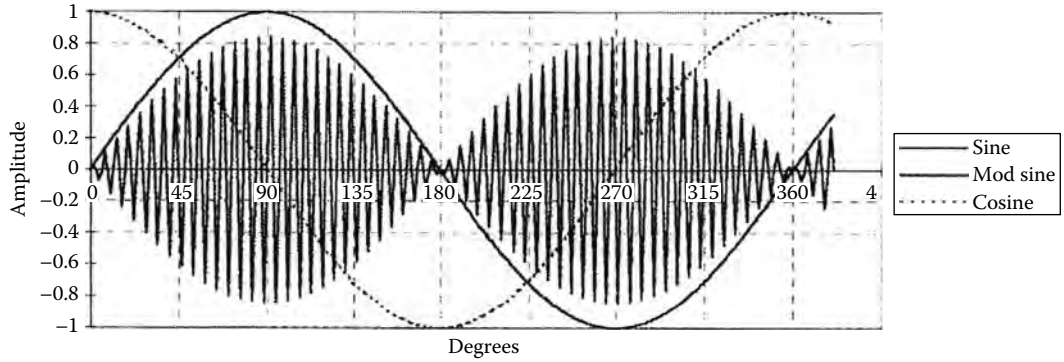


FIGURE 32.5 The single-speed resolver stator output is the sine or cosine of the angle.

and the amplitude of the output of the second stator winding (S1–S3) will be proportional to cosine θ (see Figure 32.5). This is commonly called the "control transmitter" mode and is used with most "state-of-the-art" resolver-to-digital converters.

In the control transmitter mode, electrical zero may be defined as the position of the rotor with respect to the stator at which there is minimum voltage across S2–S4 when the rotor winding R1–R3 is excited with rated voltage. Nulls will occur across S2–S4 at the 0° and 180° positions and will occur across SI–S3 at the 90° and 270° positions.

If the stator winding SI–S3 is excited with the rated input voltage and stator winding S2–S4 is excited with the rated input voltage electrically shifted by exactly 90°, then the output sensed on the rotor winding R1–R3 does not vary with rotor rotation in amplitude or frequency from the input reference signal.

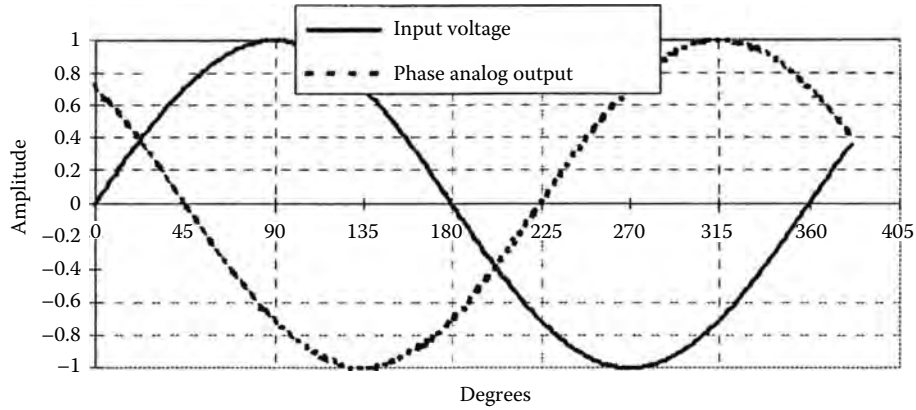


FIGURE 32.6 Resolver stator windings excited at 0° and 90° yield a phase shift with rotor rotation. Phase analog output at 225° versus input voltage.

It is the sum of both inputs. It does, however, vary in time phase from the rated input by the angle of the shaft from a referenced "zero" point (see Figure 32.6). This is a "phase analog" output and the device is termed a "control transformer." By measuring the time difference between the zero crossing of the reference voltage waveform and the output voltage waveform, the phase angle (which is the physical angular displacement of the output shaft) can be calculated.

Because the resolver is an analog device and the outputs are continuous through 360°, the theoretical resolution of a resolver is infinite. There are, however, ambiguities in output voltages caused by inherent variations in the transformation of the voltage from primary to secondary through 360° of rotation. These ambiguities result in inaccuracy when determining the true angular position. The types of error signals that are found in resolvers are shown in Figure 32.7.

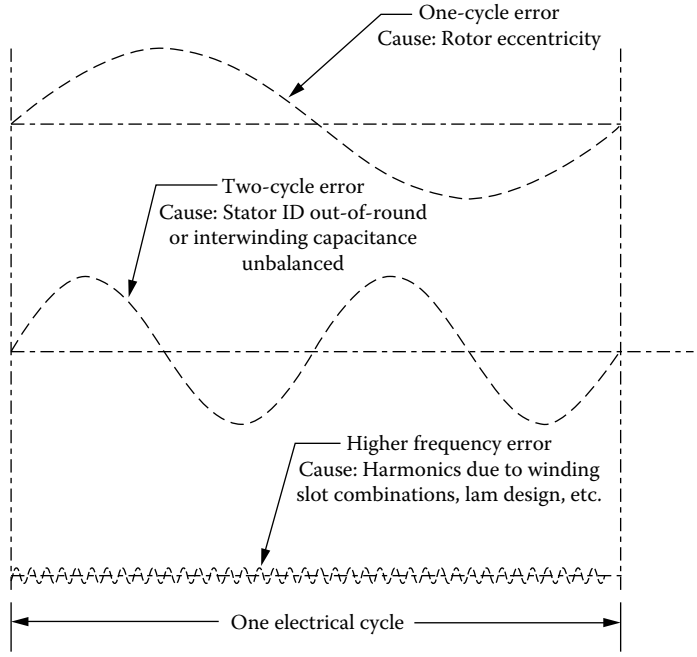


FIGURE 32.7 There are several causes of errors in resolvers and synchros.

As a rule, larger diameter stator laminations lead to a better accuracy and higher absolute resolution of the device. This is a function of the number of magnetic poles that can be fit into the device, which is a direct function of the number of slots in the stator and rotor laminations. With multispeed units (see the section on multispeed as follows), the resolution increases as a multiple of the speeds. For most angular excursions, the multispeed resolver can exceed the positioning accuracy capability of any other component in its size, weight, and price range.

32.3 Operating Parameters and Specifications for Resolvers

There are seven functional parameters that define the operation of a resolver in the analog mode. These are (1) accuracy, (2) operating voltage amplitude, (3) operating frequency, (4) phase shift of the output voltage from the referenced input voltage, (5) maximum allowable current draw, (6) transformation ratio of output voltage over input voltage, and (7) null voltage. Although impedance controls the functional parameters, it is transparent to the user. The lamination and coil design are usually developed to minimize null voltage and input current, and the impedance is the direct fallout of the inherent design of the resolver. The following procedure can be used to measure the seven values for most resolvers.

32.4 Equipment Needed for Testing Resolvers

A mechanical index stand that can position the shaft of the resolver to an angular accuracy that is an order of magnitude greater than the specified accuracy of the resolver
An ac signal generator capable of up to 24 V rms at 10 kHz
A phase angle voltmeter (PAV) capable of measuring "in phase" and "quadrature" voltage components for determining the transformation ratio and the null voltage as well as the phase angle between the output voltage and the reference input voltage
A 1 Ω resistor used to measure input current with the PAV

An angle position indicator (API) used in conjunction with the index stand to measure the accuracy of the resolver

1. Using the aforementioned equipment with the index stand in the 0° position, mount the resolver on the index stand and lock the resolver shaft in place.
2. Place the 1 Ω resistor in series with the R1 (red/wht) lead of the resolver.
3. Connect the resolver rotor leads, R1 (red/wht) and R3 (yel/wht) (see Figure 32.3) to the sine wave signal generator and set the voltage and frequency to the designed values for the resolver.
4. Connect the terminals from the sine wave signal generator to the reference input of the PAV.
5. Connect the resolver output leads S2 (yel) and S4 (blu) to the PAV input terminals and place the PAV in the total output voltage mode.
6. Rotate the index stand to the 0° position.
7. Turn the resolver housing while monitoring the total output voltage on the PAV until a minimum voltage reading is obtained on the PAV. Record this value. This is the *null voltage* of the resolver.
8. Turn the index stand to the 90° position.
9. Change the output display on the PAV to show the *phase angle* between the reference voltage and the voltage on leads S2–S4. Record this phase angle reading.
10. Change the output display on the PAV to show the *total output voltage* of the resolver.
11. Record this voltage.
12. Calculate the *transformation ratio* by dividing the reading in step 11 by the input reference voltage amplitude.
13. Connect the input leads of the PAV across the 1 Ω resistor and record the total voltage on the PAV display. Since this reading is across a 1 Ω resistor, it is also the *total current*.

14. Disconnect the resolver and the signal generator from the PAV.
15. Connect the terminals from the sine wave signal generator to the reference input of the API.
16. Connect the stator leads S1 (red), S3 (blk), S2 (yel), and S4 (blu) to the API S1, S3, S2, and S4 inputs, respectively.
17. Check the *accuracy* of the resolver by recording the display on the API every 20°–360° of mechanical rotation on the index stand. Record all values.

32.5 Resolver Benefits

Designers of motion control systems today have a variety of technologies from which to choose feedback devices. The popularity of brushless dc motors has emphasized the need for rotor position information to commutate windings as a key application in sensor products. Encoders are widely used due to the ease of interface with today's drivers but have some inherent performance limitations in temperature range, shock and vibration handling, and contamination resistance. The resolver is a much more rugged device due to its construction and materials and can be mated with modular R to D converters that will meet all performance requirements and work reliably in the roughest of industrial and aerospace environments.

The excitation signal E_x into the resolver is converted by transformer coupling into sine and cosine (quadrature) outputs that equal $E_x \sin\theta$ and $E_x \cos\theta$. The resolver-to-digital converter (shown in Figure 32.10) calculates the angle:

$$\Theta = \arctan\left(\frac{E_x \sin\Theta}{E_x \cos\Theta}\right) \tag{32.1}$$

In this ratiometric format, the output of the sine winding is divided by the output of the cosine winding, and any injected noise whose magnitude is approximately equivalent on both windings is canceled. This provides an inherent noise rejection feature that is beneficial to the resolver user. This feature also results in a large degree of temperature compensation.

32.5.1 Multispeed Units

The relationship for multispeeds is that the speed (2×, 3×, etc.) designates how many full sinusoidal cycles the resolver output electrically completes in 360° of mechanical rotation. The 2× electrical output is such that the full sinusoidal cycle for a 2× resolver occurs in 180° instead of 360°. A 2× resolver output is shown in Figure 32.8. A full 3× cycle is completed in 120°. The number of speeds selected for use is a function of the system requirements. Increasing the number of magnetic poles in the rotor and stator creates multispeed units. Each speed has several winding and slot combinations. The optimum combination is selected by the resolver designer based on system demands.

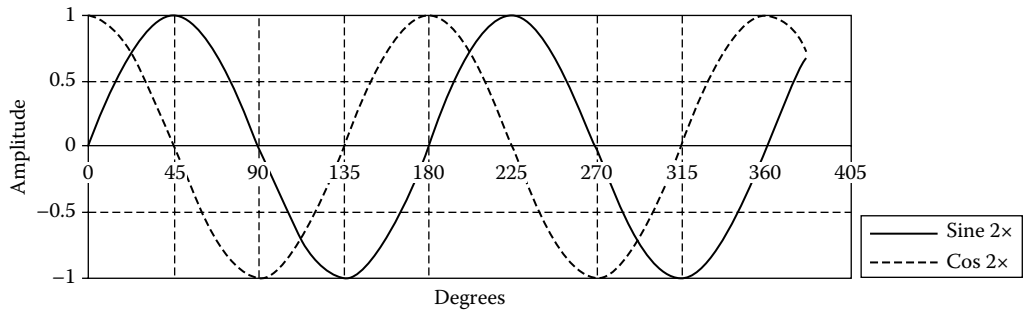


FIGURE 32.8 A two-speed resolver yields two electrical cycles for one rotation.

32.5.2 Applications

Resolvers are often used in conjunction with motors, and because of their inherent similarity of design (copper windings on iron lamination stacks), their environmental resistance is quite similar. They are ideal to design into industrial applications where dust and airborne liquids can obscure optical encoder signals. NC machines, coil winders, presses, and positioning tables are uses where resolvers excel. The resolver's inherent resistance to shock and vibration makes it uniquely suited to moving platforms, and their reliability under these conditions lends a welcome hand to the designers of robots, gantries, and automotive transfer lines.

Heat sensitivity is always a problem for motion control systems designers. Resolvers used for sensing the position of valves in high-temperature applications such as aircraft engines, petrochemical refining, and chemical processing have continually proven their reliability.

Moving devices to precise positions with smooth and accurate control can be a real challenge in the electromagnetic noise environment of the modern industrial facility. Emitted and conducted EMI from adjacent equipment and input voltage variations with unwanted current spikes on input power lines can rob digital systems of their signal integrity. The analog resolver continues to function without information loss or signal interruption. Digitizing the signal can be done at a remote interface under more controlled conditions than on the factory floor. Only robust materials can perform well in harsh environments.

32.6 Synchros

As long ago as World War II, synchros were used in analog positioning systems to provide data and to control the physical position of mechanical devices such as radar antennae, indicator needles on instrumentation, and fire control mechanisms in military equipment. The term "synchro" defines an electromagnetic position transducer that has a set of three-phase output windings that are electrically and mechanically spaced by 120° instead of the 90° spacing found in a resolver. In the rotor primary mode, the synchro is excited by a single-phase ac signal on the rotor. As the rotor moves 360°, the three amplitude-modulated sine waves on the three phases of the output have a discrete set of amplitudes for each angular position. By interpreting these amplitudes, a table can be established to decode the exact rotary position. In most applications, resolvers have replaced synchros because of the sophistication of the resolver-to-digital converters that are commercially available. Working with a sine and cosine is simpler and requires less conversion and decoding than using three 120°-spaced signals. If conversion of a synchro output is desired in resolver format, a device known as a Scott "T" transformer can be used for conversion. In most synchro-to-digital processors, the first step is to convert the signal to a resolver format with a Scott "T" device.

32.7 Modular Solution

The brushless resolver is a self-contained feedback device that, unlike optical encoders, provides an analog signal with infinite resolution. Not only can the output signal be converted to precise digital position information, but it also provides an accurate velocity signal, thus eliminating the need for using a separate tachometer. Reliability is enhanced using the same resolver for speed feedback and commutation. Piece part count can be reduced and the complexity of using Hall-effect devices for timing signals for commutation can be eliminated.

A modular approach allows the designer to easily select a single or multispeed resolver and appropriate electronics that will meet almost any desired level of resolution and accuracy. The resolvers are designed in the most commonly used frame sizes: 8, 11, 15, and 21. Housed models feature high-quality, motor-grade ball bearings. Heavy-duty industrial grade units are enclosed in rugged black-painted aluminum housings with flange, face, or servo-type mounting and utilize MS-style connectors.

32.8 Sensible Design Alternative for Shaft Angle Encoding

The requirement for velocity and position feedback plays an important role in today's motion control systems. With the development of low-cost monolithic resolver-to-digital converters, a resolver-based system provides design engineers with the building blocks to handle a wide variety of applications. A resolver's small size, rugged design, and the ability to provide a very high degree of accuracy under severe conditions make this the ideal transducer for absolute position sensing. These devices are also well suited for use in extremely hostile environments such as continuous mechanical shock and vibration, humidity, oil mist, coolants, and solvents. Absolute position sensing vs. incremental position sensing is a necessity when working in an environment where there is the possibility of power loss. Whenever power is supplied to an absolute system, it is capable of reading its position immediately; this eliminates the need for a "go home" or reference starting point.

32.8.1 Resolver-to-Digital Converters

A monolithic resolver-to-digital converter requires only six external passive components to set the bandwidth and maximum tracking rate. The bandwidth controls how quickly the converter will react to a large change in position on the resolver output. The converter can also be programmed to provide 10, 12, 14, or 16 bits of parallel data. A resolver-based system can provide high dynamic capability and high resolution for today's motion control systems where precision feedback for both position and velocity is required.

32.8.2 Closed-Loop Feedback

In a typical closed-loop servo model as in Figure 32.9, the position sensor plays an important role by constantly updating the position and velocity information. Selection of a machine control strategy will often be based on performance, total application cost, and technology comfort.

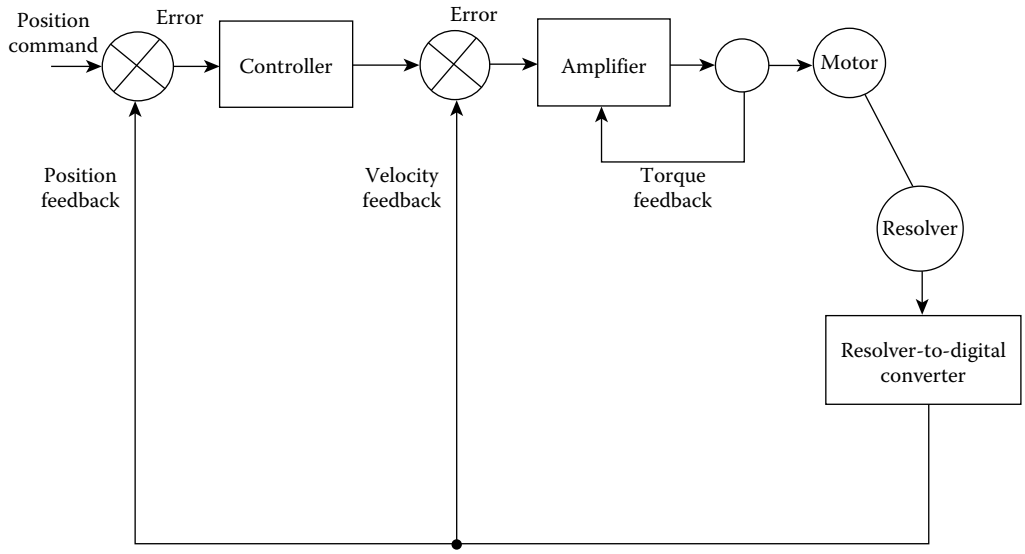


FIGURE 32.9 A closed-loop servo model uses a resolver-to-digital converter.

The accuracy of the system is determined by the smallest resolution of the position-sensing device. A resolver-to-digital converter in the 16-bit mode has 2^{16} (65,536) counts per revolution, which is equivalent to a resolution of 20 arcs. The overall accuracy of the resolver-to-digital converter is ±2.3 arcmin. An accuracy specification defines the maximum error in achieving a desired position. System accuracy must be smaller than the tolerance on the desired measurement. An important feature of the resolver-to-digital converter is repeatability. With a repeatability specification of ±1 LSB (least significant bit) in the 16-bit mode, this provides an accurate measurement when determining position from point to point. For example, moving from point A to point B and back to point A, the converter in the 16-bit mode will be accurate within 20 arcs of the original position. The error curve of a resolver-to-digital converter is repeatable within ±1 LSB. The combination of high-precision resolvers (±20 arcs) with a resolver-to-digital converter provides accurate absolute position information for precision feedback for motion control.

32.9 Type II Servo Loop

The motor speed is monitored using the velocity output signal generated by the resolver-to-digital converter. This signal is a dc voltage proportional to the rate of speed, positive for increasing angles and negative for decreasing angles, with a typical linearity specification of 0.25% and a typical reversal error of 0.75%. The error processing is performed using the industry standard technique for type II tracking, resolver-to-digital converters (see Figure 32.10).

The dc error is integrated, yielding a velocity voltage that drives a voltage-controlled oscillator (VCO). This VCO is an incremental integrator (constant voltage input to position rate output) that together with the velocity integrator forms a type II critically damped, servo feedback loop. This information allows the motor to maintain constant speeds under varying loads when it is interfaced with a programmable logic controller (PLC). The PLC-based architecture is used for integration of the motion (I/O) intensive control applications. The PLC provides a low-cost option for those developers familiar with its ladder logic programming language. I/O operator's interface and communication are usually supported through additional cards that are plugged into the backplane.

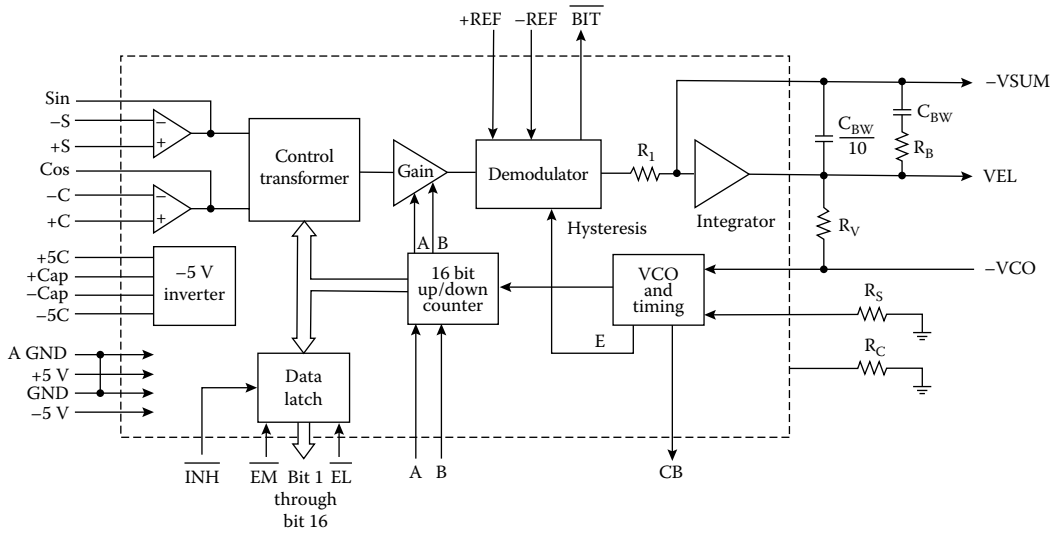


FIGURE 32.10 Error processing uses type II tracking resolver-to-digital converters.

32.10 Applications

Specific applications require unique profiles to control the speed and acceleration of the motor to perform the task at hand. By reducing the accelerations and decelerations that occur during each operation, it is possible to lower the cost and use more efficient motors. Industrial applications include the following:

- Ball-screw positioning
- Motor commutation
- Robotics positioning
- Machine vision systems
- *X*–*Y* tables
- Component insertion
- Remote video controls
- Web guides
- Pick and place machines

32.10.1 Resolver-to-Digital Conversion

For a resolver-to-digital converter, the resolver information is presented to a solid-state resolver conditioner that reduces the signal amplitude to 2 V rms sine and cosine, the amplitude of one being proportional to the sine of θ (the angle to be digitized) and the amplitude of the other being proportional to the cosine of θ. (The amplitudes referred to are, of course, the carrier amplitudes at the reference frequency, i.e., the cosine wave is actually $\cos\theta \cos\omega t$, but the carrier term, $\cos\omega t$, will be ignored in this discussion because it will be removed in the demodulator and at any rate contains no data.) A quadrant selector circuit in the control transformer enables selection of the quadrant in which θ lies and automatically sets the polarities of the $\sin\theta$ and $\cos\theta$ appropriately, for computational significance. The $\sin\theta$ and $\cos\theta$ outputs of the quadrant selector are then fed to the sine and cosine multipliers, also contained in the control transformer. These multipliers are digitally programmed resistive networks. The transfer function of each of these networks is determined by a digital input (which switches in proportioned resistors), so that the instantaneous value of the output is the product of the instantaneous value of the analog input and the sine (or cosine) of the digitally encoded angle. If the instantaneous value of the analog input of the sine multiplier is $\cos\theta$, and the digitally encoded "word" presented to the sine multiplier is ϕ, then the output code is $\cos\theta \sin\phi$. Thus, the two outputs of the multipliers are

$$\text{From the sine multiplier: } \cos\theta\sin\phi$$

$$\text{From the cosine multiplier: } \sin\theta\cos\phi$$

These outputs are fed to an operational subtractor, at the differencing junction shown, so that the input fed to the demodulator is

$$\sin\theta\cos\phi - \cos\theta\sin\phi = \sin(\theta - \phi) \tag{32.2}$$

The right-hand side of this trigonometric identity indicates that the differencing-junction output represents a carrier-frequency sine wave with an amplitude proportional to the sine of the difference between θ (the angle to be digitized) and ϕ (the angle stored in digital form in the up/down counter).

This point is the ac error signal brought out as (e). The demodulator is also presented with the reference voltage, which has been isolated from the reference source, and appropriately scaled, by the reference conditioner. The output of the demodulator is then, an analog dc level, proportional to sin ($\theta - \phi$), in other words, to the sine of the "error" between the actual angular position of the resolver and the digitally encoded angle, ϕ, which is the output of the counter. This point dc error is sometimes brought out as (D), while an addition of a threshold detector will give a built-in-test (BIT) flag. When the ac error signal exceeds 100 LSBs, the BIT flag will indicate a tracking error. This angular error signal is then fed into the error processor and VCO. This circuit consists essentially of an analog integrator whose output (the time integral to the error) controls the frequency of a VCO. The VCO produces clock pulses that are counted by the up/down counter. The "sense" of the error (ϕ too high or ϕ too low) is determined by the polarity of (ϕ) and is used to generate a control counter signal (U), which determines whether the counter increments upward or downward. Finally, note that the up/down counter, like any counter, is functionally an incremental integrator; therefore, the tracking converter constitutes in itself a closed-loop servomechanism (continuously attempting to null the error to zero) with two integrators in series. This called a "type II" servo loop, which has decided advantages over type 1 or type 0 loops. In order to appreciate the value of a type II servo behavior of this tracking converter, consider first the shaft is not moving. Ignoring inaccuracies, drifts, and the inevitable quantizing error, the error should be zero ($\theta = \phi$) and the digital output represents the true shaft angle of the resolver. Now, start the resolver shaft moving, and allow it to accelerate uniformly, from $d\theta/dt = 0$ to $d\theta/dt = V$. During the acceleration, an error will develop because the converter cannot instantaneously respond to the change of angular velocity. However, since the VCO is controlled by an integrator, the output of which is the integral of the error, the greater the lag (between θ and ϕ), the faster the counter will be called on to catch up. When the velocity becomes constant at V, the VCO will have settled to a rate of counting that exactly corresponds to the rate of change in θ per unit time and instantaneously $\theta = \phi$. Therefore, $d\phi/dt$ will always track $d\theta/dt$ without a velocity or position error, the only error will be momentary (transient) error, during acceleration or deceleration. Furthermore, the information produced by the tracking converter is always "fresh," being continually updated, and always available at the output of the counter. Since $d\theta/dt$ tracks the input velocity, it can be brought out as velocity, a dc voltage proportional to the rate of rotation, which is of sufficient linearity in modern converters to eliminate the need for a tachometer in many systems.

32.10.2 Bandwidth Optimization

When using a low-cost monolithic converter for position and velocity feedback, it is important to understand the dynamic response for a changing input. When considering what bandwidth to set the converter, several parameters must be taken into consideration. The ability to track step responses and accelerations will determine what bandwidth to select. The lower the bandwidth of the resolver-to-digital converter, the greater the noise immunity; high-frequency noise will be rejected. The relationship between the maximum tracking rate and bandwidth determines the settling time for small and large steps. For a small step input, the bandwidth determines the converter settling time. When one has a large step, the maximum velocity slew rate and bandwidth together determine the settling time.

32.10.3 Encoder Emulation

Today's resolver-to-digital converters also have the ability to emulate the output of an optical incremental encoder. By providing the outputs A, B, and zero index, the encoder can be replaced with a resolver and resolver-to-digital converter without changing the existing interface hardware.

32.10.4 Determining Position Lag Error Due to Acceleration

As the bandwidth and the maximum tracking rate are varied, one can determine the acceleration constant (K_a) and large step settling time.

Example 32.1

Resolution: 16 bit
Bandwidth: 100 Hz
Reference: 1000 Hz
Max tracking: 10 rps

$$\text{BW} = \frac{\sqrt{2}A}{\pi} \tag{32.3}$$

If the bandwidth = 100 Hz
$A = 222$
$K_a = A^2$
$K_a = 49{,}284° \text{ s}^{-2}$
The lag in degrees during acceleration is

$$\frac{\text{Acceleration}}{K_a} \tag{32.4}$$

Example 32.2

Acceleration = $19{,}000° \text{ s}^{-2}$
$K_a = 49{,}284° \text{ s}^{-2}$
LAG = 19,000/49,284 = 0.38°
In 16 bit 1 LSB = 0.0055°
Acceleration (1 LSB lag) = $K_a \times 0.0055$ (16 bit)
$49{,}284 \times 0.0055 = 270° \text{ s}^{-2}$

32.10.5 Large Step Settling Time

To determine the settling time for a large step response (179°), one must take into account the maximum tracking rate and bandwidth.

Example 32.3

A 179° step with 100 Hz BW and 10 rps max tracking
Max tracking at 10 rps = $3600° \text{ s}^{-1}$
179/3600 = 49 ms

Then one must add the settling time due to bandwidth limitations; this is approximately 11 time constants (16-bit mode).

Time Constants

Resolution	Number of Counts/Rotation	Number of Time Constants
10	1,024	7
12	4,096	8
14	16,384	10
16	65,536	11

Time constant = $1/A$
$A = 222$
$1/A = 4.5$ ms
11 time constants = 45 ms

Therefore, the approximate settling time for a large step would be 94 ms. This is an approximation. Synchros and resolvers are used in a wide variety of dynamic conditions. Understanding how the converter reacts to these input changes will allow one to optimize the bandwidth and maximum tracking rate for each application. Table 32.1 provides a list of resolver and synchro suppliers.

TABLE 32.1 List of Resolver and Synchro Suppliers

Company	Location, Phone Number	Types of Resolvers
API Harowe	West Chester, PA, (800) 566-5274, http://www.servosystems.com/harowe_resolvers.htm	Brushless frameless, housed, and heavy-duty units
Admotec, Inc.	Norwich, VT, (802) 649-5800, http://www.admotec.com/	Rotasyn solid rotor resolvers
Neotech, Inc.	Hatfield, PA, (215) 822-5520, http://neotechsystems.com/	Housed brush and brushless units
Vernitron Corp.	San Diego, CA, (800) 777-3393	Brushless, segment, brushed, housed units
Servo Systems	Montville, NJ, (973)335-1007, http://www.servosystems.com/	Brushed 8c brushless units
American Electronics	Fullerton, CA (714)871-3020	Housed units
Computer Conversions, Corp.	East Northport, NY, (516) 261-3300, http://www.computerconversions.com/index.html	Explosion proof 8c specialty units
Poltron Corp.	Gaithersburg, MD, (301)208-6597	Resolvers
Tamagawa Seiki Co., Ltd.	Tokyo, Japan, 011-81-37-383-175, http://www.tamagawa-seiki.co.jp/english/index.html	Brushless frameless 8c housed units
MPC Products	Skokie, IL, (800) 323-4302	Housed resolvers
Transicoil, Inc.	Trooper, PA, (800)323-7115, http://www.isensors.com/nav/manuf.asp?vid=35	Housed brushless 8c brushed resolvers
Litton Poly-Scientific	Blacksburg, VA, (800)336-2112	Brushed and brushless resolvers
Kearfott Guidance & Navigation Corp.	Wayne, NJ, (973) 785-6000, http://www.kearfott.com/	Brushed 8c brushless resolvers
Novatronics, Inc.	Stratford, Ontario, Canada, (519) 271-3880, http://www.novatronicsinc.com/	Resolvers
Muirhead Vactric	Lake Zurich, IL, (847) 726-0270, http://www.muirheadaerospace.com/	Resolvers

Further Readings

Analog Devices, Inc., *Analog-Digital Conversion Handbook*, 3rd edn., Englewood Cliffs, NJ: Prentice-Hall, 1986.

Moog Components, Components Groop. http://www.scribd.com/doc/68820721/Synchro-and-Resolver-Hand-Book (accessed on June 26, 2012).

Synchro & Resolver Conversion, East Molesey, U.K.: Memory Devices Ltd., 1980.

Synchro/Resolver Conversion Handbook, 4th edn., Bohemia, NY: ILC Data Device Corp., 1994.

Synchro resolver, Otomation. http://www.globalspec.com/Industrial-Directory/synchro_resolver (accessed on June 28, 2012).

Welbun, R. D., *Self corrected synchro/resolver*. http://www.devileye.net/catalog/manufacturing_metallosilicates/self-corrected_synchro_resolver.html (accessed July 1, 2012).

33

Optical Fiber Displacement Sensors

33.1 Extrinsic Fabry–Perot Interferometric Sensor **33**-2
33.2 Intrinsic Fabry–Perot Interferometric Sensor **33**-4
33.3 Fiber Bragg-Grating Sensor **33**-5
Principle of Operation • Fabrication Techniques • Bragg-Grating Sensors • Limitations of Bragg-Grating Strain Sensors
33.4 Long-Period Grating Sensor **33**-9
Principle of Operation • Fabrication Procedure • Temperature Sensitivity of Long-Period Gratings
33.5 Comparison of Sensing Schemes **33**-14
33.6 Conclusion **33**-14
References **33**-14
Further Readings **33**-15

Richard O. Claus
Virginia Tech

Vikram Bhatia
Virginia Tech

Anbo Wang
Virginia Tech

The objective of this chapter is to present a rigorous theoretical and experimental analysis of short gage length optical fiber sensors for measurement of cyclical strain on or in materials. Four different types of sensors are evaluated systematically on the basis of various performance criteria such as strain resolution, dynamic range, cross sensitivity to other ambient perturbations, simplicity of fabrication, and complexity of demodulation process. The sensing methods that would be investigated include well-established technologies (e.g., fiber Bragg gratings (FBGs)) and rapidly evolving measurement techniques such as long-period gratings. Other than the grating-based sensors, two popular versions of Fabry–Perot interferometric sensors (intrinsic and extrinsic) will be evaluated for their suitability. A theoretical study of the cross sensitivities of these sensors to an arbitrary combination of strain vectors and temperature, similar to that proposed by Sirkis in his SPIE paper in 1991 [1], will be developed.

The outline of this chapter is as follows. The principle of operation and fabrication process of each of the four sensors is discussed separately. Sensitivity to strain and other simultaneous perturbations such as temperature are analyzed. The overall cost and the performance of a sensing technique depend heavily on the signal demodulation process. The detection schemes for all four sensors are discussed and compared on the basis of their complexity. Finally, a theoretical analysis of the cross sensitivities of the four sensing schemes is presented and their performances are compared.

Strain measurements using optical fiber sensors in both embedded and surface-mounted configurations have been reported by researchers in the past [2]. Fiber-optic sensors are small in size and immune to electromagnetic interference and can be easily integrated with existing optical fiber communication links. Such sensors can typically be easily multiplexed, resulting in distributed networks that can be used for health monitoring of integrated, high-performance materials and structures. Optical fiber sensors for strain measurements should possess certain important characteristics. These sensors should either be insensitive to ambient fluctuations in temperature and pressure or should have demodulation techniques

that compensate for changes in the output signal due to the undesired perturbations. In the embedded configuration, the sensors for axial strain measurements should have minimum cross sensitivity to other strain states. The sensor signal should itself be simple and easy to demodulate. Nonlinearities in the output demand expensive decoding procedures or require precalibrating the sensor. The sensor should ideally provide an absolute and real-time strain measurement in a form that can be easily processed. For environments where large strain magnitudes are expected, the sensor should have a large dynamic range while at the same time maintaining the desired sensitivity. A discussion of each of the four sensing schemes individually, along with their relative merits and demerits, follows.

33.1 Extrinsic Fabry–Perot Interferometric Sensor

The extrinsic Fabry–Perot interferometric (EFPI) sensor, proposed by Murphy et al., is one of the most popular fiber-optic sensors used for applications in health monitoring of smart materials and structures [3]. As the name suggests, the EFPI is an interferometric sensor in which the detected intensity is modulated by the parameter under measurement. The simplest configuration of an EFPI is shown in Figure 33.1.

The EFPI system consists of a single-mode laser diode that illuminates a Fabry–Perot cavity through a fused biconical tapered coupler. The cavity is formed between an input single-mode fiber and a reflecting single-mode or multimode fiber. Since the cavity is external to the lead-in/lead-out fiber, the EFPI sensor is independent of transverse strain and small ambient temperature fluctuations. The input fiber and the reflecting fiber are aligned using a hollow-core silica fiber. For uncoated fiber ends, a 4% Fresnel reflection results at both ends. The first reflection, R_1, called the reference reflection, is independent of the applied perturbation. The second reflection, R_2, termed the sensing reflection, is dependent on the length of the cavity, d, which in turn is modulated by the applied perturbation. These two reflections interfere (provided $2d < L_c$, the laser diode's coherence length), and the intensity I at the detector varies as a function of the cavity length:

$$I = I_0 \cos\left(\frac{4\pi}{\lambda} d\right) \tag{33.1}$$

where

I_0 is the maximum value of the output intensity
λ is the laser diode center wavelength

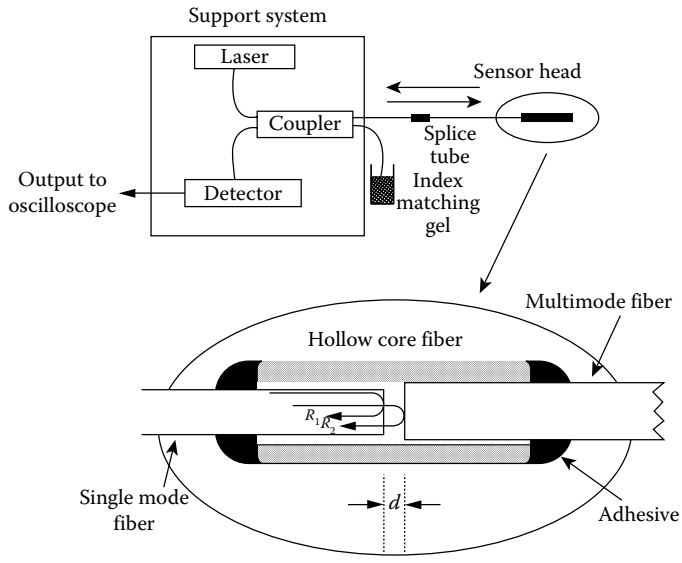


FIGURE 33.1 Simple configuration of an EFPI sensing system.

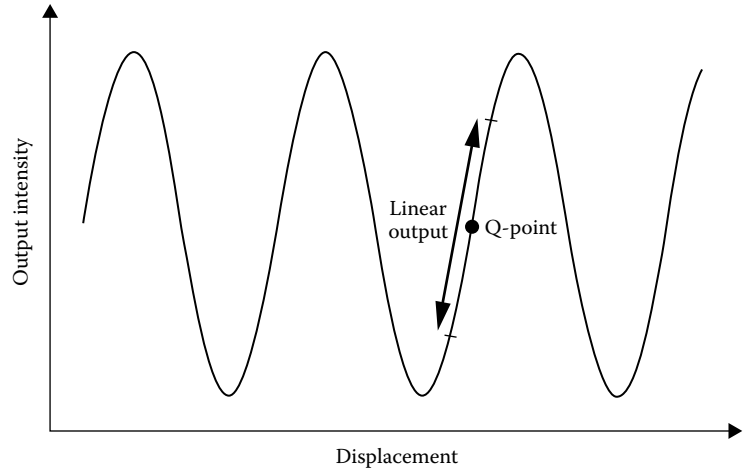


FIGURE 33.2 Typical EFPI transfer function curve.

The typical EFPI transfer function curve is shown in Figure 33.2. Small perturbations that result in operation around the quiescent-point or Q-point of the sensor lead to a linear variation in output intensity. A fringe in the output signal is defined as the change in intensity from a maximum to a maximum or from a minimum to a minimum. Each fringe corresponds to a change in the cavity length by one-half of the operating wavelength, λ. The change in the cavity length, Δd, is then employed to calculate the strain ε using the expression

$$\varepsilon = \frac{\Delta d}{L} \tag{33.2}$$

where L is defined as the gage length of the sensor and is typically the distance between two points where the input and the reflecting fibers are bonded to the hollow-core fiber. Matching of the two reflection signal amplitudes allows good fringe visibility in the output signal.

The EFPI sensor has been extensively used for measuring fatigue loading on F-15 aircraft wings, detection of crack formation and propagation in civil structures, and cure and lifetime monitoring in concrete and composite specimens [2,4]. The temperature insensitivity of this sensor makes it attractive for a large number of applications. The EFPI sensor is capable of measuring sub-angstrom displacements with strain resolution better than 1 microstrain and a dynamic range greater than 10,000 $\mu\varepsilon$. Although the change in output intensity of the EFPI is nonlinear corresponding to the magnitude of the parameter being measured, for small perturbations, its operation can be limited to that around the Q-point of the transfer function curve. Moreover, the large bandwidth available with this sensor simplifies the measurement of highly cyclical strain. The EFPI sensor is capable of providing single-ended operation and is hence suitable for applications where access to the test area is limited. The sensor requires simple and inexpensive fabrication equipment and an assembly time of less than 10 min. Additionally, since the cavity is external to the fibers, transverse strain components that tend to influence intrinsic sensors through the Poisson's effect have negligible effect on the EFPI sensor output. The sensitivity to only axial strain and the insensitivity to input polarization state have made the EFPI sensor the most preferred fiber-optic sensor for embedded applications [1]. Thus, overall, the EFPI sensing system is very well suited to measurement of small magnitudes of cyclical strain.

Although a version of the EFPI sensor that provides absolute output has been demonstrated, it lacks the bandwidth typically desired during the measurement of cyclical strain [5]. We have also recently proposed a small-cavity-length/high-finesse EFPI sensor for measurement of small perturbations [6]. This configuration has a simple output that can be demodulated using an optical filter/photodetector combination.

33.2 Intrinsic Fabry–Perot Interferometric Sensor

The intrinsic Fabry–Perot interferometric (IFPI) sensor is similar in operation to its extrinsic counterpart, but significant differences exist in the configurations of the two sensors [7]. The basic IFPI sensor is shown in Figure 33.3. An optically isolated laser diode is used as the optical source to one of the input arms of a bidirectional 2 × 2 coupler. The Fabry–Perot cavity is formed by fusing a small length of a single-mode fiber to one of the output legs of the coupler. As shown in Figure 33.3, the reference (R) and sensing (S) reflections interfere at the detector face to provide a sinusoidal intensity variation. The cavity can also be obtained by introducing two Fresnel reflectors—discontinuities in refractive index—along the length of a single fiber. Photosensitivity in germanosilicate fibers has been used in the past to fabricate broadband reflectors that enclose an IFPI cavity [8]. Since the cavity is formed within an optical fiber, changes in the refractive index of the fiber due to the applied perturbation can significantly alter the phase of the sensing signal, S. Thus, the intrinsic cavity results in the sensor being sensitive to ambient temperature fluctuations and all states of strain.

The IFPI sensor, like all other interferometric signals, has a nonlinear output that complicates the measurement of large magnitude strain. This can again be overcome by operating the sensor in the linear regime around the Q-point of the sinusoidal transfer function curve. The main limitation of the IFPI strain sensor is that the photoelastic effect–induced change in index of refraction results in a nonlinear relationship between the applied perturbation and the change in cavity length. In fact, for most IFPI sensors, the change in propagation constant of the fundamental mode dominates the change in cavity length. Thus, IFPIs are highly susceptible to temperature changes and transverse strain components [1]. In embedded applications, the sensitivity to all the strain components can result in erroneous outputs. The fabrication process of an IFPI strain sensor is more complicated than that of the EFPI sensor since the sensing cavity must be formed within the optical fiber by some special procedure. The strain resolution of the IFPIs is also expected to be around 110 $\mu\varepsilon$ with an operating range greater than 10,000 $\mu\varepsilon$. IFPI sensors also suffer from drift in the output signal due to variations in the polarization state of the input light.

Thus, the preliminary analysis shows that the extrinsic version of the Fabry–Perot optical fiber sensor seems to have an overall advantage over its intrinsic version. The extrinsic sensor has negligible cross sensitivity to temperature and transverse strain. Although the strain sensitivity, dynamic range, and bandwidth of the two sensors are comparable, the IFPIs can be expensive and cumbersome to fabricate due to the intrinsic nature of the sensing cavity.

The EFPI and the IFPI sensors possess nonlinear sinusoidal outputs that complicate the signal processing at the detection end. Although intensity-based sensors have a simple output variation, they suffer from limited sensitivity to strain or other perturbations of interest. Grating-based sensors have recently become popular as transducers that provide wavelength-encoded output signals that can typically be easily demodulated to derive information about the perturbation under investigation. The advantages and drawbacks of Bragg-grating sensing technology are discussed first. The basic operating mechanism of the Bragg-grating-based strain sensor is elucidated, and the expressions for strain resolution are obtained. These sensors are then compared to the recently developed long-period gratings in terms of fabrication process, cross sensitivity to other parameters, and simplicity of signal demodulation.

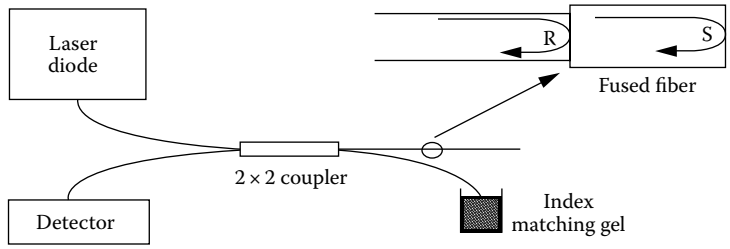


FIGURE 33.3 Intrinsic Fabry–Perot interferometric (IFPI) sensor.

33.3 Fiber Bragg-Grating Sensor

The phenomenon of photosensitivity in optical fibers was discovered by Hill and coworkers in 1978 [9]. It was found that permanent refractive index changes could be induced in fibers by exposing the germanium-doped core to intense light at 488 or 514 nm. The sinusoidal modulation of index of refraction in the core due to the spatial variation in the writing beam gives rise to a refractive index grating that can be used to couple the energy in the fundamental guided mode to various guided and lossy modes. Later, Meltz et al. proposed that photosensitivity is more efficient if the fiber is side exposed to fringe pattern at wavelengths close to the absorption wavelength (242 nm) of the germanium defects in the fiber [10]. The side-writing process simplified the fabrication of Bragg gratings, and these devices have recently emerged as highly versatile components for communication and sensing systems. Recently, loading the fibers with hydrogen has been reported to result in two orders of magnitude higher index change in germanosilicate fibers [11].

33.3.1 Principle of Operation

Bragg gratings are based on the phase-matching condition between spatial modes propagating in optical fibers. This phase-matching condition is given by

$$k_g + k_c = k_B \tag{33.3}$$

where k_g, k_c, and k_B are, respectively, the wave vectors of the coupled guided mode, the resulting coupling mode, and the grating. For a first-order interaction, $k_B = 2\pi/\Lambda$, where Λ is the grating periodicity. Since it is customary to use propagation constants while dealing with optical fiber modes, this condition reduces to the widely used equation for mode coupling due to a periodic perturbation:

$$\Delta\beta = \frac{2\pi}{\Lambda} \tag{33.4}$$

where $\Delta\beta$ is the difference in the propagation constants of the two modes involved in mode coupling (both assumed to travel in the same direction).

FBGs involve the coupling of the forward-propagating fundamental LP_{01} optical fiber waveguide propagation mode to the reverse-propagating LP_{01} mode [12]. Consider a single-mode fiber with β_{01} and $-\beta_{01}$ as the propagation constant of the forward- and reverse-propagating fundamental LP_{01} modes. To satisfy the phase-matching condition,

$$\Delta\beta = \beta_{01} - (-\beta_{01}) = \frac{2\pi}{\Lambda} \tag{33.5}$$

where $\beta_{01} = 2\pi n_{eff}/\lambda$ (n_{eff} is the effective index of the fundamental mode and λ is the free-space wavelength). Equation 33.5 reduces to [12]

$$\lambda_B = 2\Lambda n_{eff} \tag{33.6}$$

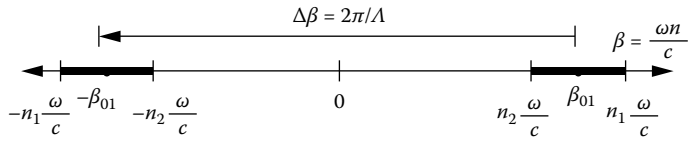


FIGURE 33.4 Mode-coupling mechanism in an FBG.

where λ_B is termed the Bragg wavelength. The Bragg wavelength is the wavelength at which the forward-propagating LP_{01} mode couples to the reverse-propagating LP_{01} mode. This coupling is wavelength dependent since the propagation constants of the two modes are a function of the wavelength. Hence, if an FBG is interrogated by a broadband optical source, the wavelength at which phase-matching occurs is found to be reflected back. This wavelength is a function of the grating periodicity (Λ) and the effective index (n_{eff}) of the fundamental mode (Equation 33.6). Since strain and temperature effects can modulate both these parameters, the Bragg wavelength shifts with these external perturbations. This spectral shift is utilized to fabricate FBGs for sensing applications.

Figure 33.4 shows the mode-coupling mechanism in FBGs using the β-plot. Since the difference in propagation constants ($\Delta\beta$) between the modes involved in coupling is large, Equation 33.4 reveals that only a small value of periodicity, Λ, is needed to induce this mode coupling. Typically for telecommunication applications, the value of λ_B is around 1.5 μm. From Equation 33.6, Λ is determined to be 0.5 μm (for $n_{eff} = 1.5$). Due to the small periodicities (of the order of 1 μm), FBGs are classified as short-period gratings.

33.3.2 Fabrication Techniques

FBGs have commonly been manufactured using two side-exposure techniques: the interferometric method and the phase-mask method. The interferometric method, depicted in Figure 33.5, comprises a UV beam at 244 or 248 nm spilt in two equal parts by a beam splitter [10]. The two beams are then focused on a portion of Ge-doped fiber (whose protective coating has been removed) using cylindrical lenses, and the periodicity of the resulting interference pattern and, hence, the Bragg wavelength are varied by altering the mutual angle, θ. The limitation of this method is that any relative vibration of the pairs of mirrors and lenses can lead to the degradation of the quality of the final grating and, hence, the entire system has a stringent stability requirement. To overcome

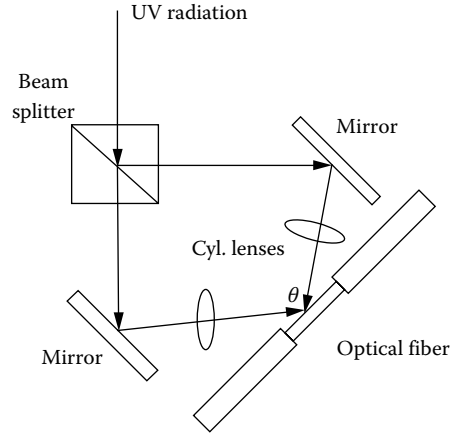


FIGURE 33.5 Interferometric FBG.

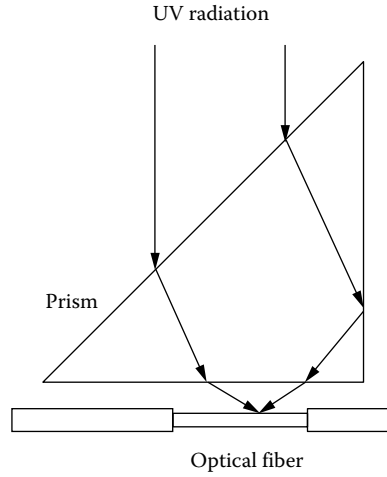


FIGURE 33.6 Novel interferometer technique.

this drawback, Kashyap et al. have proposed the novel interferometer technique where the path difference between the interfering UV beams is produced by the propagation through a right-angled prism (Figure 33.6) [12]. This technique is inherently stable because both beams are perturbed similarly by any prism vibration.

The phase-mask technique has recently gained popularity as an efficient holographic side-writing procedure for grating fabrication [13]. In this method, as shown in Figure 33.7, an incident UV beam is diffracted into –1, 0, and +1 orders by a relief grating generated on a silica plate by *e*-beam exposure and plasma etching. The two first diffraction orders undergo total internal reflection at the glass/air interface of a rectangular prism and interfere on the bare fiber surface placed directly behind the mask. This technique is wavelength specific since the periodicity of the resulting two-beam interference pattern is uniquely determined by the diffraction angle of –1 and +1 orders and, thus, the properties of the phase mask. Obviously, different phase masks are required for fabrication of gratings at different Bragg wavelengths. The setup for actively monitoring the growth of the grating in transmission during fabrication is shown in Figure 33.8.

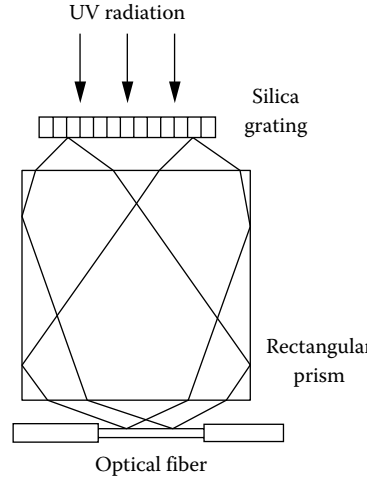


FIGURE 33.7 Phase-mask technique for grating fabrication.

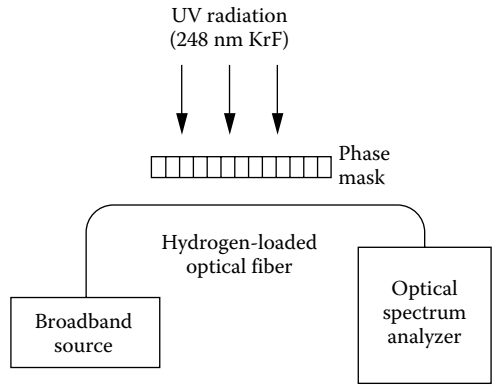


FIGURE 33.8 Diagram of the setup for monitoring the growth of the grating in transmission during fabrication.

33.3.3 Bragg-Grating Sensors

From Equation 33.6, we see that a change in the value of n_{eff} and/or Λ can cause the Bragg wavelength, λ, to shift. This fractional change in the resonance wavelength, $\Delta\lambda/\lambda$, is given by the expression

$$\frac{\Delta\lambda}{\lambda} = \frac{\Delta\Lambda}{\Lambda} + \frac{\Delta n_{\text{eff}}}{n_{\text{eff}}} \tag{33.7}$$

where $\Delta\Lambda/\Lambda$ and $\Delta n_{\text{eff}}/n_{\text{eff}}$ are the fractional changes in the periodicity and the effective index, respectively. The relative magnitudes of the two changes depend on the type of perturbation the grating is subjected to; for most applications, the effect due to change in effective index is the dominating mechanism.

Any axial strain, ε, applied to the grating changes the periodicity and the effective index and results in a shift in the Bragg wavelength, given by

$$\frac{1}{\lambda}\frac{\Delta\lambda}{\varepsilon} = \frac{1}{\Lambda}\frac{\Delta\Lambda}{\varepsilon} + \frac{1}{n_{\text{eff}}}\frac{\Delta n_{\text{eff}}}{\varepsilon} \tag{33.8}$$

The first term on the right-hand side is unity, while the second term has its origin in the photoelastic effect. An axial strain on the fiber serves to change the refractive index of both the core and the cladding. This results in the variation in the value of the effective index of glass. The photoelastic or strain-optic coefficient that relates the change in index of refraction due to mechanical displacement is about −0.27. Thus, the variations in n_{eff} and Λ due to strain have contrasting effects on the Bragg peak. The fractional change in the Bragg wavelength due to axial strain is 0.73 ε or 73% of the applied strain. At 1550 and 1300 nm, the shifts in the resonance wavelength are 11 nm $\%\varepsilon^{-1}$ and 9 nm $\%\varepsilon^{-1}$, respectively. With temperature, an FBG at 1500 nm shifts by 1.6 nm for every 100 °C rise in temperature [9].

33.3.4 Limitations of Bragg-Grating Strain Sensors

The major limitation of Bragg-grating sensors is the complex and expensive fabrication technique. Although side writing is commonly used to manufacture these gratings, the requirement of expensive phase masks increases the cost of the sensing system. In the interferometric technique, stability of the setup is a critical factor in obtaining high-quality gratings. Since index changes of the order of 10^{-3} are required to fabricate these gratings, laser pulses of high energy levels are necessary. This might reduce laser operating lifetime and lead to increased maintenance expense. Additionally, introducing hydrogen or deuterium into the fiber allows increased index modulation as a result of the irradiation process.

The second major limitation of Bragg gratings is their limited bandwidth. The typical value of the full width at half maximum (FWHM) is between 0.1 and 1 nm. Although higher bandwidths potentially can be obtained by chirping the index or periodicity along the grating length, this adds to the cost of the grating fabrication. The limited bandwidth requires high-resolution spectrum analyzers to monitor the grating spectrum. Kersey et al. have proposed an unbalanced Mach–Zehnder interferometer to detect the perturbation-induced wavelength shift [14]. Two unequal arms of the Mach–Zehnder interferometer are excited by the back reflection from a Bragg-grating sensor element. Any change in the input optical wavelength modulates the phase difference between the two arms and results in a time-varying sinusoidal intensity at the output. This interference signal can be related to the shift in the Bragg peak, and hence, the magnitude of the perturbation can be obtained. Recently, modal interferometers have also been proposed to demodulate the output of a Bragg-grating sensor [15]. The unbalanced interferometers are also susceptible to external perturbations and hence need to be isolated from the parameter under investigation. Moreover, the nonlinear output might require fringe counting equipment, which can be complex and expensive. Additionally, a change in the perturbation polarity at the maxima or minima of the transfer function curve will not be detected by this demodulation scheme. To overcome this limitation, two unbalanced interferometers can be employed for dynamic measurements.

The cross sensitivity to temperature fluctuations leads to erroneous strain measurements in applications where the ambient temperature has a temporal variation. Thus, a reference grating that measures the temperature change must be utilized to compensate for the output of the strain sensor. Recently, temperature-independent sensing has been demonstrated using chirped gratings written in tapered optical fibers [16].

Last, the sensitivity of FBG strain sensors might not be adequate for certain applications. This sensitivity of the sensor depends on the minimum detectable wavelength shift at the detection end. Although excellent wavelength resolution can be obtained with unbalanced interferometric detection techniques, standard spectrum analyzers typically provide a resolution of 0.1 nm. At 1300 nm, this minimum detectable change in wavelength corresponds to a strain resolution of 111 $\mu\varepsilon$. Hence, in applications where strain smaller than 100 $\mu\varepsilon$ is anticipated, Bragg-grating sensors might not be practical. The dynamic range of strain measurement can be as much as 15,000 $\mu\varepsilon$.

33.4 Long-Period Grating Sensor

This section discusses the use of novel long-period gratings as strain sensing devices. The principle of operation of these gratings and their fabrication process, preliminary strain tests, demodulation process, and cross sensitivity to ambient temperature are analyzed.

33.4.1 Principle of Operation

Long-period gratings that couple the fundamental guided mode to different guided modes have been demonstrated in the past [17,18]. Gratings with longer periodicities that involve coupling of a guided mode to forward-propagating cladding modes were recently proposed by Vengsarkar et al. [19,20]. As stated previously, fiber gratings satisfy the Bragg phase-matching condition between the guided and the cladding or radiation modes or another guided mode. This wavelength-dependent phase-matching condition is given by

$$\beta_{01} - \beta = \Delta\beta = \frac{2\pi}{\Lambda} \tag{33.9}$$

where

Λ is the periodicity of the grating

β_{01} and β are the propagation constants of the fundamental guided mode and the mode to which coupling occurs, respectively

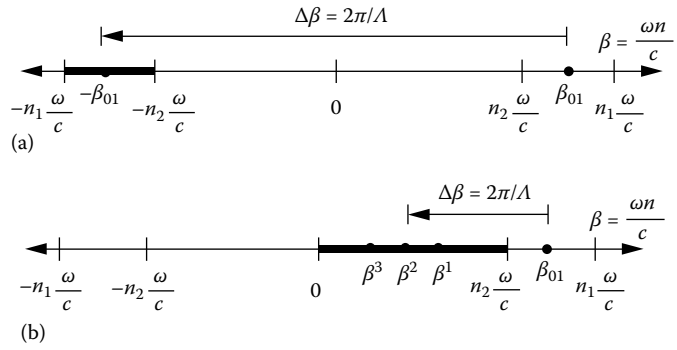


FIGURE 33.9 Blazed and unblazed cladding mode periodicities.

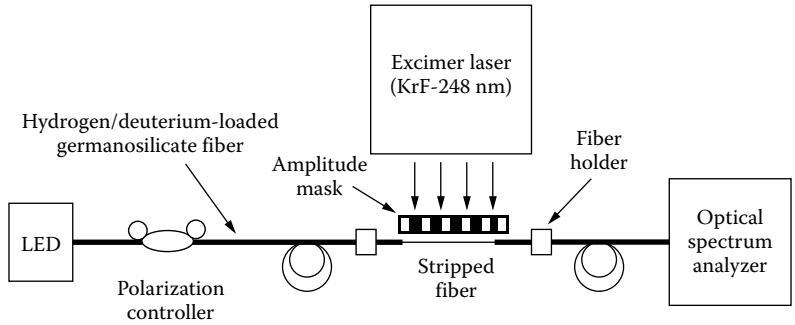


FIGURE 33.10 Setup used to fabricate long-period gratings.

For conventional FBGs, the coupling of the forward-propagating LP_{01} mode occurs to the reverse-propagating LP_{01} mode ($\beta = -\beta_{01}$). Since $\Delta\beta$ is large in this case (Figure 33.9a), the grating periodicity is small, typically of the order of 1 μm. Unblazed long-period gratings having index variations parallel to the long axis of the fiber couple the fundamental mode to the discrete and circularly symmetric, forward-propagating cladding modes ($\beta = \beta^n$), resulting in smaller values of $\Delta\beta$ (Figure 33.9b) and hence periodicities ranging in hundreds of micrometers [19]. The cladding modes attenuate rapidly as they propagate along the length of the fiber due to the lossy cladding-coating interface and bends in the fiber. Since $\Delta\beta$ is discrete and function of the wavelength, this coupling to the cladding modes is highly selective, leading to a wavelength-dependent loss. As a result, any modulation of the core and cladding guiding properties modifies the spectral response of long-period gratings, and this phenomenon can be utilized for sensing purposes. Moreover, since the cladding modes interact with the fiber jacket or any other material surrounding the cladding, changes in the properties of these ambient materials can also be detected. A typical example of such setup is illustrated in Figure 33.10.

33.4.2 Fabrication Procedure

To fabricate long-period gratings, hydrogen-loaded (3.4 mol%) germanosilicate fibers are exposed to 248 nm UV radiation from a KrF excimer laser, through a chrome-plated amplitude mask possessing a periodic rectangular transmittance function. Figure 33.10 shows the setup used to fabricate the gratings. The laser was pulsed at 20 Hz with an 8 ns pulse duration. The typical writing times for energy of 100 mJ cm^{-2} $pulse^{-1}$ and a 2.5 cm exposed length vary between 6 and 15 min for different fibers. The coupling wavelength, λ_P, shifts to higher values during exposure, due to the photoinduced enhancement of the refractive index of the fiber core and the resulting increase in β_{01}. After writing, the gratings

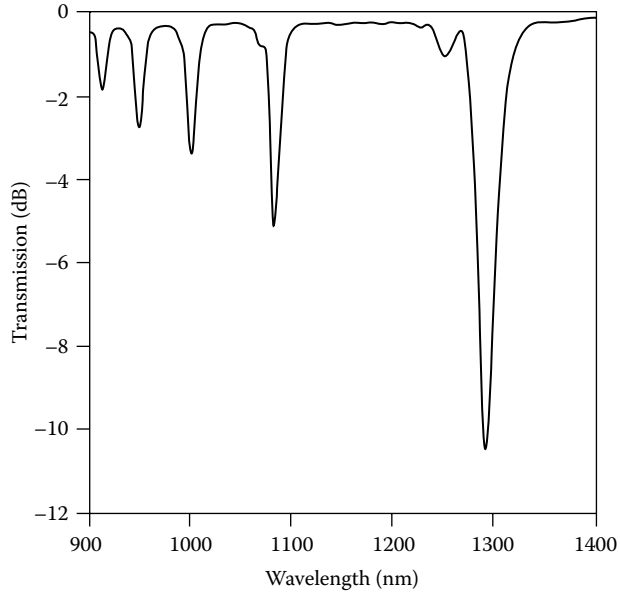


FIGURE 33.11 Typical transmission of a grating.

are annealed at 150 °C for 10 h to remove the unreacted hydrogen. This high-temperature annealing causes λ_p to move to shorter wavelengths due to the decay of UV-induced defects and diffusion of molecular hydrogen from the fiber. Figure 33.11 depicts the typical transmittance of a grating. Various attenuation bands correspond to coupling to discrete cladding modes of different orders. A number of gratings can be fabricated at the same time by placing more than one fiber behind the amplitude mask. Moreover, the stability requirements during the writing process are not as severe as those for short-period Bragg gratings.

For coupling to the highest-order cladding mode, the maximum isolation (loss in transmission intensity) is typically in the 5–20 dB range on wavelengths, depending on fiber parameters, duration of UV exposure, and mask periodicity. The desired fundamental coupling wavelength can easily be varied using inexpensive amplitude masks of different periodicities. The insertion loss, the polarization-mode dispersion, the back reflection, and the polarization-dependent loss of a typical grating are 0.2 dB, 0.01 ps, −80 dB, and 0.02 dB, respectively. The negligible polarization sensitivity and the back reflection of these devices eliminate the need for expensive polarizers and isolators.

Preliminary experiments were performed to examine the strain sensitivity of long-period gratings written in different fibers [21,22]. Gratings were fabricated in four different types of fibers: standard dispersion-shifted fiber (DSF), standard 1550 nm fiber, and conventional 980 and 1050 nm single-mode fibers, which for the sake of brevity are referred to as fibers A, B, C, and D, respectively. The strain sensitivity of gratings written in different fibers was determined by axially straining the gratings between two longitudinally separated translation stages. The shift in the peak loss wavelength of the grating in fiber D as a function of the applied strain is depicted in Figure 33.12, along with that for a Bragg grating (about 9 nm $\%\varepsilon^{-1}$, at 1300 nm) [9]. The strain coefficients of wavelength shift (β) for fibers A, B, C, and D are shown in Table 33.1. Fiber D has a coefficient 15.2 nm $\%\varepsilon^{-1}$ that gives it a strain-induced shift that is 50% larger than that for a conventional Bragg grating. The strain resolution of this fiber for a 0.1 nm detectable wavelength shift is 65.75 $\mu\varepsilon$.

The demodulation scheme of a sensor determines the overall simplicity and sensitivity of the sensing system. Short-period Bragg-grating sensors were shown to possess signal processing techniques that are complex and expensive to implement. A simple demodulation method to extract

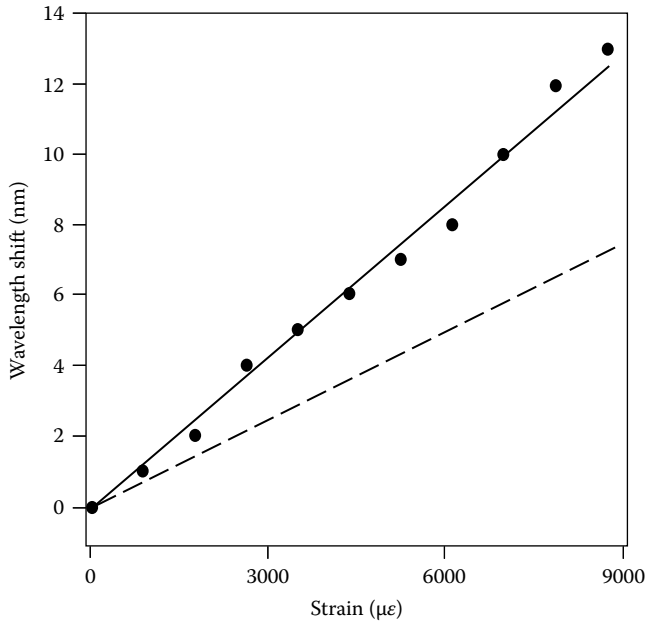


FIGURE 33.12 Shift in peak loss wavelength as a function of the applied strain.

TABLE 33.1 Strain Sensitivity of Long-Period Gratings Written in Four Different Types of Fibers

Type of Fiber	Strain Sensitivity (nm% ε^{-1})
A—DSF	−7.27
B—Standard 1550 nm communication fiber	4.73
C—Conventional 980 nm single-mode fiber	4.29
D—Conventional 1060 nm single-mode fiber	15.21

Note: The values correspond to the shift in the highest-order resonance wavelength.

information from long-period gratings is possible. The wide bandwidth of the resonance bands enables the wavelength shift due to the external perturbation to be converted into an intensity variation that can be easily detected.

Figure 33.13 shows the shift induced by strain in a grating written in fiber C. The increase in the loss at 1317 nm is about 1.6 dB. A laser diode centered at 1317 nm was used as the optical source, and the change in transmitted intensity was monitored as a function of applied strain. The transmitted intensity is plotted in Figure 33.14 for three different trials. The repeatability of the experiment demonstrates the feasibility of using this simple scheme to utilize the high sensitivity of long-period gratings. The transmission of a laser diode centered on the slope of the grating spectrum on either side of the resonance wavelength can be used as a measure of the applied perturbation. A simple detector and amplifier combination at the output can be used to determine the transmission through the detector. On the other hand, a broadband source can also be used to interrogate the grating. At the output, an optical band-pass filter can be used to transmit only a fixed bandwidth of the signal to the detector. The band-pass filter should again be centered on either side of the peak loss band of the resonance band. These schemes are easy to implement, and unlike conventional Bragg gratings, the requirement of complex and expensive interferometric demodulation schemes is not necessary [22].

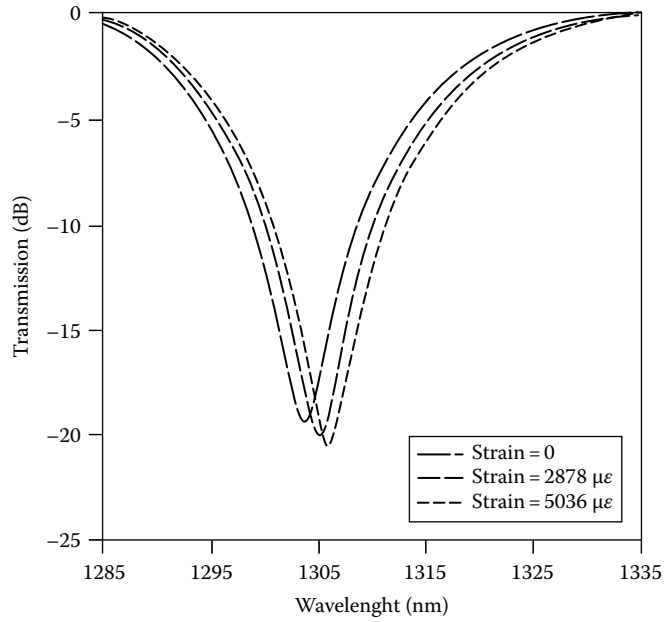


FIGURE 33.13 Shift induced by strain in a grating written in fiber C.

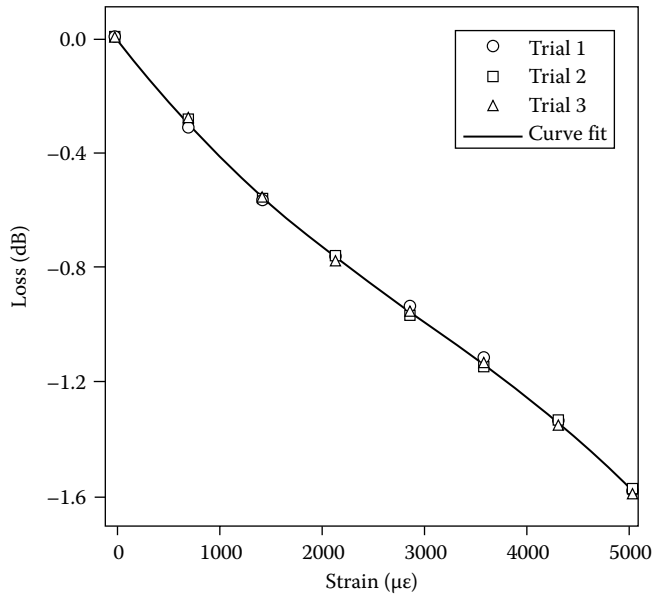


FIGURE 33.14 Plot of the change in transmitted intensity as a function of strain, for three different trials.

33.4.3 Temperature Sensitivity of Long-Period Gratings

Gratings written in different fibers were also tested for their cross sensitivity to temperature [22]. The temperature coefficients of wavelength shift for different fibers are shown in Table 33.2. The temperature sensitivity of an FBG is 0.014 nm $°C^{-1}$. Hence, the temperature sensitivity of a long-period grating is typically an order of magnitude higher than that of a Bragg grating. This large cross sensitivity to ambient temperature can degrade the strain sensing performance of the system unless the output signal

TABLE 33.2 Temperature Sensitivity of Long-Period Gratings Written in Four Different Types of Fibers

Type of Fiber	Temperature Sensitivity (nm °C^{-1})
A—DSF	0.062
B—Standard 1550 nm communication fiber	0.058
C—Conventional 980 nm single-mode fiber	0.154
D—Conventional 1060 nm single-mode fiber	0.111

Note: The values correspond to the shift in the highest-order resonance wavelength.

is adequately compensated. Multiparameter sensing using long-period gratings has been proposed to obtain precise strain measurements in environments with temperature fluctuations [21].

In summary, long-period grating sensors are highly versatile. These sensors can easily be used in conjunction with simple and inexpensive detection techniques. Experimental results prove that these methods can be used effectively without sacrificing the enhanced resolution of the sensors. Long-period grating sensors are insensitive to the input polarization and do not require coherent optical sources. The cross sensitivity to temperature is a major concern while using these gratings for strain measurements.

33.5 Comparison of Sensing Schemes

Based on the aforementioned results, the interferometric sensors have a high sensitivity and bandwidth but are limited by the nonlinearity in their output signals. Conversely, intrinsic sensors are susceptible to ambient temperature changes, while the grating-based sensors are simpler to multiplex. Each may be used in specific applications.

33.6 Conclusion

We have investigated the performance of four different interferometric and grating-based sensors. This analysis was based on the sensor-head fabrication and cost, signal processing, cross sensitivity to temperature, resolution, and operating range. The relative merits and demerits of the various sensing schemes were also discussed.

References

1. J. Sirkis, Phase-strain-temperature model for structurally embedded interferometric optical fiber strain sensors with applications, *Fiber Optic Smart Structures and Skins IV*, Boston, MA, *Proc. SPIE*, 1588, 26–43, 1991.
2. R. O. Claus, M. F. Gunther, A. Wang, and K. A. Murphy, Extrinsic Fabry-Perot sensor for strain and crack opening displacement measurements from −200 to 900°C, *J. Smart Mater. Struct.*, 1, 237–242, 1992.
3. K. A. Murphy, M. F. Gunther, A. M. Vengsarkar, and R. O. Claus, Fabry-Perot fiber optic sensors in full-scale fatigue testing on an F-15 aircraft, *Appl. Opt.*, 31, 431–433, 1991.
4. V. Bhatia, C. A. Schmid, K. A. Murphy, R. O. Claus, T. A. Tran, J. A. Greene, and M. S. Miller, Optical fiber sensing technique for edge-induced and internal delamination detection in composites, *J. Smart Mater. Struct.*, 4, 164–169, 1995.
5. V. Bhatia, M. J. de Vries, K. A. Murphy, R. O. Claus, T. A. Tran, and J. A. Greene, Extrinsic Fabry-Perot interferometers for absolute measurements, *Fiberopt. Prod. News*, 9(December), 12–13, 1994.
6. V. Bhatia, M. B. Sen, K. A. Murphy, and R. O. Claus, Wavelength-tracked white light interferometry for highly sensitive strain and temperature measurements, *Electron. Lett.*, 32, 247–249, 1996.

7. C. E. Lee and H. F. Taylor, Fiber-optic Fabry-Perot temperature sensor using a low-coherence light source, *J. Lightwave Technol.*, 9, 129–134, 1991.
8. J. A. Greene, T. A. Tran, K. A. Murphy, A. J. Plante, V. Bhatia, M. B. Sen, and R. O. Claus, Photoinduced Fresnel reflectors for point-wise and distributed sensing applications, *Proceedings of the Conference on Smart Structures and Materials, SPIE'95*, San Jose, CA, paper 2444-05, February 1995.
9. K. O. Hill, Y. Fuijii, D. C. Johnson, and B. S. Kawasaki, Photosensitivity in optical fiber waveguides: Applications to reflection filter fabrication, *Appl. Phys. Lett.*, 32, 647, 1978.
10. G. Meltz, W. W. Morey, and W. H. Glenn, Formation of Bragg gratings in optical fibers by transverse holographic method, *Opt. Lett.*, 14, 823, 1989.
11. P. J. Lemaire, A. M. Vengsarkar, W. A. Reed, V. Mizrahi, and K. S. Kranz, Refractive index changes in optical fibers sensitized with molecular hydrogen, *Proceedings of the Conference on Optical Fiber Communications, OFC'94*, San Jose, CA, Technical Digest, paper TuLl, 47, 1994.
12. R. Kashyap, Photosensitive optical fibers: Devices and applications, *Opt. Fiber Technol.*, 1, 17–34, 1994.
13. D. Z. Anderson, V. Mizrahi, T. Ergodan, and A. E. White, Phase-mask method for volume manufacturing of fiber phase gratings, *Proceedings of the Conference on Optical Fiber Communication*, San Jose, CA, post-deadline paper PD16, 1993, p. 68.
14. A. D. Kersey and T. A. Berkoff, Fiber-optic Bragg-grating differential-temperature sensor, *IEEE Photonics Technol. Lett.*, 4, 1183–1185, 1992.
15. V. Bhatia, M. B. Sen, K. A. Murphy, A. Wang, R. O. Claus, M. E. Jones, J. L. Grace, and J. A. Greene, Demodulation of wavelength-encoded optical fiber sensor signals using fiber modal interferometers, *SPIE Photonics East*, Philadelphia, PA, paper 2594-09, October 1995.
16. M. G. Xu, L. Dong, L. Reekie, J. A. Tucknott, and J. L. Cruz, Chirped fiber gratings for temperature-independent strain sensing, *Proceedings of the. First OSA Topical Meeting, Photosensitivity and Quadratic Non-linearity in Glass Waveguides: Fundamentals and Applications*, Monterey, CA, paper PMB2, 1995.
17. K. O. Hill, B. Malo, K. Vineberg, F. Bilodeau, D. Johnson, and I. Skinner, Efficient mode-conversion in telecommunication fiber using externally written gratings, *Electron. Lett.*, 26, 1270–1272, 1990.
18. F. Bilodeau, K. O. Hill, B. Malo, D. Johnson, and I. Skinner, Efficient narrowband $LP_{01} \leftrightarrow LP_{02}$ mode convertors fabricated in photosensitive fiber: Spectral response, *Electron. Lett.*, 27, 682–684, 1991.
19. A. M. Vengsarkar, P. J. Lemaire, J. B. Judkins, V. Bhatia, J. E. Sipe, and T. E. Ergodan, Long-period fiber gratings as band-rejection filters, *Proceedings of the Conference on Optical Fiber Communications, OFC '95*, Washington, DC, post-deadline paper, PD4-2, 1995.
20. A. M. Vengsarkar, P. J. Lemaire, J. B. Judkins, V. Bhatia, J. E. Sipe, and T. E. Ergodan, Long-period fiber gratings as band-rejection filters, *J. Lightwave Technol.*, 14, 58–65, 1996.
21. V. Bhatia, M. B. Burford, K. A. Murphy, and A. M. Vengsarkar, Long-period fiber grating sensors, *Proceedings of the Conference on Optical Fiber Communication*, San Jose, CA, paper ThPl, February 1996.
22. V. Bhatia and A. M. Vengsarkar, Optical fiber long-period grating sensors, *Opt. Lett.*, 21, 692–694, 1996.
23. C. D. Butter and G. B. Hocker, Fiber optics strain gage, *Appl. Opt.*, 17, 2867–2869, 1978.
24. J. S. Sirkis and H. W. Haslach, Interferometric strain measurement by arbitrarily configured, surface mounted, optical fiber, *J. Lightwave Technol.*, 8, 1497–1503, 1990.

Further Readings

Crisp, J. and Elliott, B., *Introduction to Fiber Optics*, 3rd edn., Elsevier, Oxford, U.K., 2005.

Downing, J. N., *Fiber Optic Communication*, Thomson Delmar, Clifton Park, NY, 2005.

Elsevier, Recent optical fiber technology articles. http://www.journals.elsevier.com/optical-fiber-technology/recent-articles/ (accessed on August 2, 2012).

Hecht, J., *Understanding of Fiber Optics*, 5th edn., Prentice Hall, Upper Saddle Drive, NJ, 2005.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

34

Optical Beam Deflection Sensors

34.1 Theory **34**-2
Knife-Edge Photodetector • Bicell Detector • Continuous Position Sensor
34.2 Characterization of PSDs **34**-6
34.3 Summary **34**-9
References **34**-9
Further Readings **34**-10

Grover C. Wetsel
University of Texas at Dallas

Measurements of the intensity of the light reflected and transmitted by a sample have been sources of information concerning the structure of matter for over a century. In recent decades, it has been found that measurement of the position of an optical beam that has scattered from a sample is an important and versatile means of characterizing materials and the motion of devices. Surely, the availability of a well-collimated beam from a laser has been crucial in the development of techniques and applications of *optical beam deflection* (OBD) sensing; however, the development and ready availability of various types of *position-sensing detectors* (PSDs) have also been important factors. OBD may be caused, for example, by propagation of a laser beam through a refractive-index gradient or by reflection from a displaced surface. A PSD provides an electronic signal that is a function of the laser beam position on the detector.

In this section, applications of OBD sensing are reviewed, the theories of operation of the three most common types of OBD sensors are developed, and typical operational characteristics of the devices are presented. The advantages and disadvantages of the various PSDs are also discussed.

A schematic diagram of the basic OBD sensing configuration is illustrated in Figure 34.1. In this case, the displacement, u, of the surface causes the position of the reflected beam on the PSD to move from point P to point P′; the positional change produces a change in the output voltage of the PSD. The output voltage, V, of the PSD electronics can be calibrated in terms of the actual displacement, u, by measuring V versus u for known displacements.

OBD sensing has been used in a variety of applications, including photothermal optical beam deflection (PTOBD) spectroscopy [1], absolute measurement of optical attenuation [2], PTOBD imaging of surface and subsurface structure [3], photothermal displacement spectroscopy [4], atomic-force microscopy [5], and materials characterization [6]. It has also been used as an uncomplicated, sensitive, and accurate method of measurement of surface motion for scanning tunneling microscope scanner transducers [7] and ultrasonic transducer imaging [8].

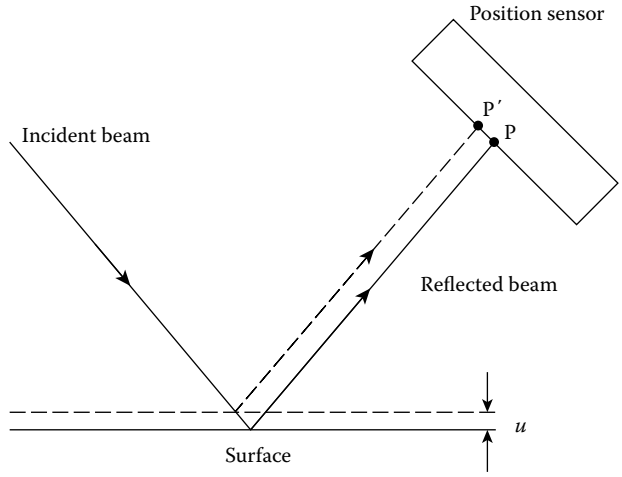


FIGURE 34.1 A schematic diagram of the basic OBD sensing configuration.

34.1 Theory

The three basic types of devices for OBD sensing are (1) a photodetector behind a sharply edged screen (a knife edge), (2) a small array of photodetectors separated by relatively small, insensitive areas (bicell, quadcell), and (3) a continuous solid-state position sensor (1-D or 2-D). Sensing characteristics of a device are determined by the effect of optical beam displacement on the photodetector power distribution. Since laser beams are commonly used in OBD sensing, the analysis involves the assumption that the spatial distribution of the intensity (I) in the plane perpendicular to the direction of wave propagation is axially symmetric with a Gaussian radial variation.

34.1.1 Knife-Edge Photodetector

The essential features of a PSD are represented by a photodetector shadowed by a semi-infinite knife edge, $y < 0$, as illustrated in Figure 34.2. As can be anticipated from the symmetry of the arrangement and proved mathematically, the maximal deflection sensitivity occurs when the undeflected beam is centered on the knife edge. The intensity of the light reaching the photodetector due to the displacement (u) of the center of the beam is given in the reference frame of the displaced beam by

$$I(r') = \frac{aP}{\pi} e^{-ar'^2} \tag{34.1}$$

where

P is the total incident beam power
$a = 2/r_1^2$
r_1 is the Gaussian beam radius
$r'^2 = x'^2 + y'^2$

In terms of the coordinates (x, y) of the undeflected beam, the rectangular coordinates of the deflected beam are $x' = x$ and $y' = y - u$. The power (P_d) on the detector is thus given by

$$P_d = \frac{aP}{\pi} \int_0^\infty e^{-a(y-u)^2} dy \int_{-\infty}^{\infty} e^{-ax^2} dx = \frac{P}{2}\left[1 + erf\left(\sqrt{2}\frac{u}{r_1}\right)\right] \tag{34.2}$$

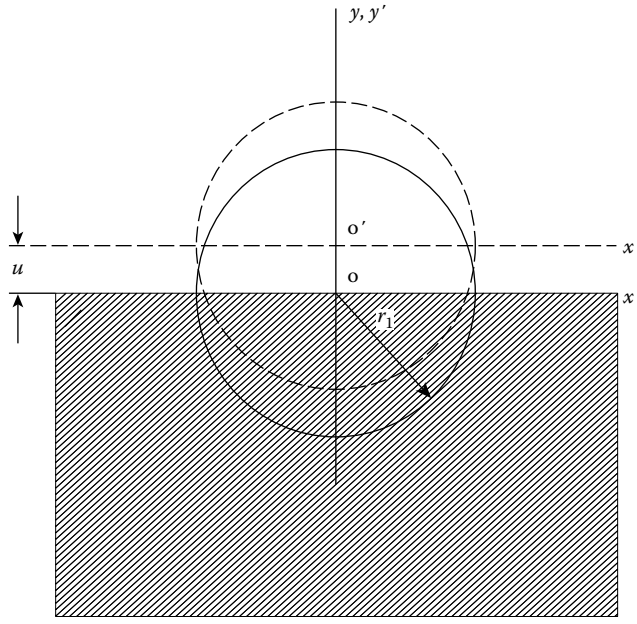


FIGURE 34.2 Essential features of a position-sending detector (PSD), as represented by a photodetector shadowed by a semi-infinite knife edge.

where *erf* is the error function. One can see by inspection of Equation 34.2 that the essential characteristics of this position sensor are determined by u/r_1. The normalized response, P_d/P, is shown in Figure 34.3 as a function of u/r_1. When $u = r_1$, then $P_d = 97.7\%\ P$.

The deflection sensitivity is given by the slope of Equation 34.2:

$$\frac{dP_d}{du} = \sqrt{\frac{2}{\pi}}\frac{P}{r_1}e^{-2(u/r_1)^2},\ \text{with}\left(\frac{dP_d}{du}\right)_{max} = \left(\frac{dP_d}{du}\right)_{u=0} = \sqrt{\frac{2}{\pi}}\frac{P}{r_1} \tag{34.3}$$

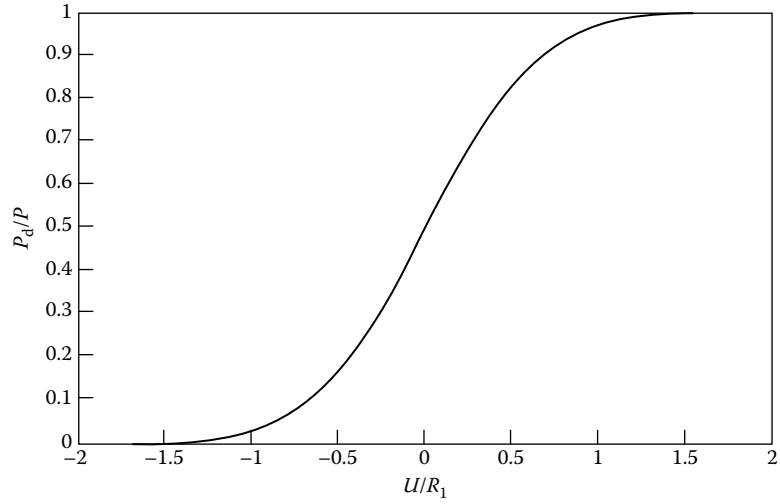


FIGURE 34.3 The normalized response P_d/P as a function of u/r_1. u, displacement of center of beam; r_1, Gaussian beam radius; P_d, power on the detector; P, total incident beam power.

Define the small-signal position sensor sensitivity (units of m^{-1}):

$$\alpha_{KE} \equiv \frac{1}{P}\left(\frac{dP_d}{du}\right)_{u=0} = \frac{1}{r_1}\sqrt{\frac{2}{\pi}} \tag{34.4}$$

The optical power reaching the photodetector for small signals is then given by

$$P_d \cong \frac{P}{2} + \left(\frac{dP_d}{du}\right)_{u=0} u = \frac{P}{2} + \alpha u P = \frac{P}{2}[1 + 2\alpha u] \tag{34.5}$$

The photodetector signal will be linear in displacement if $u \leq 0.387 r_1$.

The photocurrent is

$$I = KP_d \cong \frac{KP}{2}[1 + 2\alpha u] \tag{34.6}$$

where K is the photodetector responsivity in A W^{-1}. The position sensor voltage is obtained using a transimpedance amplifier with gain Z:

$$V = KZP_d \cong \frac{KZP}{2}[1 + 2\alpha u] \tag{34.7}$$

34.1.2 Bicell Detector

The deflection of a Gaussian beam initially centered in the insensitive gap of a bicell detector is illustrated in Figure 34.4. The power incident on the upper half of the bicell is given by

$$P_2 = \frac{aP}{\pi}\int_{y_2}^{\infty} e^{-a(y-u)^2} dy \int_{-\infty}^{\infty} e^{-ax^2} dx = \frac{P}{2}\left[1 - erf\left(\frac{\sqrt{2}}{r_1}(y_2 - u)\right)\right] \tag{34.8}$$

The power incident on the lower half of the bicell is given by

$$P_1 = \frac{aP}{\pi}\int_{-\infty}^{-y_1} e^{-a(y-u)^2} dy \int_{-\infty}^{\infty} e^{-ax^2} dx = \frac{P}{2}\left[1 - erf\left(\frac{\sqrt{2}}{r_1}(y_1 + u)\right)\right] \tag{34.9}$$

The photocurrent from each detector of the bicell is converted to voltage by identical transimpedance amplifiers: $V_2 = KZP_2$ and $V_1 = KZP_1$; a difference amplifier is then used to obtain the bicell signal voltage:

$$V = V_2 - V_1 = KZ(P_2 - P_1) = \frac{KZP}{2}\left[erf\left(\frac{\sqrt{2}}{r_1}(y_1 + u)\right) - erf\left(\frac{\sqrt{2}}{r_1}(y_2 - u)\right)\right] \tag{34.10}$$

The normalized response, $2V/(KZP)$, is shown in Figure 34.5 as a function of u/r_1 for $y_1 = y_2 = r_1/10$.

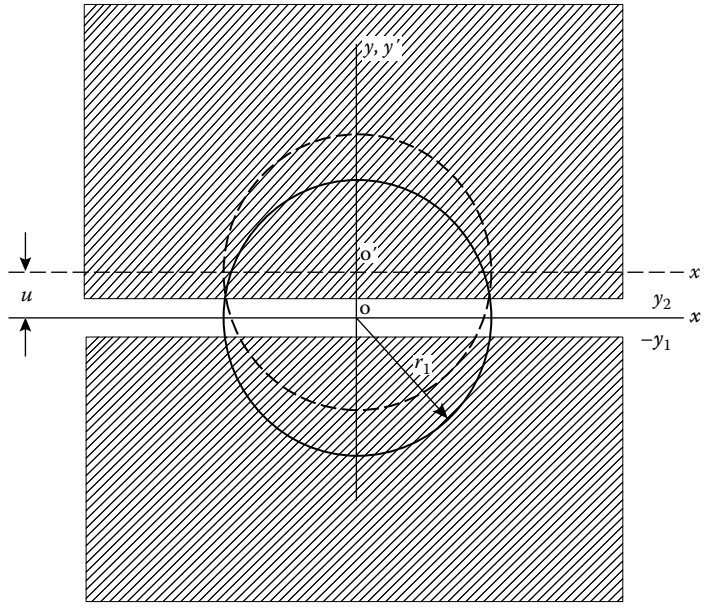


FIGURE 34.4 Deflection of a Gaussian beam initially centered in the insensitive gap of a bicell detector.

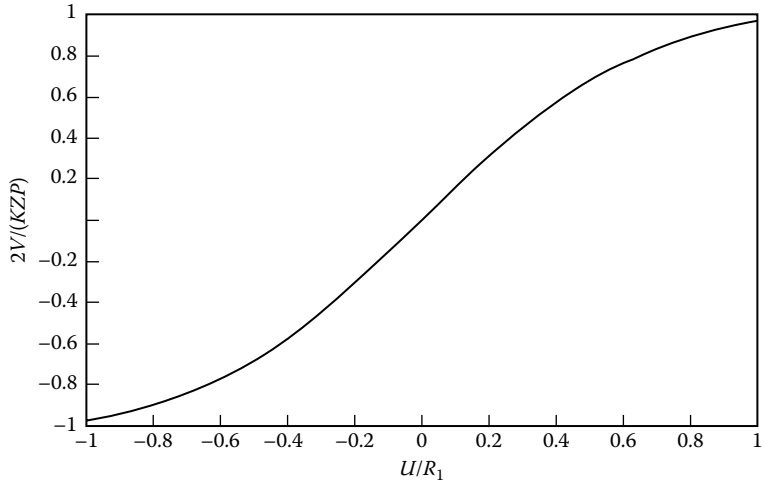


FIGURE 34.5 The normalized response $2V/KZP$ in a bicell detector as a function of u/r_1 for $y_1 = y_2 = r_1/10$.

Suppose that the beam is centered in the gap, $y_1 = y_2$; then, for small displacements, one obtains

$$V \cong 2\sqrt{\frac{2}{\pi}}\frac{KZPu}{r_1}e^{-2(y_1/r_1)^2} \tag{34.11}$$

The small-signal sensitivity is

$$\alpha_{BC} \equiv \frac{1}{KZP}\left(\frac{dV}{du}\right)_{u=0} = 2\sqrt{\frac{2}{\pi}}\frac{e^{-2(y_1/r_1)^2}}{r_1} \tag{34.12}$$

This quantity is optimized when $r_1 = 2y_1$ and the optimal sensitivity is $0.484/y_1$.

34.1.3 Continuous Position Sensor

The position information in a continuous position sensor (also known as a lateral-effect photodiode) is derived from the divided path taken by photon-generated electrons to two back electrodes on the device. For a homogeneous device, the current to an electrode depends only on the distance of the centroid of the light beam from that electrode; the currents would be equal in an ideal device when the beam is located at its electrical center.

Consider the analysis of a 1-D continuous PSD. The current signal from each electrode is converted to a voltage signal by a transimpedance amplifier with gain, Z. Operational amplifiers are then used to provide the sum signal, $V_s = KPZA_s$, and the difference signal, $V_d = KPZA_d\alpha u$, where A_s and A_d are the sum and difference amplifier gains, respectively. The position sensor sensitivity, α, is then given by

$$\alpha \equiv \frac{V_d A_s}{u V_s A_d} \tag{34.13}$$

which is determined by measuring V_d and V_s as a function of u.

34.2 Characterization of PSDs

The PSD characteristics presented here were measured by mounting the device to a translation stage with an optical encoder driven by a piezoelectric motor [9]; the position accuracy was ±0.1 μm. An appropriately attenuated He–Ne laser beam was directed at normal incidence to the PSD as it was translated past the beam. The PSD sum and difference voltages were measured with GP–IB digital voltmeters as a function of displacement, and the PSD displacements and voltage measurements were computer controlled.

The operation of a knife-edge PSD can be evaluated using the signal from one side of a bicell as well as from a photodetector behind a knife edge. The transimpedance amplifier output voltage corresponding to the current signal from one cell of a United Detector Technology [10] (UDT) SPOT2D bicell is shown as a function of displacement in Figure 34.6. The data are plotted as normalized cell voltage as a function of u/r_1. Since the signal from one cell of a bicell is equivalent to a knife-edge photodetector, the data can be compared with Figure 34.3. The data are in reasonable agreement for $u/r_1 < 1$, and a best fit allows the inference of $r_1 = 0.4$ mm; the disagreement for $u/r_1 \geq 1$ corresponds to the laser beam partially moving off the outside edge of the cell.

The operation of a bicell PSD is evaluated by the data of Figure 34.7, which shows the difference voltage as a function of displacement for the UDT SPOT2D. The deviations from the theoretical curve of

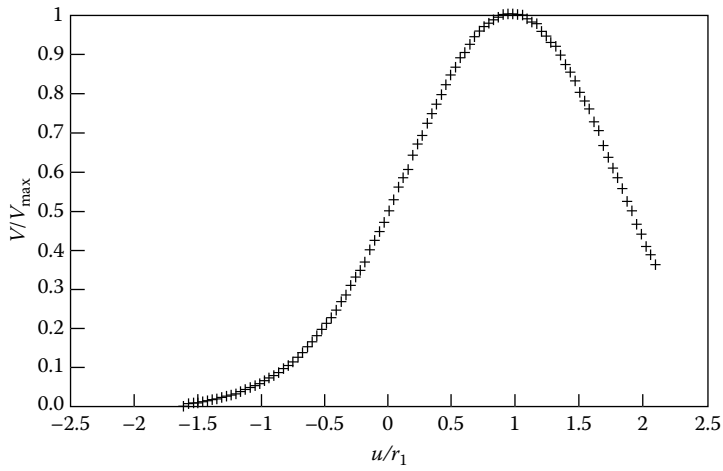


FIGURE 34.6 Transimpedance amplifier output voltage corresponding to the current signal from one cell of a UDT SPOT2D bicell as a function of displacement. Data are plotted as normalized cell voltage versus u/r_1.

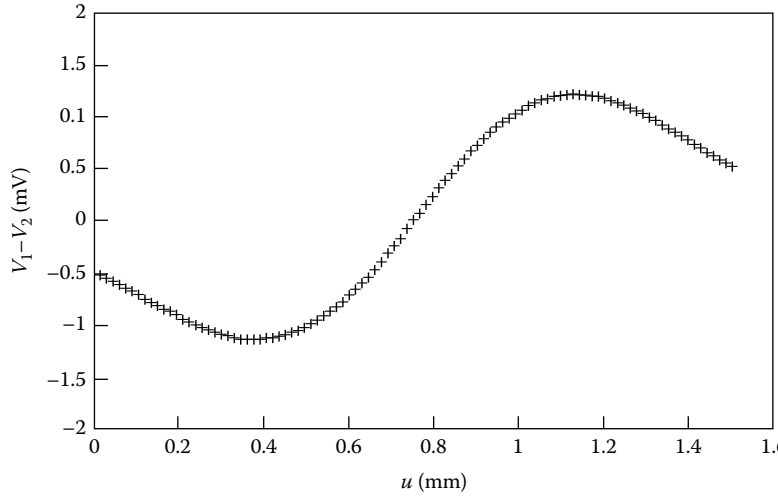


FIGURE 34.7 Difference voltage as a function of displacement for the UDT SPOT2D bicell.

Figure 34.5 for $u \leq 0.4$ mm and $u \geq 1.1$ mm are due to the beam moving off the outside edges of the bicell. The linear region of the PSD is centered at the optimal quiescent point ($V_1 - V_2 = 0$); a least-squares fit of a straight line gives a slope of $A_{12} = 4.88$ mV μm^{-1} for this device. For the data of Figures 34.6 and 34.7, $P = 0.37$ mW and $Z = 5$ kΩ; thus, the experimental value of the responsivity is determined to be $K_{exp} = 0.65$ A W^{-1}. The nominal gap size of the SPOT2D bicell is such that $y_1 = 63.5$ μm; thus, using this value of y_1 and $r_1 = 0.4$ mm in Equation 34.12, the calculated bicell sensitivity is 3.8×10^3 m^{-1}. The experimental value of sensitivity is calculated from A_{12}, K_{exp}, P, and Z to be 4.1×10^3 m^{-1}.

Evaluation of the operation of the x-axis of a UDT SC10 two-axis continuous PSD is illustrated in Figures 34.8 through 34.10. The PSD was oriented such that the y-axis was vertical; thus, ideally, V_{dy} is constant as the PSD is translated, except near the edges of the 10 mm × 10 mm device. As shown in Figure 34.8, V_{sx} is virtually constant except near the edges of the device; this means that the optical responsivity (K) does not greatly vary with the laser beam position on this detector. Variation of V_{dx} with displacement is shown in Figure 34.9; this PSD has a broad linear range. The linearity of continuous PSDs can often be improved by using the ratio, V_{dx}/V_{sx}, as shown in Figure 34.10; the sensitivity of this PSD is determined from a least-squares fit of the linear part of this characteristic to be 183 m^{-1}.

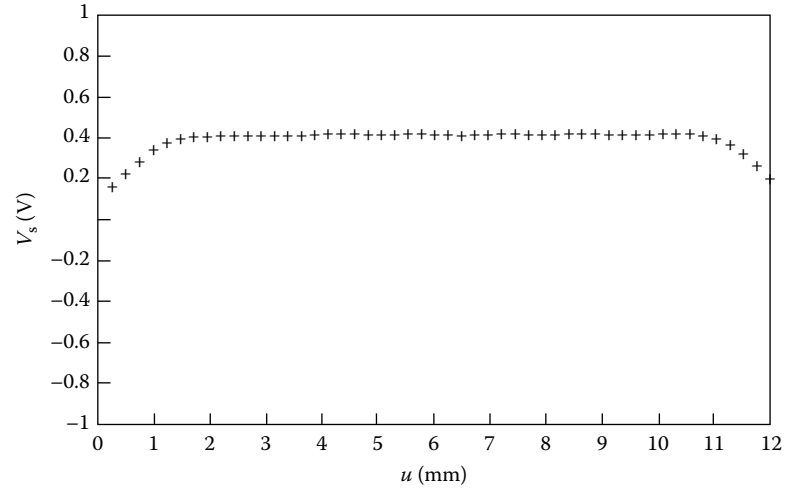


FIGURE 34.8 Sum voltage versus displacement for x-axis of UDT SCIO continuous PSD.

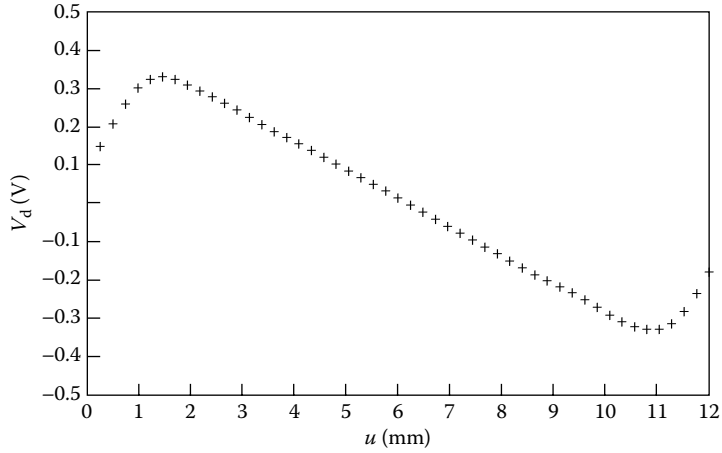


FIGURE 34.9 Difference voltage versus displacement for *x*-axis of UDT SCIO continuous PSD.

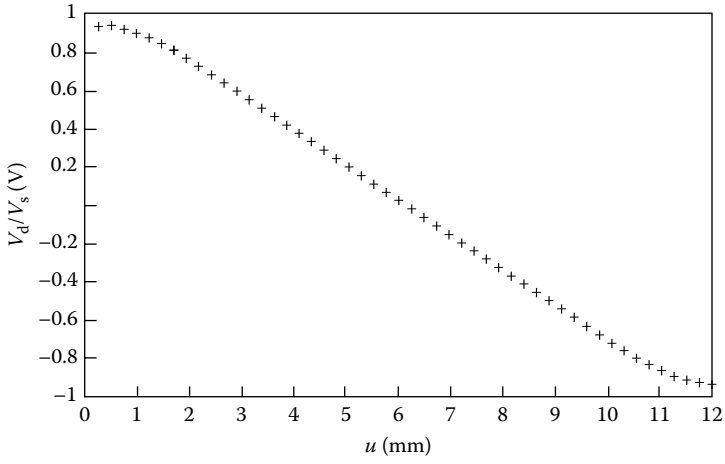


FIGURE 34.10 Ratio of difference voltage to sum voltage versus displacement for *x*-axis of UDT SCIO continuous PSD.

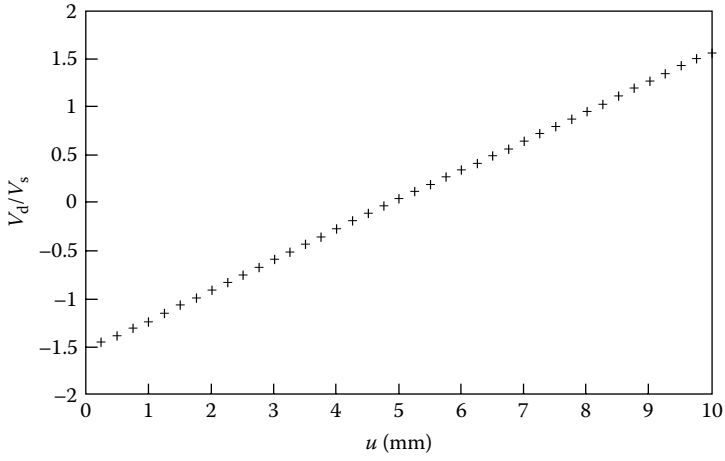


FIGURE 34.11 Ratio of difference voltage to sum voltage versus displacement for SiTek 1L10 continuous PSD.

A newer class of PSDs is represented by the SiTek [11] 1L10 single-axis continuous PSD, which is typically more linear. An example of the V_d/V_s versus u characteristic is shown in Figure 34.11. The sensitivity is determined from a least-squares fit of the linear part of the characteristic to be 307 m^{-1}.

34.3 Summary

A PSD composed of a knife-edge photodetector is simple, has a sensitivity that depends only on the radius of the laser beam, and has a response time determined by the photodetector. Thus, for fast rise-time detection of small displacements, this type of PSD is to be preferred.

The maximal sensitivity of a bicell is about 20% greater than that of a knife-edge photodetector. It has the advantages of small size and economy. The nominal risetime of the UDT SPOT2D is of the order of 10 ns. Furthermore, two-axis detection can be readily obtained with a quadcell, with an increase in risetime due to the increased capacitance of the detector.

The disadvantages due to small laser beam sizes and small displacements are overcome by the continuous PSDs. Since these devices are typically much larger (available in several-inch diameters), they typically have longer risetimes than the other PSDs. However, the Sitek 1L10, with a 10 mm linear active range, has a measured upper half-power frequency of 3 MHz.

In applications where the output signal from a PSD must be linearly proportional to the displacement of the beam, analog-divider operational amplifiers to obtain V_d/V_s in real time are used to extend the range of linearity of the device. Unfortunately, the frequency response of these amplifiers is often the frequency-response-limiting factors of the PSD system. In cases where high-frequency response is important, V_d alone can often be used if care is taken to operate in the linear range of the device. For large static displacements that are of the order of the size of the detector, V_d and V_s can be recorded with computer-controlled data acquisition, the calibration characteristic numerically fitted to a polynomial, and then any voltage from the detector can be related to beam position.

The noise limitations in OBD sensing are due to the laser, the nature of the reflecting surface, and the PSD. Lasers with good amplitude stability are to be preferred, but this is not an important contribution to noise when V_d/V_s is used to infer displacement. Laser beam-pointing stability, on the other hand, is important. If the reflecting surface is that of a typical solid, then negligible noise is introduced on reflection; this may not be true for a reflector such as a pellicle, where Brownian motion of the surface may be significant. The noise limitations of the PSD are the usual ones associated with the photodetector and the amplifiers.

References

1. A. C. Boccara, D. Fournier, and J. Badoz, *Appl. Phys. Lett.*, 30, 933, 1983.
2. G. C. Wetsel, Jr. and S. A. Stotts, *Appl. Phys. Lett.*, 42, 931, 1983.
3. D. Fournier and A. C. Boccara, In *Scanned Image Microscopy*, E. A. Ash (Ed.), London, U.K.: Academic Press, 1980, pp. 347–351; J. C. Murphy and L. C. Aaamodt, *Appl. Phys. Lett.*, 39, 519, 1981; G. C. Wetsel, Jr. and F. A. McDonald, *Appl. Phys. Lett.*, 41, 926, 1982.
4. M. A. Olmstead, S. Kohn, N. M. Amer, D. Fournier, and A. C. Boccara, *Appl. Phys. A*, 132, 68, 1983.
5. G. Meyer and N. M. Amer, *Appl. Phys. Lett.*, 53, 1045, 1988.
6. J. C. Murphy and G. C. Wetsel, Jr., *Mater. Eval.*, 44, 1224, 1986.
7. G. C. Wetsel, Jr., S. E. McBride, R. J. Warmack, and B. Van de Sande, *Appl. Phys. Lett.*, 55, 528, 1989.
8. S. E. McBride and G. C. Wetsel, Jr., Surface-displacement imaging using optical beam deflection, In *Review of Progress in Quantitative Nondestructive Evaluation*, Vol. 9A, D. O. Thompson and D. E. Chimenti (Eds.), New York: Plenum, 1990, pp. 909–916.
9. Burleigh Instruments, Inc., Fishers, NY.
10. United Detector Technology, Hawthorne, CA.
11. On-Trak Photonics, Inc., Lake Forest, CA.

Further Readings

Alwyn, V., *Fiber-Optic Technologies*, Cisco Press, Indianapolis, IN, April 2004. http://www.ciscopress.com/articles/article.asp?p=170740 (accessed on July 22, 2012).

Crisp, J. and Elliott, B., *Introduction to Fiber Optics*, 3rd edn., Elsevier, Oxford, U.K., 2005.

Elsevier, Recent optical fiber technology articles. http://www.journals.elsevier.com/optical-fiber-technology/recent-articles/ (accessed on August 2, 2012).

Hecht, J., *Understanding of Fiber Optics*, 5th edn., Prentice Hall, Upper Saddle Drive, NJ, 2005.

Oliver, B., An introduction to fiber optic technology, Ezine Articles. http://ezinearticles.com/?An-Introduction-to-Fiber-Optic-Technology&id=1363984 (accessed on July 23, 2012).

35

Velocity Measurement

35.1 Introduction **35**-1
35.2 Measurement of Linear Velocity **35**-2
35.3 Reference-Based Linear Velocity Measurement **35**-3
Conversion of Linear to Rotational Velocity • Doppler Shift
35.4 Light Interference Methods for Linear Velocity Measurement ... **35**-5
VISAR System
35.5 Seismic Devices for Linear Velocity **35**-8
35.6 Angular Velocity Measurement **35**-9
35.7 Relative Angular Sensors **35**-10
Electrical (dc and ac) Tachometer Generator • Counter Types • Optical Sensors • Hall Effect • Wiegand Effect
35.8 Absolute Angular Velocity Sensors **35**-14
Gyroscopes
35.9 Conclusion **35**-17
References **35**-17
Further Readings **35**-17
Partial List of Vendors and Suppliers **35**-18

Charles P. Pinney
Pinney Technologies, Inc.

William E. Baker
University of New Mexico

35.1 Introduction

The *linear velocity* of an object, or more correctly a particle, is defined as the time rate of change of position of the object. It is a vector quantity, meaning it has a direction as well as a magnitude, and the direction is associated with the direction of the change in position. The magnitude of velocity is called the speed (or pace), and it quantifies how fast an object is moving. This is what the speedometer in a car tells you; thus, the speedometer is well named. Linear velocity is always measured in terms of, or from, some reference object. Thus, the speedometer of a car tells how fast one is moving relative to the earth. Usually, linear velocity is identified using only the term "velocity." Common units for velocity include meters per second and miles per hour, but any similar combination of units of length per unit of time is correct.

The *rotational velocity* (or angular velocity) of an object is defined as the time rate of change of angular position, and it is a measure of how fast an object is turning. It is completely analogous to linear velocity but for angular motion. Common units are revolutions per minute, but any angular unit of measurement per unit of time can be used. Rotational velocity is a vector quantity also, with the direction of the vector being the same as the direction of the axis about which object is turning. For example, with a car stopped at a stop light with the motor running, the rotational velocity of the crankshaft of the motor is given by a magnitude (rotational speed), say, 800 rpm (rev/min), and a direction associated with the direction in which the crankshaft is pointing. The axis of rotation of the object may be moving, rather than fixed as when the car is turning a corner. The roll, yaw, or pitch velocity of an airplane would be given in terms of rotational speeds about each of the turning axes in the same manner as for a crankshaft.

Usually, the reference from which linear or rotational velocity is given is understood from the context of the problem. It is often not stated explicitly. The measurement method used defines the reference.

Applications for velocity measurement include the following:

1. Controlling the speed at which metal stock is fed into a machine tool. If the metal is fed too quickly, the result could be premature tool wear or it could even lead to machine failure. Feeding the material too slowly will reduce the yield of the machine tool.
2. Measuring the approach speed of a robotic tool onto its target.
3. Monitoring the speed of a generator in an electric power station.
4. An airport radar system measuring the speed of an approaching aircraft using the Doppler effect.
5. Measuring an automobile's wheel speed in order to provide feedback to an antilock brake system.

35.2 Measurement of Linear Velocity

The problem of velocity measurement is somewhat different from that of measurement of other quantities in that there are not a large number of transducer types and transducer manufacturers from which to choose for a given problem. Frequently, the problem is such that the person must use his/her knowledge of measurement of other quantities and ingenuity to develop a velocity measurement method suitable for the problem at hand. Velocity is often obtained by differentiation of displacement or integration of acceleration. As background information for this, the necessary equations are given later.

Figure 35.1 shows a graph that represents the position of an object as a function of time as it moves along a straight, vertical path (y direction). The quantity to be measured could be an average velocity, and its magnitude would then be defined as follows:

$$\text{Average speed} = V_{\text{avg}} = \frac{y_2 - y_1}{t_2 - t_1} = \frac{\Delta y}{\Delta t} \tag{35.1}$$

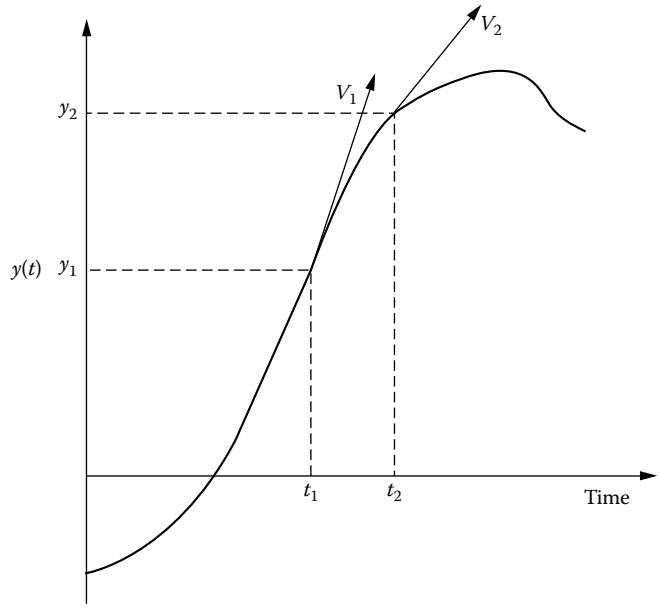


FIGURE 35.1 Position–time function for an object moving on a straight path.

for the time interval t_1 to t_2. As the time interval becomes small, the average speed becomes the instantaneous speed V_y, and the definition becomes

$$V_y = \lim_{\Delta t \to 0} \frac{\Delta y}{\Delta t} = \frac{dy}{dt} \tag{35.2}$$

which is the slope of the position–time curve. The subscript indicates the y component. This speed, when associated with the known direction, becomes the velocity.

Since acceleration is defined as the time rate of change of velocity, the speed of an object may also be given by

$$V_y(t) = V_i - \int_{t_i}^{t} a_y(t)dt \tag{35.3}$$

where

$a_y(t)$ is the acceleration in the y direction (for Figure 35.1)
V_i is the speed at time t_i

Each of the preceding equations can be used as a basis for making a velocity measurement. Note that for motion in more than one dimension, there would be more than one component, and there would be corresponding equations for the other dimensions (x and y). However, velocity measurements are always done by individual component.

It is convenient in the discussion of techniques of measuring velocity to divide the methods into two categories: one will be called "reference-based methods" and the other "seismic or inertial referenced transducers." Reference-based methods refer to measurements made for which the instrumentation has component(s) on both the moving object and the reference frame for the measurement. The seismic transducers do not require contact with the reference frame. However, they give a speed relative to the transducer speed at the start of the test. The initial motion must be determined from other considerations in the test setup and added to the relative speed.

35.3 Reference-Based Linear Velocity Measurement

Using Equation 35.1, one value of the average speed in a given direction of an object can be determined from the distance traveled in that direction and the time required. Determining the muzzle speed of a projectile is an example. Having two pickups spaced at known distance apart, and recording the time for the projectile edge to pass between them, is a common way of doing this. Typical pickups would include proximity transducers (see displacement measurement), laser or collimated light beams with diode sensors, and electric contacts closed (or opened) by the moving object. Measuring the time interval can be done with an electronic counter or displaying the output of the pickups on an oscilloscope. In designing such a system, care must be exercised to minimize or eliminate the effect to the object passing through the window on the positions of the sensors and their response. For example, pressure from a muzzle blast can move the sensors on their supports. This could distort the distance between them during the measurement, but afterward, the appearance of the setup could be unchanged, so the experimenter would be unaware of the error.

Using a series of equally spaced pickups can determine the average speed for a sequence of positions. For some applications, illumination of the path of motion of the object with a stroboscope flashing at a known rate and use of time exposure photography can give a single picture of the object at a sequence of positions. With knowledge of the length scale and the flash rate, the average speed at a sequence of

positions can be calculated. If the plane of motion is the same as the plane of the photograph, then two components of the velocity can be determined. A variation of this method is to use video recording of the motion and the time base of the video for measurement of time increments. Availability of high-speed video cameras, to 12,000 frames/s, extends the range of applicable velocities, and digital recording can enhance the ease and accuracy of making the distance measurements.

Another variation of this method is to use some type of position transducer to record the position–time function of the moving object and then differentiate this function to get speed–time. Displacement transducers were discussed in an earlier chapter, and the selection of an acceptable transducer is important. Range, resolution, and mass loading are important parameters. Because differentiation of experimental data is a noise-generating process, particular care must be exercised to reduce the electric noise in the displacement data to a minimum. Also, the calculated speed–time function might require some smoothing to reduce the numerically introduced noise.

One type of velocity transducer is based on a linear generator. When a coil cuts the magnetic field lines around a magnet, a voltage is induced in the coil, and this voltage is dependent on the following relation:

$$e_i \propto BLV \tag{35.4}$$

where

e_i is the induced voltage
B is the magnetic field strength
L is the length of wire in the coil
V is the speed of the coil relative to the magnet

This relation is used as the basis for linear velocity transducers, called LVTs, and a schematic is shown in Figure 35.2. Manufacturers of these transducers include Trans-Tek, Inc., of Ellington, CN; Robinson-Halpern Products of Valley Forge, PA; and the Macro Sensors, Div. of Howard A. Schaevitz Technologies, Inc., of Pennsauken, NJ. The working displacement ranges are from 0.5 to 24 in., and typical sensitivities are from 40 mV/ips (inches per second) to 600 mV/ips.

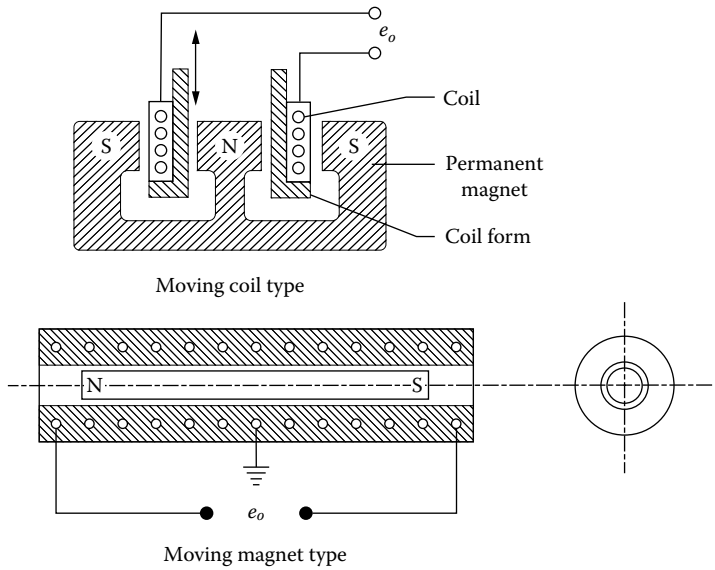


FIGURE 35.2 Velocity transducers (LVT).

35.3.1 Conversion of Linear to Rotational Velocity

A rotational dc generator (discussed in the next section) can also be used to measure linear velocities by placing a rack on the moving object and having the rack drive the generator through a pinion gear. This is the same principle by which a speedometer converts the linear velocity of an automobile to an angular velocity gage on the dashboard of a car.

35.3.2 Doppler Shift

The Doppler shift is an apparent change in the frequency of waves occurring when the source and observer are in motion relative to each other. This phenomenon is applicable to waves in general, for example, sound, light, and microwaves. It was first observed for sound waves, and it is named after the Austrian mathematician and physicist Christian Doppler (1803–1853) who first published a paper on it for light waves in 1842. The frequency will increase when the source and observer approach each other (red shift) and decrease when they move apart (blue shift). This phenomenon was illustrated by having people listen to the pitch of an oncoming train. The high-pitched whistle would transition to a lower pitch as the train passed the observer.

Radar, which is named for radio detection and ranging, is another technique for detecting the position, motion, and nature of a remote object by means of radio waves reflected from its surface. Pulse radar systems use a single directional antenna to transmit and receive the waves. They transmit pulses of electromagnetic waves (usually microwaves), some of which are reflected by objects in the path of the beam. Reflections are received by the radar unit, processed electronically, and converted into images on a cathode-ray tube. The antenna must be connected only to the transmitter when sending and only to the receiver while receiving; this is accomplished by switching from one to the other and back again in the fraction of a microsecond between pulses. The distance of the object from the radar source is determined by measuring the time required for the radar signal to reach the target and return. The direction of the object with respect to the radar unit is determined from the direction in which the pulses were transmitted. In most units, the beam of pulses is continuously rotated at a constant speed, or it is scanned (swung back and forth) over a sector at a constant rate. Pulse radar is used primarily for aircraft and naval navigation and for military applications. In Doppler radar, or continuous-wave radar, two antennas are used—one to transmit and the other to receive. Because the time a continuous-wave signal takes to reach the target and return cannot be measured, Doppler radar cannot determine distance. The velocity of the object is determined using the Doppler effect. If the object is approaching the radar unit, the frequency of the returned signal is greater than the frequency of the transmitted signal. If the object is receding, the returned frequency is less, and if the object is not moving relative to the radar unit, the frequency of the returned signal is the same as the frequency of the transmitted signal.

One value of this Doppler technology is shown on the evening weather broadcast. Radar can measure wind rotation inside a thunderstorm and identify possible tornadoes. The VORAD system by Eaton is an onboard system for vehicle safety. It detects when a dangerous approach to another vehicle is taking place. It will automatically apply the brakes in an emergency situation.

35.4 Light Interference Methods for Linear Velocity Measurement

Velocity measurements can be made using light interference principles. Figure 35.3 shows the setup used by Michelson in the 1890s to demonstrate light interference. A beam of monochromatic light is split into two beams. One beam is directed onto a stationary mirror. The other beam is directed onto a moving target. The observer sees the superposition of the two beams. As the mirror moves in one direction, summation of the waves of the two beams will alternately reinforce and cancel each other. The amount of motion for one cycle of light intensity variation is the wavelength of the light being used.

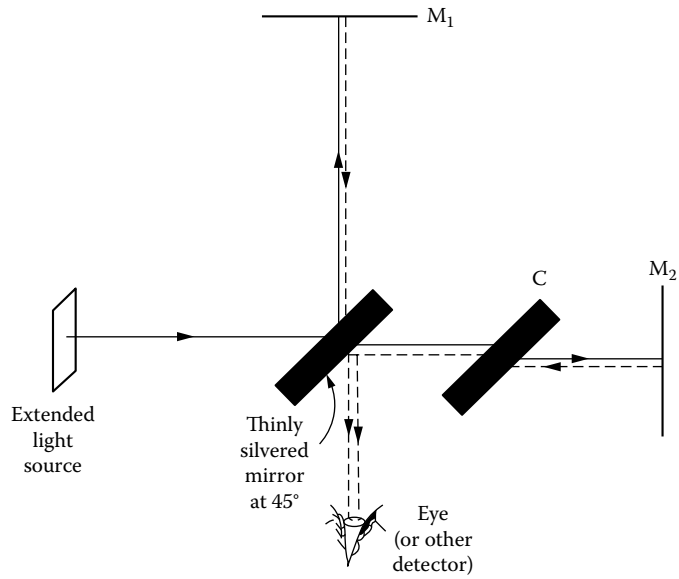


FIGURE 35.3 Basic components of a Michelson interferometer. The clear glass slab C is called a compensating plate. It has the same dimensions and orientation as the 45° mirror in order to make the light paths in glass equal along the two arms, a condition necessary when a white-light source is used. (From Hudson, A. and Nelson, R., *University Physics*, Harcourt Brace Jovanovich, New York, 1982. With permission.)

The frequency of these light-to-dark transitions is proportional to the velocity of the moving target. Highly accurate measurements are available with interferometer techniques. For example, 1 m is 1,650,763.73 fringe counts for the orange light emitted by krypton-86.

Refinements of this principle are needed for convenience of use. Lasers are used as the light source, for example. One commercial supplier of this type of device, commonly called a laser Doppler vibrometer, is Polytec PI, Inc., of Costa Mesa, CA. The basic principle gives velocity parallel to the laser beam, but Polytec PI also has a unit that utilizes scattered laser light that permits measurement of the in-plane velocity. It is called a laser surface velocimeter.

35.4.1 VISAR System

Another application of interferometry to the measurement of velocity–time profiles is a device called VISAR, for "velocity interference system for any reflector." Earlier interferometer systems had required that the target have a highly polished reflecting surface and that there be very little surface tilt during a test. The VISAR system functions with either specularly or diffusely reflecting surfaces and is quite insensitive to tilting of the target. It was developed for shock wave research work and is useful for measurement of very high speeds. Reference [3] gives a detailed description of the principles of operation.

The signal from the VISAR is generated with a photodiode or other light-sensitive device and is basically a measure of the rate of fringe variation. Additional data reduction is required to obtain speeds. The sensitivities of the devices are specified in "fringes per meter per second." Typical sensitivities are in the range of 100–4000 m/s per fringe.

The first VISARs were laboratory devices, individually assembled from the needed optical components. Commercial units are now available. Valyn International of Albuquerque, NM, makes VISARs and components. Figure 35.4 shows a schematic of a test setup. The standard measurement range, given as depth of field, of the VISAR is 12 mm, but systems measuring over 10 m have been used.

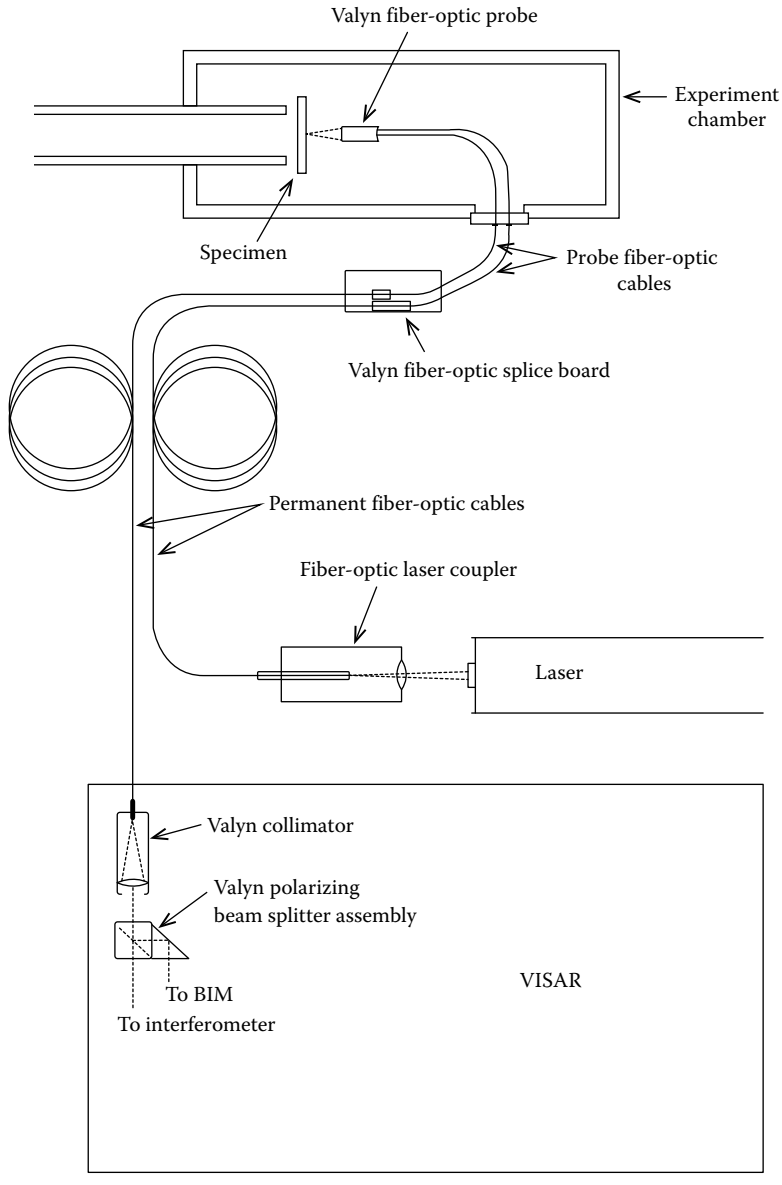


FIGURE 35.4 Schematic diagram showing how fiber-optic components, available from Valyn, can transport laser light to and from a shock experiment, minimizing any laser light beam hazards. (Courtesy of Valyn International, Albuquerque, NM.)

Applications for the VISAR include the following:

In-bore projectile velocity
Flyer plate velocity
Flying foil velocity
Hugoniot equation of state
Structural response to shock loading

35.5 Seismic Devices for Linear Velocity

The devices discussed in the previous section required a link of some type between the reference and the moving object. Seismic devices do not have this requirement. A seismic device refers to a transducer, which is based on a mass attached to the transducer base, usually with a linear spring. The base is attached to the surface whose motion is desired, and the motion of the seismic mass relative to the base is recorded with a motion transducer. Figure 35.5 shows the principal components of this transducer type. Use of the governing equation of motion for the seismic mass permits the determination of the motion of the base from the relative motion function.

If the motion transducer in the seismic instrument measures the displacement of the mass relative to the base, the output of the transducer is proportional to the acceleration of the transducer base for a specific frequency range, and the device is called an *accelerometer*. Acceleration measurement using this type of device (or with other types of accelerometers) permits the determination of the velocity–time function by integration through the application of Equation 35.3. In this equation, $a_y(t)$ would be determined from the output of the accelerometer.

The simplicity of this concept is evident, and because integration is a smoothing process, the numerically introduced noise encountered with a "differentiation of displacement" method of speed measurement does not occur. However, other errors can be introduced. First, any error in the acceleration measurement will be carried over. However, additional precautions are necessary to obtain good results for speed measurement. The problem areas include the following:

- The initial speed, V_i, must be known at the beginning of the time of interest. Because this quantity is added to the change in speed, an error in it will be a constant on each value.
- A bias, or zero shifts, in the accelerometer signal will be included as a constant acceleration, and thus introduce a linearly increasing error in the calculated values throughout the time interval of interest.

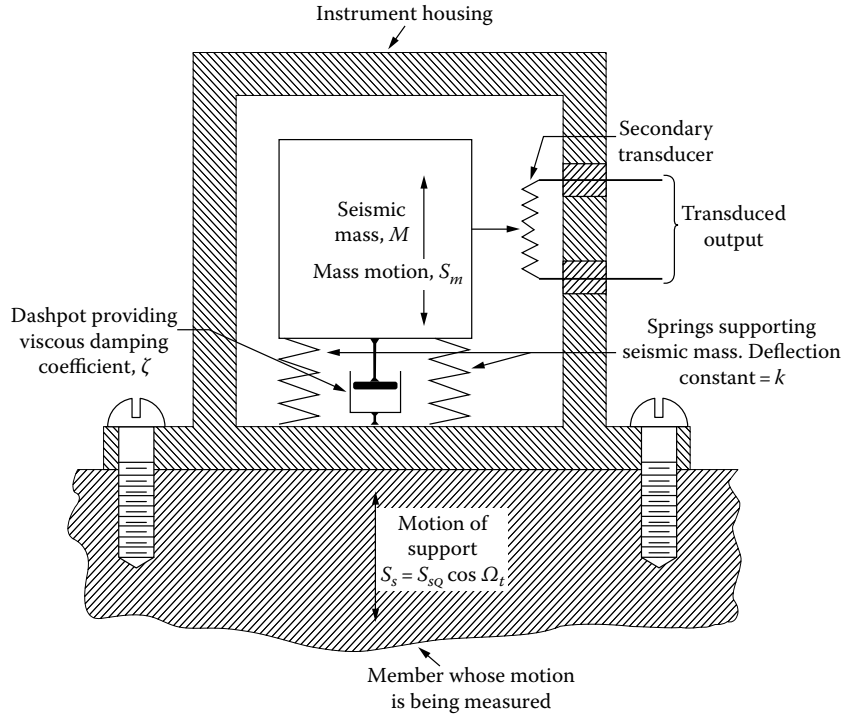


FIGURE 35.5 Seismic type of motion-measuring instrument. (From Beckwith, T. et al., *Mechanical Measurements*, 4th edn., Addison-Wesley, Reading, MA. With permission.)

This bias may be introduced from electrical or thermal characteristics of the circuit or, in the case of measurement of accelerations during impact after a free fall, by the 1 *g* acceleration of gravity.
- If the frequency content of the acceleration falls outside the usable bandwidth of the accelerometer and recording circuit, errors in acceleration occur. The low-frequency cutoff depends on the recording equipment and circuit, and the high-frequency cutoff depends on the natural frequency and damping of the accelerometer, as well as the bandwidth of each piece of equipment in the recording circuit.
- Accelerometer theory is based on harmonic excitation of the system. For many velocity measurement applications, the input is a transient. Combination of these two factors can result in inaccurate accelerometer data; for example, ringing may occur and cause errors in the calculated speeds. This problem is accentuated for lightly damped accelerometers.

When this method of speed measurement must be used, a series of check tests should be conducted to evaluate the accuracy of the method for that particular system.

A variation of the aforementioned method is to put an integrating circuit in the accelerometer and perform the integration with an analog circuit. Then, the output of the "velocity" transducer is proportional to the change in speed. This type of device is subject to all of the potential error sources discussed previously. A manufacturer of this type of transducer is Wilcoxon Research of Gaithersburg, MD.

It can be shown that if the electromechanical transducer in a seismic instrument gives an output that is proportional to the speed on one end relative to the other end, then the output of the seismic transducer is proportional to the speed of the transducer in an inertial reference frame, that is, relative to the earth, for input motion frequencies well above the natural period of the seismic mass. Thus, use of a linear velocity transducer as the motion measurement transducer in a seismic instrument makes it a "seismic velocity transducer." This type of device is called several different names, including seismometer, geophone, and vibrometer, as well as velocity transducer.

The natural frequency and damping in these instruments are selected to match the application. As with an accelerometer, the usable bandwidth depends on these two characteristics. The low-frequency limit for this type of transducer is dependent on the accuracy required in the measurement. The governing equation is given in Doebelin [3]. As an example, it can be used to show that if an accuracy of 5% is required, the lowest data frequency must be 4.6 times the natural frequency of the transducer and that the upper data frequency is not limited. In fact, the higher the upper frequency is, the more accurate the results.

Seismometers are used for recording and studying motion from earthquakes, and these devices can be quite large. Natural periods can be in the range of 10–50 s, and damping is normally selected as 0.7 of critical to extend the frequency range as much as possible. Geophones are commonly used for oil well logging and related work. Their natural periods are in the vicinity of 10 s. Manufacturers of these devices include Teledyne Brown Engineering and GeoSpace Corporation of Houston, TX.

35.6 Angular Velocity Measurement

Measurement of angular velocity is often applied to rotating machinery such as pumps, engines, and generators. The most familiar unit of measurement in rotating machinery applications is revolutions per minute (rpm). In most cases, the measurement of rpm involves the generation of a pulse train or sine wave whose frequency is proportional to angular velocity. The measuring technologies using pulse trains and waves include ac and dc generator tachometers, optical sensors, variable reluctance sensors, rotating magnet sensors, Wiegand effect sensors, and stroboscopy.

These types of measurements are taken with respect to the base of the item being measured. They are relative measurements because one is measuring the motion between two bodies.

Another class of measurement problem is that of moving or inertial bodies. In this case, a measurement of absolute motion is performed. Some fixed reference must be stated or implied. This reference is often the Earth. A universal reference is sometimes required for celestial measurements. These inertial measurements are typically taken with gyroscope-type devices.

35.7 Relative Angular Sensors

35.7.1 Electrical (dc and ac) Tachometer Generator

A rotating generator produces a voltage signal proportional to the rotational velocity of the input shaft. A dc generator produces a voltage level proportional to speed, as in Figure 35.6. The ac generator produces an ac voltage output with a frequency proportional to rotational speed, as shown in Figure 35.7. In a simple two-phase motor, the ac voltage is applied to one phase of the motor and the measurement is taken off the other. Typical operating frequencies are 60 and 400 Hz. This carrier frequency should be 5–10 times the required frequency response of the ac generator tachometer. The direction of travel is determined by the phase of the signal with opposite directions being 180° out of phase. The basic dc generator is shown in Figure 35.8.

Sources of tachometer generators include the GE Company of Fairfield, CT; Kollmorgen Motion Technologies Group of Radford, VA; Sierracin/Magnedyne of Vista, CA; and MicroMo Electronics of Clearwater, FL.

35.7.2 Counter Types

An entire class of angular velocity measuring techniques exists that uses pulses generated by electromechanical interaction. The common thread is a pulse-to-voltage converter giving a voltage output proportional to velocity.

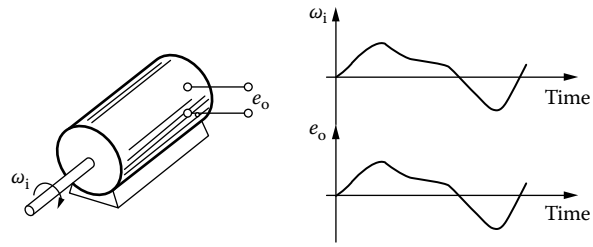


FIGURE 35.6 Permanent-magnet dc tachogenerator. (With permission.)

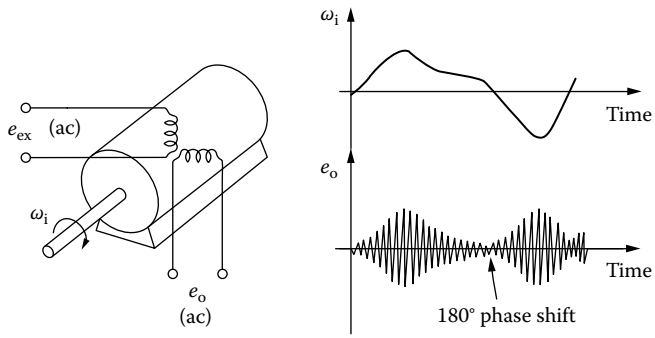


FIGURE 35.7 Ac tachogenerator. (With permission.)

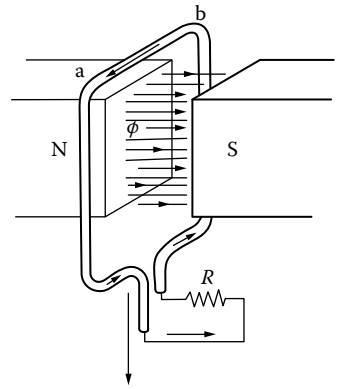


FIGURE 35.8 Generated electromotive force. Moving conductor. (From Loew, E.A., *Direct and Alternating Currents: Theory and Machinery*, McGraw-Hill, New York, 1946. With permission.)

Rotating magnet sensors: Passive speed sensors convert mechanical motions to ac voltages without the need for an external power source. These self-contained magnetic sensors produce a magnetic field that, when in the proximity of ferrous objects in motion, generates a voltage.

When a magnetic sensor is mounted in the proximity of a ferrous target, such as gear teeth on a rotating shaft, the voltage output frequency is directly proportional to the rotational speed of the target. A frequency-to-voltage converter can then convert the signal to a voltage. An engineering unit conversion from voltage to velocity then provides an actual velocity measurement.

Typical applications for these types of sensors include the following:

Transmission speed
Engine rpm
Over/under speed
Wheel speed
Pump shaft speed
Multiple engine synchronization
Feedback for speed control
Crankshaft position/engine timing
Computer peripheral speeds

The typical specifications for magnetic speed sensors are given by a graph of output voltage vs. surface speed in inches per second, as in Figure 35.9.

Sources for magnetic sensors include Smith Systems of Brevard, NC; Optek Technology of Carrolton, TX; AlliedSignal of Morristown, NJ; and Baluff of Florence, KY.

35.7.3 Optical Sensors

Optical methods of angular velocity detection employ a light emitter and a light detector. A light-emitting diode (LED) paired with a light-sensitive diode is the most common arrangement.

A slotted disk is placed in the axis of a rotating shaft. Each slot or slit will allow the light to pass through the disk. Figure 35.10 shows a typical arrangement. The detector will generate a pulse train with a rate proportional to the angular velocity.

The effects of external light sources must be considered in the application of optical sensors.

Sources of optical sensor systems include Scientific Technologies of Fremont, CA; Banner Engineering Corp. of Minneapolis, MN; and Aromat Corp. of New Providence, NJ.

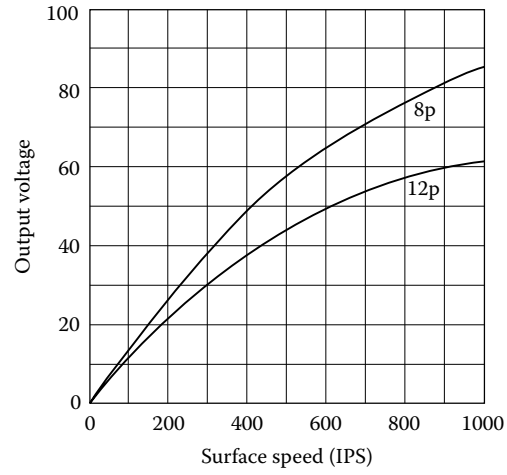


FIGURE 35.9 Magnetic speed sensor output voltage vs. speed. (Courtesy of Smith Systems, Inc., Brevard, NC.)

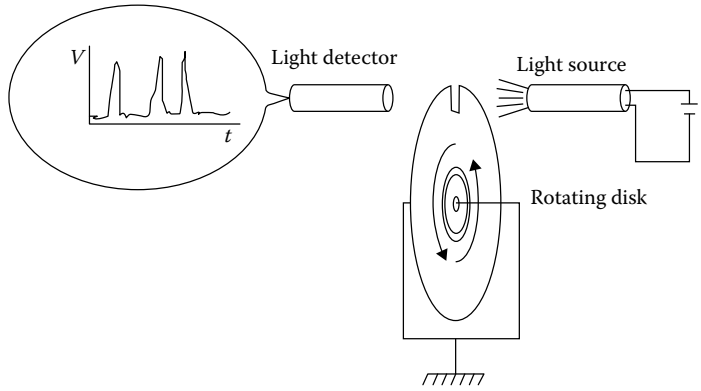


FIGURE 35.10 A slotted disk provides one pulse output for each rotation.

35.7.4 Hall Effect

The Hall effect describes the potential difference that develops across the width of a current-carrying conductor. E.H. Hall first used this effect in 1879 to determine the sign of current carriers in conductors. Hall-effect devices are finding their way into many sensing applications. A typical Hall-effect sensor application is the wheel speed sensor for antilock braking systems in automobiles. The Allegro ATS632LSC gear-tooth sensor, shown in Figure 35.11, is an optimized Hall-effect IC/magnet combination. The sensor consists of a high-temperature plastic shell that holds together a compound samarium–cobalt magnet, a single-element self-calibrating Hall-effect IC, and a voltage regulator. The operation of this circuit is shown in Figure 35.12.

35.7.5 Wiegand Effect

The Wiegand effect is useful for proximity sensing, tachometry, rotary shaft encoding, and speed sensing in applications such as

Electronic indexing for water, gas, and electric meters and remote metering systems
Measuring shaft speed in engines and other machinery
Tachometers, speedometers, and other rotational counting devices

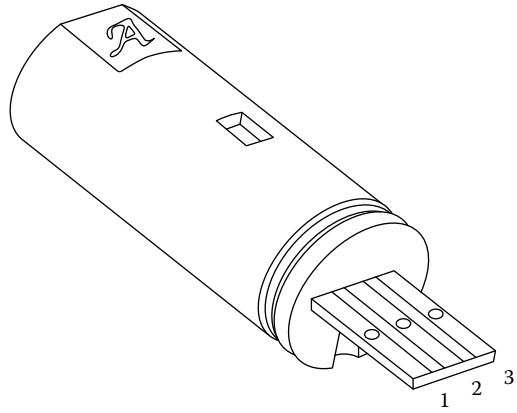


FIGURE 35.11 Hall-effect gear-tooth sensor. *Note*: 1, supply; 2, output; 3, ground. (Courtesy of Allegro Microsystems, Inc., Worcester, MA.)

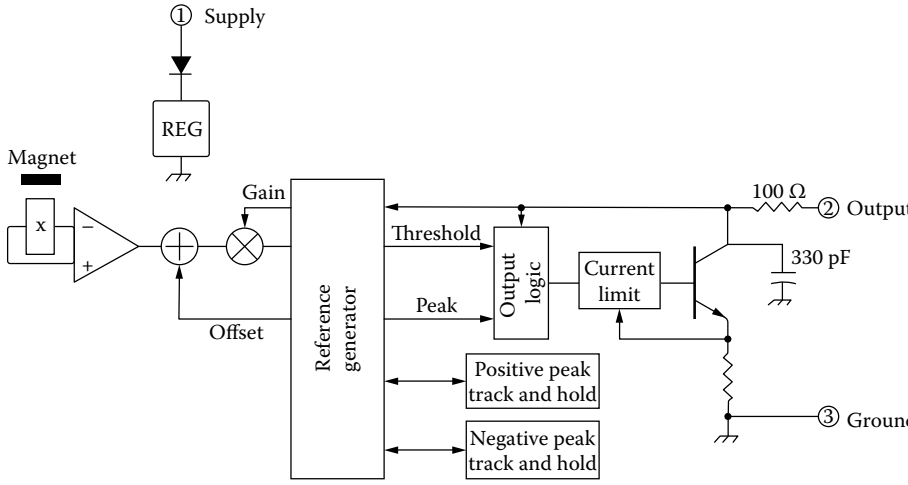


FIGURE 35.12 Functional block diagram of Hall-effect sensor circuit in a gear-tooth application. (Courtesy of Allegro Microsystems, Inc., Worcester, MA.)

Wiegand effect technology employs unique magnetic properties of specially processed, small-diameter ferromagnetic wire. By causing the magnetic field of this wire to suddenly reverse, a sharp, uniform voltage pulse is generated. This pulse is referred to as a Wiegand pulse. Sensors utilizing this effect require only a few simple components to produce sharply defined voltage pulses in response to changes in the applied magnetic field. These sensors consist of a short length of Wiegand wire, a sensing coil, and alternating magnetic fields that generally are derived from small permanent magnets.

The major advantages of the Wiegand effect–based sensors are

No external power requirement
Two-wire operation
Noncontact with no wear
20 kHz pulse rate
High-level voltage output pulse
Wide operating temperature range (e.g., −40 °C to +125 °C)

When an alternating magnetic field of proper strength is applied to the Wiegand wire, the magnetic field of the core switches polarity and then reverses, causing the Wiegand pulse to be generated, as shown in

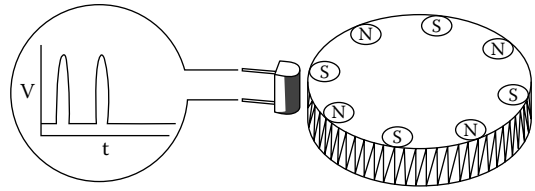


FIGURE 35.13 Small magnets cause sudden reversal in the ferromagnetic wire in a Wiegand sensor. (Courtesy of HID Corporation, North Haven, CT.)

Figure 35.13. The magnetic switching action of the Wiegand wire induces a voltage across the pickup coil of approximately 10 μs duration. These alternating magnetic fields are typically produced by magnets that are affixed to the rotating or moving equipment, by a stationary read head and moving Wiegand wires or by an ac-generated field.

35.8 Absolute Angular Velocity Sensors

35.8.1 Gyroscopes

Many absolute angular rate-measuring devices fall under the designation of gyroscope. A mechanical gyroscope is a device consisting of a spinning mass, typically a disk or wheel, mounted on a base so that its axis can turn freely in one or more directions and thereby maintain its orientation regardless of any movement of the base. It is important to make an initial distinction between angular velocity gyros and rate-integrating gyros. Angular velocity gyros are used to measure motion and as signal inputs to stabilization systems. Rate-integrating gyros are used as the basis for highly accurate inertial navigation systems. They allow a stable platform to maintain a fixed attitude with reference. These devices can be very complex. Three gyros are often teamed with three double-integrated accelerometers to provide an accurate measurement of absolute vehicle motion.

Ricardo Dao of Humphrey Inc. provided an excellent comparison of angular rate sensors in an article in *Measurements & Control* [14]. The five different technologies are summarized as follows.

Spinning mass: The traditional gyro consists of a spinning wheel in a gimbaled frame. The principle of conservation of angular momentum provides the measurement tool.

Fluidic: A stream of helium gas flows past two thin tungsten wires [14]. The tungsten wires act as two arms of a Wheatstone bridge. At rest, the gas flow cools the sensing wires equally and the transducer bridge is balanced with zero output. When angular motion is applied to the sensor, one sensor wire will be subjected to increased flow while the other will see less flow. The resistance of the two wires will change, and the bridge will be unbalanced. The sensor will produce a voltage output proportional to the angular velocity.

A pump is used to circulate the helium gas. This pump is a piezoelectric crystal circular disk that is excited with an external circuit. The pump produces a laminar flow of relatively high-velocity gas across the two parallel sensing wires.

Piezoelectric vibration: A number of angular velocity sensors have been developed that use micromachined quartz elements. A number of shapes are used, but the operating principle is similar for each. The quartz element vibrates at its natural frequency. Angular motion causes a secondary vibration that, when demodulated, is proportional to angular vibration. A description of one design follows.

The QRS and GyroChip™ family products use vibrating quartz tuning fork to sense angular velocity [15,16]. Using the Coriolis effect, a rotational motion about the sensor's longitudinal axis produces a dc voltage proportional to the rate of rotation. Figure 35.14 shows that the sensor consists of a microminiature double-ended quartz tuning fork and supporting structure, all fabricated chemically from a single wafer of monocrystalline piezoelectric quartz (similar to quartz watch crystals).

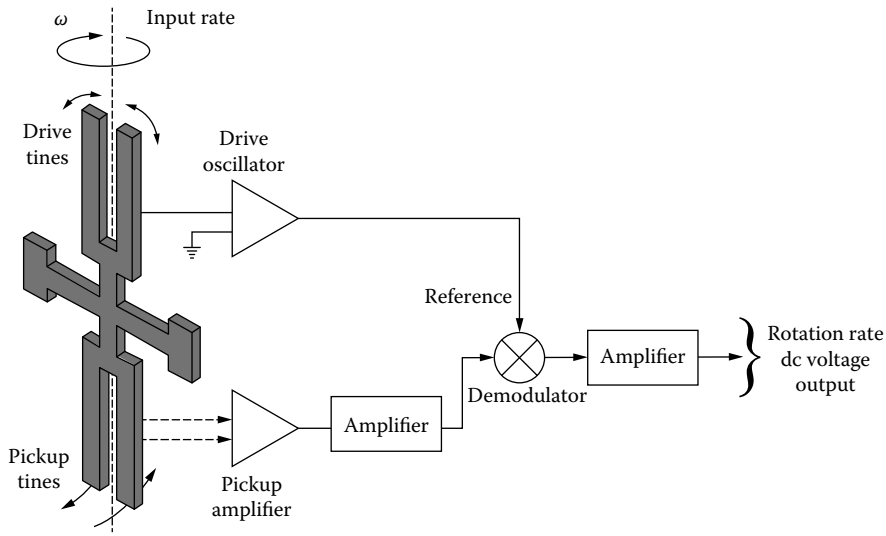


FIGURE 35.14 A vibrating quartz tuning fork uses the Coriolis effect to sense angular velocity. (Courtesy of BEI Sensors and Systems Co., Concord, CA.)

Use of piezoelectric quartz material simplifies the active element, resulting in exceptional stability over temperature and time. The drive tines, being the active portion of the sensor, are driven by an oscillator circuit at precise amplitude, causing the tines to move toward and away from each another at a high frequency.

Each tine will have Coriolis force acting on it of: $\{F = 2m\ W_i \times V_r\}$ where the tine mass is m, the instantaneous radial velocity is V_r, and the input rate is W_i. This force is perpendicular to both the input rate and the instantaneous radial velocity.

The two drive tines move in opposite directions, and the resultant forces are perpendicular to the plane of the fork assembly and in opposite directions. This produces a torque that is proportional to the input rotational rate. Since the radial velocity is sinusoidal, the torque produced is also sinusoidal at the same frequency of the drive tines and in-phase with the radial velocity of the tine.

The pickup tines, being the sensing portion of the sensor, respond to the oscillating torque by moving in and out of plane, producing a signal at the pickup amplifier. After amplification, those signals are demodulated into a dc signal that is proportional to the rotation of the sensor.

The output signal of the GyroChip™ reverses sign with the reversal of the input rate since the oscillating torque produced by the Coriolis effect reverses phase when the direction of rotation reverses. The GyroChip™ will generate a signal only with rotation about the axis of symmetry of the fork, that is, the only motion that will, by Coriolis sensing, produce an oscillating torque at the frequency of the drive tines. This also means that the GyroChip™ can truly sense a zero rate input.

MHD effect: The magnetohydrodynamic angular rate sensor is used to measure angular vibrations in the frequency range of 1–1000 Hz. It is used where there is a high-shock environment and a high rate of angular motion such as 10–250 rad s^{-1}. It does not measure a constant or dc velocity. It is used to measure impacts shorter than 1 s duration and vibrations between 1 and 1000 Hz.

The principle of operation is illustrated in Figure 35.15 [17,18]. A permanent magnet is attached to the outer case of the sensor. When the case turns, a moving magnetic field is produced (B). There is also a conductive fluid inside the sensor. When the sensor case turns, the fluid tends to stay in one place, according to Newton's first law. This produces a relative motion (U) between a magnetic field and conductor. This motion will produce a voltage (E) across the conductor proportional to relative velocity according to Faraday's law.

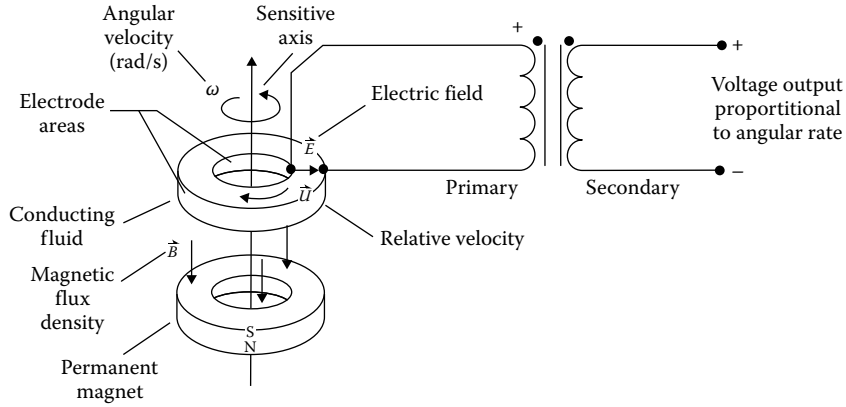


FIGURE 35.15 Magnetohydrodynamic angular rate sensor. (Courtesy of ATA Sensors, Albuquerque, NM.)

Since the fluid is constrained to move in an angular path, the voltage signal will be proportional to angular velocity about the center axis of the sensor. Due to this constraint, the sensor is insensitive to linear motion. The voltage signal is amplified through a transformer or an amplifier for output to a measuring device.

Fiber optic/laser: A beam of light is directed around the axis of rotation. A phase shift of the optical or laser beam is detected to measure angular velocity. The principle of operation is similar to the Doppler shift.

Differenced and integrated accelerometers: An array of accelerometers can be used to measure angular motion. The output of the accelerometers is differenced when they are aligned or summed when they are mounted in opposite directions. This differencing will eliminate the linear component of motion. As shown in Figure 35.16, the magnitude of the differenced signals, a_1, and a_2, is divided by the distance between the two sensors, l. This gives a measure of angular acceleration. The angular acceleration is integrated over time to give angular velocity. It is important to address the same concerns in this process as when integration was discussed in the linear section. It is assumed that there is a rigid mounting structure between the two accelerometers.

This technique is commonly applied to crash testing of anthropomorphic test devices (ATDs). The ATDs are used in automotive crash testing and aerospace egress system testing.

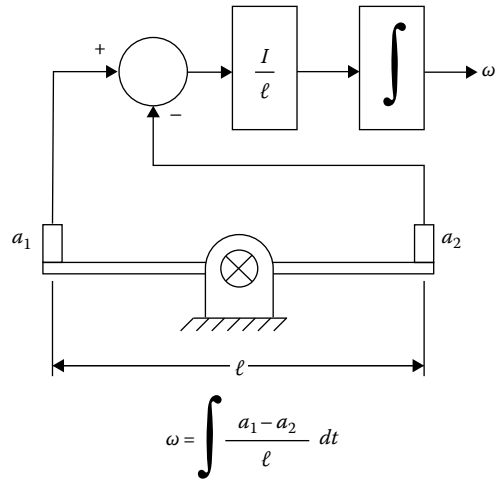


FIGURE 35.16 Angular acceleration by differencing accelerometers and integration.

35.9 Conclusion

But alas, as Poincaré [20] stated, there is no way to determine absolute velocity.

References

1. J. W. Dally, W. F. Riley, and K. G. McConnell, *Instrumentation for Engineering Measurements*, New York: John Wiley & Sons, 1984.
2. A. Hudson and R. Nelson, *University Physics*, New York: Harcourt Brace Jovanovitch, 1982.
3. E. O. Doebelin, *Measurement Systems,* 4th edn., New York: McGraw-Hill, 1990.
4. Trans-Tek, Inc., [Online], 1998. Available at: http://transtekinc.com/lvt.htm (accessed on March 15, 2013).
5. L. M. Barker and R. E. Hollenbach, Laser interferometer for measuring high velocities of any reflecting surface, *J. Appl. Phys*., 43(11), 4669–4675, 1972.
6. Valyn International, [Online], 1998. Available at: www.valynvisar.com (accessed on March 15, 2013).
7. L. C. Rodman, J. H. Bell, and R. D. Mehta, A 3-component laser-Doppler velocimeter data acquisition and reduction system, NASA contractor report; NASA CR-177390.
8. E. A. Loew, *Direct and Alternating Currents,* 3rd edn., New York: McGraw-Hill, 1946.
9. P. Emerald, Low duty cycle operation of hall effect sensors for circuit power conservation, *Sensors*, 15(3), 38, 1998.
10. Allegro MicroSystems, Inc., [Online], 1998. Available at: http://www.allegromicro.com/ (accessed on March 15, 2013).
11. Electronic Tutorials, Tutorial 6, Hall Effect Sensors, http://www.electronics-tutorials.ws/electromagnetism/hall-effect.html (accessed on March 15, 2014).
12. J. B. Scarborough, *The Gyroscope: Theory and Applications*, New York: Interscience, 1958.
13. R. F. Deimel, *Mechanics of the Gyroscope: The Dynamics of Rotation*, New York: Dover Publications, 1950.
14. E. Dao, Fluidic angular rate sensors, *Measurements & Control*, February 1994, Vol. 16, No. 1, 126–131.
15. Systron Donner Inertial Division, [Online], 1998. Available at: http://www.systron.com (accessed on March 15, 2013).
16. S. Orloski and B. Zimmerman, Dead reckoning in an emergency vehicle location system using a quartz rotation sensor and GPS, *Proceedings of Sensor Expo*, Chicago, IL, September 1995, pp. 103–109, Peterborough, NH: Helmers Publishing, 1995.
17. G. Kulikovskiy and G. A. Lyubimov, *Magnetohydrodynamics*, Reading, MA: Addison-Wesley, 1965.
18. Applied Technology Associates, [Online], 1998. Available at: http://www.aptec.com/ (accessed on March 15, 2013).
19. G. Unger et al., NASA's first in-space optical gyroscope: A technology experiment on the X-ray timing Explorer spacecraft, NASA technical memorandum; 109242.
20. R.P. Feynman, R.B. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Reading, MA: Addison-Wesley, 1963, p. 15-5.

Further Readings

Gu, J. H., Song, C. H., and Lee, S. Y., A new velocity measurement method using linear type Hall-effect sensor for an electro-mechanical fin actuator, *SICE '07. 46th SICE Annual Conference*, Takamatsu, Japan, pp. 2179–2183, 2007.

Limin, Y., A current statistical model adopted velocity measurement, *International Conference on Computer Science and Service System (CSSS)*, Nanjing, China, pp. 1914–1917, 2011.

Liu, L., Hu, J., He, Z., Han, C., Li, H., and Li, J., A velocity measurement method based on scaling parameter estimation of a chaotic system, *Metrology and Measurement Systems*, 18(2), 275–281, 2011.

Ma, C. and Xu, X., A novel technique for accurate velocity measurement using LFM radar, *Proceedings of the SPIE—The International Society for Optical Engineering*, 8186, 81860L (11 pp.), 2011.

Nabavi, M. and Siddiqui, K., A critical review on advanced velocity measurement techniques in pulsating flows, *Measurement Science & Technology*, 21(4), 042002 (19 pp.), April 2010.

Partial List of Vendors and Suppliers

Apex instruments, Inc., 204 Technology Park Lane, Fuquay Varina, NC 27526, Phone: 919-557-7300, 800-882-3214 (toll free), Fax: 919-557-7110, www.apexinst.com

Archangel Systems, Inc., 1635 Pumphrey Ave., Auburn, AL 36832, Phone: 334-826-8008, Fax: 334-826-8038, www.archangel.com

Davis/Inotek, 1236 Mark St., Bensenville, IL 60106, Phone: 410-358-3900, 800-548-9409 (toll free), Fax: 410-358-0252, www.davis.com

Furness Controls, 2020 Younts Rd., Indian Trail, NC 28079, Phone: 704-882-3311, 800-898-5325 (toll free), Fax: 704-882-3811, www.furnesscontrols.com/

Neal system Inc., 122 Terry Dr., Newtown, PA 18940, Phone: 215-968-7577, Fax: 215-968-6480, www.nealsystems.com/newsite/

Siemens Corp, 300 New Jersey Avenue, N.W., Suite 1000, Washington, DC 20004-2611, Phone: 202-434-4800, 800-SIEMENS (toll free), Fax: 202-347-4015, www.usa.siemens.com/entry/en/

Wilcoxon research Inc., 20511 Seneca Meadows Pkwy., Germantown, MD 20876, Phone: 301-330-8811, 800-945-2696 (toll free), Fax: 301-330-8873, www.wilcoxon.com

Zemarc Corp., 6431 Flotilla St., Los Angeles, CA 90040, Phone: 323-721-5598, 800-275-8698 (toll free), Fax: 323-722-2220, www.zemarc.com

IV

Mechanical Variables

36 **Acceleration, Vibration, and Shock Measurement** *Halit Eren* **36**-1
Accelerometer Dynamics: Frequency Response, Damping, Damping Ratio, and Linearity • Electromechanical Force-Balance (Servo) Accelerometers • Piezoelectric Accelerometers • Piezoresistive Accelerometers • Differential-Capacitance Accelerometers • Strain-Gage Accelerometers • Seismic Accelerometers • Inertial Types, Cantilever, and Suspended-Mass Configuration • Electrostatic Force Feedback Accelerometers • Microaccelerometers • Cross-Axis Sensitivity • Selection, Full-Scale Range, and Overload Capability • Selection, Full-Scale Range, and Overload Capability (Frequency Range): Sensitivity, Mass, and Dynamic Range • Signal Conditioning • Bibliography • Partial List of Vendors and Suppliers

37 **Strain Measurement** *Christopher S. Lynch* **37**-1
Fundamental Definitions of Strain • Principles of Operation of Strain Sensors • References

38 **Tactile Sensing** *Ricardo E. Saad, A. Bonen, K.C. Smith, and Ben Benhabib* **38**-1
Sensing Classification • Mechanical Effects of Contact • Theory for Tactile Sensing • Requirements for Tactile Sensors • Technologies for Tactile Sensing • References • Further Readings • Partial List of Vendors and Suppliers

39 **Pressure Measurement** *Kevin H.L. Chau* **39**-1
Basic Definitions • Sensing Principles • Silicon Micromachined Pressure Sensors • References • Further Information

40 **Vacuum Measurement** *Ron Goehner, Emil Drubetsky, Howard M. Brady, and William H. Bayles, Jr.* **40**-1
Background and History of Vacuum Gages • Direct Reading Gages • Indirect Reading Gages • Cathode Ionization Gages • Resonance Gages • Molecular Drag (Spinning Rotor) Gages • Partial Pressure Measurements and Mass Spectrometers • References • Further Readings • Partial List of Vendors and Suppliers

41 **Force Measurement** *M.A. Elbestawi* **41**-1
General Considerations • Hooke's Law • Force Sensors • Strain-Gage Load Cell • Piezoelectric Methods • Capacitive Force Transducer • Force-Sensing Resistors (Conductive Polymers) • Magnetoelastic Force Sensors • Torsional Balances • Tactile Sensors • References • Further Information • Partial List of Vendors and Suppliers

42 Angle Measurement *Robert J. Sandberg* **42**-1
Angle Gage Blocks • Clinometers • Optical Comparator • Protractor • Sine Bar • Sine Plate • Taper • Further Information

43 Mass, Weights, and Instrumentation *Emil Hazarian* **43**-1
Introduction • Etymology • Definitions of Terms • Theory • Test Instrumentation • Weighing Methods • Calibration Standards • Range and Accuracy • Units of Measurement • References

44 Torque and Power Measurement *Ivan J. Garshelis* **44**-1
Fundamental Concepts • Arrangement of Apparatus for Torque and Power Measurement • Torque Transducer Technologies • Torque Transducer Constructions, Operation, and Application • Apparatus for Power Measurement • References

45 Density Measurement *Halit Eren* **45**-1
Introduction • Solid Density • Fluid Density • Bibliography • Partial List of Manufacturers and Suppliers

46 Fluid Viscosity Measurement *R.A. Secco, M. Kostic, and J.R. deBruyn* **46**-1
Shear Viscosity • Newtonian and Non-Newtonian Fluids • Viscometer Types • Pressure and Gravity Flow Methods • Falling Body Methods • Oscillation Methods • Acoustic Methods • Microrheology • High-Pressure Rheology • References • Further Information

47 Surface Tension Measurement *David B. Thiessen and Kin F. Man* **47**-1
Mechanics of Fluid Surfaces • Standard Methods and Instrumentation • Specialized Methods • Defining Terms • Acknowledgments • References • Further Readings

36

Acceleration, Vibration, and Shock Measurement

36.1 Accelerometer Dynamics: Frequency Response, Damping, Damping Ratio, and Linearity **36**-3
Periodic Vibrations • Stationary Random Vibrations • Transients and Shocks • Nonstationary Random Vibrations
36.2 Electromechanical Force-Balance (Servo) Accelerometers **36**-8
Coil- and Magnetic-Type Accelerometers • Induction-Type Accelerometers
36.3 Piezoelectric Accelerometers **36**-11
36.4 Piezoresistive Accelerometers **36**-14
36.5 Differential-Capacitance Accelerometers **36**-14
36.6 Strain-Gage Accelerometers **36**-17
36.7 Seismic Accelerometers **36**-17
36.8 Inertial Types, Cantilever, and Suspended-Mass Configuration **36**-21
36.9 Electrostatic Force Feedback Accelerometers **36**-23
36.10 Microaccelerometers **36**-24
3D Accelerometers
36.11 Cross-Axis Sensitivity **36**-26
36.12 Selection, Full-Scale Range, and Overload Capability **36**-27
Frequency Range
36.13 Selection, Full-Scale Range, and Overload Capability (Frequency Range): Sensitivity, Mass, and Dynamic Range **36**-28
Transient Response • Full-Scale Range and Overload Capability • Environmental Conditions
36.14 Signal Conditioning **36**-29
Signal-Conditioning Piezoelectric Accelerometers • Signal Conditioning of Piezoresistive Transducers • Microaccelerometers
Bibliography **36**-32
Partial List of Vendors and Suppliers **36**-33

Halit Eren
Curtin University

Acceleration is an important parameter for general-purpose absolute motion measurements and vibration and shock sensing. Acceleration is measured accelerometers, which are commercially available in a wide variety, types, and shapes to meet the requirements of diverse range applications. They are manufactured in small sizes, light in weight, rugged, and robust to be able to operate in harsh environment. They can be configured as active or passive sensors. An active accelerometer (e.g., piezoelectric) gives an output without the need for an external power supply, while a passive accelerometer only changes its electric properties (e.g., capacitance) and requires an external electrical

power source. In applications, the selection of active- or passive-type accelerometer is important, since active sensors cannot measure static or DC mode operations. In applications where static measurements are involved, passive sensors must be used. In general, accelerometers are preferred over displacement and velocity sensors for the following reasons:

1. They have a wide frequency range from zero to very high values. Steady accelerations can easily be measured.
2. Acceleration is more frequently needed since destructive forces are often related to acceleration rather than to velocity or displacement.
3. Measurement of transients and shocks can readily be made, easier than displacement or velocity sensing.
4. Displacement and velocity can be obtained by simple integration of acceleration by electronic circuitry. Integration is preferred over differentiation.

Accelerometers can be classified in a number of ways, such as *deflection* or *null-balance* types, mechanical or electrical types, and dynamic or kinematic types. The majority of industrial accelerometers can be classified as either deflection or null-balance type. Those used in vibration and shock measurements are usually the deflection types, whereas those used for measurements of motions of vehicles, aircraft, etc., for navigation purposes may be either type. In general, null-balance types are used when extreme accuracy is needed.

A large number of practical accelerometers are of the deflection type; a general configuration is shown in Figure 36.1. There are many different deflection-type accelerometers. Although their principles of operation are similar, they only differ in minor details, such as the spring elements used, types of damping provided, and types of relative motion transducers employed. These types of accelerometers behave as second-order systems; the detailed mathematical analysis will be given in this chapter in the following.

Accelerometers can be classified as *dynamic*, meaning that the operation is based on measuring the force required to constrain a seismic mass to track the motion of the accelerated base, such as spring-constrained slug types. Another type is the *kinematic* accelerometer, which is based on timing of the passage of an unconstrained proof mass from spaced points marked on the accelerated base. These types of accelerometers are used in highly specific applications such as interspace spacecrafts and gravimetric measurements.

For practical purposes, accelerometers can also be classified as *mechanical* or *electrical*, depending on whether the restoring force or other measuring mechanism is based on mechanical properties (e.g., the law of motion, distortion of a spring or fluid dynamics, etc.) or on electrical or magnetic forces.

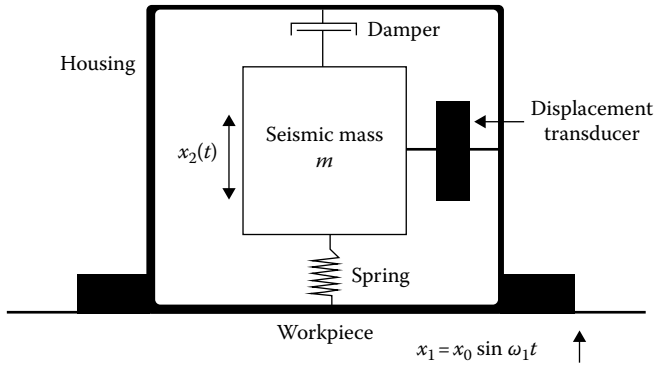


FIGURE 36.1 A typical deflection-type seismic accelerometer. In this basic accelerometer, the seismic mass is suspended by a spring or cantilever inside a rigid frame. The frame is connected to the vibrating structure; as vibration takes place, the mass tends to remain fixed such that the relative displacements can be picked up. They are manufactured in many different types and sizes and they exhibit diverse characteristics.

The calibration of accelerometers is very important and necessary in acceleration, vibration, and shock sensing applications. The calibration methods can broadly be classified to be *static* or *dynamic*. Static calibration is conducted at one or several levels of constant acceleration. For example, if a tilting table calibration method is selected, the vertical component of the free fall is used without a choice of magnitude. On the other hand, if a centrifuge is selected, it produces a constant acceleration as a function of the speed of rotation, and the magnitudes can be selected in a wide range from 0 to well over 50,000 *g*. The dynamic calibration is usually done using an electrodynamics shaker. The electrodynamics shaker is designed to oscillate in a sinusoidal motion with variable frequencies and amplitudes. They are stabilized at selected levels of calibration. This is an absolute method that consists of measuring the displacement with a laser interferometer and a precise frequency meter for accurate frequency measurements. The shaker must be driven by a power amplifier, thus giving a sinusoidal output with minimal distortion. The National Bureau of Standards uses this method as a reference standard. Precision accelerometers, mostly of the piezoelectric type, are calibrated by the absolute method and then used as the working standard. A preferred method is back-to-back calibration, where the test specimen is directly mounted on the working standard that, in turn, is mounted on an electrodynamics shaker.

Before providing details of different types of accelerometers, the common features such as accelerometer dynamics, velocity, distance, and shock frequency responses will be introduced next.

36.1 Accelerometer Dynamics: Frequency Response, Damping, Damping Ratio, and Linearity

This section concerns the physical properties of acceleration, vibration, and shock measurements in which accelerometers are commonly used. A full understanding of accelerometer dynamics is necessary in relation to characteristics of acceleration, vibration, and shock. For instance, the vibrations can be periodic, stationary random, nonstationary random, or transient, all of which affect the operational nature of the accelerometers and the arrangements to be made for correct measurements.

36.1.1 Periodic Vibrations

In periodic vibrations, the motion of an object repeats itself in an oscillatory manner. This can be represented by a sinusoidal waveform:

$$x(t) = X_{\text{peak}} \sin(\omega t) \tag{36.1}$$

where

$x(t)$ is the time-dependent displacement
$\omega = 2\pi ft$ is the angular frequency
X_{peak} is the maximum displacement from a reference point

The velocity of the object is the time rate of change of displacement:

$$u(t) = \frac{dx}{dt} = \omega X_{\text{peak}} \cos(\omega t) = U_{\text{peak}} \sin\left(\omega t + \frac{\pi}{2}\right) \tag{36.2}$$

where

$u(t)$ is the time-dependent velocity
$U_{\text{peak}} = \omega X_{\text{peak}}$ is the maximum velocity

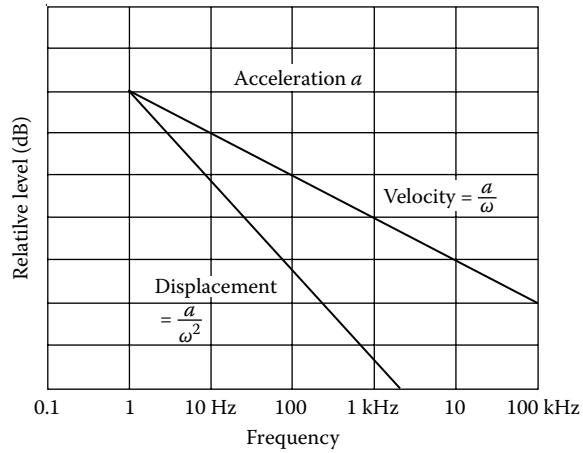


FIGURE 36.2 The logarithmic relationship between acceleration, velocity, and displacement. Velocity at a particular frequency can be obtained by dividing acceleration by a factor proportional to frequency. For displacement, acceleration must be divided by another factor proportional to the square of the frequency. Phase angles need to be determined separately, but they can be neglected in measurements in which time averages are involved.

Then the acceleration of the object is the time rate change of the velocity:

$$a(t) = \frac{du}{dt} = \frac{d^2u}{dt^2} = -\omega^2 X_{\text{peak}} \sin(\omega t) = A_{\text{peak}} \sin(\omega t + \pi) \tag{36.3}$$

where

$a(t)$ is the time-dependent acceleration

$A_{\text{peak}} = \omega^2 X_{\text{peak}} = \omega U_{\text{peak}}$ is the maximum acceleration

From the earlier equations, it can be seen that the basic form and the period of vibration remain the same in the case of acceleration, velocity, and displacement. But velocity leads displacement by a phase angle of 90° and acceleration leads velocity by another 90°. The amplitudes of the three quantities are related as a function of frequency, as shown in Figure 36.2.

In nature, vibrations are often periodic but not necessarily sinusoidal. If they are periodic but non-sinusoidal, they can be expressed as a combination of a number of pure sinusoidal signals that can be described by Fourier analysis as

$$x(t) = X_0 + X_1 \sin(\omega_1 t + \phi_1) + X_2 \sin(\omega_2 t + \phi_2) + \cdots X_n \sin(\omega_n t + \phi_n) \tag{36.4}$$

Where

$\omega_1, \omega_2, \ldots, \omega_n$ are the frequencies (rad s^{-1})

$X_0, X_1, \ldots, X_n$ are the maximum amplitudes of respective frequencies

$\phi_1, \phi_2, \ldots, \phi_n$ are the phase angles

The number of terms may be infinite: the higher the number of terms, the better the approximation. These terms constitute the *frequency spectrum*. This indicates that the vibrations can be represented in time domain or frequency domain, both of which are extremely useful in the analysis. As an example, in Figure 36.3, the time response of the seismic mass of an accelerometer is given against a rectangular pattern of excitation of the base.

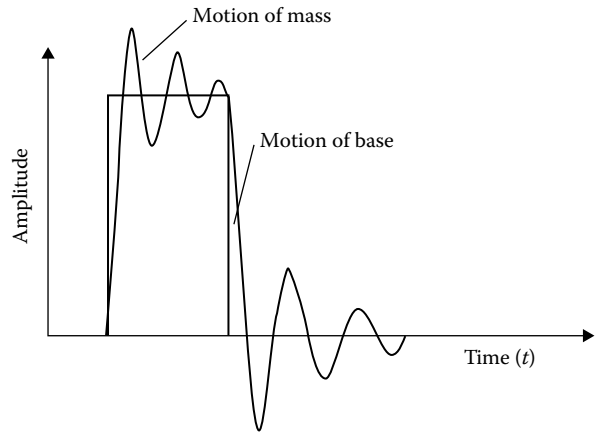


FIGURE 36.3 Time response of a shock excitation of a single degree-of-freedom system. As the duration of the shock pulse increases, sustained oscillations get shorter in time but become larger in amplitude. The maximum system response may be as high as twice the magnitude of the shock pulse.

36.1.2 Stationary Random Vibrations

Random vibrations are often met in nature where they constitute irregular cycles of motions that never repeat themselves exactly. Theoretically, an infinitely long time record is necessary to obtain a complete description of these vibrations. However, statistical methods and probability theory can be used for mathematical analysis by taking representative samples. Tools such as probability distributions, probability densities, frequency spectra, cross-correlation, autocorrelations, digital Fourier transforms (DFT), fast Fourier transforms (FFT), autospectral analysis, RMS values, and digital filter analysis are some of the techniques that are commonly employed.

36.1.3 Transients and Shocks

Often, short-duration and sudden-occurrence vibrations need to be measured. Shock and transient vibrations may be described in terms of force, acceleration, velocity, or displacement. As in the case of random transients and shocks, statistical methods and Fourier transforms are frequently used in the analysis.

36.1.4 Nonstationary Random Vibrations

In this case, the statistical properties of vibrations vary in time. Methods such as time averaging and other statistical techniques can be employed. A majority of accelerometers described here can be viewed and analyzed as seismic instruments consisting of a mass, a spring, and a damper arrangement, as shown in Figure 36.1. Taking only the mass–spring system, if the system behaves linearly in a time invariant manner, the basic second-order differential equation for the motion of the mass alone under the influence of a force can be written as

$$f(t) = \frac{md^2x}{dt^2} + c\frac{dx}{dt} + kx \tag{36.5}$$

where
$f(t)$ is the force
m is the mass
c is the velocity constant
k is the spring constant

Nevertheless, in seismic accelerometers, the base of the arrangement is also in motion. Therefore, Equation 36.5 can be generalized by taking the effect motion into account. Then, it can be expressed as

$$\frac{md^2z}{dt^2} + c\frac{dz}{dt} + kz = mg\cos(\theta) - \frac{md^2x_1}{dt^2} \tag{36.6}$$

where

$z = x_2 - x_1$ is the relative motion between the mass and the base
x_1 is the displacement of the base
x_2 is the displacement of the mass
θ is the angle between sense axis and gravity

In order to lay a background for further analysis, taking the simple case, the complete solution to Equation 36.5 can be obtained by applying the superposition principle. The superposition principle states that if there are simultaneously superimposed actions on a body, the total effect can be obtained by summing the effects of each individual action.

Equation 36.5 describes essentially a second-order system that can be expressed in Laplace transform as

$$\frac{X(s)}{F(s)} = \frac{1}{ms^2} + cs + k \tag{36.7}$$

or

$$\frac{X(s)}{F(s)} = \frac{K}{s^2/\omega_n^2 + 2\zeta s/\omega_n + 1} \tag{36.8}$$

where

s is the Laplace operator
$K = 1/\underline{k}$ is the static sensitivity
$\omega_n = \sqrt{k/m}$ is the undamped critical frequency, rad/s
$\zeta = c/2\sqrt{km}$ is the damping ratio

As can be seen, in the performance of accelerometers, important parameters are the static sensitivity, the natural frequency, and the damping ratio, which are functions of mass, velocity, and spring constants. Accelerometers are designed to have different performance characteristics by suitable selection of these parameters.

Once the response is expressed in the form of Equations 36.7 and 36.8, analysis can be taken further, either in the time domain or in the frequency domain. The time response of a typical second-order system for a unit-step input is given in Figure 36.4. The Bode plot gain phase responses are depicted in Figure 36.5. Detailed discussions about frequency response, damping, damping ratio, and linearity are made in relevant sections.

Systems in which a single structure moves in more than one direction are termed *multidegree-of-freedom systems*. In this case, the accelerations become functions of dimensions as d^2x/dt^2, d^2y/dt^2, and d^2z/dt^2. Hence, in multichannel vibration tests, multiple transducers must be used to create uniaxial, biaxial, or triaxial sensing points for measurements. Mathematically, a linear multidegree-of-freedom system can be described by a set of coupled second-order linear differential equations, and when the frequency response is plotted, it normally shows one resonance peak per degree of freedom.

Frequently, acceleration and vibration measurements of thin plates or small masses are required. Attaching an accelerometer with a comparable mass onto a thin plate or a small test piece can cause "mass loading." Since acceleration is dependent on the mass, the vibration characteristics of the loaded

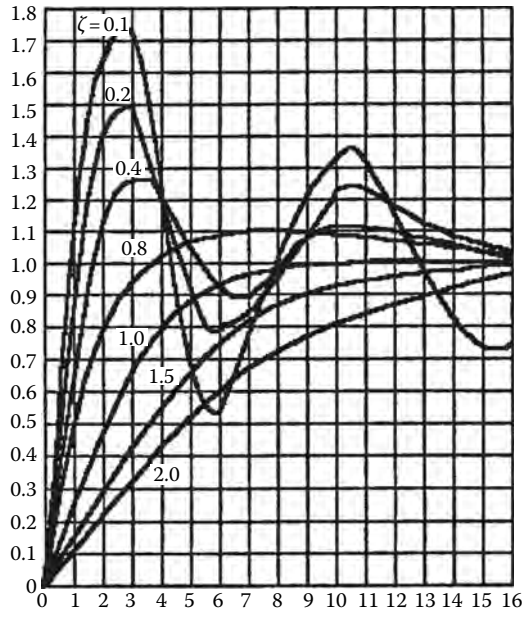


FIGURE 36.4 Unit-step time responses of a second-order system with various damping ratios. The maximum overshoot, delay, rise, settling times, and frequency of oscillations depend on the damping ratio. Smaller damping ratios give faster response but larger overshoot. In many applications, a damping ratio of 0.707 is preferred.

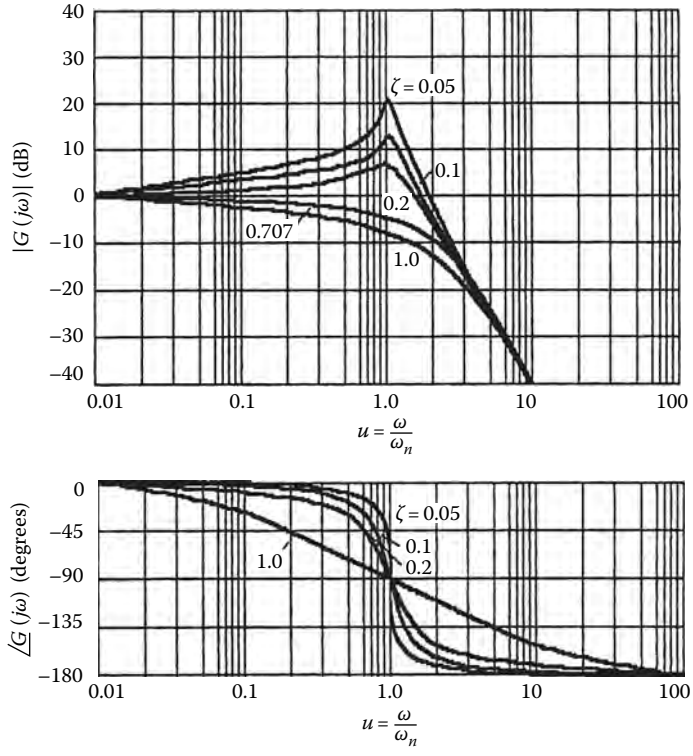


FIGURE 36.5 Bode plots of gains and phase angles against frequency of a second-order system. Curves are functions of frequencies as well as damping ratios. These plots can be obtained theoretically or by practical tests conducted in specified frequency ranges.

test piece could be altered, thus yielding wrong measurements. In such cases, a correct interpretation of the results of the measuring instruments must be made. Some experimental techniques are also available for the correction of the test results by performing repetitive tests conducted by sequentially adding small known masses and by observing the differences.

The following sections discuss different types of accelerometers.

36.2 Electromechanical Force-Balance (Servo) Accelerometers

Electromechanical accelerometers, essentially servo- or null-balance types, rely on the principle of feedback. In these devices, an acceleration-sensitive mass is kept very close at a neutral position or zero displacement point by sensing the displacement and feeding back this displacement. A proportional magnetic force is generated to oppose the motion of the mass displaced from the neutral, thus restoring neutral position—just as a mechanical spring in a conventional accelerometer would do. The advantages of this approach are (1) better linearity and (2) elimination of hysteresis effects when compared to mechanical spring countertypes. Also, in some cases, an electric damping can be provided so that the system becomes less sensitive to variations in temperature.

One very important feature of null-balance types is the capability of testing the static and dynamic performances of the devices by introducing electrically excited test forces into the system. This self-checking feature can be quite convenient in complex and expensive tests where it is extremely critical that the system operates correctly before the test commences. They are also useful in acceleration control systems, since the reference value of acceleration can be introduced by means of a proportional current from an external source. They are usually used for general-purpose motion measurements and monitoring low-frequency vibrations. They are specifically used in applications requiring better accuracy than achieved by those accelerometers based on mechanical springs used as force-to-displacement transducers.

There are a number of different types of electromechanical accelerometers: coil and magnetic types and induction types.

36.2.1 Coil- and Magnetic-Type Accelerometers

These accelerometers are based on Ampere's law; that is, "a current carrying conductor disposed within a magnetic field experiences a force proportional to the current, the length of the conductor within the field, the intensity of the magnetic field, and the sine of the angle between the conductor and the field."

Figure 36.6 illustrates one form of accelerometer making use of the earlier principle. The coil is located within the cylindrical gap defined by a permanent magnet and a cylindrical soft iron flux return path. It is mounted by means of an arm situated on a minimum friction bearing so as to constitute an acceleration-sensitive seismic mass. A pick-off mechanism senses the displacement of the coil under acceleration and causes the coil to be supplied with a direct current via a suitable servo controller to restore or maintain a null condition.

Assuming a downward acceleration with the field being radial (90°) and using Ampere's law, the force experienced by the coil may be written as

$$F = ma = ilB \tag{36.9}$$

or the current

$$i = \frac{ma}{lB} \tag{36.10}$$

where

B is the effective flux density

l is the total effective length of the conductor in the magnetic field

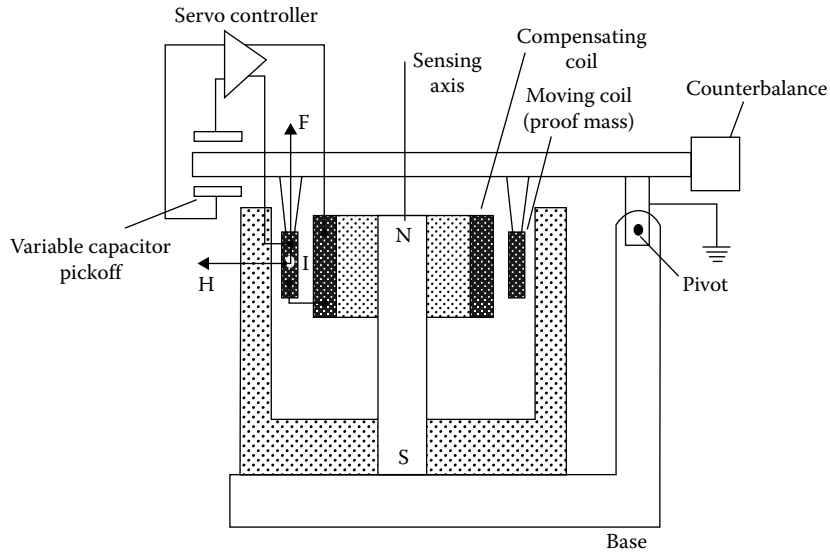


FIGURE 36.6 A basic coil and permanent magnet accelerometer. The coil is supported by an arm with minimum friction bearings to form a proof mass in a magnetic field. Displacement of the coil due to acceleration induces an electric potential in the coil to be sensed and processed. A servo system maintains the coil in a null position.

Current in the restoring circuit is linearly proportional to acceleration, if (1) the armature reaction effects are negligible and fully neutralized by the compensating coil in opposition to the moving coil and (2) the gain of the servo system is large enough to prevent displacement of the coil from the region in which the magnetic field is constant.

In these accelerometers, the magnetic structure must be shielded adequately to make the system insensitive to external disturbances and Earth's magnetic field. Also, in the presence of acceleration, there will be a temperature rise due to i^2R losses. The effects of these i^2R losses are determined by the appropriate thermal design and the heat transfer properties of the accelerometer. In many applications, special care must be exercised in choosing an appropriate accelerometer such that the temperature rises caused by unexpected accelerations cannot affect excessively the scale factors or the bias conditions.

A simplified version of another type of servo accelerometer is given in Figure 36.7. The acceleration a of the instrument case causes an inertial force F on the sensitive mass m, tending to make it pivot in its bearings or flexure mount. The rotation θ from neutral is sensed by an inductive pickup to be amplified, demodulated, and filtered to produce a current i_a, which is directly proportional to the motion from the null. This current is passed through a precision stable resistor R to give an output voltage signal and is applied to a coil suspended in a magnetic field. The current through the coil produces magnetic torque on the coil, which takes action to return the mass to neutral. The current required to produce magnetic torque that just balances the inertial torque due to acceleration is directly proportional to acceleration a. Therefore, the output voltage e_o becomes a measure of acceleration a. Since a nonzero displacement θ is necessary to produce the current i_a, the mass is not exactly returned to null but becomes very close to zero because of the high-gain amplifier. Analysis of the block diagram reveals that

$$\frac{e_o}{R} = \left(mra - \frac{e_o K_c}{R} \right) \times \left(\frac{K_p K_a / K_s}{s^2/\omega_{nl}^2 + 2\zeta_1 s/\omega_{nl} + 1} \right) \tag{36.11}$$

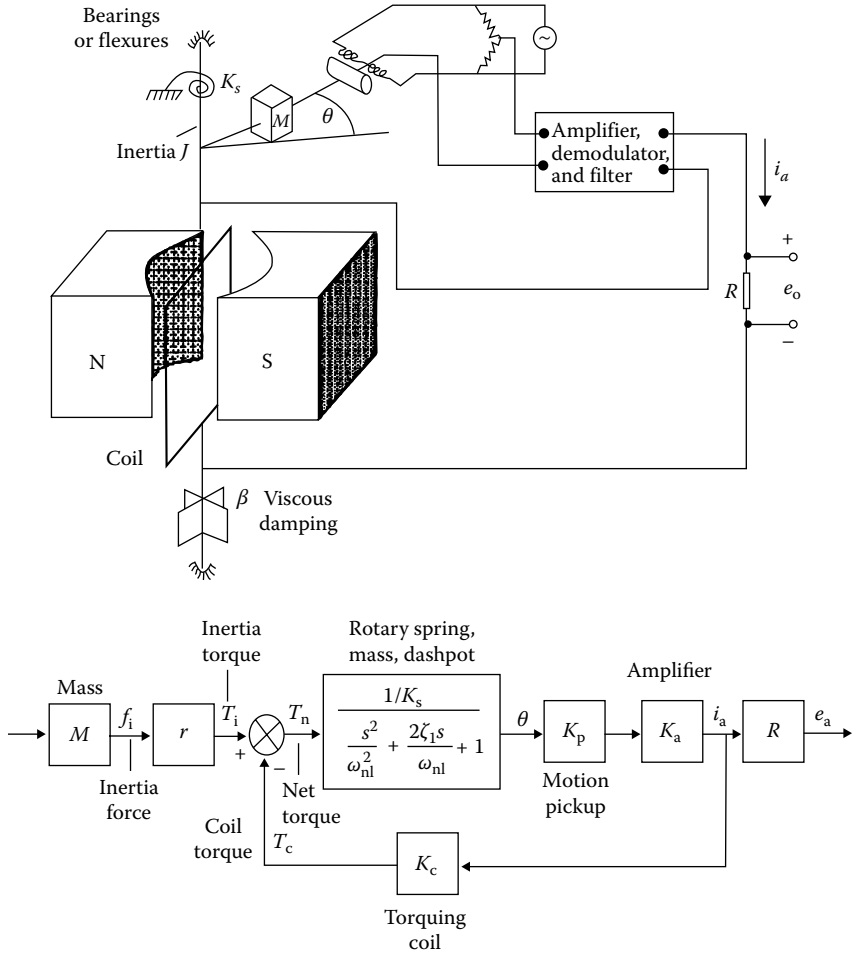


FIGURE 36.7 A simplified version of a rotational-type servo accelerometer. Acceleration of the instrument case causes an inertial force on the sensitive mass, tending to make it pivot in its bearings or flexure mounts. The rotation from neutral is sensed by inductive sensing and amplified, demodulated, and filtered to produce an electric current that is directly proportional to the motion from the null. The block diagram representation is useful in analysis.

Rearranging this expression gives

$$\frac{mrRK_pK_aa}{K_s}=\frac{s^2}{\omega_{nl}^2}+\frac{2\zeta_1 s}{\omega_{nl}}+1+\left(\frac{K_cK_pK_aa}{K_s}\right)e_o \tag{36.12}$$

By designing the amplifier gain, K_a is made large enough so that $K_cK_pK_aa/K_s \gg 1.0$; then

$$\frac{e_o}{a(s)}=\frac{K}{(s^2/\omega_{nl}^2+2\zeta_1 s/\omega_{nl}+1+K_cK_pK_aa/K_s)e_o} \tag{36.13}$$

where

$$K\equiv\frac{MrR}{K_c},\quad (\mathrm{V\,m^{-1}s^{-2}}) \tag{36.14}$$

$$\omega_n \cong \omega_{nl}\sqrt{\frac{K_c K_p K_a}{K_s}} \text{ rad/s} \tag{36.15}$$

$$\zeta \equiv \frac{\zeta_1}{\sqrt{K_c K_p K_a / K_s}} \tag{36.16}$$

Equation 36.14 shows that the sensitivity depends on the values of m, r, R, and K_c, all of which can be made constant. In this case, a high-gain feedback is useful in shifting the requirements for accuracy and stability from mechanical components to a selected few parameters where the requirements can easily be met. As in all feedback systems, the gain cannot be made arbitrarily high because of dynamic instability; however, a sufficiently high gain can be achieved to obtain good performance. An excellent comprehensive treatment of this topic is given by Doebelin (1990), and the detail of the book is given in the bibliography.

36.2.2 Induction-Type Accelerometers

The cross-product relationship of current, magnetic field, and force gives the basis for induction-type electromagnetic accelerometers, which are essentially generators rather than motors. One type of instrument, cup and magnet, includes a pendulous element with a pick-off and a servo controller driving a tachometer coupling and a permanent magnet and a flux return ring, closely spaced with respect to an electrically conductive cylinder attached to the pendulous element. A rate proportional drag force is obtained by electromagnetic induction effects between magnet and conductor. The pick-off senses pendulum deflection under acceleration and causes the servo controller to turn the rotor in a sense to drag the pendulous element toward null. Under steady-state conditions, motor speed is a measure of the acceleration acting on the instrument. Stable servo operation is achieved employing a time-lead network to compensate the inertial time lag of the motor and magnet combination. The accuracy of servo-type accelerometers is ultimately limited by consistency and stability of the scale factors of coupling devices and magnet and cup as a function of time and temperature.

Another accelerometer based on induction types is the *eddy current induction torque generation*. It was noted that the force-generating mechanism of an induction accelerometer consists of a stable magnetic field, usually supplied by a permanent magnet, which penetrates orthogonally through a uniform conduction sheet. The movement of the conducting sheet relative to the magnetic field in response to an acceleration results in a generated electromotive potential in each circuit in the conductor. This action is in accordance with Faraday's law induction. In these accelerometers, the induced eddy currents are confined to the conductor sheets, making the system essentially a drag coupling.

A typical commercial instrument based on the servo-accelerometer principle might have a micromachined quartz flexure suspension, differential capacitance angle pick-off, and air squeeze film, plus servo lead compensation for system damping. Of the various available models, a 30 g range unit has threshold and resolution of 1 μg, frequency response flat within 0.05% at 10 Hz and 2% at 100 Hz, natural frequency 500 Hz, damping ratio 0.3–0.8, and transverse or cross-axis sensitivity 0.1%. If the output current is about 1.3 mA g^{-1}, then a 250 Ω readout resistor would give about ±10 V full scale at 30 g. These accelerometers are good with respect to precision and are used in many applications, such as aircrafts, missiles, measurement of tilt angles, and axle angular bending in weight and balance systems.

36.3 Piezoelectric Accelerometers

Piezoelectric accelerometers are very popular and are used extensively for general-purpose acceleration, shock, and vibration measurements. They are based on the piezoelectric properties of some materials. When a physical force is exerted on the piezoelectric material, it generates a voltage proportional to the force.

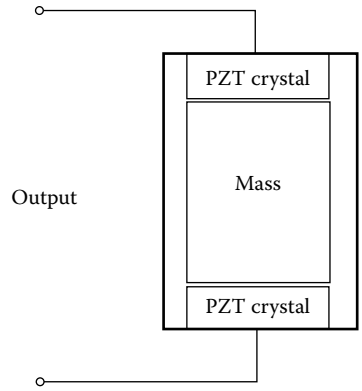


FIGURE 36.8 A compression-type piezoelectric accelerometer. The crystals are under compression at all times, either by a mass or mass and spring arrangement. Acceleration causes a deformation of the crystal, thus producing a proportional electric signal. They are small in size and are widely used. They demonstrate poor performance at low frequencies.

Piezoelectric accelerometers are basically motion transducers with large output signals and comparatively small sizes. They are available with very high natural frequencies and are therefore suitable for a wide range of applications.

These devices utilize a mass in direct contact with the piezoelectric component or crystal, as shown in Figure 36.8. When a varying motion is applied to the accelerometer, the crystal experiences a varying force excitation ($F = ma$), causing a proportional electric charge q to be developed across it:

$$q = d_{ij}F = d_{ij}ma \tag{36.17}$$

where

q is the charge developed

d_{ij} is the material's piezoelectric coefficient

As Equation 36.17 shows, the output from the piezoelectric material is dependent on its mechanical properties, d_{ij}. Two commonly used piezoelectric crystals are lead–zirconate–titanate ceramic (PZT) and quartz. They are both self-generating materials and produce a large electric charge compared to their sizes. The piezoelectric strain constant of PZT is about 150 times that of quartz. As a result, PZTs are much more sensitive and smaller in size than their quartz counterparts. In accelerometers, the mechanical spring constants for the piezoelectric components are high, and the inertial masses attached to them are small. Therefore, these accelerometers are useful for high-frequency applications. Figure 36.9 illustrates a typical frequency response for a PZT device. Since piezoelectric accelerometers have comparatively low mechanical impedances, their effects on the motion of most structures are negligible. They are also manufactured to be rugged and they have outputs that are stable with time and environment.

Mathematically, their transfer function approximates to a third-order system as

$$\frac{e_o(s)}{a(s)} = \frac{(K_q/C\omega_n{}^2)\tau s}{[(\tau s+1)(s^2/\omega_n{}^2 + 2\zeta s/\omega_n + 1)]} \tag{36.18}$$

where

K_q is the piezoelectric constant related to charge (Coulomb cm)

τ is the time constant of the crystal

It is worth noting that the crystal itself does not have a time constant τ, but the time constant is observed when the accelerometer is connected into an electric circuit (e.g., an *RC* circuit).

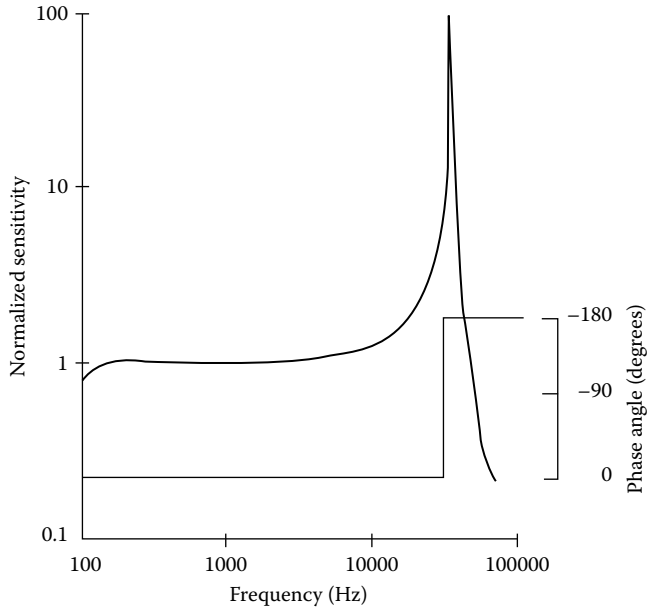


FIGURE 36.9 Frequency response of a typical piezoelectric accelerometer. Measurements are normally confined to the linear portion of the response curve. The upper frequency of the accelerometer is limited by the resonance of the PZT crystal. The phase angle is constant up to the resonance frequency.

The low-frequency response is limited by the piezoelectric characteristic $\tau s/(\tau s + 1)$, while the high-frequency response is related to mechanical response. The damping factor ζ is very small, usually less than 0.01 or near zero. Accurate low-frequency response requires large τ, which is usually achieved by the use of high-impedance voltage amplifiers. At very low frequencies, thermal effects can have severe influences on the operation characteristics.

In piezoelectric accelerometers, two basic design configurations are used: compression types and shear stress types. In compression types, the crystal is held in compression by a preload element; therefore, the vibration varies the stress in compressed mode. In the shear accelerometer, vibration simply deforms the crystal in shear mode. The compression type has a relatively good mass/sensitivity ratio and hence exhibits better performance. But, since the housing acts as an integral part of the spring mass system, it may produce spurious interfaces in the accelerometer output, if excited in its proper natural frequency.

Microelectronic circuits have allowed the design of piezoelectric accelerometers with charge amplifiers and other signal-conditioning circuits built into the instrument housing. This arrangement allows greater sensitivity, high-frequency response, and smaller-size accelerometers, thus lowering the initial and implementation costs.

Piezoelectric accelerometers are available in a wide range of specifications and are offered by many manufacturers. For example, the specifications of a shock accelerometer may have 0.004 pC g^{-1} sensitivity and a natural frequency of up to 250,000 Hz, while a unit designed for low-level seismic measurements might have 1,000 pC g^{-1} in sensitivity and with only 7,000 Hz natural frequency. They are manufactured as small as 3 × 3 mm in dimensions with about 0.5 g in mass, including cables. They have excellent temperature ranges, and some of them are designed to survive intensive radiation environment of nuclear reactors. However, piezoelectric accelerometers tend to have larger cross-axis sensitivity than other types, about 2%–4%. In some cases, large cross-axis sensitivity can be used during installation for the correct orientation of the device. These accelerometers can be mounted with threaded studs, with cement or wax adhesives, or with magnetic holders.

36.4 Piezoresistive Accelerometers

Piezoresistive accelerometers are essentially semiconductor strain gages with large gage factors. High gage factors are obtained because the material resistivity is dependent primarily on the stress, not only on dimensions. The increased sensitivity is critical in vibration measurement because it allows the miniaturization of the accelerometer. Most piezoresistive accelerometers use two or four active gages arranged in a Wheatstone bridge. Extraprecision resistors are used, as part of the circuit, in series with the input to control the sensitivity, balancing, and offsetting temperature effects. The mechanical construction of a piezoresistive accelerometer is shown in Figure 36.10.

In some applications, overload stops are necessary to protect the gages from high-amplitude inputs. These instruments are useful for acquiring vibration information at low frequencies (e.g., below 1 Hz). In fact, the piezoresistive sensors are inherently true static acceleration measurement devices. Typical characteristics of piezoresistive accelerometers are 100 mV g^{-1} in sensitivity, 0–750 Hz in frequency range, 2500 Hz in resonance frequency, 25 *g* in amplitude range, 2000 *g* in shock rating, and 0 °C–95 °C in temperature range and has a total mass of about 25 g.

36.5 Differential-Capacitance Accelerometers

Differential-capacitance accelerometers are based on the principle of change of capacitance in proportion to applied acceleration. These accelerometers come in different shapes and sizes. In one type, the seismic mass of the accelerometer is made as the movable element of an electrical oscillator as shown in Figure 36.11. The seismic mass is supported by a resilient parallel-motion beam arrangement from the base. The system is characterized to have a certain defined nominal frequency when undisturbed. If the device is accelerated, the frequency varies above and below the nominal value, depending on the direction of the acceleration.

The seismic mass carries an electrode located in opposition to a number of base-fixed electrodes that define variable capacitors. The base-fixed electrodes are resistance coupled in the feedback path of a wideband, phase-inverting amplifier. The gain of the amplifier is made of such a value to ensure

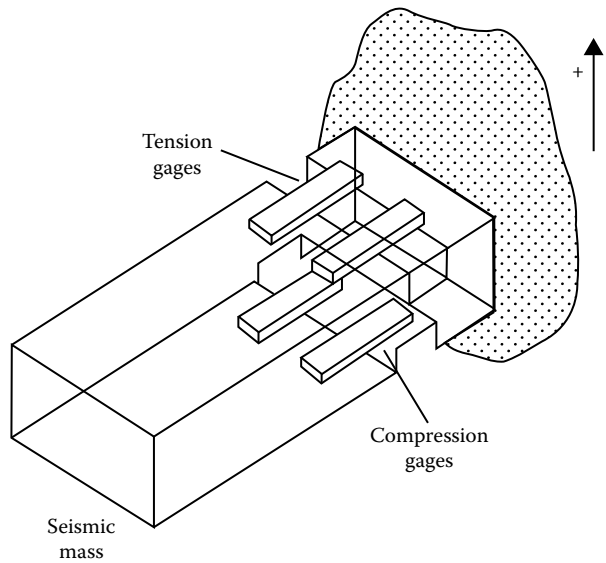


FIGURE 36.10 Bonding of piezoelectric and piezoresistive elements onto an inertial system. As the inertial member vibrates, deformation of the tension and compression gages causes the resistance to change. The change in resistance is picked up and processed further. Accelerometers based on PZTs are particularly useful in medium- to high-frequency applications.

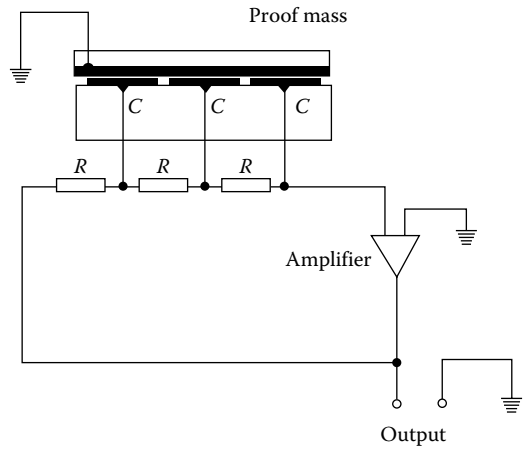


FIGURE 36.11 A typical differential capacitive accelerometer. The proof mass is constrained in its null position by a spring. Under acceleration, variable frequencies are obtained in the electric circuit. In a slightly different version, the proof mass may be constrained by an electrostatic feedback force, thus resulting in a convenient mechanical simplicity.

maintenance of oscillations over the range of variation of capacitance determined by the applied acceleration. The value of the capacitance C for each of the variable capacitor is given by

$$C = \frac{\varepsilon k S}{h} \tag{36.19}$$

where
- k is the dielectric constant
- ε is the permittivity of free space
- S is the area of electrode
- h is the variable gap

Denoting the magnitude of the gap for zero acceleration as h_0, the value of h in the presence of acceleration a may be written as

$$h = h_0 + \frac{ma}{K} \tag{36.20}$$

where m is the value of the proof mass and K is the spring constant. Thus,

$$C = \frac{\varepsilon k S}{(h_0 + ma/K)} \tag{36.21}$$

If the frequency of oscillation of the resistance-capacitance-type circuit is given by the expression, then

$$f = \frac{\sqrt{6}}{2\pi RC} \tag{36.22}$$

Substituting this value of C in Equation 36.21 gives

$$f = \left(h_0 + \frac{ma}{K} \right) \frac{\sqrt{6}}{2\pi R\varepsilon kS} \tag{36.23}$$

Denoting the constant quantity $(\sqrt{6}/2\pi R\varepsilon kS)$ as B and rewriting Equation 36.23 give

$$f = Bh_0 + \left(\frac{Bma}{K} \right) \tag{36.24}$$

The first term on the right-hand side expresses the fixed bias frequency f_0, and the second term denotes the change in frequency resulting from acceleration, so that the expression may be written as

$$f = f_0 + f_a \tag{36.25}$$

If the output frequency is compared with an independent source of constant frequency f_0, f_a can be determined.

A commonly used example of a capacitive-type accelerometer is based on a thin diaphragm with spiral flexures that provide the spring, proof mass, and moving plate necessary for the differential capacitor, as shown in Figure 36.12. Plate motion between the electrodes pumps air parallel to the plate surface and through holes in the plate to provide squeeze film damping. Since air viscosity is less temperature sensitive than oil, the desired damping ratio of 0.7 hardly changes more than 15%. A family of such instruments is readily available, having full-scale ranges from ±0.2 *g* (4 Hz flat response) to ±1000 *g* (3000 Hz), cross-axis sensitivity less than 1%, and full-scale output of ±1.5 V. The size of a typical device is about 25 mm^3 with a mass of 50 g.

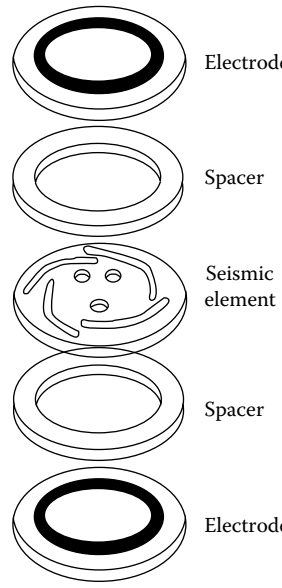


FIGURE 36.12 A diaphragm-type capacitive accelerometer. The seismic element is cushioned between the electrodes. Motion of the mass between the electrodes causes air movement through the holes, which provides a squeeze film damping. In some cases, oil can be used as the damping element.

36.6 Strain-Gage Accelerometers

Strain-gage accelerometers are based on resistance properties of electrical conductors. If a conductor is stretched or compressed, its resistance alters due to two reasons: dimensional changes and the changes in the fundamental property of material called *piezoresistance*. This indicates that the resistivity ρ of the conductor depends on the mechanical strain applied onto it. The dependence is expressed as the gage factor:

$$\left(\frac{dR/R}{dL/L}\right) = 1 + 2\nu + \left(\frac{d\rho/\rho}{dL/L}\right) \tag{36.26}$$

where

1 is the resistance change due to length
2ν is the resistance change due to area
$(d\rho/\rho)/(dL/L)$ is the resistance change due to piezoresistivity

In acceleration measurements, the resistance strain gages can be selected from different types, including unbonded metal-wire gages, bonded metal-wire gages, bonded metal-foil gages, vacuum-deposited thin-metal-film gages, bonded semiconductor gages, and diffused semiconductor gages. However, usually, bonded and unbonded metal-wire gages find wider applications in accelerometers. Occasionally, bonded semiconductor gages, known as piezoresistive transducers, are used but suffer from high-temperature sensitivities, nonlinearity, and some mounting difficulties. Nevertheless, in recent years, they have found new application areas with the development of micromachine transducer technology, which is discussed in detail in Section 36.14.3.

Unbonded strain-gage accelerometers use the strain wires as the spring element and as the motion transducer, using similar arrangements as in Figure 36.10. They are useful for general-purpose motion and vibration measurements from low to medium frequencies. They are available in wide ranges and characteristics, typically ±5–±200 *g* full scale, natural frequency 17–800 Hz, 10 V excitation voltage AC or DC, full-scale output ±20–±50 mV, resolution less than 0.1%, inaccuracy less than 1% full scale, and cross-axis sensitivity less than 2%. Their damping ratio (using silicone oil damping) is 0.6–0.8 at room temperature. These instruments are small and lightweight, usually with a mass of less than 25 g.

Bonded strain-gage accelerometers generally use a mass supported by a thin flexure beam. The strain gages are cemented onto the beam to achieve maximum sensitivity, temperature compensation, and a good sensitivity to both cross-axis and angular accelerations. Their characteristics are similar to unbonded strain-gage accelerometers but have larger sizes and weights. Often, silicone oil is used for damping. Semiconductor strain gages are widely used as strain sensors in cantilever beam/mass types of accelerometers. They allow high outputs (0.2–0.5 V full scale). Typically, a ±25 *g* acceleration unit has a flat response from 0 to 750 Hz, a damping ratio of 0.7, a mass of 28 g, and an operational temperature of –18 °C to +93 °C. A triaxial ±20,000 *g* model has a flat response from 0 to 15 kHz, a damping ratio 0.01, a compensation temperature range of 0°C–45°C, 13 × 10 × 13 mm in size, and 10 g in mass.

36.7 Seismic Accelerometers

These accelerometers make use of a seismic mass that is suspended by a spring or a lever inside a rigid frame. The schematic diagram of a typical instrument is shown in Figure 36.1. The frame carrying the seismic mass is connected firmly to the vibrating source whose characteristics are to be measured. As the system vibrates, the mass tends to remain fixed in its position so that the motion can be registered as a relative displacement between the mass and the frame. This displacement is sensed by an appropriate transducer and the output signal is processed further. Nevertheless, the seismic mass does not remain absolutely steady, but for selected frequencies, it can satisfactorily act as a reference position.

By proper selection of mass, spring, and damper combinations, the seismic instruments may be used for either acceleration or displacement measurements. In general, a large mass and soft spring are suitable for vibrations and displacements, while a relatively small mass and stiff spring are used in accelerometers.

The following equation may be written by using Newton's second law of motion to describe the response of seismic arrangements similar to what is shown in Figure 36.1:

$$\frac{md^2x_2}{dt^2} + \frac{cdx_2}{dt} + kx_2 = \frac{cdx_1}{dt} + kx_1 + mg\ \cos(\theta) \tag{36.27}$$

where

x_1 is the displacement of the vibration frame
x_2 is the displacement of the seismic mass
c is the velocity constant
k is the spring constant

Taking md^2x_1/dt^2 from both sides of the equation and rearranging give

$$m\frac{d^2z}{dt^2} + c\frac{dz}{dt} + kz = mg\ \cos(\theta) - m\frac{d^2x_1}{dt^2} \tag{36.28}$$

where

$z = x_2 - x_1$ is the relative motion between the mass and the base
θ is the angle between sense axis and gravity

In Equation 36.27, it is assumed that the damping force on the seismic mass is proportional to velocity only. If a harmonic vibratory motion is impressed on the instrument such that:

$$x_1 = x_0 \sin \omega_1 t \tag{36.29}$$

where ω_1 is the frequency of vibration, in rad s^{-1}.

Writing

$$\frac{md^2x_1}{dt^2} = mx_0{\omega_1}^2 \sin\omega_1 t$$

modifies Equation 36.28 as

$$m\frac{d^2z}{dt^2} + c\frac{dz}{dt} + kz = mg\cos(\theta) + ma_1 \sin\omega_1 t \tag{36.30}$$

where $a_1 = mx_0\omega_1^2$.

Equation 36.30 will have transient and steady-state solutions. The steady-state solution of the differential Equation 36.30 can be determined as

$$z = \left(\frac{mg\cos(\theta)}{k}\right) + \left(\frac{ma_1 \sin\omega_1 t}{(k - m{\omega_1}^2 + jc\omega_1)}\right) \tag{36.31}$$

Rearranging Equation 36.31 results in

$$z = \left(\frac{mg\cos(\theta)}{\omega_n}\right) + \left\{\frac{a_1 \sin(\omega_1 - \phi)}{[\omega_n{}^2(1-r^2)^2 + (2\zeta r)^2]^{1/2}}\right\} \tag{36.32}$$

where

$\omega_n = \sqrt{k/m}$ is the natural frequency of the seismic mass

$\zeta = c/2\sqrt{km}$ is the damping ratio and also can be written in terms of critical damping ratio as $\zeta = c/c_c$, where $\left(c_c = 2\sqrt{km}\right)$

$\phi = \tan^{-1}(c\omega_1/(k - m\omega_1^2))$ is the phase angle

$r = \omega_1/\omega_n$ is the frequency ratio

A plot of Equation 36.32, $(x_1 - x_2)_0/x_0$ against frequency ratio ω_1/ω_n, is illustrated in Figure 36.13. This figure shows that the output amplitude is equal to the input amplitude when $c/c_c = 0.7$ and $\omega_1/\omega_n > 2$. The output becomes essentially a linear function of the input at high-frequency ratios. For satisfactory system performance, the instrument constant c/c_c and ω_n should carefully be calculated or obtained from calibrations. In this way, the anticipated accuracy of measurement can be predicted for frequencies of interest.

If the seismic instrument has a low natural frequency and a displacement sensor is used to measure the relative motion z, then the output is proportional to the displacement of the transducer case. If the velocity sensor is used to measure the relative motion, the signal is proportional to the velocity of the transducer. This is valid for frequencies significantly above the natural frequency of the transducer. However, if the instrument has a high natural frequency and the displacement sensor is used, the measured output is proportional to the acceleration:

$$kz = \frac{m\, d^2x_1}{dt^2} \tag{36.33}$$

This equation is true since displacement x_2 becomes negligible in comparison to x_1.

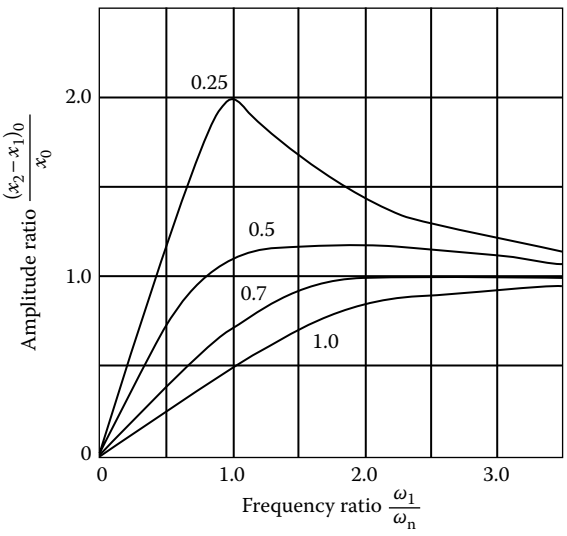


FIGURE 36.13 A typical displacement of a seismic instrument. Amplitude becomes large at low damping ratios. The instrument constants should be selected such that, in measurements, the frequency of vibration is much higher than the natural frequency (e.g., greater than by a factor of 2). Optimum results are obtained when the value of instrument constant c/c_c is about 0.7.

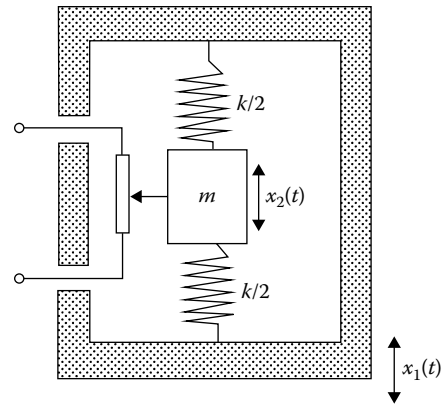


FIGURE 36.14 A potentiometer accelerometer. The relative displacement of the seismic mass is sensed by a potentiometer arrangement. The potentiometer adds extra weight, making these accelerometers relatively heavier. Suitable liquids filling the frame can be used as damping elements. These accelerometers are mainly used in low-frequency applications.

In these instruments, the input acceleration a_0 can be calculated by simply measuring $(x_1 - x_2)_0$. Generally, in acceleration measurements, unsatisfactory performance is observed at frequency ratios above 0.4. Thus, in such applications, the frequency of acceleration must be kept well below the natural frequency of the instrument. This can be accomplished by constructing the instrument to have a low natural frequency by selecting soft springs and large masses.

Seismic instruments are constructed in a variety of ways. Figure 36.14 illustrates the use of a voltage divider potentiometer for sensing the relative displacement between the frame and the seismic mass. In the majority of potentiometric instruments, the device is filled with a viscous liquid that interacts continuously with the frame and the seismic mass to provide damping. These accelerometers have low frequency of operation (less than 100 Hz) and are mainly intended for slow varying acceleration and low-frequency vibrations. A typical family of such instrument offers many different models, covering the range of ±1 to ±50 *g* full scale. The natural frequency ranges from 12 to 89 Hz, and the damping ratio ζ can be kept between 0.5 and 0.8, by using a temperature-compensated liquid damping arrangement. Potentiometer resistance can be selected in the range of 1,000–10,000 Ω, with a corresponding resolution of 0.45%–0.25% of full scale. The cross-axis sensitivity is less than ±1%. The overall accuracy is ±1% of full scale or less at room temperatures. The size is about 50 mm^3, with a mass of about 1/2 kg.

Linear variable differential transformers (LVDTs) offer another convenient means to measure the relative displacement between the seismic mass and the accelerometer housing. These devices have higher natural frequencies than potentiometer devices, up to 300 Hz. Since the LVDT has lower resistance to the motion, hence, it offers much better resolution. A typical family of liquid-damped differential-transformer accelerometers exhibits the following characteristics: full-scale range from ±2 to ±700 *g*, natural frequency from 35 to 620 Hz, nonlinearity 1% of full scale, the full-scale output about 1 V with an LVDT excitation of 10 V at 2000 Hz, damping ratio 0.6–0.7, residual voltage at null less than 1%, and hysteresis less than 1% full scale; the size is 50 mm^3, with a mass of about 120 g.

Electric resistance strain gages are also used for displacement sensing of the seismic mass as shown in Figure 36.15. In this case, the seismic mass is mounted on a cantilever beam rather than on springs. Resistance strain gages are bonded on each side of the beam to sense the strain in the beam resulting from the vibrational displacement of the mass. Damping for the system is provided by a viscous liquid that entirely fills the housing. The output of the strain gages is connected to an appropriate electric

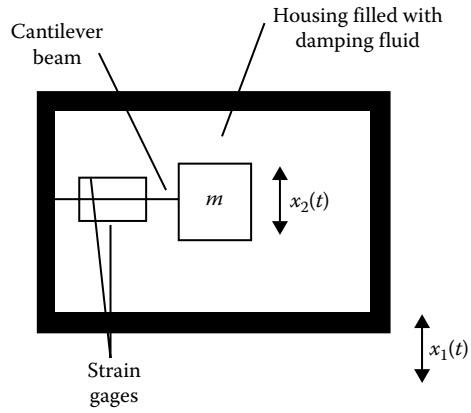


FIGURE 36.15 A strain-gage seismic instrument. The displacement of the proof mass is sensed by piezoresistive strain gages. The natural frequency of the system is low, due to the need of a long lever beam to accommodate the strain gages. The signal is processed by bridge circuits.

bridge circuit. The natural frequency of such a system is about 300 Hz. The low natural frequency is due to the need for a sufficiently large cantilever beam to accommodate the mounting of the strain gages. Other types of seismic instruments using piezoelectric transducers and seismic masses are discussed in detail in the section dedicated to piezoelectric-type accelerometers.

Seismic vibration instruments are affected seriously by the temperature changes. Devices employing variable resistance displacement sensors require correction factors to account for resistance change due to temperature. The damping of the instrument may also be affected by changes in the viscosity of the fluid due to temperature. For example, the viscosity of silicone oil, often used in these instruments, is strongly dependent on temperature. One way of eliminating the temperature effect is by using an electrical resistance heater in the fluid to maintain the temperature at a constant value regardless of surrounding temperatures.

36.8 Inertial Types, Cantilever, and Suspended-Mass Configuration

There are a number of different inertial accelerometers, including gyropendulum, reaction rotor, vibrating string, and centrifugal-force-balance types. In many of them, the force required to constrain the mass in the presence of the acceleration is supplied by an inertial system.

A vibrating string-type instrument, Figure 36.16, makes use of proof mass supported longitudinally by a pair of tensioned, transversely vibrating strings with uniform cross section and having equal lengths and masses. The frequency of vibration of the strings is set to several thousand cycles per second. The proof mass is supported radially in such a way that the acceleration normal to strings does not affect the tension of the string. In the presence of acceleration along the sensing axis, a differential tension exists on the two strings, thus altering the frequency of vibration. From the second law of motion, the frequencies can be written as

$$f_1^2 = \frac{T_1}{4m_s l} \quad \text{and} \quad f_2^2 = \frac{T_2}{4m_s l} \tag{36.34}$$

where

T is the tension

m_s and l are the masses and lengths of strings, respectively

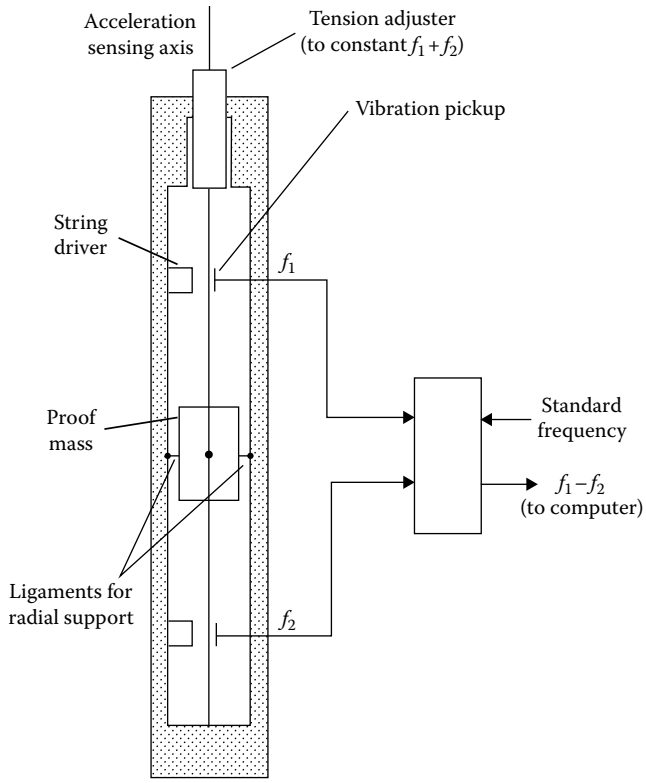


FIGURE 36.16 A vibrating string accelerometer. A proof mass is attached to two strings of equal mass and length and supported radially by suitable bearings. The vibration frequencies of strings are dependent on the tension imposed by the acceleration of the system in the direction of the sensing axis.

Quantity $(T_1 - T_2)$ is proportional to ma, where a is the acceleration along the axis of the strings. An expression for the difference of the frequency-squared terms may be written as

$$f_1^2 - f_2^2 = \left(\frac{T_1 - T_2}{4ml_s}\right) = \left(\frac{ma}{4m_s l}\right) \tag{36.35}$$

Hence,

$$f_1 - f_2 = \frac{ma}{\left[(f_1 + f_2)4m_s l\right]} \tag{36.36}$$

The sum of frequencies $(f_1 + f_2)$ can be held constant by servoing the tension in the strings with reference to the frequency of a standard oscillator. Then, the difference between the frequencies becomes linearly proportional to acceleration. In some versions, the beam-like property of the vibratory elements is used by gripping them at nodal points corresponding to the fundamental mode of vibration of the beam. Improved versions of these devices lead to cantilever-type accelerometers, as discussed next.

In cantilever-type accelerometers, a small cantilever beam mounted on the block is placed against the vibrating surface, and an appropriate mechanism is provided for varying the beam length. The beam length is adjusted such that its natural frequency is equal to the frequency of the vibrating surface—hence the resonance condition obtained. Slight variations of cantilever beam-type arrangements are finding new applications in microaccelerometers.

In a different type of suspended-mass configuration, a pendulum is used that is pivoted to a shaft rotating about a vertical axis. The pick-off is provided for the pendulum and the shaft speed. The system is servo controlled to maintain it at null position. Gravitational acceleration is balanced by the centrifugal acceleration. Shaft speed is proportional to the square root of the local value of the acceleration.

All inertial force accelerometers described earlier have the property of absolute instruments. That is, their scale factors can be predetermined solely by establishing mass, length, and time quantities, as distinguished from voltage, and spring stiffness.

36.9 Electrostatic Force Feedback Accelerometers

Electrostatic accelerometers are based on Coulomb's law between two charged electrodes. They measure the voltage in terms of force required to sustain a movable electrode of known area, mass, and separation from an affixed electrode. The field between the electrodes is given by

$$E = \frac{Q}{\varepsilon k S} \tag{36.37}$$

where

E is the intensity or potential gradient (dV/dx)
Q is the charge
S is the area of the conductor
k is the dielectric constant of the space outside the conductor

Using this expression, it can be shown that the force per unit area of the charged conductor (in N m^{-2}) is given by

$$\frac{F}{S} = \frac{Q^2}{2\varepsilon k S^2} = \frac{\varepsilon k E^2}{2} \tag{36.38}$$

In an electrostatic force feedback-type accelerometer (similar in structure to that in Figure 36.10), an electrode of mass m and area S is mounted on a light pivoted arm for moving relative to the fixed electrodes. The nominal gap, h, between the pivoted and fixed electrodes is maintained by means of a force balancing servo system capable of varying the electrode potential in response to signals from a pick-off that senses relative changes in the gaps.

Considering one movable electrode and one stationary electrode and assuming that the movable electrode is maintained at a bias potential V_1 and the stationary one at a potential V_2, the electrical intensity E in the gap can be expressed as

$$E_1 = \frac{(V_1 - V_2)}{h} \tag{36.39}$$

so that the force of attraction may be found as

$$F_1 = \frac{\varepsilon k E^2 S}{(2h^2)} = \frac{\varepsilon k (V_1 - V_2)^2 S}{(2h^2)} \tag{36.40}$$

In the presence of acceleration, if V_2 is adjusted to restrain the movable electrode to null position, the expression relating acceleration and electrical potential may be given by

$$a = \frac{F_1}{m} = \frac{\varepsilon k (V_1 - V_2)^2 S}{(2h^2 m)} \tag{36.41}$$

The device thus far described can measure acceleration in one direction only, and the output is of quadratic character; that is,

$$(V_1 - V_2) = D\sqrt{a} \tag{36.42}$$

where D is the constant of proportionality.

The output can be linearized in a number of ways, for example, by quarter-square method. If the servo controller applies a potential $-V_2$ to other fixed electrode, the force of attraction between this electrode and the movable electrode becomes

$$a = \frac{F_1}{m} = \frac{\varepsilon k(V_1 + V_2)^2 S}{(2h^2 m)} \tag{36.43}$$

and the force-balance equation of the movable electrode when the instrument experiences a downward acceleration a is

$$ma = F_1 - F_2 = \frac{[(V_1 + V_2)^2 - (V_1 - V_2)^2]\varepsilon kS}{(2h^2 m)}$$

or

$$= \frac{\varepsilon kS(4V_1V_2)}{(2h^2 m)} \tag{36.44}$$

Hence, if the bias potential V_1 is held constant and the gain of the control loop is high so that variations in the gap are negligible, the acceleration becomes a linear function of the controller output voltage V_2 as

$$a = V_2\left[\frac{\varepsilon kS(2V_1)}{(h^2 m)}\right] \tag{36.45}$$

The principal difficulty in mechanizing the electrostatic force accelerometer is the relatively high electric field intensity required to obtain adequate force. Also, extremely good bearings are necessary. Damping can be provided electrically or by viscosity of gaseous atmosphere in the interelectrode space if the gap h is sufficiently small.

The main advantages of the electrostatic accelerometers include extreme mechanical simplicity, low-power requirements, absence of inherent sources of hysteresis errors, zero temperature coefficients, and ease of shielding from stray fields.

36.10 Microaccelerometers

By the end of the 1970s, it became apparent that the essentially planar processing IC (integrated circuit) technology could be modified to fabricate 3D electromechanical structures, called micromachining. Accelerometers and pressure sensors were among the first IC sensors and today they are produced in large numbers by a diverse range of manufacturers. The first accelerometer was developed in 1979. Since then, technology has been progressing rapidly to the point where an extremely diverse range of accelerometers are cost-effectively available for mass consumption. Most sensors use

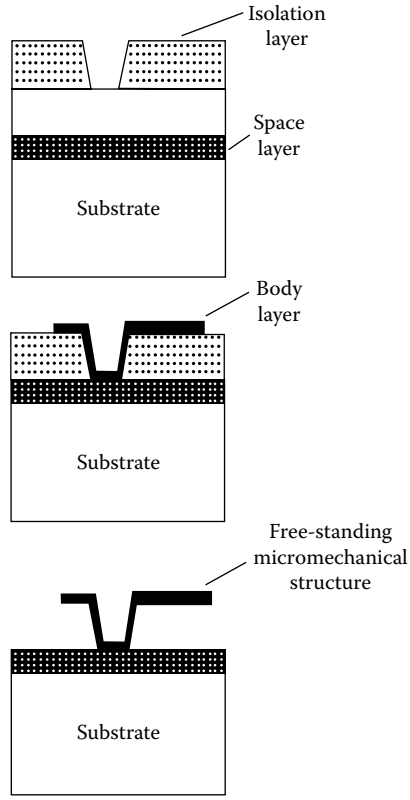


FIGURE 36.17 Steps of surface micromachining. The acceleration-sensitive, 3D structure is formed on a substrate and a sacrificial element. The sacrificial element is etched to leave a free-standing structure. The spacing between the structure and the substrate is about 2 μm.

bulk micromachining rather than surface micromachining techniques. In bulk micromachining, the flexures, resonant beams, and all other critical components of the accelerometer are made from bulk silicon in order to exploit the full mechanical properties of single-crystal silicon. With proper design and film process techniques, the bulk micromachining yields in manufacturing of extremely stable and robust accelerometers.

The selective etching of multiple layers of deposited thin films, or surface micromachining, allows movable microstructures to be fabricated on silicon wafers. With surface micromachining, layers of structure material are disposed and patterned, as shown in Figure 36.17. These structures are formed by polysilicon and a sacrificial material such as silicon dioxide. The sacrificial material acts as an intermediate spacer layer and is etched away to produce a free-standing structure. Surface machining technology also allows smaller and more complex structures to be built in multiple layers on a single substrate.

The operational principles of microaccelerometers are very similar to capacitive force-balance-type accelerometers or vibrating beam types, as discussed earlier. Nevertheless, manufacturing techniques may change from one manufacturer to another. In general, the vibrating beam accelerometers are preferred because of better air-gap properties and improved bias performance characteristics.

The vibrating beam accelerometers, also called resonant beam force transducers, are made in such a way that the acceleration along a positive input axis places the vibrating beam in tension. Thus, the resonant

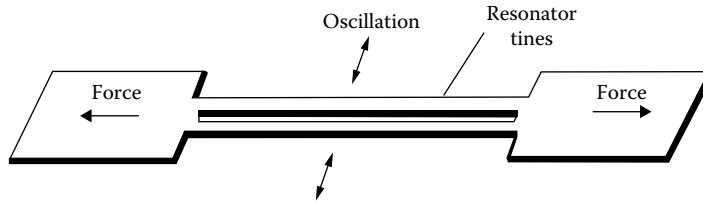


FIGURE 36.18 A DETF acceleration sensor. Two beams are vibrated 180° out of phase to eliminate reaction forces at the beam ends. The resonant frequency of the beam is altered by the acceleration. The signal processing circuits are also integrated in the same chip.

frequencies of the vibrating beam increase or decrease with the applied acceleration. Another version known as the mechanically coupled beam structure, also known as a double-ended tuning fork (DETF), is shown in Figure 36.18.

In DETF, an electronic oscillator capacitively couples energy into two vibrating beams to keep them oscillating at their resonant frequency. The beams vibrate 180° out of phase to cancel reaction forces at the ends. The dynamic cancellation effect of the DETF design prevents energy from being lost through the ends of the beam. Hence, the dynamically balanced DETF resonator has a high Q factor, which leads to a stable oscillator circuit. The acceleration signal is an output from the oscillator as a frequency-modulated square wave that can be used for digital interface.

The frequency of resonance of the system must be much higher than any input acceleration, and this limits the measurable range. In a typical military micromachine accelerometer, the following characteristics are given: range ±1,200 g, sensitivity 1.11 Hz g^{-1}, bandwidth 2,500 Hz, unloaded DETF frequency 9,952 Hz, frequency at +1,200 g 11,221 Hz, frequency at –1,200 g 8,544 Hz, and temperature sensitivity 5 mg°C^{-1}. Accelerometer size is 6 mm diameter × 4.3 mm length, with a mass of about 9 g, and it has a turn on time less than 60 s. The accelerometer is powered with +9 to +16 V DC and the nominal output is a 9000 Hz square wave.

Surface micromachining has also been used to manufacture specific application accelerometers, such as air-bag applications in automotive industry. In one type, a three-layer differential capacitor is created by alternate layers of polysilicon and phosphosilicate glass (PSG) on a 0.38 mm thick and 100 mm long wafer. A silicon wafer serves as the substrate for the mechanical structure. The trampoline-shaped middle layer is suspended by four supporting arms. This movable structure is the seismic mass for the accelerometer. The upper and lower polysilicon layers are fixed plates for the differential capacitors. The glass is sacrificially etched by hydrofluoric acid.

36.10.1 3D Accelerometers

Micromachined IC accelerometers are produced in a single chip for measuring accelerations in three dimensions: X, Y, and Z axes. They offer integrated features including multiple analog to digital converters (ADCs), digital low-pass filters, and other signal processors. They have selectable sensitivity ranges, typically ±2, ±4, ±8, or ±16 g. Many of these devices can be configured to detect motion pulses as single taps, double taps, and 0 g (free fall) conditions on any or all axes. They also can be integrated with devices such as gyroscopes and temperature sensors that require sensitive signal processing techniques. They are small and can be housed in 3 × 3 × 1 mm packages.

36.11 Cross-Axis Sensitivity

A vibrational structure may have been subjected to different forms of vibrations, such as compressional, torsional, and transverse, or a combination of all these vibrations may take place simultaneously, which makes the analysis and measurements difficult and complex. It was discussed

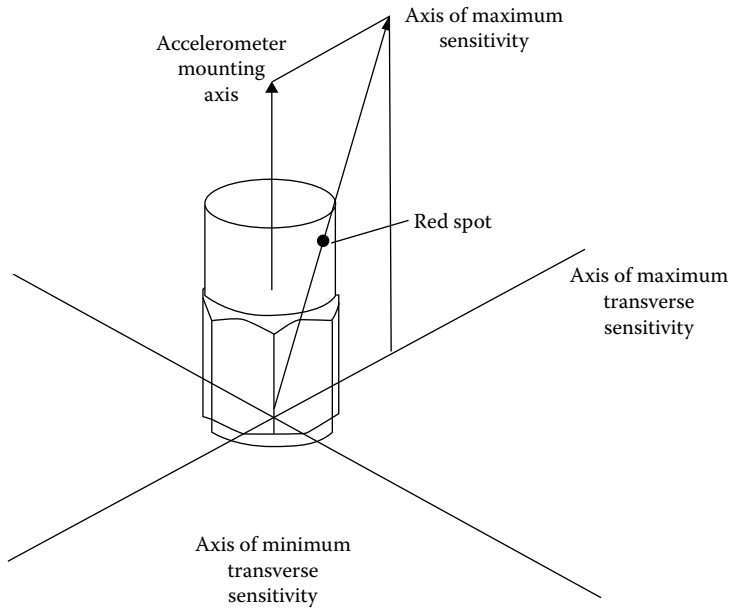


FIGURE 36.19 Vectorial illustration of cross-axis sensitivity. Accelerometers may sense vibrations not only in the direction of main axis but also perpendicular to the main axis. These cross-axis responses are minimized in many accelerometers to a value less than 1%. Sometimes, this sensitivity is used to determine the correct orientation of the device.

earlier that the differential equations governing the vibrational motion of a structure were dependent on the number of degrees of freedom, which can be described as a function of the space coordinates $f(x, y, z, t)$. For example, the transverse vibrations of structures may be a fourth-order differential equation.

Fortunately, most common acceleration and vibration measurements are simple in nature, being either compressional or torsional types. They can easily be expressed as second-order differential equations, as explained in the frequency response section. However, during measurements, most accelerometers are affected by transverse vibrations and their sensitivity can play a major role in the accuracy of the measurements.

The transverse, also known as cross-axis sensitivity, of an accelerometer is its response to acceleration in a plane perpendicular to the main accelerometer axis, as shown in Figure 36.19. The cross-axis sensitivity is normally expressed in percent of the main axis sensitivity and should be as low as possible. There is not a single value of cross-axis sensitivity, but it varies depending on the direction. The direction of minimum sensitivity is usually supplied by the manufacturer.

The measurement of the maximum cross-axis sensitivity is part of the individual calibration procedure and should always be less than 3%–4%. If high levels of transverse vibration are present, this may result in erroneous overall results. In this case, separate arrangements should be made to establish the level and frequency contents of the cross-axis vibrations. Cross-axis sensitivities of typical accelerometers are mentioned in the relevant sections: 2%–3% for piezoelectric types and less than 1% in most others.

36.12 Selection, Full-Scale Range, and Overload Capability

Ultimate care must be exercised for the selection of the correct accelerometer to meet the requirements of a particular application. At first glance, there may seem to be a confusingly large repertoire of accelerometers available; however, they can be classified into two main groups. The first group is

the general-purpose accelerometers offered in various sensitivity, frequency, full-scale range, and overload capabilities together with different mechanical and electrical connection options. The second group of accelerometers is the special types that have characteristics targeted toward a particular application.

In deciding the application type (e.g., general purpose or special) and the accelerometer to be employed, the following characteristics need to be considered: transient response or cross-axis sensitivity; frequency range; sensitivity, mass, and dynamic range; cross-axis response; and environmental conditions such as temperature and cable noise. Some useful hints about these characteristics are given in the following.

36.12.1 Frequency Range

Acceleration measurements are normally confined to using the linear portion of the response curve. The response is limited at the low frequencies as well as at the high frequencies by their natural resonances. As a rule of thumb, the upper frequency limit for the measurement can be set to one-third of the accelerometer's resonance frequency such that the vibrations measured will be less than 1 dB in linearity. It should be noted that an accelerometer's useful frequency range can significantly be higher, that is, to 1/2 or 2/3 of its resonant frequency. The measurement frequencies may be set to higher values in applications where lower linearity (e.g., 3 dB) may be acceptable, as in the case of monitoring internal conditions of machines since the reputability is more important than the linearity. The lower measuring frequency limit is determined by two factors. The first is the low-frequency cut-off of the associated preamplifiers. The second is the effect of ambient temperature fluctuations that an accelerometer could be sensitive to.

36.13 Selection, Full-Scale Range, and Overload Capability (Frequency Range): Sensitivity, Mass, and Dynamic Range

Ideally, the higher the transducer sensitivity, the better, but compromises might have to be made for sensitivity versus frequency, range, overload capacity, and size.

Accelerometer mass becomes important when using it on small and light test objects. The accelerometer should not load the structural member, since additional mass can significantly change the levels and frequency presence at measuring points and invalidate the results. As a general rule, the accelerometer mass should not be greater than one-tenth the effective mass of the part or the structure that it is mounted onto for measurements.

The dynamic range of the accelerometer should match the high or low acceleration levels of the measured objects. General-purpose accelerometers can be linear up to 5,000 *g*–10,000 *g*, which is well into the range of most mechanical shocks. Special accelerometers can measure up to 100,000 *g*.

An important point in the practical application of accelerometers is that if mechanical damping is a problem, air damping is preferable to oil damping, since oil damping is extremely sensitive to viscosity changes. If the elements are temperature stable, electronic damping may be sufficient.

36.13.1 Transient Response

Shocks are characterized as sudden releases of energy in the form of short-duration pulses exhibiting various shapes and rise times. They have high magnitudes and wide frequency contents. In applications where transients and shock measurements are involved, the overall linearity of the measuring system may be limited to high and low frequencies by phenomena known as *zero shift*

and *ringing*, respectively. The zero shift is caused by both the phase nonlinearity in the preamplifiers and the accelerometer not returning to steady-state operating conditions after being subjected to high shocks. Ringing is caused by high-frequency components of the excitation near resonance frequency preventing the accelerometer to return to its steady state. To avoid measuring errors due to these effects, the operational frequency of the measuring system should be limited to the linear range.

36.13.2 Full-Scale Range and Overload Capability

Most accelerometers are able to measure acceleration in both positive and negative directions. They are also designed to be able to accommodate overload capacity. Appropriate discussions are made on full-scale range and overload capacity of accelerometers in the relevant sections. Manufacturers usually supply information on these characteristics.

36.13.3 Environmental Conditions

In the selection and implementation of accelerometers, environmental conditions such as temperature ranges, temperature transients, cable noise, magnetic field effects, humidity, and acoustic noise need to be considered. Manufacturers also supply information on the environmental conditions for effective usage.

36.14 Signal Conditioning

Most of the common signal conditioners are appropriate to use for interfacing accelerometers to computers or other instruments. Particular caution needs to be exercised to provide appropriate electric loads when self-generating accelerometers are used. Generally, the generated raw signals are amplified and filtered suitably by the circuits within the accelerometer casing supplied by the manufacturers. Nevertheless, piezoelectric and piezoresistive sensors require special signal conditioners with certain characteristics, as will be discussed in Sections 36.14.1 and 36.14.2. Examples of signal-conditioning circuits are also given for microaccelerometers.

36.14.1 Signal-Conditioning Piezoelectric Accelerometers

The piezoelectric accelerometers supply very small energy to the signal conditioners. They have high capacitive source impedances. The equivalent circuit of a piezoelectric accelerometer can be regarded as an active capacitor that charges itself when loaded mechanically. The configuration of external signal-conditioning elements is dependent on the equivalent circuit selected. The charge amplifier design of the conditioning circuit is the most common approach since the system gain and low-frequency responses are well defined. The performance of the circuit is independent of cable length and capacitance of the accelerometer.

A charge amplifier consists of a charge converter output voltage that occurs as a result of the charge input signal returning through the feedback capacitor to maintain the input voltage at the input level close to zero, as shown in Figure 36.20. An important point about charge amplifiers is that their sensitivities can be standardized. They basically convert the input charge to voltage first and then amplify this voltage. With the help of basic operational-type feedback, the amplifier input is maintained essentially at zero volts; therefore, it looks like a short circuit to the input. The charge converter output voltage that occurs as a result of a charge input signal is returned through

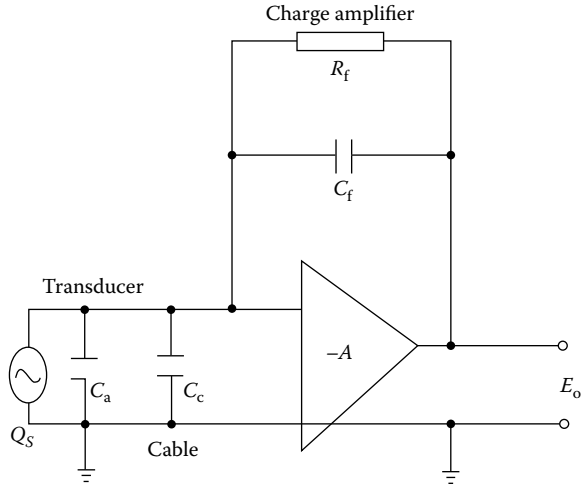


FIGURE 36.20 A typical charge amplifier. The transducer charge, proportional to acceleration, is first converted to voltage form so that it can be amplified. The output voltage is a function of the input charge. The response of the amplifier can be approximated by a first-order system. In PZT transducers, the preamplifier is integrated within the same casing.

the feedback capacitor to maintain the voltage at the input level near zero. Thus, the charge input is stored in the feedback capacitor, producing a voltage across it, that is, equal to the value of the charge input divided by the capacitance of the feedback capacitor. The complete transfer function of the circuit describing the relationship between the output voltage and the input acceleration magnitude can be determined by

$$\frac{E_o}{a_0} = \frac{S_a j R_f C_f \omega}{\left\{1 + jR_f C_f \left[1 + (C_a + C_c)/(1+G) \times C_f\right]\omega\right\}} \tag{36.46}$$

where

E_o is the charge converter output (V)
a_0 is the magnitude of acceleration (m s^{-2})
S_a is the accelerometer sensitivity (mV g^{-1})
C_a is the accelerometer capacitance (F)
C_c is the cable capacitance (F)
C_f is the feedback capacitance (F)
R_f is the feedback loop resistance
G is the amplifier open loop gain

In most applications, since C_f is selected to be large compared to $(C_a + C_c)/(1+G)$, the system gain becomes independent of the cable length. In this case, the denominator of the equation can be simplified to give a first-order system with roll-off at

$$f_{-3\,\mathrm{dB}} = \frac{1}{(2\pi R_f C_f)} \tag{36.47}$$

with a slope of 20 dB per decade. For practical purposes, the low-frequency response of this system is a function of well-defined electronic components and does not vary with cable length. This is an important feature when measuring low-frequency vibrations.

Many accelerometers are manufactured with preamplifiers and other signal-conditioning circuits integrated with the transducer enclosed in the same casing. Some accelerometer preamplifiers include integrators to convert the acceleration proportional outputs to either velocity or displacement proportional signals. To attenuate noise and vibration signals that lie outside the frequency range of interest, most preamplifiers are equipped with a range of high-pass and low-pass filters. This avoids interference from electric noise or signals inside the linear portion of the accelerometer frequency range. Nevertheless, it is worth mentioning that these devices usually have two time constants, external and internal. The mixture of these two time constants can lead to problems, particularly at low frequencies. The internal time constant is usually fixed by the manufacturer in design and construction. Special care must be observed to take care of the effect of external time constants by mainly observing strict impedance matching.

36.14.2 Signal Conditioning of Piezoresistive Transducers

Piezoresistive sensors have high-amplitude outputs, low output impedances, and low intrinsic noise. Most of them are designed for constant voltage excitations. They are usually calibrated for constant current excitations to minimize external influences. Many piezoresistive accelerometers are configured as part of a full bridge. In some cases, all four arms of the bridge are made from the sensors, as shown in Figure 36.21.

36.14.3 Microaccelerometers

In microaccelerometers, signal-conditioning circuitry is integrated within the same chip with the sensor as shown in Figure 36.22. A typical example of the signal-conditioning circuitry is given in Figure 36.23 in block diagram form. In this type of accelerometer, the electronic system is essentially a crystal-controlled oscillator circuit and the output signal of the oscillator is a frequency-modulated acceleration signal. Some circuits provide a buffered square-wave output that can directly be interfaced digitally. In this case, the need for analog to digital conversion is eliminated, thus removing one of the major sources of error. In other types of accelerometers, signal-conditioning circuits such as ADCs are retained within the chip.

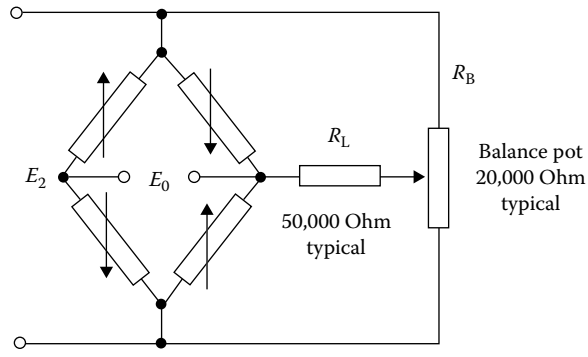


FIGURE 36.21 Bridge circuit for piezoresistive and strain-gage accelerometers. The strain gages form the four arms of the bridge. The two extra resistors are used for balancing and for the purpose of precision adjustment. This type of arrangement reduces temperature effects.

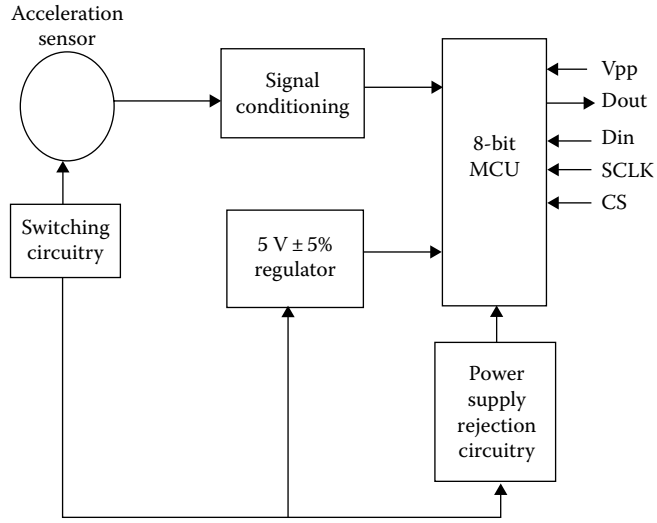


FIGURE 36.22 A block diagram of an accelerometer combined with MCU. The signal-conditioning, switching, and power supply circuits are integrated to form a microaccelerometer. The device can directly be interfaced with a digital signal processor or a computer. In some cases, ADCs and memory are also integrated within the chip.

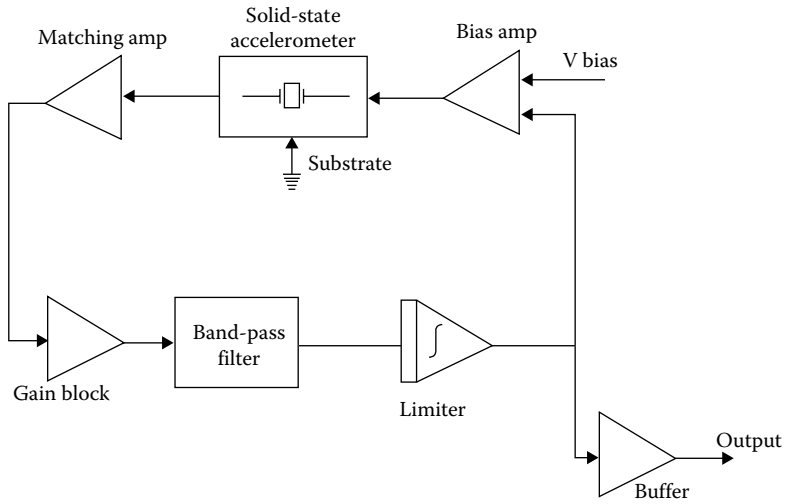


FIGURE 36.23 Block diagram of a signal-conditioning circuit of a microaccelerometer. The output signal of the oscillator is a frequency-modulated acceleration signal. The circuit provides a buffered square wave so that the output can directly be read into a digital device.

Bibliography

Bernstein, J., An overview of MEMS inertial sensing technology. http://www.sensorsmag.com/sensors/acceleration-vibration/an-overview-mems-inertial-sensing-technology-970 (accessed on September 9, 2011).

Connolly, C., Vibration isolation theory and practice, In: Conneely, K. (ed.), *Assembly Automation*, Emerald Publisher, U.K., Vol. 29, No. 1, pp. 8–13, 2009.

Doebelin, E. O., *Measurement Systems: Application and Design*, 4th edn., New York: McGraw-Hill, 1990.

Fleming, W. E., Overview of automotive sensors, *IEEE Sensors Journal*, 1(4), 296–308, 2001.

Hong, J. and Park, K., Design and control of six degree-of-freedom active vibration isolation table, *Review of Scientific Instruments*, American Institute of Physics, Vol. 81, No. 3, Article number 035106, 2010.

McConnell, K. G., *Vibration Testing: Theory and Practice*, New York: John Wiley & Sons, 1995.

Omega Engineering, Inc., Force Acceleration and Torque. http://www.omega.com/literature/transactions/volume3/force2.html (accessed on September 9, 2011).

Ripper, G. P., Dias, R. S., and Garcia, G. A., Primary accelerometer calibration problems due to vibration exciters, *Measurement*, 42(9), 1363–1369, 2009.

Silicon Design Inc., MEMS Accelerometers. http://www.silicondesigns.com/tech.html (accessed on September 15, 2011).

Stauffer, J. M., Dutoit, B., and Arbab, B., Standard MEMS sensor technologies for harsh environment, *Institution of Engineering and Technology Seminar on MEMS Sensors and Actuators*, London, U.K., pp. 93–96, 2006.

Partial List of Vendors and Suppliers

Allied International, 7 Hill St. Bedford Hills, NY 10507-1811, Phone: 914-241-6900, Fax: 914-241-6985, www.alliedinter.com

Allied Signal Aerospace, 15001-T N.E. 36th St. Redmond, WA 98073-9701, Phone: 425-885-3711, Fax: 425-885-8866, www.alliedinter.com

Automation Products Group, Inc.(APGSENSORS), 1025 W. 1700 N. Logan, UT 84321, Phone: 435-753-7300, 888-525-7300 (toll free), Fax: 435-753-7490, www.apgsensors.com

CEC Vibration Products Inc., 746 Arrow Grand Circle Covina, CA 91722, Phone: 626-468-1345, 800-468-1345 (toll free), Fax: 626-938-0202, www.cecvp.com

Davis/Inotek, 300 Garden Oaks Houston, TX 77018, Phone: 800-548-9409, Fax: 410-358-0252, www.davis.com

Electro-Sensors, Inc., 6111 Blue Circle Dr. Minnetonka, MN 55343-9108, Phone: 952-930-0100, 800-328-6170 (toll free), Fax: 952-930-0130, www.electro-sensors.com

Honeywell Sensing and Control, 1985 Douglas Dr. North, MN10-192B Golden Valley, MN 55422, Phone: 763-954-4818, 800-446-6555 (toll free), sensing.honeywell.com/index.cfm/ci_id/15…

Inertia Switch, Inc., 70 S. Greenbush Rd. Orangeburg, NY 10962, Phone: 845-359-8300, Fax: 845-359-6227, www.inertiaswitch.com

Instrumented Sensor Technology, Inc., 4704 Moore Street Okemos, MI 48864, Phone: +1 517 349-8487, Fax: +1 517 349-8469, www.isthq.com

Integrated Dynamics Engineering, Inc., 377 T-University Ave Westwood, MA 02090, Phone: 781-326-5700, Fax: 781-326-3004, www.ideworld.com

Jewell Instruments, LLC, 850 Perimeter Rd. Manchester, NH 03103, Phone: 603-669-6400, 800-638-3771 (toll free), Fax: 603-669-5962, www.jewellinstruments.com

Kearfott Guidance & Navigation, US Hwy. 70 Black Mountain, NC 28711, Phone: 828-686-3811, Fax: 828-686-3567, www.kearfott.com

Kinetic Ceramics, Inc., 26242 Industrial Blvd. Hayward, CA 94545-5012, Phone: 510-264-2140, 800-806-6016 (toll free), Fax: 510-264-2159, www.kineticceramics.com

Kistler North America, 75 John Glenn Dr. Amherst, NY 14228-2119, Phone: 716-691-5100, 888-KISTLER, 888-547-8537 (toll free), Fax: 716-691-5226, www.kistler.com

Measurement Specialties, Inc., 1000 Lucas Way Hampton, VA 23666, Phone: 757-766-1500, 800-745-8008 (toll free), Fax: 757-766-4297, www.meas-spec.com

Motion Sensors, Inc., 786 Pitts Chapel Rd. Elizabeth City, NC 27909, Phone: 252-331-2080, Fax: 252-331-1666, www.motionsensors.com

MTI Instruments, Inc., 325 Washington Ave., Extension Albany, NY 12205-5505, Phone: 518-218-2550, Fax: 518-218-2506, www.mtiinstruments.com

PCB Piezotronics, Inc., 3425 Walden Ave. Depew, NY 14043-2417, Phone: 716-684-0001, 800-828-8840 (toll free), Fax: 716-684-0987, www.pcb.com

Piezo Technologies, 8431 Georgetown Rd., Suite 300 Indianapolis, IN 46268, Phone: 317-876-4670, Fax: 317-876-4681, www.piezotechnologies.com

Schneider Electric-Hyde Park, 1875 Founders Dr. Dayton, OH 45420, Phone: 937-514-7207, Fax: 937-258-5830, www.sesensors.com

Select Controls, Inc., 45 Knickerbocker Ave. Bohemia, NY 11716-3110, Phone: 866-929-5312 (toll free), Fax: 631-567-9013, www.select-controls.com

Shock Watch, 1111 W. Mockingbird Lane, Suite 1050 Dallas, TX 75247, Phone: 214-630-9625, 800-527-9497 (toll free), Fax: 214-638-4512, www.shockwatch.com/

Silicon Designs, Inc., 1445 NW Mall St. Issaquah, WA 98027, Phone: (425) 391-8329, Fax: (425) 391-0446, www.silicondesigns.com/contact.html

Spectrum Sensors & Controls, 3939 West Ridge Rd. Suite A—108 Erie, PA 16505, Phone: 814-474-1571, www.specsensors.com

Vitec, Inc., 24755 Highpoint Rd. Cleveland, OH 44122-5915, Phone: 216-464-4670, 800-321-6384 (toll free), Fax: 216-464-5324, www.vitec-inc.com

Warren-Knight Instrument Co., 2045 Bennett Dr. Philadelphia, PA 19116-3019, Phone: 215-464-9300, Fax: 215-464-9303, www.warrenind.com

Wilcoxon Research, Inc., 20511 Seneca Meadows Pkwy. Germantown, MD 20876, Phone: 301-330-8811, 800-945-2696 (toll free), Fax: 301-330-8873, www.wilcoxon.com

37

Strain Measurement

37.1 Fundamental Definitions of Strain **37**-1
37.2 Principles of Operation of Strain Sensors **37**-6
Piezoresistive Foil Gages • Piezoresistive Semiconducting Gages • Piezoelectric Gages • Fiber-Optic Strain Gages • Birefringent Film Strain Sensing • Moiré Strain Sensing
References **37**-17

Christopher S. Lynch
Georgia Tech

This chapter begins with a review of the fundamental definitions of strain and ways it can be measured. This is followed by a review of the many types of strain sensors and their application and sources for strain sensors and signal conditioners. Next, a more detailed look is taken at operating principles of various strain measurement techniques and the associated signal conditioning.

37.1 Fundamental Definitions of Strain

Stress and strain are defined in many elementary textbooks about the mechanics of deformable bodies [1,2]. The terms *stress* and *strain* are used to describe loads on and deformations of solid materials. The simplest types of solids to describe are homogeneous and isotropic. *Homogeneous* means the material properties are the same at different locations and *isotropic* means the material properties are independent of direction in the material. An annealed steel bar is homogeneous and isotropic, whereas a human femur is not homogeneous because the marrow has very different properties from the bone, and it is not isotropic because its properties are different along the length and along the cross section.

The concepts of stress and strain are introduced in the context of a long homogeneous isotropic bar subjected to a tensile load (Figure 37.1). The stress σ is the applied force F, divided by the cross-sectional area A. The resulting strain ε is the length change ΔL, divided by the initial length L. The bar elongates in the direction the force is pulling (longitudinal strain ε_L) and contracts in the direction perpendicular to the force (transverse strain ε_t).

When the strain is not too large, many solid materials behave like linear springs; that is, the displacement is proportional to the applied force. If the same force is applied to a thicker piece of material, the spring is stiffer and the displacement is smaller. This leads to a relation between force and displacement that depends on the dimensions of the material. Material properties, such as the density and specific heat, must be defined in a manner that is independent of the shape and size of the specimen. Elastic material properties are defined in terms of stress and strain. In the linear range of material response, the stress is proportional to the strain (Figure 37.2). The ratio of stress to strain for the bar under tension is an elastic constant called the Young's modulus E. The negative ratio of the transverse strain to longitudinal strain is the Poisson's ratio ν.

Forces can be applied to a material in a manner that will cause distortion rather than elongation (Figure 37.3). A force applied tangent to a surface divided by the cross-sectional area is described as a shear stress τ. The distortion can be measured by the angle change produced. This is the shear strain γ

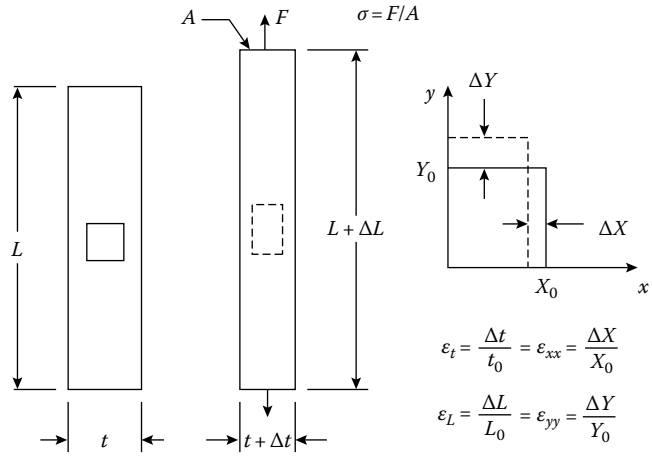


FIGURE 37.1 When a homogeneous isotropic bar is stretched by a uniaxial force, it elongates in the direction of the force and contracts perpendicular to the force. The relative elongation and contraction are defined as the longitudinal and transverse strains, respectively.

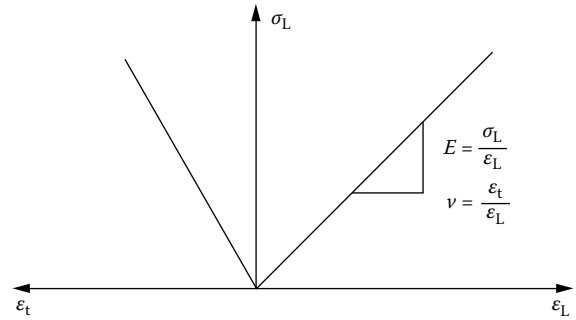


FIGURE 37.2 The uniaxial force shown in Figure 37.1 produces uniaxial stress in the bar. When the material response is linear, the slope of the stress versus strain curve is the Young's modulus. The negative ratio of the transverse to longitudinal strain is the Poisson's ratio.

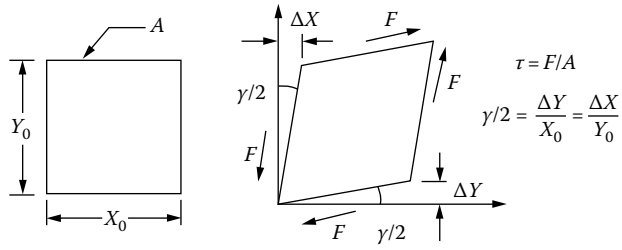


FIGURE 37.3 When a block of material is subjected to forces parallel to the sides as shown, it distorts. The force per unit area is the shear stress τ, and the angle change is the shear strain γ.

when the angle change is small. When the relation between shear stress and shear strain is linear, the ratio of the shear stress to shear strain is the shear modulus G.

Temperature change also induces strain. This is thermal expansion. In most materials, thermal strain increases with temperature. Over a limited temperature range, the relationship between thermal strain and temperature is linear. In this case, the strain divided by the temperature change is the thermal expansion coefficient α. In isotropic materials, thermal expansion only produces elongation strain, no shear strain.

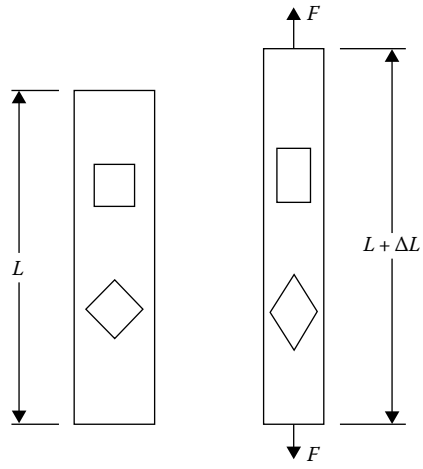


FIGURE 37.4 Some elements of a bar under uniaxial tension undergo elongation and contraction. These elements lie in principal directions. Other elements undergo distortion as well.

Principal directions in a material are directions that undergo elongation but no shear. On any particular surface of a solid, there are always at least two principal directions in which the strain is purely elongation. This is seen if two squares are drawn on the bar under uniform tension (Figure 37.4). When the bar is stretched, the square aligned with the load is elongated, whereas the square at 45° is distorted (angles have changed) and elongated. If the principal directions are known, as with the bar under tension, then strain gages can be applied in these directions. If the principal directions are not known, such as near a hole or notch, in an anisotropic specimen, or in a structure with complicated geometry, then additional strain gages are needed to fully characterize the strain state.

The elastic and thermal properties can be combined to give Hooke's law, Equations 37.1 through 37.6:

$$\varepsilon_{xx} = \frac{\sigma_{xx}}{E} - \frac{\nu}{E}(\sigma_{yy} + \sigma_{zz}) \tag{37.1}$$

$$\varepsilon_{yy} = \frac{\sigma_{yy}}{E} - \frac{\nu}{E}(\sigma_{xx} + \sigma_{zz}) \tag{37.2}$$

$$\varepsilon_{xx} = \frac{\sigma_{zz}}{E} - \frac{\nu}{E}(\sigma_{xx} + \sigma_{yy}) \tag{37.3}$$

$$\gamma_{xy} = \frac{\tau_{xy}}{G} \tag{37.4}$$

$$\gamma_{xz} = \frac{\tau_{xz}}{G} \tag{37.5}$$

$$\gamma_{yz} = \frac{\tau_{yz}}{G} \tag{37.6}$$

Several types of sensors are used to measure strain. These include piezoresistive gages (foil or wire strain gages and semiconductor strain gages), piezoelectric gages (polyvinylidene fluoride [PVDF] film and quartz), fiber-optic gages, birefringent films and materials, and moiré grids. Each type of sensor requires its own specialized signal conditioning. Selection of the best strain sensor for a given measurement is based on many factors, including specimen geometry, temperature, strain rate, frequency, magnitude, as well as cost, complexity, accuracy, spatial resolution, time resolution, sensitivity to transverse strain, sensitivity to temperature, and complexity of signal conditioning. Table 37.1 describes typical

TABLE 37.1 Comparison of Strain Sensors

Description	Longitudinal Strain Sensitivity	Transverse Strain Sensitivity	Temperature Sensitivity	Strain Resolution	Spatial Resolution	Time Resolution (μs)	Measurable Strain Range (%)
Piezoresistive constantan foil	$\Delta R/R/\Delta\varepsilon_L = 2.1$	$\Delta R/R/\Delta\varepsilon_t = <0.02$	$\Delta R/R/\Delta T = 2 \times 10^{-6}\ °C^{-1}$	<1 μstrain[a]	5–100 mm[b]	<1[c]	0–3
Annealed constantan foil[d]	$\Delta R/R/\Delta\varepsilon_L = 2.1$	$\Delta R/R/\Delta\varepsilon_t = <0.02$	$\Delta R/R/\Delta T = 2 \times 10^{-6}\ °C^{-1}$	<11 μstrain	5–100 mm	<1	0–10
Piezoresistive semiconductor	$\Delta R/R/\Delta\varepsilon_L = 150$	$\Delta R/R/\Delta\varepsilon_t = ???$	$\Delta R/R/\Delta T = 1.7 \times 10^{-3}\ °C^{-1}$	<0.1 μstrain	1–15 mm	<1	0–0.1
Piezoelectric PVDF	$\Delta Q/A/\Delta\varepsilon_L = 120\ nC/m^2/\mu\varepsilon$	$\Delta Q/A/\Delta\varepsilon_t = 60\ nC/m^2/\mu\varepsilon$	$\Delta Q/A/\Delta T = -27\ \mu C\ m^{-2}\ °C^{-1}$	1–10 μstrain	Gage size	<1	0–30
Piezoelectric quartz	$\Delta Q/A/\Delta\varepsilon_L = 150\ nC/m^2/\mu\varepsilon$ bonded to steel		$\Delta Q/A/\Delta T = 0$	<0.01 μstrain 20 mm gage	Gage size	<10	0–0.1
Fiber-optic Fabry–Perot	2–1000 μstrain V^{-1}	Near zero		<1 μstrain	2–10 mm	<20	
Birefringent film	K[e] = 0.15–0.002				0.5 mm[f]	<5	0.05–5
Moiré	1 fringe order/417 nm displ.	1 fringe order/417 nm displ.	Not defined	41.7 με over 10 mm	Full fields[g]	Limited by signal conditioning	0.005–5

[a] With good signal conditioning.
[b] Equal to grid area.
[c] Gage response is within 100 ns. Most signal conditioning limits response time to far less than this.
[d] Annealed foil has a low yield stress and a large strain to failure. It also has hysteresis in the unload and a zero shift under cyclic load.
[e] This technique measures a difference in principal strains. $\varepsilon_2 - \varepsilon_1, = N\lambda/2tK$.
[f] Approximately the film thickness.
[g] The spatial strain resolution depends on the strain level. This is a displacement measurement technique.

TABLE 37.2 Sources and Prices of Strain Sensors

Supplier	Address	Sensor Types	Sensor Cost	Signal Conditioning	Cost
Micro-Measurements	P.O. Box 27777, Raleigh, NC 27611	Piezoresistive foil	From $5.00	Wheatstone bridge	From $500
		Birefringent film	From $10.00	Polariscope	$5,000–10,000
Texas Measurements	P.O. Box 2618, College Station, TX 77841	Piezoresistive foil and wire Load cells	From $5.00		
Omega Engineering	P.O. Box 4047, Stamford, CT 06907-0047	Piezoresistive foil	From $5.00	Strain meter Wheatstone bridge	From $550 From $2700
Dynasen, Inc.	20 Arnold Pl., Goleta, CA 93117	Piezoresistive foil	From $55.00	2-Channel pulsed Wheatstone bridge	$5000
		Specialty gages for shock wave measurements	From $55.00		$250.00
		Piezoelectric PVDF (calibrated)		Passive charge integrator	
Entran Sensors and Electronics, Entran Devices, Inc.	10 Washington Ave., Fairfield, NJ 07004-3877	Piezoresistive semiconductor	From $15.00		
Amp, Inc.	Piezo Film Sensors P.O. Box 799, Valley Forge, PA 19482	Piezoelectric PVDF (not calibrated)	From $5.00		
Kistler Instrument Corp.	Amherst, NY 14228-2171	Piezoelectric quartz		Electrometer Charge amplifier	
F&S, Inc.	Fiber and Sensor Technologies P.O. Box 11704, Blacksburg, VA 24062-1704	Fabry–Perot strain sensors	From $75	Electronics	From $3500.00
Photomechanics, Inc.	512 Princeton Dr., Vestal, NY 13850-2912	Moiré interferometer			$60,000

characteristics of several sensors. Table 37.2 lists some manufacturers and the approximate cost of the sensors and associated signal conditioning electronics.

The data in Table 37.1 are to be taken as illustrative and by no means complete. The sensor description section describes only the type of sensor, not the many sizes and shapes. The longitudinal strain sensitivity is given as sensor output per unit longitudinal strain in the sensor direction. If the signal conditioning is included, the sensitivities can all be given in volts out per unit strain [3,4], but this is a function of amplification and the quality of the signal conditioner. The temperature sensitivity is given as output change due to a temperature change. In many cases, higher strain resolution can be achieved, but resolving smaller strain is more difficult and may require vibration and thermal isolation. For the moiré technique, the strain resolution is a function of the length of the viewing area. This technique can resolve a displacement of 100 nm (1/4 fringe order). This is divided by the viewing length to obtain the strain resolution. The spatial resolution corresponds to the gage length for most of the sensor types. The measurable strain range listed is the upper limit for the various sensors. Accuracy and reliability are usually reduced when sensors are used at the upper limit of their capability.

Manufacturers of the various sensors provide technical information that includes details of using the sensors, complete calibration or characterization data, and details of signal conditioning. The extensive

technical notes and technical tips provided by Measurements Group, Inc. address such issues as thermal effects [5], transverse sensitivity corrections [6], soldering techniques [7], rosettes [8], and gage fatigue [9]. Strain gage catalogs include information about gage materials, sizes, and selection. Manufacturers of other sensors provide similar information.

37.2 Principles of Operation of Strain Sensors

37.2.1 Piezoresistive Foil Gages

Piezoresistive foil and wire gages comprise a thin insulating substrate (usually polyimide film), a foil or wire grid (usually constantan) bonded to the substrate, lead wires to connect the grid to a resistance measuring circuit, and often an insulating encapsulation (another sheet of polyimide film) (Figure 37.5). The grid is laid out in a single direction so that strain will stretch the legs of the grid in the length direction. The gages are designed so that strain in the width or transverse direction separates the legs of the grid without straining them. This makes the gage sensitive to strain only along its length. There is always some sensitivity to transverse strain and almost no sensitivity to shear strain. In most cases, the transverse sensitivity can be neglected.

When piezoresistive foil or wire strain gages are bonded to a specimen and the specimen is strained, the gage strains as well. The resistance change is related to the strain by a gage factor, Equation 37.7:

$$\frac{\Delta R}{R} = G_L \varepsilon_L \tag{37.7}$$

where

$\Delta R/R$ is the relative resistance change
G is the gage factor
ε is the strain

These gages respond to the average strain over the area covered by the grid [10]. The resistance change is also sensitive to temperature. If the temperature changes during the measurement period, a correction must be made to distinguish the strain response from the thermal response. The gage response to longitudinal strain, transverse strain, and temperature change is given by Equation 37.8:

$$\frac{\Delta R}{R} = G_L \varepsilon_L + G_t \varepsilon_t + G_T \Delta T \tag{37.8}$$

where G_L, G_t, and G_T are the longitudinal, transverse, and temperature sensitivity, respectively. Micro-Measurements, Inc., uses a different notation. Their gage data are provided as $G_L = F_G$, $G_t = K_t F_G$, $G_T = \beta_g$.

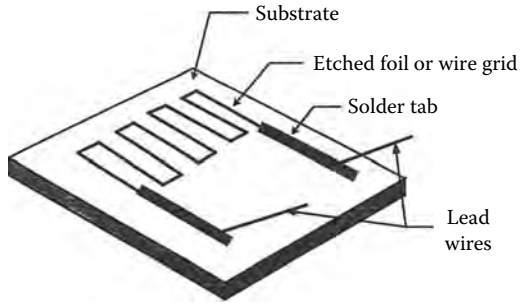


FIGURE 37.5 Gage construction of a foil or wire piezoresistive gage.

When a strain gage is bonded to a specimen and the temperature changes, the strain used in Equation 37.8 is the total strain, thermal plus stress induced, as given by Equation 37.7.

The temperature contribution to gage output must be removed if the gages are used in tests where the temperature changes. A scheme referred to as self-temperature compensation (STC) can be used. This is accomplished by selecting a piezoresistive material whose thermal output can be canceled by the strain induced by thermal expansion of the test specimen. Gage manufacturers specify STC numbers that match the thermal expansion coefficients of common specimen materials.

Strain of piezoresistive materials produces a relative resistance change. The resistance change is the result of changes in resistivity and dimensional changes. Consider a single leg of the grid of a strain gage with a rectangular cross section (Figure 37.6). The resistance is given by Equation 37.9:

$$R = \rho \frac{L}{A} \tag{37.9}$$

where

R is the resistance
ρ is the resistivity
L is the length
A is the area of the cross section

A small change in resistance is given by the first-order terms of a Taylor's series expansion, Equation 37.10:

$$\Delta R = \frac{\partial R}{\partial \rho}\Delta\rho + \frac{\partial R}{\partial L}\Delta L + \frac{\partial R}{\partial A}\Delta A \tag{37.10}$$

Differentiating Equation 37.9 to obtain each term of Equation 37.10 and then dividing by the initial resistance lead to Equation 37.11:

$$\frac{\Delta R}{R_0} = \frac{\Delta\rho}{\rho_0} + \frac{\Delta L}{L_0} - \frac{\Delta A}{A_0} \tag{37.11}$$

The relative resistance change is due to a change in resistivity, a change in length strain, and a change in area strain.

The strain gage is a composite material. The metal in a strain gage is like a metal fiber in a polymer matrix. When the polymer matrix is deformed, the metal is dragged along in the length direction, but in the width and thickness directions, the strain is not passed to the metal. This results in a stress state called uniaxial stress. This state was discussed in the examples earlier. The mathematical details involve an inclusion problem [11,12]. Accepting that the stress state is uniaxial, the relationship between the area change and the length change in Equation 37.11 is found from the Poisson's ratio. The area strain is the sum of the width and thickness strain, Equation 37.12:

$$\frac{\Delta A}{A_0} = \frac{\Delta w}{w_0} + \frac{\Delta t}{t_0} \tag{37.12}$$

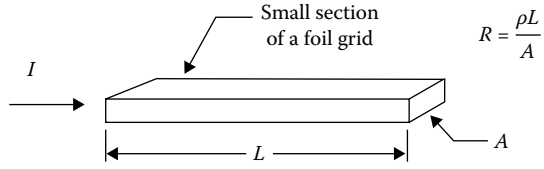


FIGURE 37.6 A single leg of a piezoresistive gage is used to explain the source of the relative resistance change that occurs in response to strain.

The definition of the Poisson's ratio gives Equation 37.13:

$$\frac{\Delta A}{A_0} = -2\nu\frac{\Delta L}{L_0} \tag{37.13}$$

where

$\Delta w/w$ is the width strain

$\Delta t/t$ is the thickness strain

Substitution of Equation 37.13 into 37.11 gives Equation 37.14 for the relative resistance change:

$$\frac{\Delta R}{R_0} = \frac{\Delta\rho}{\rho_0} + \frac{\Delta L}{L_0}(1+2\nu) \tag{37.14}$$

The relative resistivity changes in response to stress. The resistivity is a second-order tensor [13], and the contribution to the overall resistance change can be found in terms of strain using the elastic constitutive law [14]. The results lead to an elastic gage factor just over 2 for constantan gages. If the strain is large, the foil or wire in the gage will experience plastic deformation. When the deformation is plastic, the resistivity change is negligible and the dimensional change dominates. In this case, Poisson's ratio is 0.5 and the gage factor is 2. This effect is utilized in manufacturing gages for measuring strains larger than 1.5%. In this case, annealed foil is used. The annealed foil undergoes plastic deformation without failure. These gages are capable of measuring strain in excess of 10%. When metals undergo plastic deformation, they do not unload to the initial strain. This shows up as hysteresis in the gage response, that is, on unload, the resistance does not return to its initial value.

Foil and wire strain gages can be obtained in several configurations. They can be constructed with different backing materials and left open faced or fully encapsulated. Backing materials include polyimide and glass fiber-reinforced phenolic resin. Gages can be obtained with solder tabs for attaching lead wires or with lead wires attached. They come in many sizes and in multiple gage configurations called rosettes.

Strain gages are mounted to test specimens with adhesives using a procedure that is suitable for bonding most types of strain sensors. This is accomplished in a step-by-step procedure [15] that starts with surface preparation. An overview of the procedure is briefly described. To successfully mount strain gages, the surface is first degreased. The surface is abraded with a fine emery cloth or 400 grit paper to remove any loose paint, rust, or deposits. Gage layout lines are drawn (not scribed) on the surface in a cross pattern with pen or pencil, one line in the grid direction and one in the transverse direction. The surface is then cleaned with isopropyl alcohol. This can be done with an ultrasonic cleaner or with wipes. If wiped, the paper or gauze wipe should be folded and a small amount of alcohol applied. The surface should be wiped with one pass and the wipe discarded. This should be repeated, wiping in the other direction. The final step is to neutralize the surface, bringing the alkalinity to a pH of 7–7.5. A surface neutralizer is available from most adhesive suppliers. The final step is to apply the gage.

Gage application is accomplished with cellophane tape, quick-set glue, and a catalyst. The gage is placed on a clean glass or plastic surface with bonding side down, using tweezers. (Never touch the gage. Oils from skin prevent proper adhesion.) The gage is then taped down with a 100 mm piece of cellophane tape. The tape is then peeled up with the gage attached. The gage can now be taped onto its desired location on the test specimen. Once the gage has been properly aligned, the tape is peeled back from one side, lifting the gage from the surface. The tape should remain adhered to the surface about 1 cm from the gage. Note that one side of the tape is still attached to the specimen so that the gage can be easily returned to its desired position. A thin coating of catalyst is applied to the exposed gage surface. A drop of glue is placed at the joint of the tape and the specimen. Holding the tape at about a 30° angle from the surface, the tape can be slowly wiped down onto the surface. This moves the glue line forward.

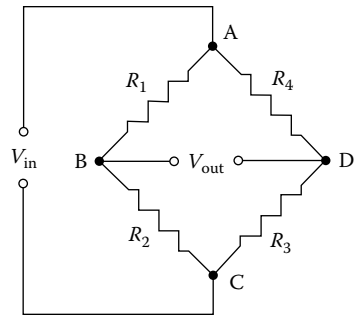


FIGURE 37.7 The Wheatstone bridge is used to measure the relative resistance change of piezoresistive strain gages.

After the glue line has passed the gage, the gage should be pressed in place and held for approximately 1 min. The tape can now be peeled back to expose the gage, and lead wires can be attached.

The relative resistance change of piezoresistive gages is usually measured using a Wheatstone bridge [16]. This allows a small change of resistance to be measured relative to an initial zero value, rather than relative to a large resistance value, with a corresponding increase in sensitivity and resolution. The Wheatstone bridge is a combination of four resistors and a voltage source (Figure 37.7). One to four of the resistors in the bridge can be strain gages. The output of the bridge is the difference between the voltage at points B and D. Paths ABC and ADC are voltage dividers so that V_B and V_D are given by Equations 37.15a and 37.15b:

$$V_B = V_{in}\frac{R_2}{R_1 + R_2} \tag{37.15a}$$

$$V_D = V_{in}\frac{R_3}{R_3 + R_4} \tag{37.15b}$$

The bridge output, Equation 37.16, is zero when the balance condition, Equation 37.17, is met:

$$V_{out} = V_B - V_D \tag{37.16}$$

$$R_1R_3 = R_2R_4 \tag{37.17}$$

Wheatstone bridge signal conditioners are constructed with a way to "balance" the bridge by adjusting the ratio of the resistances so that the bridge output is initially zero.

The balance condition is no longer met if the resistance values undergo small changes ΔR_1, ΔR_2, ΔR_3, ΔR_4. If the values $R_1 + \Delta R_1$ are substituted into Equation 37.15, the results substituted into Equation 37.16, condition (37.17) used, and the higher-order terms neglected, the result is Equation 37.18 for the bridge output:

$$V_{out} = V_{in}\frac{R_1R_3}{(R_1 + R_2)(R_3 + R_4)}\left(-\frac{\Delta R_1}{R_1} + \frac{\Delta R_2}{R_2} - \frac{\Delta R_3}{R_3} + \frac{\Delta R_4}{R_4}\right) \tag{37.18}$$

The Wheatstone bridge can be used to directly cancel the effect of thermal drift. If R_1 is a strain gage bonded to a specimen and R_2 is a strain gage held onto a specimen with heat sink compound (a thermally conductive grease available at any electronics store), then R_1 will respond to strain plus temperature, and R_2 will only respond to temperature. Since the bridge subtracts the output of R_1 from that of R_2, the temperature effect cancels.

The sensitivity of a measuring system is the output per unit change in the quantity to be measured. If the resistance change is from a strain gage, the sensitivity of the Wheatstone bridge system is proportional to the input voltage. Increasing the voltage increases the sensitivity. There is a practical limitation to increasing the voltage to large values. The power dissipated (heat) in the gage is $P = I^2R$, where I, the current through the gage, can be found from the input voltage and the bridge resistances. This heat must go somewhere or the temperature of the gage will continuously rise and the resistance will change due to heating. If the gage is mounted on a good thermal conductor, more power can be conducted away than if the gage is mounted on an insulator. The specimen must act as a heat sink.

Heat sinking ability is proportional to the thermal conductivity of the specimen material. A reasonable temperature gradient to allow the gage to induce in a material is 40 °C m^{-1} (about 1 °C per 25 mm). For thick specimens (thickness several times the largest gage dimension), this can be conducted away to the grips or convected to the surrounding atmosphere. If the four bridge resistances are approximately equal, the power to the gage in terms of the bridge voltage is given by Equation 37.19:

$$P_g = \frac{V_{in}^2}{4R} \tag{37.19}$$

The power per unit grid area, A_g, or power density to the gage can be equated to the thermal conductivity of the specimen and the allowable temperature gradient in the specimen by Equation 37.20:

$$\frac{P_g}{A_g} = \frac{V_{in}^2}{4\ RA_g} = K\nabla T \tag{37.20}$$

Thermal conductivities of most materials can be found in tables or on the web. Some typical values are Al, K = 204 W m^{-1} °C^{-1}; steel, K = 20–70 W m^{-1} °C^{-1}; and glass, K = 0.78 W m^{-1} °C^{-1} [17].

The acceptable bridge voltage can be calculated from Equation 37.21:

$$V_{in} = \sqrt{K\nabla T 4RA_g} \tag{37.21}$$

A sample calculation shows that for a 0.010 m × 0.010 m 120 Ω grid bonded to a thick piece of aluminum with a thermal conductivity of 204 W m^{-1} °C^{-1} and an acceptable temperature gradient of 40 °C m^{-1}, the maximum bridge voltage is 19 V. If thin specimens are used, the allowable temperature gradient will be smaller. If smaller gages are used for better spatial resolution, the bridge excitation voltage must be reduced with a corresponding reduction in sensitivity.

A considerably higher bridge voltage can be used if the bridge voltage is pulsed for a short duration. This dissipates substantially less energy in the gage and thus increases the sensitivity by a factor of 10–100. Wheatstone bridge pulse power supplies with variable pulse width from 10 μs and excitation of 350 V are commercially available [18].

The strain measurement required is often in a complex loading situation where the directions of principal strain are not known. In this case, three strain gages must be bonded to the test specimen at three angles. This is called a strain rosette. The angle between the rosette and the principal directions, as well as the magnitude of the principal strains, can be determined from the output of the rosette gages. This is most easily accomplished using the construct of a Mohr's circle (Figure 37.8).

A common rosette is the 0°–45°–90° pattern. The rosette is bonded to the specimen with relative rotations of 0°, 45°, and 90°. These will be referred to as the x, x', and y directions. The principal directions are labeled the 1 and 2 directions. The unknown angle between the x direction and the 1 direction is labeled θ. The Mohr's circle is drawn with the elongational strain on the horizontal axis and the shear strain on the vertical axis. The center of the circle is labeled C and the radius R. The principal directions correspond to zero shear strain. The principal values are given by Equations 37.22 and 37.23:

$$\varepsilon_1 = C + R \tag{37.22}$$

$$\varepsilon_2 = C - R \tag{37.23}$$

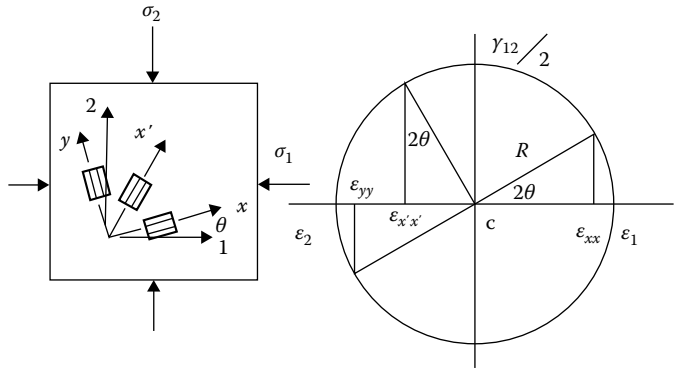


FIGURE 37.8 A three-element rosette is used to measure strain when the principal directions are not known. The Mohr's circle is used to find the principal directions and the principal strain values.

A rotation through an angle 2θ on the Mohr's circle corresponds to a rotation of the rosette of θ relative to the principal directions. The center of the circle is given by Equation 37.24 and the output of the strain gages is given by Equations 37.25 through 37.27:

$$C = \frac{\varepsilon_{xx} + \varepsilon_{yy}}{2} \tag{37.24}$$

$$\varepsilon_{xx} - C = R\cos 2\theta \tag{37.25}$$

$$\varepsilon_{x'x'} - C = -R\sin 2\theta \tag{37.26}$$

$$\varepsilon_{yy} - C = -R\cos 2\theta \tag{37.27}$$

Dividing Equation 37.25 by Equation 37.26 leads to θ and then to R, Equations 37.28 and 37.29:

$$R^2 = (\varepsilon_{xx} - C)^2 + (\varepsilon_{x'x'} - C)^2 \tag{37.28}$$

$$\tan 2\theta = \frac{C - \varepsilon_{x'x'}}{\varepsilon_{xx} - C} \tag{37.29}$$

The principal directions and principal strain values have been found from the output of the three rosette gages.

37.2.2 Piezoresistive Semiconducting Gages

Piezoresistive semiconductor strain gages, like piezoresistive foil and wire gages, undergo a resistance change in response to strain but with nearly an order of magnitude larger gage factor [19]. The coupling between resistance change and temperature is very large, so these gages have to be temperature compensated. The change of resistance with a small temperature change can be an order of magnitude larger than that induced by strain. Semiconductor strain gages are typically used to manufacture transducers such as load cells. They are fragile and require great care in their application.

37.2.3 Piezoelectric Gages

Piezoelectric strain gages are, effectively, parallel plate capacitors whose dielectric changes polarization in response to strain [14]. When the polarization changes, a charge proportional to the strain is produced on the electrodes. PVDF film strain gages are inexpensive, but not very accurate and subject

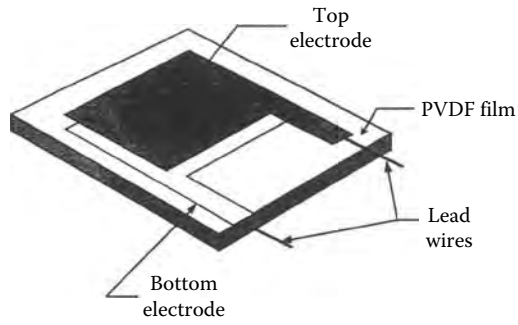


FIGURE 37.9 Typical gage construction of a piezoelectric gage.

to depoling by moderate temperature. They make good sensors for dynamic measurements such as frequency and logarithm decrement, but not for quantitative measurements of strain. When used for quasistatic measurements, the charge tends to drain through the measuring instrument. This causes the signal to decay with a time constant dependent on the input impedance of the measuring instrument. Quartz gages are very accurate but also lose charge through the measuring instrument. Time constants can be relatively long (seconds to hours) with electrometers or charge amplifiers.

The PVDF gage consists of a thin piezoelectric film with metal electrodes (Figure 37.9). Lead wires connect the electrodes to a charge measuring circuit. Gages can be obtained with the electrodes encapsulated between insulating layers of polyimide.

The gage output can be described in terms of a net dipole moment per unit volume. If the net dipole moment is the total charge, Q, on the electrodes multiplied by spacing, d, between the electrodes, then the polarization is given by Equation 37.30:

$$P = \frac{Qd}{V} \tag{37.30}$$

From Equation 37.30, it is seen that the polarization P (approximately equal to the electric displacement D) is the charge per unit electrode area (Figure 37.10).

A Taylor's series expansion of Equation 37.30 gives Equation 37.31:

$$\Delta P = \frac{\partial P}{\partial V}\Delta V + \frac{\partial P}{\partial (Qd)}\Delta Qd, \tag{37.31}$$

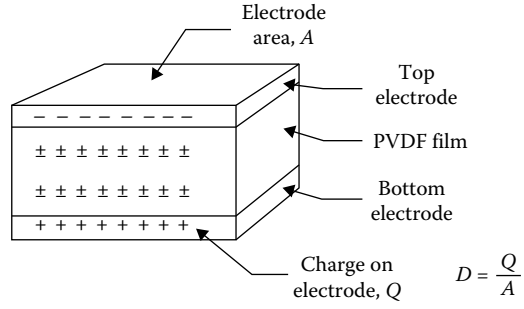


FIGURE 37.10 A representative cross section of a piezoelectric material formed into a parallel plate capacitor. The piezoelectric material is polarized. This results in charge on the electrodes. When the material is strained, the polarization changes and charge flows.

which, after differentiating Equation 37.30 and substituting in Equation 37.31, gives

$$\Delta P = \frac{\Delta Q d}{V_0} - \frac{\Delta V}{V_0} P_0 \tag{37.32}$$

For PVDF film, the second term in Equation 37.32 dominates. The output is proportional to the remanent polarization P_0. The remanent polarization slowly decays with time, has a strong dependence on temperature, and decays rapidly at temperatures around 50 °C. This makes accuracy a problem. If the sensors are kept at low temperature, accuracy can be maintained within ±3%.

Strain sensors can also be constructed from piezoelectric ceramics like lead zirconate titanate (PZT) or barium titanate. Ceramics are brittle and can be depoled by strain so should only be used at strains less than 200 μstrain. PZT loses some of its polarization with time and thus has accuracy problems but remains polar to temperatures of 150 °C or higher. "Hard" PZT (usually iron doped) is the best composition for polarization stability and low hysteresis. Quartz has the best accuracy. It is not polar but polarization is induced by strain. Quartz has excellent resolution and accuracy over a broad temperature range but is limited to low strain levels. It is also brittle and so is limited to small strain.

Two circuits are commonly used for piezoelectric signal conditioning: the electrometer and the charge amplifier (Figure 37.11). In the electrometer circuit, the piezoelectric sensor is connected to a capacitor with a capacitance value C, at least 1000 times that of the sensor C_g. There is always some resistance in the cable that connects the sensor to the capacitor. The circuit is simply two capacitors in parallel connected by a resistance. The charge equilibrates with a time constant given by R_gC_g. This time constant limits the fastest risetime that can be resolved to about 50 ns, effectively instantaneous for most applications. The charge is measured by measuring the voltage on the capacitor, then using Equation 37.33:

$$Q = CV \tag{37.33}$$

The difficulty is that measuring devices drain the charge, causing a time decay with a time constant RC. This causes the signal to be lost rapidly if conventional op amps are used. FET input op amps have very high input impedance and can extend this time constant to many hours. The charge amplifier is another circuit used to measure charge. This is usually an FET input op amp with a capacitor feedback. This does not really amplify charge but produces a voltage proportional to the input charge. Again, the time constant can be many hours, allowing use of piezoelectric sensors for near static measurements.

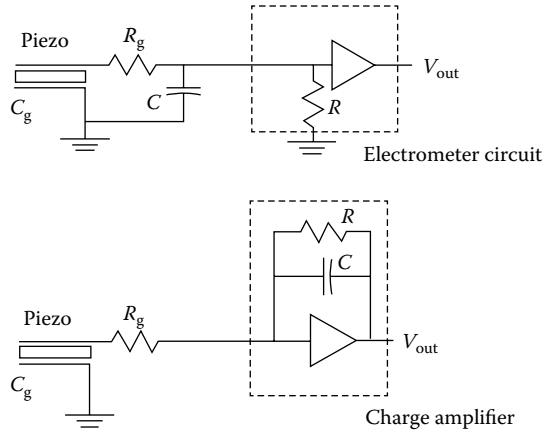


FIGURE 37.11 The electrometer and charge amplifier are the most common circuits used to measure charge from piezoelectric transducers.

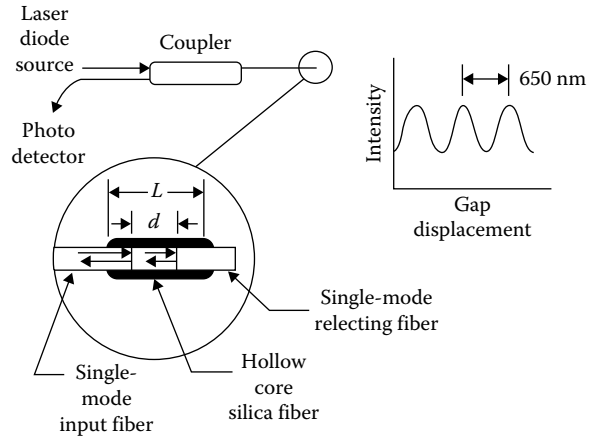


FIGURE 37.12 A schematic of the Fabry–Perot fiber-optic strain gage. When the cavity elongates, alternating constructive and destructive interferences occur.

High input impedance electrometer and charge amplifier signal conditioners for near static measurements are commercially available [20] as well as low-cost capacitive terminators for high-frequency (high kilohertz to megahertz) measurements [18]. An advantage of piezoelectric sensors is that they are active sensors that do not require any external energy source.

37.2.4 Fiber-Optic Strain Gages

Fiber-optic strain gages are miniature interferometers [21,22]. Many commercially available sensors are based on the Fabry–Perot interferometer. The Fabry–Perot interferometer measures the change in the size of a very small cavity.

Fabry–Perot strain sensors (Figure 37.12) comprise a laser light source, single-mode optical fibers, a coupler (the fiber-optic equivalent of a beam splitter), a cavity that senses strain, and a photodetector. Light leaves the laser diode. It passes down the fiber, through the coupler, and to the cavity. The end of the fiber is the equivalent of a partially silvered mirror. Some of the light is reflected back up the fiber and some is transmitted. The transmitted light crosses the cavity and then is reflected from the opposite end back into the fiber where it recombines with the first reflected beam. The two beams have a phase difference related to twice the cavity length. The recombined beam passes through the coupler to the photodetector. If the two reflected beams are in phase, there will be constructive interference. If the two beams are out of phase, there will be destructive interference. The cavity is bonded to a specimen. When the specimen is strained, the cavity stretches. This results in a phase change of the cavity beam, causing a cycling between constructive and destructive interference. For a 1.3 μm light source, each peak in output corresponds to a 650 nm gap displacement. The gap displacement divided by the gap length gives the strain. The output is continuous between peaks so that a 3 mm gage can resolve 1 μstrain.

37.2.5 Birefringent Film Strain Sensing

Birefringent film strain sensors give a full field measurement of strain. A nice demonstration of this effect can be achieved with two sheets of inexpensive Polaroid film, a 6 mm thick, 25 mm × 200 mm bar of Plexiglas (polymethylmethacrylate or PMMA), and an overhead projector. Place the two Polaroid sheets at 90° to one another so that the light is blocked. Place the PMMA between the Polaroid sheets. Apply a bending moment to the bar and color fringes will appear. Birefringent materials have a different

speed of light in different directions. This means that if light is polarized in a particular direction and passed through a birefringent specimen, if the fast direction is aligned with the electric field vector, the light passes through faster than if the slow direction is aligned with the electric field vector. This effect can be used to produce optical interference. In some materials, birefringence is induced by strain. The fast and slow directions correspond to the directions of principal strain, and the amount of birefringence corresponds to the magnitude of the strain. One component of the electric field vector travels through the specimen faster than the other. They emerge with a phase difference. This changes the relative amplitude and thus rotates the polarization of the light. If there is no birefringence, no light passes through the second polarizer. As the birefringence increases with strain, light passes through. As it further increases, the polarization rotation will be a full 180°, and again no light will pass through. This produces a fringe that corresponds to a constant difference in principal strains. The difference in principal strains is given by Equation 37.34:

$$\varepsilon_2 - \varepsilon_1 = \frac{N\lambda}{tK} \tag{37.34}$$

where

ε_1, ε_2 are the principal strains
N is the fringe order
λ is the wavelength
t is the specimen thickness
K is the strain-optical coefficient of the photoelastic material

A similar technique can be used with a birefringent plastic film with a silvered backing laminated to the surface of a specimen. Polarized light is passed through the film; it reflects from the backing, passes back through the film, and through the second polarizer. In this case, because light passes twice through the film, the equation governing the difference in principal strains is Equation 37.35:

$$\varepsilon_2 - \varepsilon_1 = \frac{N\lambda}{2tK} \tag{37.35}$$

If the polarizers align with principal strain directions, no birefringence is observed. Rotation of both polarizers allows the principal directions to be found at various locations on the test specimen. If a full view of the fringes is desired, quarter-wave plates are used (Figure 37.13). In this arrangement, light is passed through the first polarizer, resulting in plane polarization; through the quarter-wave plate, resulting in circular polarization; through the test specimen, resulting in phase changes; through the second quarter-wave plate to return to plane polarization; and then through the final polarizer.

The optical systems for viewing birefringence are commercially available as "polariscopes" [23]. Optical components to construct custom systems are available from many optical component suppliers.

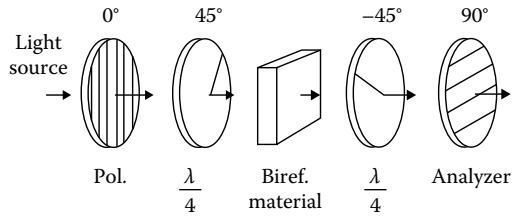


FIGURE 37.13 A schematic of the polariscope, a system for measuring birefringence. This technique gives a full field measure of the difference in principal strains.

37.2.6 Moiré Strain Sensing

Moiré interference is another technique that gives a full field measurement, but it measures displacement rather than strain. The strain field must be computed from the displacement field. This technique is based on the interference obtained when two transparent plates are covered with equally spaced stripes. If the plates are held over one another, they can be aligned so that no light will pass through or so that all light will pass through. If one of the plates is stretched, the spacing of the lines is wider on the stretched plate. Now, if one plate is placed over the other, in some regions light will pass through and in some regions it will not (Figure 37.14). The dark and light bands produced give information about the displacement field.

Moiré is defined as a series of broad dark and light patterns formed by the superposition of two regular gratings [24]. The dark or light regions are called fringes. Examples of pure extension and pure rotation are shown. In both cases, some of the light that would emerge from the first grating is obstructed by the superimposed grating. At the centers of the dark fringes, the bar of one grating covers the space of the other and no light comes through. The emergent intensity, I, is zero. Proceeding from there toward the next dark fringe, the amount of obstruction diminishes linearly and the amount of light increases linearly until the bar of one grating falls above the bar of the other. There, the maximum amount of light passes through the gratings.

Both geometric interference and optical interference are used. This discussion is restricted to geometric interference. Geometric moiré takes advantage of the interference of two gratings to determine displacements and rotations in the plane of view. In-plane moiré is typically conducted with two gratings, one applied to the specimen (specimen grating) and the other put in contact with the specimen grating (reference grating). When the specimen is strained, interference patterns or fringes occur. N is the moiré fringe order. Each fringe corresponds to an increase or decrease of specimen displacement by one grating pitch. The relationship between displacement and fringes is $\delta = gN$,

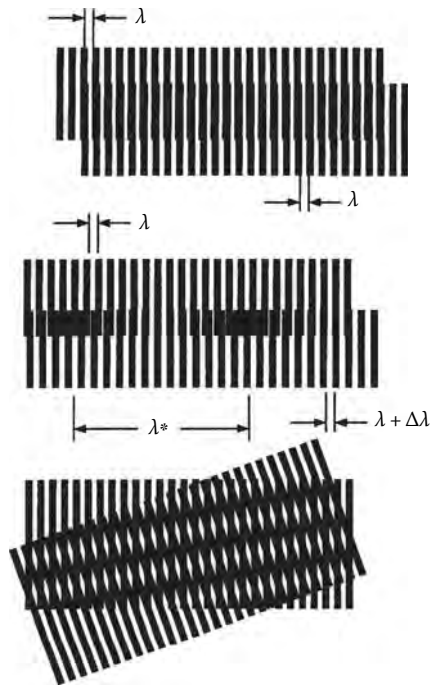


FIGURE 37.14 A demonstration of moiré fringes formed by overlapping gratings. The fringes are the result of stretching and relative rotation of the gratings. The fringe patterns are used to determine displacement fields.

where δ is the component of the displacement perpendicular to the reference grating lines, g is the reference grating pitch, and N is the fringe order.

For convenience, a zero-order fringe is designated assuming the displacement there is zero. With the reference grating at 0° and 90°, the fringe orders N_x and N_y are obtained. The displacements in x, y directions are then obtained from Equations 37.36 and 37.37:

$$u_x(x,y) = gN_x(x,y) \tag{37.36}$$

$$u_y(x,y) = gN_y(x,y) \tag{37.37}$$

Differentiation of Equations 37.36 and 37.37 gives the strains, Equations 37.38 through 37.40:

$$\varepsilon_x = \frac{\partial u_x}{\partial x} = g\frac{\partial N_x}{\partial x} \tag{37.38}$$

$$\varepsilon_{xy} = \frac{1}{2}\left(\frac{\partial u_y}{\partial x} + \frac{\partial u_x}{\partial y}\right) = \frac{1}{2}\left(g\frac{\partial N_y}{\partial x} + g\frac{\partial N_x}{\partial y}\right) \tag{37.39}$$

$$\varepsilon_y = \frac{\partial u_y}{\partial_y} = g\frac{\partial N_y}{\partial y} \tag{37.40}$$

In most cases, the sensitivity of geometric moiré is not adequate for determination of strain distributions. Strain analysis should be conducted with high-sensitivity measurement of displacement using moiré interferometry [24,25]. Moiré interferometers are commercially available [26]. Out-of-plane measurement can be conducted with one grating (the reference grating). The reference grating is made to interfere with either its reflection or its shadow [27,28].

References

1. N. E. Dowling, *Mechanical Behavior of Materials*, Englewood Cliffs, NJ: Prentice-Hall, 1993, pp. 99–108.
2. R. C. Craig, *Mechanics of Materials*, New York: John Wiley & Sons, 1996.
3. A. Vengsarkar, Fiber optic sensors: A comparative evaluation, In *The Photonics Design and Applications Handbook*, Pittsfield, MA: Laurin Publ., 1991, pp. 114–116.
4. H. U. Eisenhut, Force measurement on presses with piezoelectric strain transducers and their static calibration up to 5 MN, *New Industrial Applications of the Piezoelectric Measurement Principle*, July 1992, pp. 1–16.
5. TN-501-4, Strain gauge temperature effects, Measurements Group, Inc., Raleigh, NC.
6. TN-509, Transverse sensitivity errors, Measurements Group, Inc., Raleigh, NC.
7. TT-609, Soldering techniques, Measurements Group, Inc., Raleigh, NC.
8. TN-515, Strain gage rosettes, Measurements Group, Inc., Raleigh, NC.
9. TN-08-1, Fatigue of strain gages, Measurements Group, Inc., Raleigh, NC.
10. C. C. Perry and H. R. Lissner, *The Strain Gage Primer*, New York: McGraw-Hill, 1962.
11. J. B. Aidun and Y. M. Gupta, Analysis of Lagrangian gauge measurements of simple and nonsimple plain waves, *J. Appl. Phys.*, 69, 6998–7014, 1991.
12. Y. M. Gupta, Stress measurement using piezoresistance gauges: Modeling the gauge as an elastic-plastic inclusion, *J. Appl. Phys.*, 54, 6256–6266, 1983.
13. D. Y. Chen, Y. M. Gupta, and M. H. Miles, Quasistatic experiments to determine material constants for the piezoresistance foils used in shock wave experiments, *J. Appl. Phys.*, 55, 3984, 1984.

14. C. S. Lynch, Strain compensated thin film stress gauges for stress wave measurements in the presence of lateral strain, *Rev. Set. Instrum*., 66(11), 1–8, 1995.
15. B-129-7 M-Line Accessories Instruction Bulletin, Measurements Group, Inc., Raleigh, NC.
16. J. W. Dally, W. F. Riley, and K. G. McConnell, *Instrumentation for Engineering Measurements*, 2nd edn., New York: John Wiley & Sons, 1993.
17. J. P. Holman, *Heat Transfer*, 7th edn., New York: McGraw Hill, 1990.
18. Dynasen, Inc., 20 Arnold Pl., Goleta, CA 93117.
19. M. Dean (Ed.) and R. D. Douglas (assoc. Ed.), *Semiconductor and Conventional Strain Gages*, New York: Academic Press, 1962.
20. Kistler, Instruments Corp., Amhurst, NY, 14228–2171.
21. J. S. Sirkis, Unified approach to phase strain temperature models for smart structure interferometric optical fiber sensors. 1. Development, *Opt. Eng*., 32(4), 752–761, 1993.
22. J. S. Sirkis, Unified approach to phase strain temperature models for smart structure interferometric optical fiber sensors. 2. Applications, *Opt. Eng*., 32(4), 762–773, 1993.
23. Photoelastic Division, Measurements Group, Inc., P.O. Box 27777, Raleigh, NC 27611.
24. T. Valis, D. Hogg, and R. M. Measures, Composite material embedded fiber-optic Fabry-Perot strain rosette, *SPIE*, 1370, 154–161, 1990.
25. D. Post, B. Han, and P. Lfju, *High Sensitivity Moiré*, New York: Springer-Verlag, 1994.
26. V. J. Parks, Geometric Moiré, In *Handbook on Experimental Mechanics*, A. S. Kobayashi (ed.), New York: VCH Publisher, Inc., 1993.
27. Photomechanics, Inc., 512 Princeton Dr. Vestal, NY, 13850–2912.
28. T. Y. Kao and F. P. Chiang, Family of grating techniques of slope and curvature measurements for static and dynamic flexure of plates, *Opt. Eng*., 21, 721–742, 1982.
29. D. R. Andrews, Shadow moiré contouring of impact craters, *Opt. Eng*., 21, 650–654, 1982.

38

Tactile Sensing

Ricardo E. Saad
Harmonics Lightwaves

A. Bonen
University of Toronto

K.C. Smith
University of Toronto

Ben Benhabib
University of Toronto

38.1 Sensing Classification **38**-1
38.2 Mechanical Effects of Contact **38**-2
38.3 Theory for Tactile Sensing **38**-2
38.4 Requirements for Tactile Sensors **38**-4
38.5 Technologies for Tactile Sensing **38**-5
Resistive • Capacitive • Piezoelectric • Optical • Photoelastic
References **38**-15
Further Readings **38**-16
Partial List of Vendors and Suppliers **38**-17

Robots in industrial settings perform repetitive tasks, such as machine loading, parts assembly, painting, and welding. Only in rare instances can these autonomous manipulators modify their actions based on sensory information. Although, thus far, a vast majority of research work in the area of robot sensing has concentrated on computer vision, contact sensing is an equally important feature for robots and has received some attention as well. Without tactile-perception capability, a robot cannot be expected to effectively grasp objects. In this context, robotic-tactile sensing is the focus of this chapter.

38.1 Sensing Classification

Robotic sensing can be classified as either of the noncontact or of the contact type [1]. *Noncontact sensing* involves interaction between the robot and its environment by some physical phenomenon, such as acoustic or electromagnetic waves that interact without contact. The most important types of robotic sensors of the noncontact type are vision and proximity sensors. *Contact sensing*, on the other hand, implies the measurement of the general interaction that takes place when the robot's end effector is brought into contact with an object. Contact sensing is further classified into force and tactile sensing.

Force sensing is defined as the measurement of the global mechanical effects of contact, while *tactile sensing* implies the detection of a wide range of local parameters affected by contact. Most significant among those contact-based effects are contact stresses, slippage, heat transfer, and hardness.

The properties of a grasped object that can be derived from tactile sensing can be classified into geometric and dynamometric types [2]. Among the geometric properties are presence, location in relation to the end effector, shape and dimensions, and surface condition [3–7]. Among the dynamometric parameters associated with grasping are force distribution, slippage, elasticity and hardness, and friction [8–12].

Tactile sensing requires sophisticated transducers; yet the availability of these transducers alone is not a sufficient condition for successful tactile sensing. It is also necessary to accurately control the modalities through which the tactile sensor interacts with the explored objects (including contact forces, as well as end-effector position and orientation) [13–15]. This leads to active tactile sensing, which requires a high degree of complexity in the acquisition and processing of the tactile data [16].

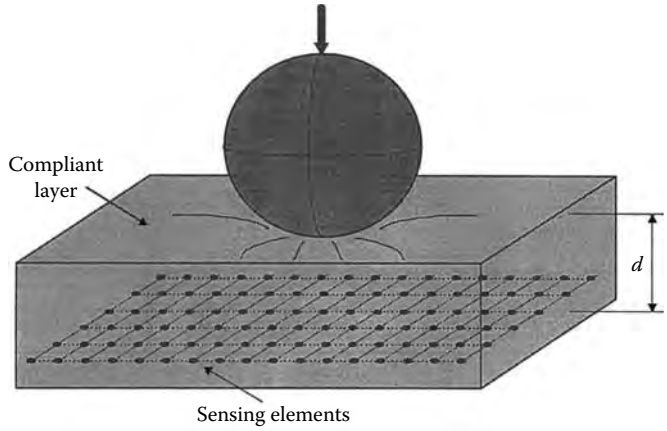


FIGURE 38.1 Object indenting a compliant layer, where an array of force-sensing elements is placed at a distance d from the surface.

38.2 Mechanical Effects of Contact

Tactile sensing normally involves a rigid object indenting the compliant cover layer of a tactile sensor array [17] (Figure 38.1). The indentation of a compliant layer due to contact can be analyzed from two conceptually different points of view [1]. The first one is the measurement of the actual contact stresses (force distribution) in the layer, which is usually relevant to controlling manipulation tasks. The second one is the deflection profile of the layer, which is usually important for recognizing geometric object features. Depending on the approach adopted, different processing and control algorithms must be utilized.

There exists a definite relationship between the local shape of a contacting body and a set of subsurface strains (or displacements); however, this relationship is quite complex. Thus, it requires the use of the theory of elasticity and contact mechanics to model sensor–object interaction [18] and the use of finite element analysis (FEA) as a practical tool for obtaining a more representative model of the sensor [19].

In general, the study of tactile sensors comprises two steps: (1) the *forward analysis*, related to the acquisition of data from the sensor (changes on the stress or strains, induced by the indentation of an object on the compliant surface of the transducer), and (2) the *inverse problem*, normally related to the recovery of force distribution or, in some cases, the recovery of the indentor's shape.

38.3 Theory for Tactile Sensing

For simplicity, the general 2D tactile problem is reduced herein to a 1D one. Figure 38.2 shows a 1D transducer that consists of a compliant, homogeneous, isotropic, and linear layer subjected to a normal stress $q_v(x)$ created by the indentation of an object.

For modeling purposes, it is assumed that the compliant layer is an elastic half-space. This simplification yields closed-form equations for the analysis and avoids the formation of a more complex problem, in which the effect of the boundary conditions at x_{min} and x_{max} must be taken into account. It has been proven that the modeling of the sensor by an elastic half-space represents a reasonable approximation to the real case [18]. Under these conditions, it can be shown that the normal strain, at a depth $y = d$, due to the normal stress $q_v(y)$ is given by [20]

$$\varepsilon_z(x) = \int_{-\infty}^{\infty} q_v(x - x_0) h_z(x_0, d) dx_0 \tag{38.1}$$

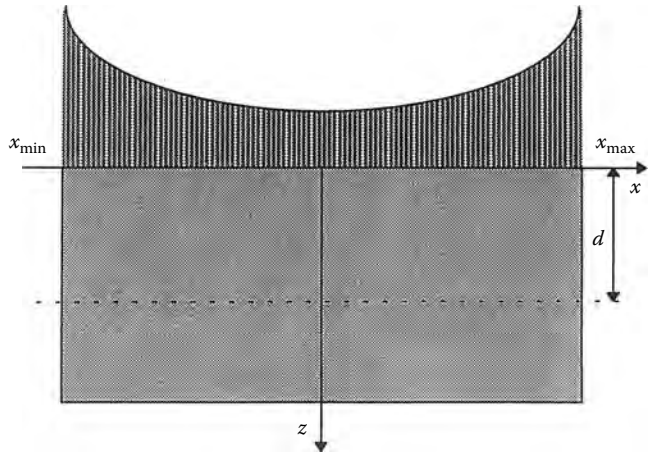


FIGURE 38.2 Ideal 1D transducer subjected to a normal stress.

where

ε_z is the strain at x

$z = d$ due to the normal stress on the surface

$$h_z(x) = -\frac{2d(1+\nu)[d^2(1-\nu) - \nu x^2]}{\pi r E(x^2 + d^2)^2} \tag{38.2}$$

E and ν are, respectively, the modulus of elasticity and the Poisson's coefficient of the compliant layer. In obtaining Equation 38.2, it is assumed that the analysis is performed under *planar strain* conditions. It should be noted that a similar analysis can be performed for tangential contact stresses or strains.

The normal displacement at the surface, w, is given by

$$w(x) = \int_{-\infty}^{\infty} q_v(x - x_0)k(x_0)dx_0 \tag{38.3}$$

where

$$k(x) = \frac{-2(1-\nu^2)}{\pi E}\log\left|\frac{x}{x_a}\right| \tag{38.4}$$

The singularity at $x = 0$ is expected due to the singularity of stress at that point. Note that $k(x)$ is the deformation of the surface when a singular load of 1 N is applied at $x = 0$. The constant x_a should be chosen such that at $x = x_a$, the deformation is zero. In this case, zero deformation should occur at $x \to \infty$ (note that it has been assumed that the sensor is modeled by an elastic half-space), namely, $x_a \to \infty$. This problem is associated with the 2D deformation of an elastic half-space. To eliminate this difficulty, the boundary conditions of the transducer must be taken into account (i.e., a finite transducer must be analyzed), which requires, in general, the use of FEA.

Since measurements of strain (or stress) are usually done by a discrete number of sensing elements, Equation 38.2 must be discretized (Figure 38.3). Correspondingly, the force distribution must be reconstructed at discrete positions as shown in Figure 38.3. Let Δx_q be the distance between points, where the force distribution must be reconstructed from strain (or stress) measurements carried out by strain (or stress) sensing elements uniformly distributed at intervals Δx_p, at $z = d$. Also assume, even though it is

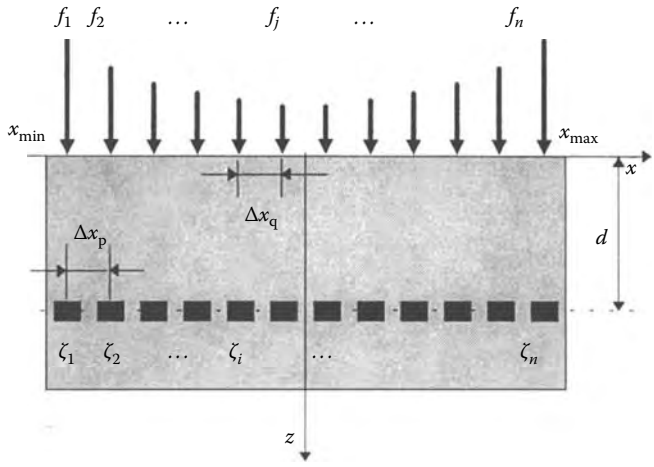


FIGURE 38.3 1D transducer with discrete-sensing elements located at $z = d$.

not necessary, that $\Delta x_q = \Delta x_p = \Delta x$ and that the forces are applied at positions immediately above the sensor elements. One can now define the strain (stress)-sample vector, ζ, whose components are given by $\zeta_i = \varepsilon_x(x_i)$, $i = 1, 2, \ldots, n$, and the force distribution vector, **F**, whose components are given by $f_i = q_v(x_j)$, $j = 1, 2, \ldots, n$. Then, the discrete form of Equation 38.1 is given by

$$\zeta = \mathbf{TF} \tag{38.5}$$

where the elements of the matrix **T** are given by $T_{ij} = k_v(x_i - x_j)$, $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, n$ [23]. A similar relation to Equation 38.5 can be obtained discretizing Equation 38.3 in the general case, where $\Delta x_q \neq \Delta x_p$ and **T** is not square. Furthermore, in the general case, the vector **F** comprises both vertical and tangential components.

Equations 38.1 and 38.3 represent the regular *forward problem*, while Equation 38.5 represents the discretized version of the forward problem. The *inverse problem*, in most cases, consists of recovering the applied force profile from the measurements of strain, stress, or deflection. (Note that the surface displacement can also be used to recover the indentor's shape.)

In [20], it was shown that the inverse problem is ill posed because the operators h and k, of Equations 38.1 and 38.3, respectively, are ill conditioned. Consequently, the inverse problem is susceptible to noise. To solve this problem, regularization techniques must be utilized [20].

It has been proven that in order to avoid aliasing in determining the continuous strain (stress) at a depth d using a discretized transducer, the elements have to be separated by one-tenth of the compliant layer's thickness. However, good results were obtained, without much aliasing, by separating the sensing elements by a distance equal to the sensor's depth [18].

38.4 Requirements for Tactile Sensors

In 1980, Harmon conducted a survey to determine general specifications for tactile sensors [21]. Those specifications have been used subsequently as guidelines by many tactile sensor designers:

1. Spatial resolution of 1–2 mm
2. Array sizes of 5 × 10 to 10 × 20 points
3. Sensitivity of 0.5×10^{-2} to 1×10^{-2} N for each force-sensing element (tactel)
4. Dynamic range of 1000:1
5. Stable behavior and with no hysteresis

6. Sampling rate of 100 Hz–1 kHz
7. Monotonic response, though not necessarily linear
8. Compliant interface, rugged and inexpensive

While properties (5), (7), and (8) earlier should apply to any practical sensor, the others are merely suggestions, particularly with respect to the number of array elements and spatial resolution.

Developments on tactile sensing following [21] have identified additional desirable qualities, namely, reliability, modularity, speed, and availability of multisensor support [16].

38.5 Technologies for Tactile Sensing

The technologies associated with tactile sensing are quite diverse: extensive surveys of the state of the art of robotic-tactile-transduction technologies have been presented in [2,3,16,17]. Some of these technologies will be briefly discussed.

38.5.1 Resistive

The transduction method that has received the most attention in tactile sensor design is concerned with the change in resistance of a conductive material under applied pressure. A basic configuration of a resistive transducer is shown in Figure 38.4. Each resistor, whose value changes with the magnitude of the force, represents a resistive cell of the transducer. Different materials have been utilized to manufacture the basic cell.

Conductive elastomers were among the first resistive materials used for the development of tactile sensors. They are insulating, natural, or silicone-based rubbers made conductive by adding particles of conductive or semiconductive materials (e.g., silver or carbon). The changes in resistivity of the elastomers under pressure are produced basically by two different physical mechanisms. In the first approach, the change in resistivity of the elastomer under pressure is associated with deformation that alters the particle density within it. Two typical designs of this kind are given in [22,23]. In the second approach, while the bulk resistance of the elastomer changes slightly when it is compressed, the design allows the

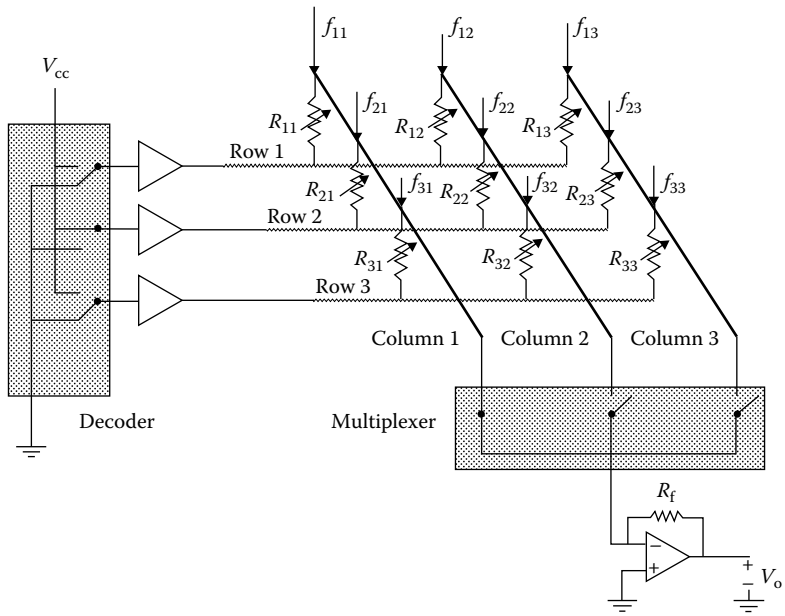


FIGURE 38.4 General configuration of a resistive transducer.

increase of the area of contact between the elastomer and an electrode and correspondingly a change in the contact resistance. A typical design of this kind is given in [24]. In [25], a newer tactile sensor is reported with both three-axis force-sensing and slippage-sensing functions. In the former case, the pressure-sensing function is achieved utilizing arrays of pressure transducers that measure a change in contact resistance between a specially treated polyimide film and a resistive substrate.

Piezoresistive elements have also been used in several tactile sensors. This technology is specifically attractive at present because, with micromachining, the piezoresistive elements can be integrated together with the signal-processing circuits in a single chip [26]. A 32 × 32-element silicon pressure sensor array incorporating CMOS processing circuits for the detection of a high-resolution pressure distribution was reported in [8]. The sensor array consists of an *x*–*y*-matrix-organized array of pressure cells with a cell spacing of 250 μm. CMOS processing circuits are formed around the array on the same chip. The fabrication of the sensor array was carried out using a 3 mm CMOS process combined with silicon micromachining techniques. The associated diaphragm size is 50 μm × 50 μm. The overall sensor-array chip size is 10 mm × 10 mm.

In Figure 38.4, a circuit topology, to scan a 3 × 3 array of piezoresistive elements, is shown. The basic idea was originally proposed in [24] and adapted on several occasions by different researchers. Using this method, the changes in resistance are converted into voltages at the output. With the connections as shown in Figure 38.4, the resistance R_{21} can be determined from

$$V_o = \frac{R_f}{R_{21}} V_{cc} \tag{38.6}$$

where

V_o is the output voltage
V_{cc} is the bias voltage
R_f is the feedback resistance of the output amplifier stage

One problem with the configuration shown in Figure 38.4 is the difficulty in detecting small changes in resistance due to the internal resistance of the multiplexer as well as changes in the voltage of power source, which have a great influence at the output. Other methods utilized to scan resistive transducer arrays are summarized in [3].

When piezoresistors and circuits are fabricated on the same silicon substrate, the sensor array can be equipped with a complex switching circuit, next to the sensing elements, which allows a better resolution in the measurements [9].

38.5.2 Capacitive

Tactile sensors within this category are concerned with measuring capacitance, which varies under applied load. The capacitance of a parallel-plate capacitor depends on the separation of the plates and their areas. A sensor using an elastomeric separator between the plates provides compliance such that the capacitance will vary according to the applied normal load (Figure 38.5a).

Figure 38.5b shows the basic configuration of a capacitive tactile sensor. The intersections of rows and columns of conductor strips form the capacitors. Each individual capacitance can be determined by measuring the corresponding output voltage at the selected row and column. To reduce cross talk and electromagnetic interference, the rows and columns that are not connected are grounded. Figure 38.5c shows an equivalent circuit when the sensor is configured to measure the capacitance formed at the intersection of rows *i* and *j*, C_{ij}. R_d is the input resistance of the detector, and C_d represents the effects of the stray capacitances, including the detector–amplifier input capacitance, the stray capacitance due to the unselected rows and columns, and the capacitance contributed by the cable that connects the transducer to the detector. Since the stray capacitance due to the unselected rows and columns changes with the applied forces, the stray capacitance due to the cable is designed to be predominant [18].

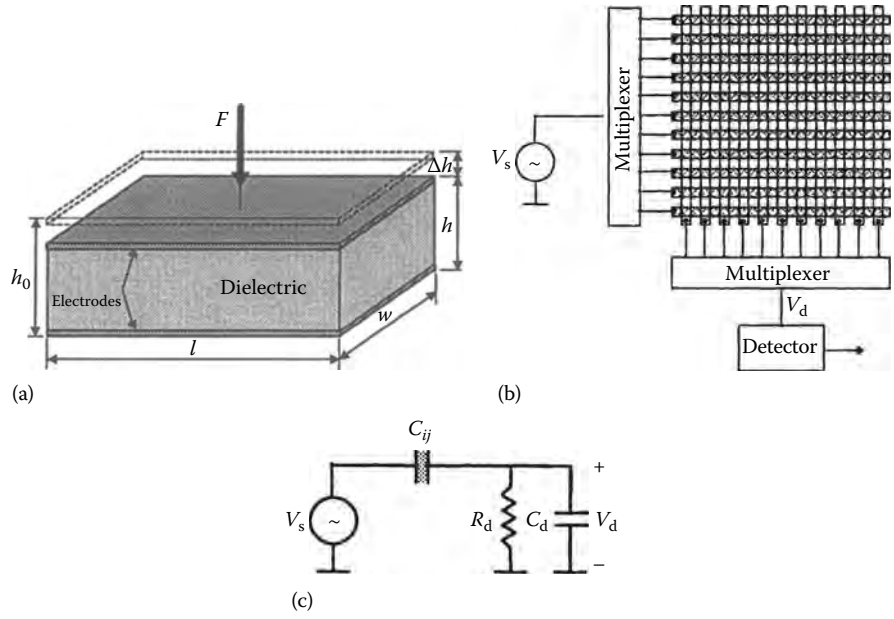


FIGURE 38.5 (a) Basic cell of a capacitor tactile sensor. (b) Typical configuration of a capacitive tactile sensor. (c) Equivalent circuit for the measurement of the capacitance C_{ij}.

The magnitude of voltage at the input of the detector, $|V_d|$, is given by

$$|V_d| = \frac{C_{ij} R_d \omega}{\sqrt{1 + [\omega R_d (C_{ij} + C_d)]^2}} |V_s| \tag{38.7}$$

Assuming that $C_d \gg C_{ij}$ and ω is sufficiently large,

$$|V_d| \cong \frac{C_{ij}}{C_d} |V_s| \tag{38.8}$$

When a load is applied to the transducer, the capacitor is deformed as shown in Figure 38.5a. For modeling purposes, it is assumed that the plate capacitor is only under compression. When no load is applied, the capacitance due to the element in the ith row and the jth column, C_{ij}^0, is given by

$$C_{ij}^0 = \varepsilon \frac{wl}{h_0} \tag{38.9}$$

where

ε is the permittivity of the dielectric
w and l are the width and the length of the plate capacitor, respectively
h_0 is the distance between plates when no load is applied

The voltage at the input of the detector for this particular case is indicated by V_{d0}; then from Equation 38.8, one obtains

$$|V_{d0}| \cong \frac{C_{ij}^0}{C_d} |V_s| \tag{38.10}$$

When a load is applied, the capacitor is under compression and the capacitance is given by

$$C_{ij} = \varepsilon \frac{wl}{h_0 - \Delta h} \tag{38.11}$$

The strain in this case is given by

$$\zeta_z \cong \frac{\Delta h}{h_0} \tag{38.12}$$

where

Δh is the displacement of the top metal plate
$\Delta h \ll h_0$

The strain can be measured by

$$\frac{|V_d| - |V_{d0}|}{|V_d|} = \frac{(C_{ij}/C_d) - (C_{ij}^0/C_d)}{C_{ij}/C_d} = 1 - \frac{C_{ij}^0}{C_{ij}} = 1 - \frac{h_0 - \Delta h}{h_0} = \frac{\Delta h}{h_0} = \frac{\Delta h}{h_0} \cong \zeta_z \tag{38.13}$$

Consequently, the strain at each tactel can be determined by measuring the magnitudes of V_d and V_{d0} for each element.

Note that the presence of a tangential force would offset the plates tangentially and change the effective area of the capacitor plates. An ideal capacitive pressure sensor can quantify basic aspects of touch by sensing normal forces and can detect slippage by measuring tangential forces. However, distinguishing between the two forces at the output of a single sensing element is a difficult task and requires a more complex transducer than the one presented in Figure 38.5a [27].

Micromachined, silicon-based capacitive devices are especially attractive due to their potential for high accuracy and low drift. A sensor with 1024 elements and a spatial resolution of 0.5 mm was reported in [28]. Several possible structures for implementing capacitive high-density tactile transducers in silicon have been reported in [29]. A cylindrical finger-shaped transducer was reported in [18].

The advantages of capacitive transducers include wide dynamic range, linear response, and robustness. Their major disadvantages are susceptibility to noise, sensitivity to temperature, and the fact that capacitance decreases with physical size, ultimately limiting the spatial resolution. Research is progressing toward the development of electronic processing circuits for the measurement of small capacitances using charge amplifiers [30] and the development of new capacitive structures [29].

38.5.3 Piezoelectric

A material is called piezoelectric if, when subjected to a stress or deformation, it produces electricity. Longitudinal piezoelectric effect occurs when the electricity is produced in the same direction of the stress (Figure 38.6). In Figure 38.6a, a normal stress σ (=F/A) is applied along the direction 3, and the charges are generated on the surfaces perpendicular to direction 3. A transversal piezoelectric effect occurs when the electricity is produced in the direction perpendicular to the stress.

The voltage V generated across the electrodes by the stress σ is given by

$$V = d_{33} \frac{h}{\varepsilon} \sigma \tag{38.14}$$

where

d_{33} is the piezoelectric constant associated with the longitudinal piezoelectric effect
ε is the permittivity
h is the thickness of the piezoelectric material

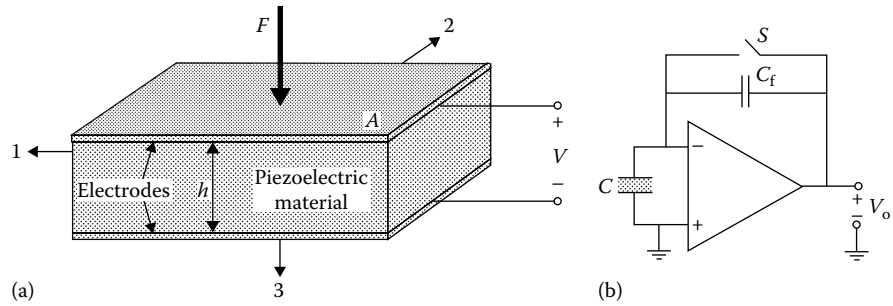


FIGURE 38.6 (a) Basic cell of a piezoelectric transducer. (b) Charge amplifier utilized for the measurement of the applied force.

Since piezoelectric materials are insulators, the transducer shown in Figure 38.6a can be considered as a capacitor, from an electric point of view. Consequently,

$$V = \frac{Q}{C} = \frac{Q}{\varepsilon A} h \tag{38.15}$$

where

Q is the charge induced by the stress σ
C is the capacitance of the parallel capacitor
A is the area of each electrode

A comparison of Equations 38.14 and 38.15 leads to

$$Q = d_{33} A \sigma \tag{38.16}$$

It is concluded that the force applied to the photoelastic material can be determined by finding the charge Q. Charge amplifiers are usually utilized for determining Q. The basic configuration of a charge amplifier is shown in Figure 38.6b. The charge generated in the transducer is transferred to the capacitor C_f, and the output voltage, V_0, is given by

$$V_0 = -\frac{Q}{C_f} \tag{38.17}$$

The circuit must periodically discharge the feedback capacitor C_f to avoid saturation of the amplifier by stray charges generated by the offset voltages and currents of the operational amplifier. This is achieved by a switch as shown in Figure 38.6b or by a resistor parallel to C_f.

The piezoelectric material most widely used in the implementation of tactile transducers is PVF2. It shows the largest piezoelectric effect of any known material. Its flexibility, small size, sensitivity, and large electric output offer many advantages for sensor applications in general and tactile sensors in particular. The examples of tactile sensors implemented with this technology can be found in [1,31].

The major advantages of the piezoelectric technology are its wide dynamic range and durability. Unfortunately, the response of available materials does not extend down to dc, and therefore, steady loads cannot be measured directly. Also, the PVF2 material produces a charge output that is prone to electric interference and is temperature dependent.

The possibility of measuring transient phenomenon using piezoelectric material has recently encouraged some researchers to use the piezoelectric effect for detecting vibrations that indicate incipient slip, occurrence of contact, local change in skin curvature, and estimating friction and hardness of

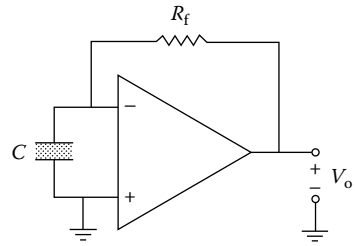


FIGURE 38.7 Current-to-voltage converter.

the object [7,10,11]. If the piezoelectric transducer shown in Figure 38.6a is connected to an FET-input operational amplifier configured as a current-to-voltage converter as shown in Figure 38.7, the output voltage is given by

$$V_o = \frac{dQ}{dt} R_f = AR_f d_{33} \frac{d\sigma}{dt} \tag{38.18}$$

where R_f is the feedback resistor. Correspondingly, the circuit configuration provides the mean to measure of changes in the contact stress. A detailed explanation of the behavior of this sensor can be found in [7].

38.5.4 Optical

Recent developments in fiber-optic technology and solid-state cameras have led to numerous novel tactile sensor designs [32,33]. Some of these designs employ flexible membranes incorporating a reflecting surface (Figure 38.8). Light is introduced into the sensor via a fiber-optic cable. A wide cone of light propagates out of the fiber, reflects back from the membrane, and is collected by a second fiber. When an external force is applied onto the elastomer, it shortens the distance between the reflective side of the membrane and the fibers, h. Consequently, the light gathered by the receiving fiber changes as a function of h (Figure 38.8b). To recover univocally the distance from the light intensity, a monotonic function

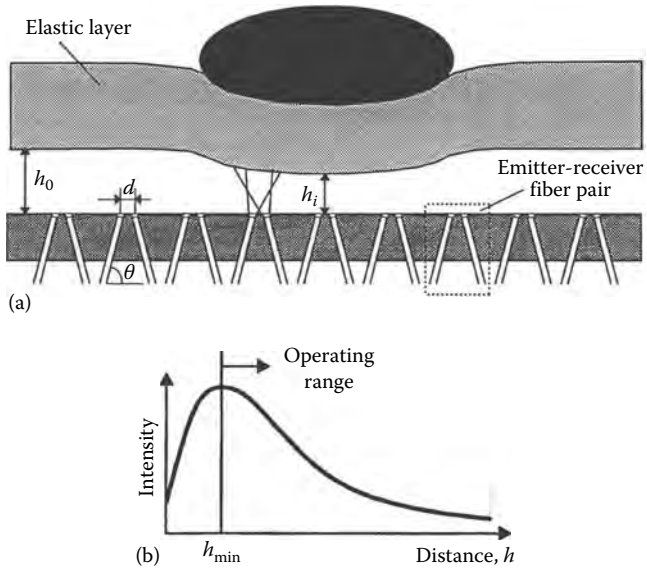


FIGURE 38.8 (a) Reflective transducer. (b) Light intensity as a function of the distance h.

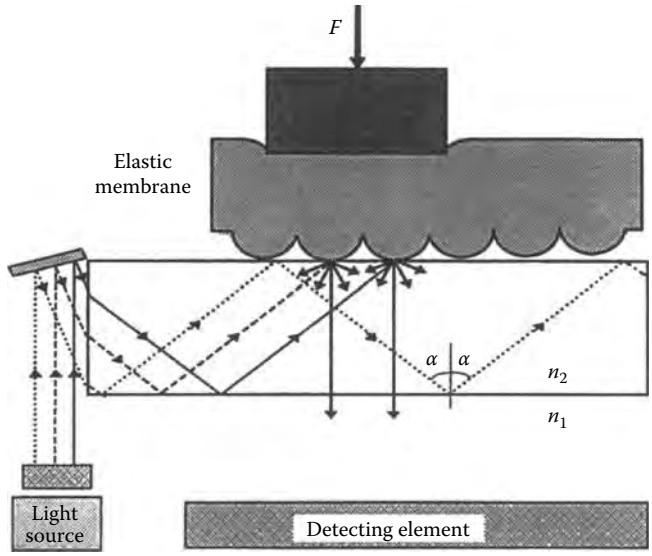


FIGURE 38.9 Tactile transducer based on the principle of internal reflection.

is needed. This can be achieved by designing the transducer such it operates for $h > h_{min}$, where h_{min} is indicated in Figure 38.8b. (The region $h > h_{min}$ is preferred to the $h < h_{min}$ for dynamic range reasons.)

Another optical effect that can be used is that of frustrated total internal reflection [5,34]. With this technique, an elastic rubber membrane covers, without touching, a glass plate (waveguide); light entering the side edge of the glass is totally reflected by the top and bottom surfaces and propagates along it (Figure 38.9).

The condition for total internal reflection occurs when

$$n_2 \sin\alpha \leq n_1 \tag{38.19}$$

where

n_1 is the index of refraction of the medium surrounding the waveguide (in this case air, $n_1 \cong 1$)
n_2 is the index of refraction of the waveguide
α is the angle of incidence at the interface glass–air

Objects in contact with the elastic membrane deform it and induce contact between the bottom part of the membrane and the top surface of the waveguide, disrupting the total internal reflection. Consequently, the light in the waveguide is scattered at the contact location. Light that escapes through the bottom surface of the waveguide can be detected by an array of photodiodes and a solid-state sensor or, alternatively, transported away from the transducer by fibers [3]. The detected imaged is stored in a computer for further analysis. A rubber membrane with a flat surface yields a high-resolution binary (contact or noncontact) image [5]. If the rubber sheet is molded with a textured surface (Figure 38.9), then an output proportional to the area of contact is obtained and, consequently, the applied forces can be detected [3]. Shear forces can also be detected using special designs [35]. Sensors based on frustrated internal reflection can be molded into a finger shape [5] and are capable of forming very high-resolution tactile images. Such sensors are commercially available. An improved miniaturized version of a similar sensor was proposed in [34].

Other types of optical transducers use "occluder" devices. One of the few commercially available tactile sensors uses this kind of transducer [36]. In one of the two available designs, the transducer's surface is made of a compliant material, which has on its underside a grid of elongated pins. When force is applied to the compliant surface, the pins on the underside undergo a mechanical motion normal to

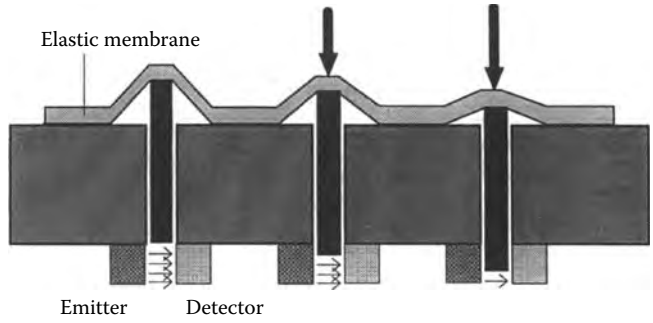


FIGURE 38.10 Principle of operation of an occluder–transducer.

the surface, blocking the light path of a photoemitter–detector pair. The amount of movement determines the amount of light reaching the photoreceiver. Correspondingly, the more force applied, the less amount of light is collected by the photoreceiver (Figure 38.10). The major problems with this specific device are associated with creep, hysteresis, and temperature variation. This scheme also requires individual calibration of each photoemitter–photodetector pair.

Fibers have also been used directly as transducers in the design of tactile sensors. Their use is based on two properties of fiber-optic cables: (1) If a fiber is subjected to a significant amount of bending, then the angle of incidence at the fiber wall can be reduced sufficiently for light to leave the core [37], and (2) if two fibers pass close to one another and both have roughened surfaces, then light can pass between the fibers. Light coupling between adjacent fibers is a function of their separation [3].

An example of an optical fiber tactile sensor, whose sensing mechanism is based on the increase of light attenuation due to the microbend in the optical fibers, is shown in Figure 38.11 [37]. The transducer consists of a four-layer, 2D fiber-optic array constructed by using two layers of optical fibers as a corrugation structure, through which microbends are induced in two orthogonal layers of active fibers. Each active fiber uses LED as the emitter and PIN photodiode as a detector. When an object is forced into contact with the transducer, a light distribution is detected at each detector. This light distribution is related to the applied force and the shape of the object. Using complex algorithms and active sensing (moving the object in relation to the transducer), the object position, orientation, size, and contour information can be retrieved [37]. However, the recovery of the applied force profiles was not reported in [37].

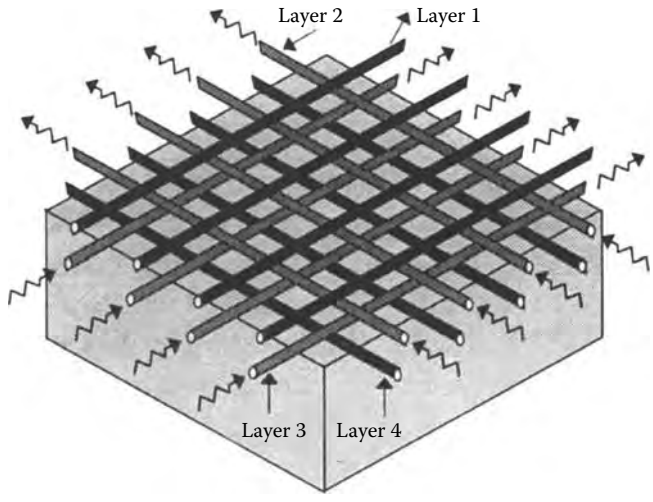


FIGURE 38.11 Four-layer tactile transducer.

38.5.5 Photoelastic

An emerging technology in optical tactile sensing is the development of photoelastic transducers. When a light ray propagates into an optically anisotropic medium, it splits into two rays that are linearly polarized at right angles to each other and propagate at different velocities. This splitting of a ray into two rays that have mutually perpendicular polarizations results from a physical property of crystalline material that is called *optical birefringence* or simply *birefringence*. The direction in which light propagates with the higher velocity is called the *fast axis*, and the one in which it propagates more slowly is called the *slow axis*. Some optically isotropic materials—such as glass, celluloid, bakelite, and transparent plastics in general—become birefringent when they are subjected to a stress field. The birefringent effect lasts only during the application of loads. Thus, this phenomenon is called *temporary* or *artificial birefringence* or, more commonly, the *photoelastic phenomenon*.

Figure 38.12a shows a photoelastic transducer proposed in [38]. It consists of a fully supported two-layer beam with a mirrored surface sandwiched in between. Normal line forces are applied to the top

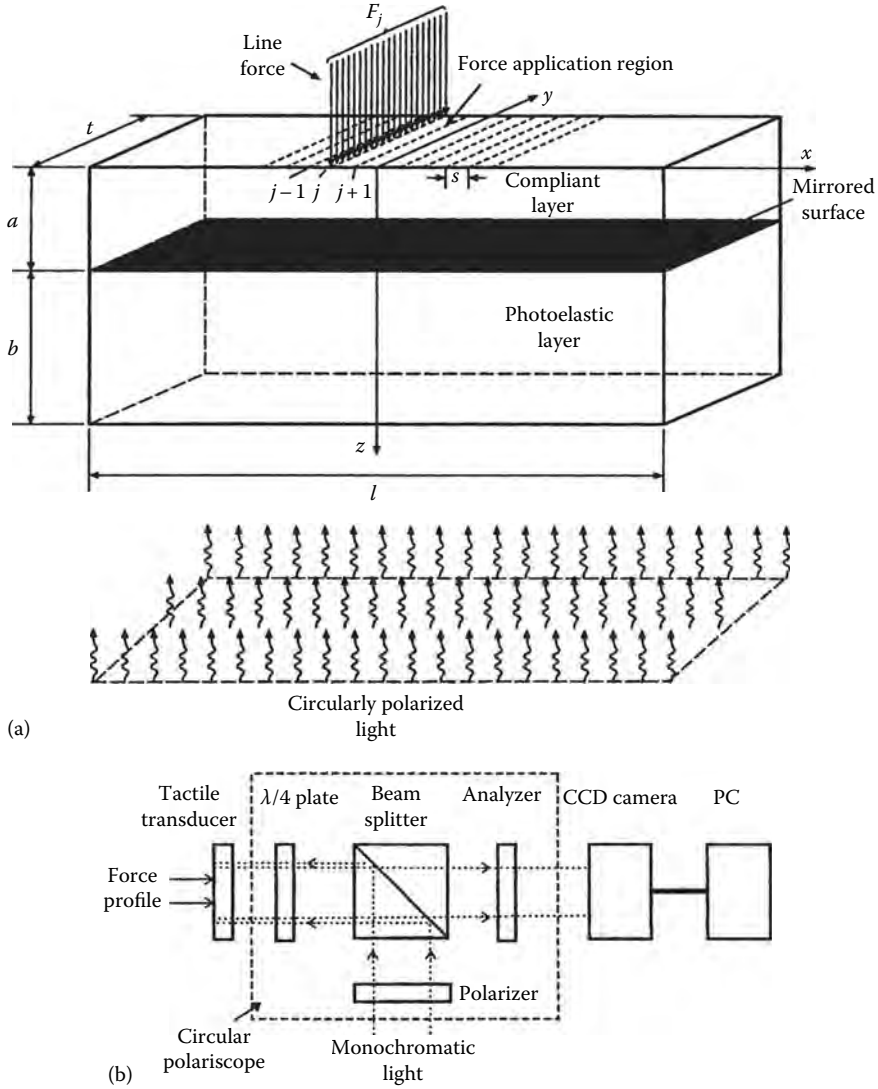


FIGURE 38.12 (a) Photoelastic transducer. (b) Circular reflective polariscope.

surface of the beam at discrete tactels, separated by equal distances, *s*, along the beam. The upper compliant layer is for the protection of the mirror, while the lower one is the photoelastic layer.

Circularly polarized monochromatic light, incident along the *z*-axis, illuminates the bottom surface of the transducer. The light propagates parallel to the *z*-axis, passes through the photoelastic layer, and then reflects back from the mirror. If no force is applied to the transducer, the returning light is circularly polarized because unstressed photoelastic material is isotropic. If force is applied, stresses are induced in the photoelastic layer, making the material birefringent. This introduces a certain phase difference between the components of the electric field associated with the light-wave propagation. The two directions of polarization are in the plane perpendicular to the direction of propagation (in this case, the *x*–*y* plane). As a consequence of this effect, the output light is elliptically polarized, creating a phase difference distribution, *p*, between the input light and the output light at each point in the *x*–*y* plane. The phase difference distribution carries the information of the force distribution applied to the transducer.

A *polariscope* is a practical method to observe the spatial variation on light intensity (fringes) due to the effect of induced phase difference distribution. Polariscopes can be either linear or circular, depending on the required polarization of the light. They can also be characterized as a reflective or a transparent type, depending on whether the photoelastic transducer reflects or transmits the light.

A circular, reflective polariscope, shown in Figure 38.12b, is utilized to illuminate the transducer shown in Figure 38.12a. The input light is linearly polarized and is directed toward the photoelastic transducer by a beam splitter. Before reaching the transducer, the light is circularly polarized by a quarter-wave plate. The output light is elliptically polarized when a force is applied. This light is directed toward a detector passing through the quarter-wave plate, the beam splitter, and an analyzer. Finally, it is detected by a camera linked to a frame grabber connected to a PC, for further data processing. The light that illuminates the camera consists of a set of fringes from where the force distribution applied to the transducer must be recovered. A technique for the recovery of the forces from the fringes is described in [38]. A model of the transducer using FEA is reported in [39].

One of the earlier applications of photoelasticity to tactile sensing dates back to the era of development phase of the Utah/MIT dextrous hand [40]. The researchers proposed the use of the photoelastic phenomenon as a transduction method for the recovery of the force profile applied to the fingers of the hand. They limited their application to the development of a single-touch transducer, although they claimed that an array of such devices could be implemented. However, the construction of a large array of their devices would be difficult. To overcome this difficulty, another research group proposed a different transducer [41]. Although an analytical model was developed for the sensor, a systematic method for recovering the 2D force profile from the light intensity distribution was not reported. Thus, the sensor was used mainly for the study of the forward analysis, namely, observing the light intensity distribution for different touching objects brought into contact with the sensor. This sensor could eventually be used for determining some simple geometric properties of a touching object.

A tactile sensor reported in [42] is capable of detecting slippage. The output light intensity (the fringe pattern) is captured by a camera interfaced to a PC. The light intensity distribution changes when an object moves across the surface of the transducer. A direct analysis of the fringes is used to detect movement of the grasped object; a special technique was reported to optimize the comparison process for detecting differences between two fringe patterns occurring due to the slippage of the object in contact with the sensor [42]. It is important to note that such an analysis of the fringes does not require the recovery of the applied force profile.

Photoelasticity offers several attractive properties for the development of tactile sensors: good linearity, compatibility with vision-base sensing technologies, and high spatial resolution associated with the latter that could lead to the development of high-resolution tactile imagers needed for object recognition and fine manipulation. Also, photoelastic sensors are compatible with fiber-optic technology that allows remote location of electronic processing devices and avoidance of interference problems.

Other technologies for tactile sensing include acoustic, magnetic, and microcavity vacuum sensors [43,44].

References

1. P. Dario, Contact sensing for robot active touch, in *Robotic Science*, M. Brady (ed.), Cambridge, MA: MIT Press, 1989, Chapter 3, pp. 138–163.
2. P. P. L. Regtien, Tactile imaging, *Sens. Actuat. A*, 31, 83–89, 1992.
3. R. A. Russell, *Robot Tactile Sensing*, Brunswick, Victoria, Australia: Prentice Hall, 1990.
4. A. D. Berger and P. K. Khosla, Using tactile data for real-time feedback, *Int. J. Rob. Res.*, 10(2), 88–102, 1991.
5. S. Begej, Planar and finger-shaped optical tactile sensors for robotic applications, *IEEE J. Rob. Autom.*, 4, 472–484, 1988.
6. R. A. Russell and S. Parkinson, Sensing surface shape by touch, *Proceedings of the IEEE International Conference on Robotics Automation*, Atlanta, GA, 1993, pp. 423–428.
7. R. D. Howe, A tactile stress rate sensor for perception of fine surface features, *Proceedings of the IEEE International Conference on Solid-State Sensors Actuators*, San Francisco, CA, 1991, pp. 864–867.
8. S. Sugiyama, K. Kawahata, H. Funabashi, M. Takigawa, and I. Igarashi, A 32 × 32 (1K)-element silicon pressure-sensor array with CMOS processing circuits, *Electron. Commun. Jpn.*, 75(1), 64–76, 1992.
9. J. S. Son, E. A. Monteverde, and R. D. Howe, A tactile sensor for localizing transient events in manipulation, *Proceedings of the IEEE International Conference on Robotics Automation*, San Diego, CA, 1994, pp. 471–476.
10. M. R. Tremblay and M. R. Cutkosky, Estimating friction using incipient slip sensing during manipulation task, *Proceedings of the IEEE International Conference on Robotics Automation*, Atlanta, GA, 1993, pp. 429–434.
11. S. Omata and Y. Terubuna, New tactile sensor like the human hand and its applications, *Sens. Actuat. A*, 35, 9–15, 1992.
12. R. Bayrleithner and K. Komoriya, Static friction coefficient determination by force sensing and its applications, *IROS'94*, Munich, Germany, 1994, pp. 1639–1646.
13. M. A. Abidi and R. C. Gonzales, The use of multisensor data for robotic applications, *IEEE Trans. Rob. Autom.*, 6, 159–177, 1990.
14. A. A. Cole, P. Hsu, and S. S. Sastry, Dynamic control of sliding by robot hands for regrasping, *IEEE Trans. Rob. Autom.*, 8, 42–52, 1992.
15. P. K. Allen and P. Michelman, Acquisition and interpretation of 3-D sensor data from touch, *IEEE Trans. Rob. Autom.*, 6, 397–404, 1990.
16. H. R. Nicholls (ed.), *Advanced Tactile Sensing for Robotics*, Singapore: World Scientific Publishing, 1992.
17. J. G. Webster (ed.), *Tactile Sensors for Robotics and Medicine*, New York: John Wiley & Sons, 1988.
18. R. S. Fearing, Tactile sensing mechanism, *Int. J. Rob. Res.*, 9(3), 3–23, 1990.
19. T. H. Speeter, Three-dimensional finite element analysis of elastic continua for tactile sensing, *Int. J. Rob. Res.*, 11(1), 1–19, 1992.
20. Y. C. Pati, R. S. Krishnaprasad, and M. C. Peckerar, An analog neural network solution to the inverse problem of early taction, *IEEE Trans. Rob. Autom.*, 8(2), 196–212, 1992.
21. L. D. Harmon, Automated tactile sensing, *Int. J. Rob. Res.*, 1(2), 3–32, 1982.
22. W. E. Snyder and J. St. Clair, Conductive elastomers as a sensor for industrial parts handling equipment, *IEEE Trans. Instrum. Meas.*, 27(1), 94–99, 1991.
23. M. Shimojo, M. Ishikawa, and K. Kanaya, A flexible high resolution tactile imager with video signal output, *Proceedings of the IEEE International Conference on Robotics Automation*, Sacramento, CA, 1991, pp. 384–391
24. W. D. Hillis, A high resolution imaging touch sensor, *Int. J. Rob. Res.*, 1(2), 33–44, 1982.
25. Y. Yamada and M. R. Cutkosky, Tactile sensor with 3-axis force and vibration sensing functions and its applications to detect rotational slip, *Proceedings of the IEEE International Conference on Robotics Automation*, San Diego, CA, 1994, pp. 3550–3557.
26. K. Njafi and C. H. Mastrangelo, Solid-state microsensors and smart structure, *Ultrasonic Symposium*, Baltimore, MD, 1993, pp. 341–350.

27. F. Zhu and J. W. Spronck, A capacitive tactile sensor for shear and normal force measurements, *Sens. Actuat. A*, 31, 115–120, 1992.
28. K. Suzuki, K. Najafi, and K. D. Wise, A 1024-element high-performance silicon tactile imager, *IEEE Trans. Electron Devices*, 17(8), 1852–1860, 1990.
29. M. R. Wolffenbuttel and P. L. Regtien, The accurate measurement of a micromechanical force using force-sensitive capacitances, *Conference on Precision Electromagnetic Measurement*, Boulder, CO, 1994, pp. 180–181.
30. M. R. Wolffenbuttel, R. F. Wolffenbuttel, and P. P. L. Regtien, An integrated charge amplifier for a smart tactile sensor, *Sens. Actuat. A*, 31, 101–109, 1992.
31. E. D. Kolesar, Jr. and C. S. Dyson, Object imaging with piezoelectric robotic tactile sensor, *J. Microelectromech. Syst*., 4(2), 87–96, 1995.
32. J. L. Scheiter and T. B. Sheridan, An optical tactile sensor for manipulators, *J. Rob. Comput. Integr. Manuf.*, 1, 65–71, 1989.
33. R. Ristic, B. Benhabib, and A. A. Goldenberg, Analysis and design of a modular electrooptical tactile sensor, *IEEE Trans. Rob. Autom*., 5(3), 362–368, 1989.
34. H. Maekawa, K. K. Tanie, K. Komoriya, M. Kaneko, C. Horiguchi, and T. Sugawara, Development of a finger shaped tactile sensor and it evaluation by active touch, *Proceedings of the IEEE International Conference on Robotics Automation*, Nice, France, 1992, pp. 1327–1334.
35. M. Ohka, Y. Mitsurya, S. Takeuchi, and O. Kamekawa, A three-axis optical tactile sensor (fern contact analyses and sensing experiments using a large-sized tactile sensor), *Proceedings of the IEEE International Conference on Robotics Automation*, Nagoya, Aichi, Japan, 1995, pp. 817–824.
36. J. Rebman and K. A. Morris, A tactile sensor with electro-optical transduction, in *Robots Sensors, Tactile and Non-Vision*, Vol. 2, A. Pugh (ed.), Berlin, Germany: EFS Publications, 1986, pp. 145–155.
37. S. R. Emge and C. L. Chen, Two dimensional contour imaging with a fiber optic microbend tactile sensor array, *Sens. Actuat. B*, 3, 31–42, 1991.
38. R. E. Saad, A. Bonen, K. C. Smith, and B. Benhabib, Distributed-force recovery for a planar photoelastic tactile sensor, *IEEE Trans. Instrum. Meas*., 45, 541–546, 1996.
39. R. E. Saad, A. Bonen, K. C. Smith, and B. Benhabib, Finite-element analysis for photoelastic tactile sensors, *Proceedings of the IEEE International Conference on Industrial Electronics, Control, and Instrumentation*, Orlando, FL, 1995, pp. 1202–1207.
40. S. C. Jacobsen, J. E. Wood, D. F. Knutti, and B. Biggers, The Utah/MIT dexterous hand: Work in progress, in *Robotics Research: The First International Symposium*, M. Brady and R. Paul (eds.), Cambridge, MA: MIT Press, 1983, pp. 601–653.
41. A. Cameron, R. Daniel, and H. Durrant-Whyte, Touch and motion, *Proceedings of the IEEE International Conference on Robotics Automation*, Philadelphia, PA, 1988, pp. 1062–1067.
42. S. H. Hopkins, F. Eghtedari, and D. T. Pham, Algorithms for processing data from a photoelastic slip sensor, *Mechatronics*, 2(1), 15–28, 1992.
43. S. Ando and H. Shinoda, Ultrasonic emission tactile sensing, *IEEE Trans. Control Syst*., 15(1), 61–69, 1996.
44. J. C. Jiang, V. Faynberg, and R. C. White, Fabrication of micromachined silicon tip transducer for tactile sensing, *J. Vac. Sci. Technol. B*, 11, 1962–1967, 1993.

Further Readings

Amazeen, C. A. and Bishop, S. S., Triaxial tactile sensor, http://www.devileye.net/catalog/motor-vehicle_seat_headrest/triaxial_tactile_sensor.html (accessed on June 22, 2012).

Dahiya, R. S. and Valle, M., *Robotic Tactile Sensing*, Dordrecht, the Netherlands: Springer, 2013. Online version available http://www.springerlink.com/content/978-94-007-0579-1#section=1102532&page=1&locus=71 (accessed August 10, 2012).

Najarian, S., Dargahi, J., and Mehrizi, A., *Artificial Tactile Sensing in Biomedical Engineering*, New York: McGraw-Hill, 2009,

Tiwana, M. I., Redmond, S. J., and Lovell, N. H., A review of tactile sensing technologies with applications in biomedical engineering, *Sens. Actuat.*, 179, 17–31, 2012. Also available at http://www.sciencedirect.com/science/article/pii/S0924424712001641 (accessed on July 24, 2012).

Partial List of Vendors and Suppliers

Beijer Electronics, Inc., 2212 South West Temple No. 50, Salt Lake City, UT 84115, Phone: 801-466-8770, 866-648-3767 (toll free), Fax: 801-466-8792, www.beijerelectronicsinc.com/?cid=ndc~th…

Densitron Corp, 2330 Pomona Rincon Rd., Corona, CA 92880, Phone: 951-284-7600, Fax: 951-284-7699, www.densitron.com

Ercon, Inc., 7 Kendrick Rd., Wareham, MA 02571, Phone: 508-291-1400, Fax: 508-291-2400, www.erconinc.com

Honeywell Sensing and Control, 1985 Douglas Dr. North, MN10-192B, Golden Valley, MN 55422, Phone: 763-954-4818, 800-446-6555 (toll free), sensing.honeywell.com/index.cfm/ci_id/15…

IFM Efector, Inc., 782 Springdale Dr., Exton, PA 19341-2843, Phone: 610-524-2000, 800-441-8246 (toll free), Fax: 800-329-0436, www.ifm.com/ifmus/web/home.htm

Kanson Electronics, 245 Forrest Ave., Hohenwald, TN 38462, Phone: 931-796-3050, Fax: 931-796-3956, www.issc-kanson.com

Microstrain, Inc., 459 Hurricane Lane, Suite 102, Williston, VT 05495, Phone: 802-862-6629, 800-449-3878 (toll free), Fax: 802-863-4093, www.microstrain.com

PICS INC, 81 Lancaster Ave., Suite 202, Malvern, PA 19355, Phone: 610-644-7701, Fax: 610-644-9297, www.picscontrols.com

Steven Engineering, 230 Ryan Way, South San Francisco, CA 94080, Phone: 888-790-0022 (toll free), Fax: 800-314-0716, 800-314-0716 (toll free), stevenengineering.rtrk.com/?scid=1975164…

Tekscan, 307 W. First St., South Boston, MA 02127, Phone: 617-464-4500, 800-248-3669 (toll free), Fax: 617-464-4266, www.tekscan.com



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

39

Pressure Measurement

39.1 Basic Definitions **39**-1
Units of Pressure and Conversion
39.2 Sensing Principles **39**-2
Sensing Elements • Detection Methods
39.3 Silicon Micromachined Pressure Sensors **39**-6
Silicon Piezoresistive Pressure Sensor Limitations
References **39**-9
Further Information **39**-9

Kevin H.L. Chau
Analog Devices, Inc.

39.1 Basic Definitions

Pressure is defined as the normal force per unit area exerted by a fluid (liquid or gas) on any surface. The surface can be either a solid boundary in contact with the fluid or, for purposes of analysis, an imaginary plane drawn through the fluid. Only the component of the force normal to the surface needs to be considered for the determination of pressure. Tangential forces that give rise to shear and fluid motion will not be a relevant subject of discussion here. In the limit that the surface area approaches zero, the ratio of the differential normal force to the differential area represents the pressure at a point on the surface. Furthermore, if there is no shear in the fluid, the pressure at any point can be shown to be independent of the orientation of the imaginary surface under consideration. Finally, it should be noted that pressure is not defined as a vector quantity and is therefore nondirectional.

Three types of pressure measurements are commonly performed:

Absolute pressure is the same as the pressure defined earlier. It represents the pressure difference between the point of measurement and a perfect vacuum where pressure is zero.

Gage pressure is the pressure difference between the point of measurement and the ambient. In reality, the ambient (atmospheric) pressure can vary, but only the pressure difference is of interest in gage pressure measurements.

Differential pressure is the pressure difference between two points, one of which is chosen to be the reference. In reality, both pressures can vary, but only the pressure difference is of interest here.

39.1.1 Units of Pressure and Conversion

The SI unit of pressure is the *pascal* (Pa), which is defined as the newton per square meter (N/m); 1 Pa is a very small unit of pressure. Hence, decimal multiples of the pascal (e.g., kilopascals [kPa] and megapascals [MPa]) are often used for expressing higher pressures. In weather reports, the hectopascal (1 hPa = 100 Pa) has been adopted by many countries to replace the millibar (1 bar = 10^5 Pa; hence, 1 millibar = 10^{-3} bar = 1 hPa) as the unit for atmospheric pressure. In the United States, pressure is commonly expressed in pound force per square inch (psi), which is about 6.90 kPa. In addition, the absolute,

TABLE 39.1 Pressure Unit Conversion Table

Key
1. kPa = kilopascal
2. psi = pound force per square inch
3. in. H_2O = inch of water at 4 °C
4. cm H_2O = centimeter of water at 4 °C
5. in. Hg = inch of mercury at 0 °C
6. mm Hg = millimeter of mercury at 0 °C
7. mbar = millibar

gage, and differential pressures are further specified as psia, psig, and psid, respectively. However, no such distinction is made in any pressure units other than the psi. There is another class of units, for example, millimeter of mercury at 0 °C (mm Hg, also known as the *torr*) or inch of water at 4 °C (in H_2O), which expresses pressure in terms of the height of a static liquid column. The actual pressure p referred to is one that will be developed at the base of the liquid column due to its weight, which is given by Equation 39.1:

$$p = \rho g h \tag{39.1}$$

where
- ρ is the density of the liquid
- g is the acceleration due to gravity
- h is the height of the liquid column

A conversion table for the most popular pressure units is provided in Table 39.1.

39.2 Sensing Principles

39.2.1 Sensing Elements

Since pressure is defined as the force per unit area, the most direct way of measuring pressure is to isolate an area on an elastic mechanical element for the force to act on. The deformation of the sensing element produces displacements and strains that can be precisely sensed to give a calibrated measurement of the pressure. This forms the basis for essentially all commercially available pressure sensors today. Specifically, the basic requirements for a pressure-sensing element are a means to isolate two fluidic pressures (one to be measured and the other one as the reference) and an elastic portion to convert the pressure difference into a deformation of the sensing element. Many types of pressure-sensing elements are currently in use. These can be grouped as diaphragms, capsules, bellows, and tubes, as illustrated in Figure 39.1. Diaphragms are by far the most widely used of all sensing elements. A special form of tube, known as the *Bourdon tube*, is curved or twisted along its length and has an oval cross section. The tube is sealed at one end and tends to unwind or straighten when it is subjected to a pressure applied to the inside. In general, Bourdon tubes are designed for measuring high pressures, while capsules and bellows are usually for measuring low pressures. A detailed description of these sensing elements can be found in [1].

39.2.2 Detection Methods

A detection means is required to convert the deformation of the sensing element into pressure readout. In the simplest approach, the displacements of a sensing element can be amplified mechanically by lever and flexure linkages to drive a pointer over a graduated scale, for example, in the moving pointer barometers. Some of the earliest pressure sensors employed a Bourdon tube to drive the wiper arm over a potentiometric resistance element. In *linear-variable differential-transformer* (LVDT) pressure sensors, the displacement of a Bourdon tube or capsule is used to move a magnetic core inside a coil assembly

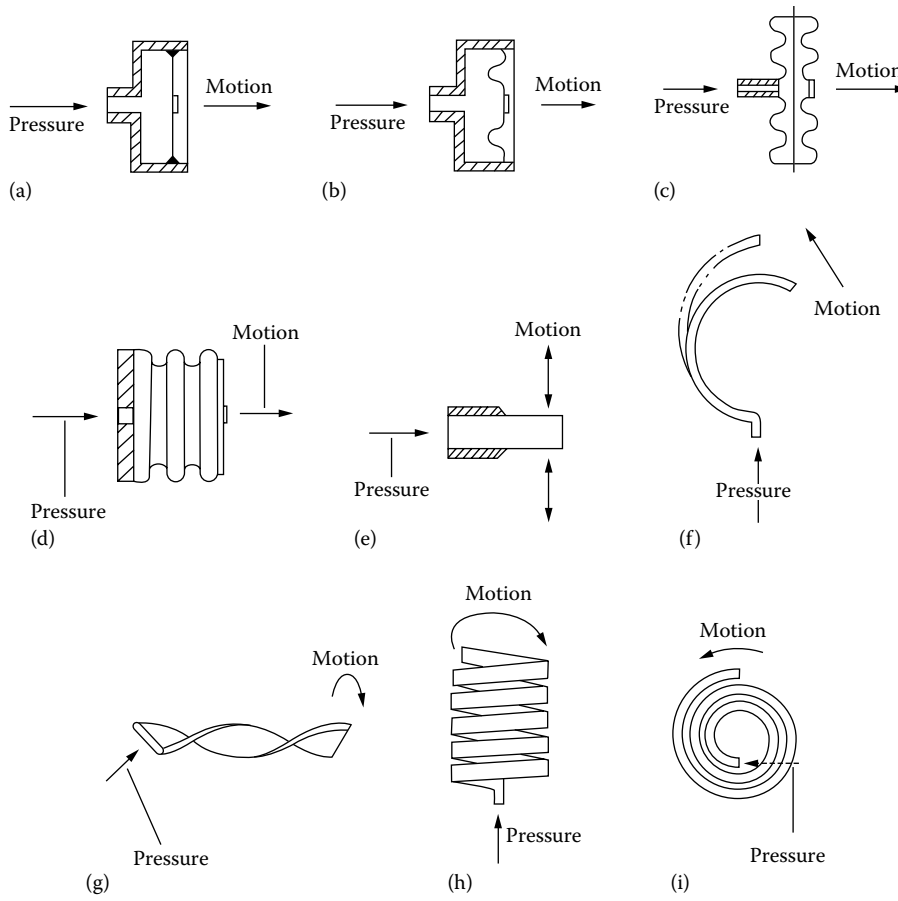


FIGURE 39.1 Pressure-sensing elements: (a) flat diaphragm, (b) corrugated diaphragm, (c) capsule, (d) bellows, (e) straight tube, (f) C-shaped Bourdon tube, (g) twisted Bourdon tube, (h) helical Bourdon tube, and (i) spiral Bourdon tube. (From Norton, H.N., *Handbook of Transducers*, Prentice Hall, Englewood Cliffs, NJ, 1989. Reprinted with permission.)

to vary its inductance. In *piezoelectric* pressure sensors, the strains associated with the deformation of a sensing element are converted into an electric charge output by a piezoelectric crystal. Piezoelectric pressure sensors are useful for measuring high-pressure transient events, for example, explosive pressures. In *vibrating-wire* pressure sensors, a metal wire (typically tungsten) is stretched between a fixed anchor and the center of a diaphragm. The wire is located near a permanent magnet and is set into vibration at its resonant frequency by an ac current excitation. A pressure-induced displacement of the diaphragm changes the tension and therefore the resonant frequency of the wire, which is measured by the readout electronics. A detailed description of these and other types of detection methods can be found in [1].

39.2.2.1 Capacitive Pressure Sensors

Many highly accurate (better than 0.1%) pressure sensors in use today have been developed using the capacitive detection approach. Capacitive pressure sensors can be designed to cover an extremely wide pressure range. Both high-pressure sensors with full-scale pressures above 10^7 Pa (a few thousand psi) and vacuum sensors (commonly referred to as capacitive *manometers*) usable for pressure measurements below 10^{-3} Pa (10^{-5} torr) are commercially available. The principle of *capacitive pressure sensors* is illustrated in Figure 39.2. A metal or silicon diaphragm serves as the pressure-sensing element and

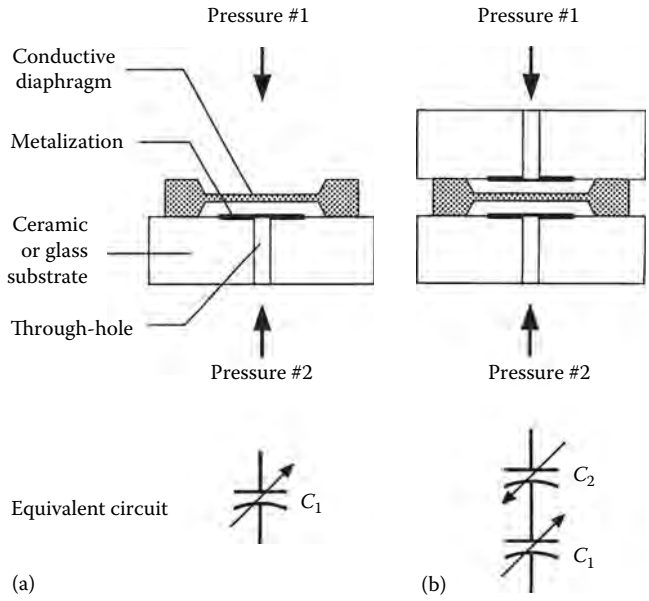


FIGURE 39.2 Operating principle of capacitive pressure sensors, (a) single capacitor design and (b) differential capacitor design.

constitutes one electrode of a capacitor. The other electrode, which is stationary, is typically formed by a deposited metal layer on a ceramic or glass substrate. An applied pressure deflects the diaphragm, which in turn changes the gap spacing and the capacitance [2]. In the differential capacitor design, the sensing diaphragm is located in between two stationary electrodes. An applied pressure will cause one capacitance to increase and the other one to decrease, thus resulting in twice the signal while canceling many undesirable common-mode effects. Figure 39.3 shows a practical design of a differential capacitive sensing cell that uses two isolating diaphragms and an oil fill to transmit the differential pressure to the

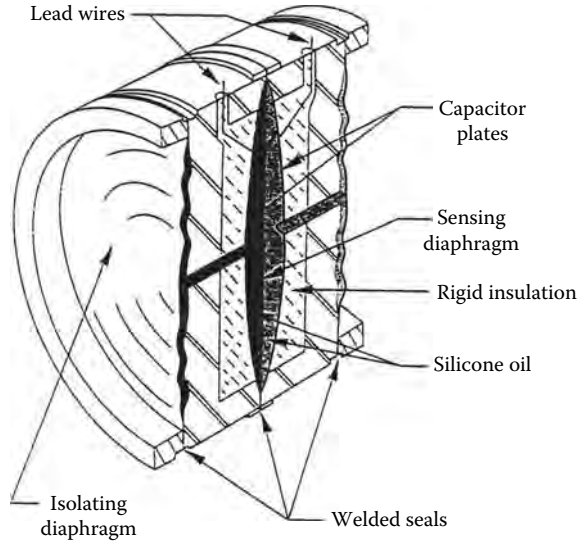


FIGURE 39.3 Differential capacitive sensing cell that is equipped with isolating diaphragms and silicone oil transfer fluid suitable for measuring pressure in corrosive media. (Courtesy of Rosemount, Inc., Eden Prairie, MN.)

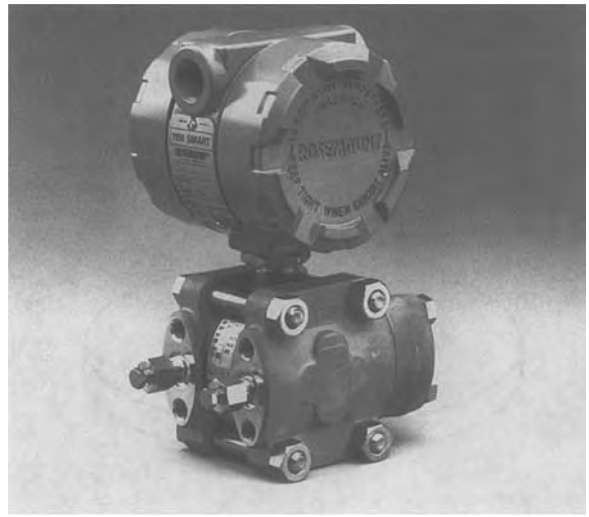


FIGURE 39.4 Rugged capacitive pressure sensor product for industrial applications. It incorporates the sensing cell shown in Figure 39.3. Readout electronics are contained in the housing at the top. (Courtesy of Rosemount, Inc., Eden Prairie, MN.)

sensing diaphragm. The isolating diaphragms are made of special metal alloys that enable them to handle corrosive fluids. The oil is chosen to set a predictable dielectric constant for the capacitor gaps while providing adequate damping to reduce shock and vibration effects. Figure 39.4 shows a rugged capacitive pressure sensor for industrial applications based on the capacitive sensing cell shown in Figure 39.3. The capacitor electrodes are connected to the readout electronics housing at the top. In general, with today's sophisticated electronics and special considerations to minimize stray capacitances (which can degrade the accuracy of measurements), a capacitance change of 10 aF (10^{-18} F) provided by a diaphragm deflection of only a fraction of a nanometer is resolvable.

39.2.2.2 Piezoresistive Pressure Sensors

Piezoresistive sensors (also known as *strain-gage* sensors) are the most common type of pressure sensor in use today. *Piezoresistive effect* refers to a change in the electric resistance of a material when stresses or strains are applied. Piezoresistive materials can be used to realize strain gages that, when incorporated into diaphragms, are well suited for sensing the induced strains as the diaphragm is deflected by an applied pressure. The sensitivity of a strain gage is expressed by its *gage factor*, which is defined as the fractional change in resistance, $\Delta R/R$, per unit strain:

$$\text{Gage factor} = \frac{\Delta R/R}{\varepsilon} \tag{39.2}$$

where strain ε is defined as $\Delta L/L$, or the extension per unit length. It is essential to distinguish between two different cases in which (1) the strain is parallel to the direction of the current flow (along which the resistance change is to be monitored) and (2) the strain is perpendicular to the direction of the current flow. The gage factors associated with these two cases are known as the *longitudinal gage factor* and the *transverse gage factor*, respectively. The two gage factors are generally different in magnitude and often opposite in sign. Typical longitudinal gage factors are ~2 for many useful metals, 10–35 for polycrystalline silicon (polysilicon), and 50–150 for single-crystalline silicon [3–5]. Because of its large piezoresistive effect, silicon has become the most commonly used material for strain gages. There are several ways to incorporate strain gages into pressure-sensing diaphragms. For example, strain gages can be

directly bonded onto a metal diaphragm. However, hysteresis and creep of the bonding agent are potential issues. Alternatively, the strain-gage material can be deposited as a thin film on the diaphragm. The adhesion results from strong molecular forces that will not creep, and no additional bonding agent is required. Today, the majority of piezoresistive pressure sensors are realized by integrating the strain gages into the silicon diaphragm using integrated circuit fabrication technology. This important class of silicon pressure sensors will be discussed in detail in the next section.

39.3 Silicon Micromachined Pressure Sensors

Silicon micromachined pressure sensors refer to a class of pressure sensors that employ integrated circuit batch processing techniques to realize a thinned-out diaphragm sensing element on a silicon chip. Strain gages made of silicon diffused resistors are typically integrated on the diaphragm to convert the pressure-induced diaphragm deflection into an electric resistance change. Over the past 20 years, silicon micromachined pressure sensors have gradually replaced their mechanical counterparts and have captured over 80% of the pressure sensor market. There are several unique advantages that silicon offers. Silicon is an ideal mechanical material that does not display any hysteresis or yield and is elastic up to the fracture limit. It is stronger than steel in yield strength and comparable in Young's modulus [6]. As mentioned in the previous section, the piezoresistive effect in single-crystalline silicon is almost two orders of magnitude larger than that of metal strain gages. Silicon has been widely used in integrated circuit manufacturing for which reliable batch fabrication technology and high-precision dimension control techniques have been well developed. A typical silicon wafer yields hundreds of identical pressure sensor chips at very low cost. Further, the necessary signal-conditioning circuitry can be integrated on the same sensor chip no more than a few millimeters in size [7]. All these are key factors that contributed to the success of silicon micromachined pressure sensors.

Figure 39.5 shows a typical construction of a silicon piezoresistive pressure sensor. An array of square or rectangular diaphragms is "micromachined" out of an (100) oriented single-crystalline silicon wafer by selectively removing material from the back. An anisotropic silicon etchant (e.g., potassium hydroxide) is typically employed; it etches fastest on (100) surfaces and much slower on (111) surfaces. The result is a pit formed on the backside of the wafer bounded by (111) surfaces and a thinned-out diaphragm section on the front at every sensor site. The diaphragm thickness is controlled by a timed etch or by using suitable etch-stop techniques [6,8]. To realize strain gages, a *p*-type dopant, typically boron, is diffused into the front of the *n*-type silicon diaphragm at stress-sensitive

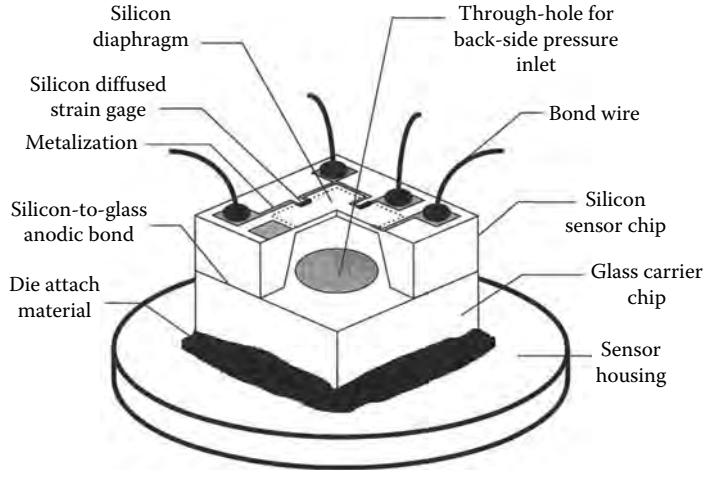


FIGURE 39.5 Cutaway view showing the typical construction of a silicon piezoresistive pressure sensor.



FIGURE 39.6 Examples of commercially available packages for silicon pressure sensors. Shown in the photo are surface-mount units, dual-in-line (DIP) units, TO-8 metal cans, and stainless steel units with isolating diaphragms. (Courtesy of EG&G IC Sensors, Eden Prairie, MN.)

locations to form resistors that are electrically isolated from the diaphragm and from each other by reverse biased *p*–*n* junctions. The strain gages, the diaphragm, and the rest of the supporting sensor chip all belong to the same single-crystalline silicon. The result is a superb mechanical structure that is free from creep, hysteresis, and thermal expansion coefficient mismatches. However, the sensor die must still be mounted to a sensor housing, which typically has mechanical properties different from that of silicon. It is crucial to ensure a high degree of stress isolation between the sensor housing and the sensing diaphragm that may otherwise lead to long-term mechanical drifts and undesirable temperature behavior. A common practice is to bond a glass wafer or a second silicon wafer to the back of the sensor wafer to reinforce the overall composite sensor die. This way, the interface stresses generated by the die mount will also be sufficiently remote from the sensing diaphragm and will not seriously affect its stress characteristics. For gage or differential pressure sensing, holes must be provided through the carrier wafer prior to bonding that are aligned to the etch pits of the sensor wafer leading to the back of the sensing diaphragms. No through holes are necessary for absolute pressure sensing. The wafer-to-wafer bonding is performed in a vacuum to achieve a sealed reference vacuum inside the etch pit [6,9]. Today's silicon pressure sensors are available in a large variety of plastic, ceramic, metal can, and stainless steel packages (some examples are shown in Figure 39.6). Many are suited for printed circuit board mounting. Others have isolating diaphragms and transfer fluids for handling corrosive media. They can be readily designed for a wide range of industrial, medical, automotive, aerospace, and military applications.

39.3.1 Silicon Piezoresistive Pressure Sensor Limitations

Despite the relatively large piezoresistive effects in silicon strain gages, the full-scale resistance change is typically only 1%–2% of the resistance of the strain gage (which yields an unamplified voltage output of 10–20 mV/V). To achieve an overall accuracy of 0.1% of full scale, for example, the combined effects of mechanical and electric repeatability, hysteresis, linearity, and stability must be controlled or compensated to within a few parts per million (ppm) of the gage resistance. Furthermore, silicon strain gages are also very temperature sensitive and require careful compensations. There are two primary sources of temperature drifts: (1) the temperature coefficient of resistance of the strain gages (from 0.06%/°C to 0.24%/°C) and (2) the temperature coefficient of the gage factors (from −0.06%/°C to −0.24%/°C), which will cause

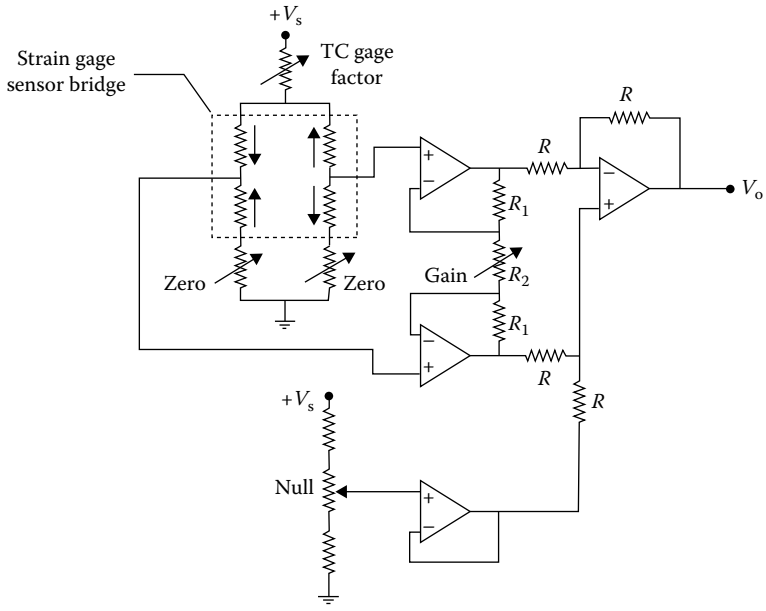


FIGURE 39.7 Signal-conditioning circuit for silicon piezoresistive pressure sensor.

a decrease in pressure sensitivity as the temperature rises. Figure 39.7 shows a circuit configuration that can be used to achieve offset (resulting from gage resistance mismatch) and temperature compensations as well as providing signal amplification to give a high-level output. Four strain gages that are closely matched in both their resistances and temperature coefficients of resistance are employed to form the four active arms of a Wheatstone bridge. Their resistor geometry on the sensing diaphragm is aligned with the principal strain directions so that two strain gages will produce a resistance increase and the other two a resistance decrease on a given diaphragm deflection. These two pairs of strain gages are configured in the Wheatstone bridge such that an applied pressure will produce a bridge resistance imbalance while the temperature coefficient of resistance will only cause a common-mode resistance change in all four gages, keeping the bridge balanced. As for the temperature coefficient of the gage factor, because it is always negative, it is possible (e.g., with the voltage divider circuit in Figure 39.7) to utilize the positive temperature coefficient of the bridge resistance to increase the bridge supply voltage, compensating for the loss in pressure sensitivity as temperature rises. Another major limitation in silicon pressure sensors is the nonlinearity in the pressure response that usually arises from the slight nonlinear behavior in the diaphragm mechanical and the silicon piezoresistive characteristics. The nonlinearity in the pressure response can be compensated by using analog circuit components. However, for the most accurate silicon pressure sensors, digital compensation using a microprocessor with correction coefficients stored in memory is often employed to compensate for all the predictable temperature and nonlinear characteristics. The best silicon pressure sensors today can achieve an accuracy of 0.08% of full scale and a long-term stability of 0.1% of full scale per year. Typical compensated temperature range is from −40 °C to 85 °C, with the errors of compensation on span and offset both around 1% of full scale. Commercial products are currently available for full-scale pressure ranges from 10 kPa to 70 MPa (1.5–10,000 psi). The 1998 prices are U.S. $5–$20 for the most basic uncompensated sensors, $10–$50 for the compensated (with additional laser-trimmed resistors integrated either on a chip or on a ceramic substrate) or signal-conditioned (compensated with amplified output) sensors, and $60–$300 for sensors with isolating diaphragms in stainless steel housings. Table 39.2 provides contact information for selected companies making pressure sensors.

TABLE 39.2 Selected Companies That Make Pressure Sensors and Pressure Calibration Systems

1. Silicon micromachined piezoresistive pressure sensor
   Druck, Inc., 4 Dunham Drive, New Fairfield, CT 06812, Tel: (203) 746-0400, http://www.druck.com
   EG&G IC Sensors, 1701 McCarthy Blvd., Milpitas, CA 95035-7416, Tel: (408) 432-1800
   Foxboro ICT, 199 River Oaks Pkwy., San Jose, CA 95134-1996, Tel: (408) 432-1010, http://www.foxboro-eckardt.com/company/company_en.html
   Honeywell, Inc., Micro Switch Div., 11 W. Spring St., Freeport, IL 61032-4353, Tel: (815) 235-5500, http://www.honeywell.com/sensing
   Lucas NovaSensor, 1055 Mission Ct., Fremont, CA 94539, Tel: (800) 962-7364, http://www.novasensor.com
   Motorola, Inc., Sensor Products Div., 5005 E. McDowell Rd., Phoenix, AZ 85008, Tel: (602) 244-3381, http://mot-sps.com/senseon
   SenSym, Inc., 1804 McCarthy Blvd., Milpitas, CA 95035, Tel: (408) 954-1100, http://www.sensym.com
2. Bonded strain-gage pressure sensors
   Gefran, Inc., 122 Terry Dr., Newtown, PA 18940, Tel: (215) 968-6238, http://www.gefran.it
3. Capacitive pressure sensors
   Kavlico Corp., 14501 Los Angeles Ave., Moorpark, CA 93021, Tel: (805) 523-2000, http://www.kavlico.com/
   Rosemount, Inc., Measurement Div., 12001 Technology Drive, Eden Prairie, MN 55344, Tel: (800) 999-9307, http://www.rosemount.com
4. Pressure calibration systems
   Mensor Corp., 2230 IH-35 South, San Marcos, TX 78666-5917, Tel: (512) 396-4200, http://www.mensor.com
   Ruska Instruments, 10311 Westpark Drive, Houston, TX 77042, Tel: (713) 975-0547, http://www.ruska.com

Note: This is not intended to be an exhaustive list of all manufacturers.

References

1. H. N. Norton, *Handbook of Transducers*, Englewood Cliffs, NJ: Prentice Hall, 1989, pp. 294–330.
2. W. H. Ko, Solid-state capacitive pressure transducers, *Sens. Actuat*., 10, 303–320, 1986.
3. C. S. Smith, Piezoresistance effect in germanium and silicon, *Phys. Rev*., 94, 42–49, 1954.
4. O. N. Tufte and E. L. Stelzer, Piezoresistive properties of silicon diffused layers, *J. Appl. Phys*., 34, 313–318, 1963.
5. D. Schubert, W. Jenschke, T. Uhlig, and F. M. Schmidt, Piezoresistive properties of polycrystalline and crystalline silicon films, *Sens. Actuat*., 11, 145–155, 1987.
6. K. E. Petersen, Silicon as a mechanical material, *IEEE Proc*., 70, 420–457, 1982.
7. R. F. Wolffenbuttel (ed.), *Silicon Sensors and Circuits: On-Chip Compatibility*, London, U.K.: Chapman & Hall, 1996, pp. 171–210.
8. H. Seidel, The mechanism of anisotropic silicon etching and its relevance for micromachining, Tech. Dig., *Transducers '87*, Tokyo, Japan, June 1987, pp. 120–125.
9. E. P. Shankland, Piezoresistive silicon pressure sensors, *Sensors*, 22, 22–26, 1991.

Further Information

ISA Directory of Instrumentation On-Line (http://www.isa.org) from the Instrument Society of America maintains a list of product categories and active links to many sensor manufacturers.

Muller, R. S., Howe, R. T., Senturia, S. D., Smith, R. L., and White, R. M. (eds.), *Microsensors*, New York: IEEE Press, 1991, provides an excellent collection of papers on silicon microsensors and silicon micromachining technologies.

Pressure Measurement Glossary, Sensors One, http://www.sensorsone.co.uk/pressure-measurement-glossary.html (accessed on June 25, 2012).

Pressure Measurements, National Instruments, http://www.ni.com/pressure/ (accessed on June 2, 2012).

Van Nostrand, J., Pressure measurement in industrial applications, Intech, http://www.isa.org/InTechTemplate.cfm?Section=Automation_Basics&template=/ContentManagement/ContentDisplay.cfm&ContentID=82299 (accessed on June 23, 2012).

Wolffenbuttel, R. F. (ed.), *Silicon Sensors and Circuits: On-Chip Compatibility*, London, U.K.: Chapman & Hall, 1996, provides a thorough discussion on sensor and circuit integration.

40

Vacuum Measurement

Ron Goehner
The Fredericks Company

Emil Drubetsky
The Fredericks Company

Howard M. Brady
The Fredericks Company

William H. Bayles, Jr.
The Fredericks Company

40.1 Background and History of Vacuum Gages **40**-1
40.2 Direct Reading Gages **40**-1
Liquid Wall Gages • Solid Wall Gages
40.3 Indirect Reading Gages **40**-6
Thermal Conductivity Gages
40.4 Cathode Ionization Gages **40**-11
Hot Cathode Ionization Gages • Cold Cathode Ionization Gages
40.5 Resonance Gages **40**-15
40.6 Molecular Drag (Spinning Rotor) Gages **40**-16
40.7 Partial Pressure Measurements and Mass Spectrometers **40**-18
References **40**-19
Further Readings **40**-21
Partial List of Vendors and Suppliers **40**-21

40.1 Background and History of Vacuum Gages

To make measurements in the vacuum region, one must possess knowledge of the expected pressure range required by the processes taking place in the vacuum chamber as well as the accuracy and/or repeatability of the measurement required for the process. Typical vacuum systems require that many orders of magnitude of pressures must be measured. In many applications, the pressure range may be eight orders of magnitude or from atmospheric 1.01×10^5 Pa (760 torr) to 1×10^{-3} Pa (7.5×10^{-6} torr).

For semiconductor lithography, high-energy physics experiments, and surface chemistry, ultimate vacuum of 7.5×10^{-9} torr and much lower is required (a range of 11 orders of magnitude below atmospheric pressure). One gage will not give reasonable measurements over such large pressure ranges. Over the past 50 years, vacuum measuring instruments (commonly called gages) have been developed that used transducers (or sensors) that can be classified as either direct reading (usually mechanical) or indirect reading [1] (usually electronic). Figure 40.1 shows vacuum gages typically in current use. When a force on a surface is used to measure pressure, the gages are mechanical and are called *direct reading* gages, whereas when any property of the gas that changes with density is measured by electronic means, they are called *indirect reading* gages. Figure 40.2 shows the range of operating pressure for various types of vacuum gages.

40.2 Direct Reading Gages

A subdivision of direct reading gages can be made by dividing them into those that utilize a liquid wall and those that utilize a solid wall. The force exerted on a surface from the pressure of thermally agitated molecules and atoms is used to measure the pressure.

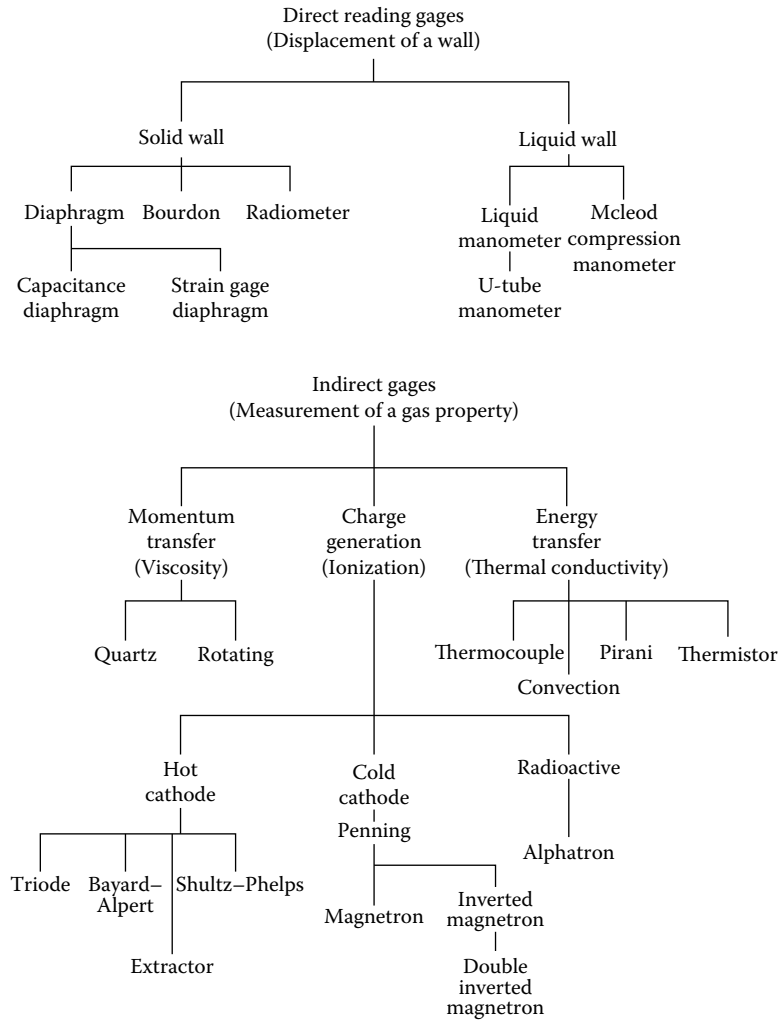


FIGURE 40.1 Classification of pressure gages. (From Hoffman, D.M. et al., *The Handbook of Vacuum Science and Technology*, Academic Press, Orlando, FL, 1998. With permission.)

40.2.1 Liquid Wall Gages

The two common gages that use a liquid wall are the manometer and the McLeod gage. The liquid column *manometer* is the simplest type of vacuum gage. It consists of a straight or U-shaped glass tube evacuated and sealed at one end and filled partly with mercury or a low vapor pressure liquid such as diffusion pump oil (see Figure 40.3). In the straight tube manometer, as the space above the mercury is evacuated, the length of the mercury column decreases. In the case of the U-tube, as the free end is evacuated, the two columns approach equal height. The pressure at the open end is measured by the difference in height of the liquid columns. If the liquid is mercury, the pressure is directly measured in mm of Hg (torr). The manometer is limited to pressures equal to or greater than ~1 torr (133 Pa). If the liquid is a low-density oil, the U-tube is capable of measuring a pressure as low as ~0.1 torr. This is an absolute, direct reading gage, but the use of mercury or low-density oils that will in time contaminate the vacuum system precludes its use as a permanent vacuum gage.

Due to the pressure limitation of the manometer, the McLeod gage [2,3] was developed to significantly extend the range of vacuum measurement (see Figure 40.4). This device is essentially a mercury

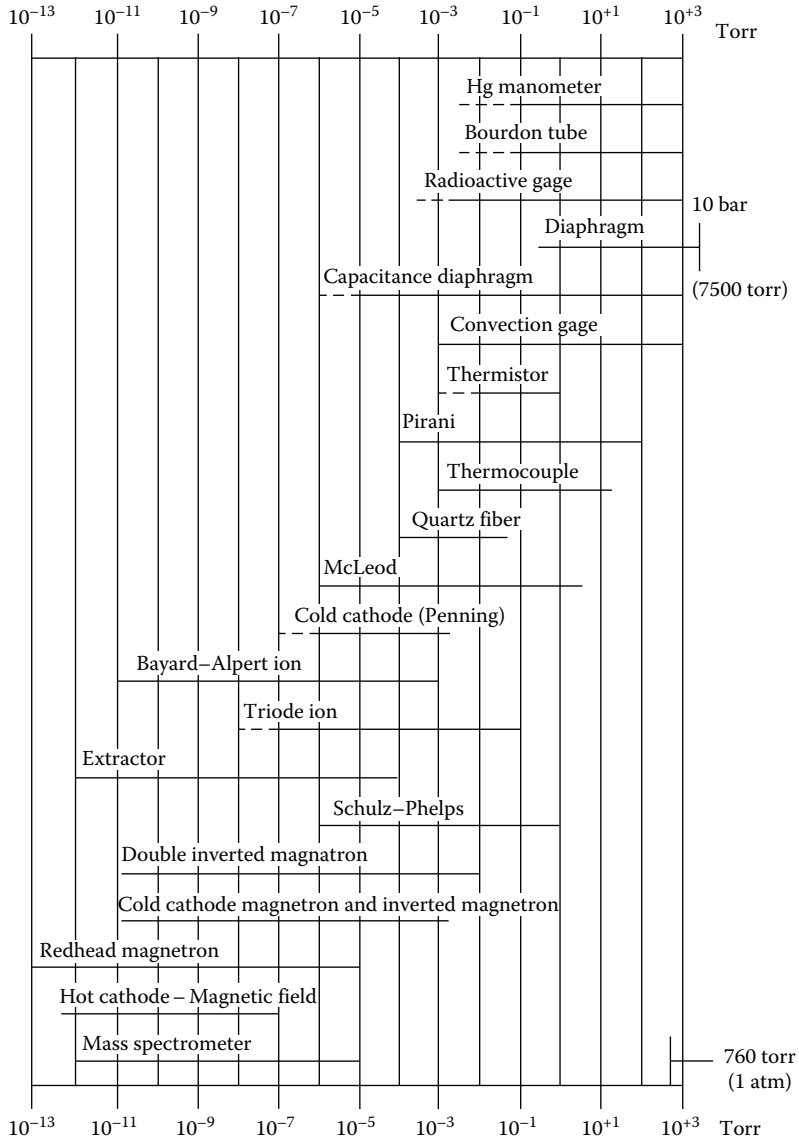


FIGURE 40.2 Pressure ranges for various gages. (From Hoffman, D.M. et al., *The Handbook of Vacuum Science and Technology*, Academic Press, Orlando, FL, 1998. With permission.)

manometer in which a volume of gas is compressed before measurement. This can be used as a primary standard device when a liquid nitrogen trap is used on the vacuum system. Figure 40.4 shows gas at 10^{-6} torr and a compression ratio of 10^{7}. In this example, the difference of the columns will be 10 mm. Extreme care must be taken not to break the glass and expose the surroundings to the mercury. The McLeod gage is an inexpensive standard but should only be used by skilled and careful technicians. The gage will give a false low reading unless precautions are taken to ensure that any condensable vapors present are removed by liquid nitrogen trapping.

40.2.2 Solid Wall Gages

There are two major mechanical solid wall gage types: capsule and diaphragm.

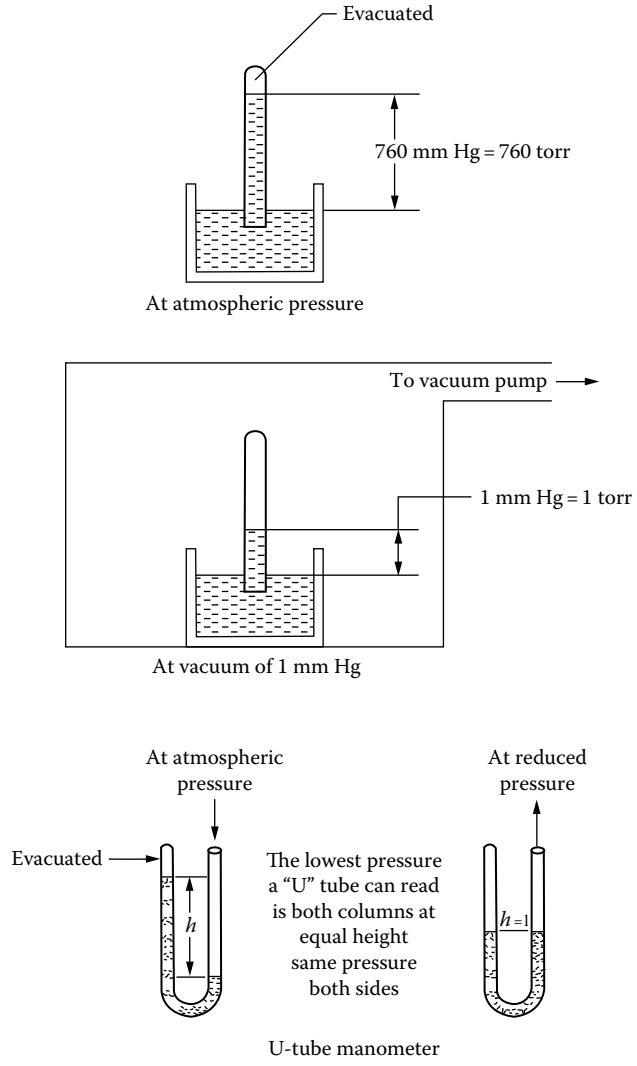


FIGURE 40.3 Mercury manometers. (From Bayles, W.H. Jr., *Ind. Heat.*, 1992. With permission.)

40.2.2.1 Bourdon Gages

The capsule-type gages depend on the deformation of the capsule with changing pressure and the resultant deflection of an indicator. Pressure gages using this principle measure pressures above atmospheric to several thousand psi and are commonly used on compressed gas systems. This type of gage is also used at pressures below atmospheric, but the sensitivity is low. The Bourdon gage (Figure 40.5) is used as a moderate vacuum gage. In this case, the capsule is in the form of a thin-walled tube bent in a circle, with the open end attached to the vacuum system with a mechanism and a pointer attached to the other end. The atmospheric pressure deforms the tube; a linear indication of the pressure is given that is independent of the nature of the gas. Certain manufacturers supply capsule gages capable of measuring pressures as low as 1 torr. These gages are rugged, inexpensive, and simple to use and can be made of materials inert to corrosive vapors. Since changing atmospheric pressure causes inaccuracies in the readings, compensated versions of the capsule and Bourdon gage have been developed that improve the accuracy [4].

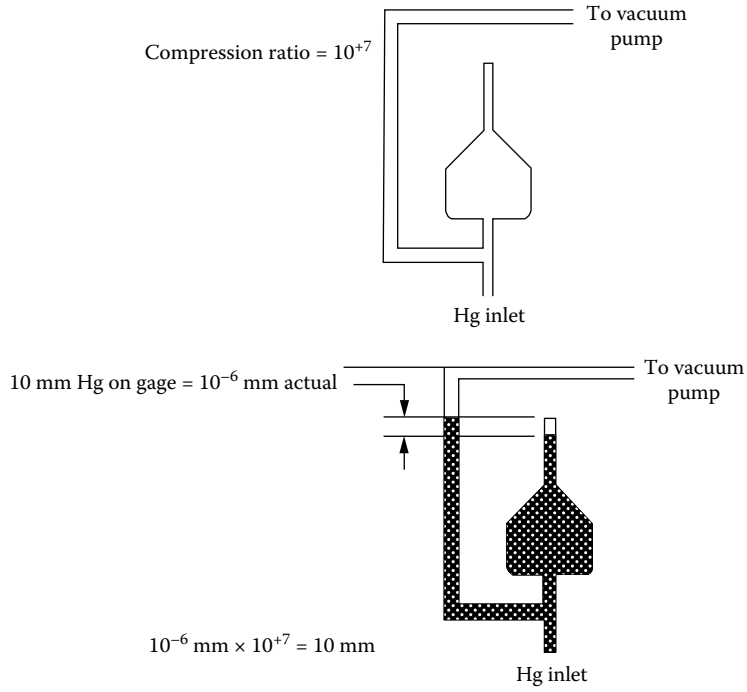


FIGURE 40.4 McLeod gage. (From Hoffman, D.M. et al., *The Handbook of Vacuum Science and Technology*, Academic Press, Orlando, FL, 1998. With permission.)

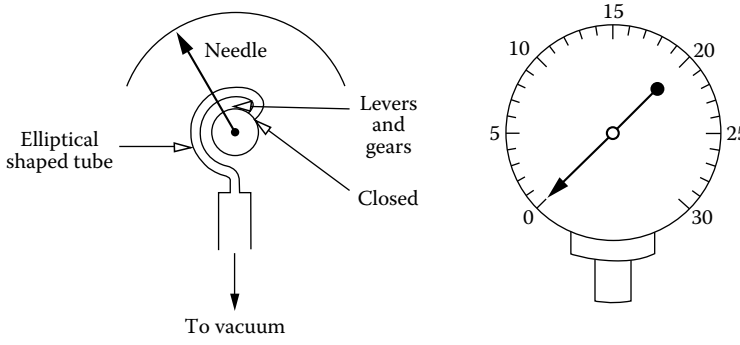


FIGURE 40.5 Bourdon gage. (From Varian Associates, *Basic Vacuum Practice*, Varian Associates, Inc., Lexington, MA, 02173, 1992. With permission.)

40.2.2.2 Diaphragm Gages

If compensated capsule or diaphragm mechanisms are combined with sensitive and stable electronic measuring circuits, performance is improved. One such gage is the capacitance diaphragm gage (also referred to as the capacitance manometer).

The capacitance diaphragm gage is shown in Figure 40.6. A flexible diaphragm forms one plate of a capacitor and a fixed probe the other. The flexible diaphragm deforms due to even slight changes in pressure, resulting in a change in the capacitance. The capacitance is converted to a pressure reading. The sensitivity, repeatability, and simplicity of this gage enables this type of direct reading gage to be a standard from 10^{-6} torr to atmospheric pressure, provided multiple heads designed for each pressure range are used. A single head can have a dynamic range of 4 or 5 orders of magnitude [5].

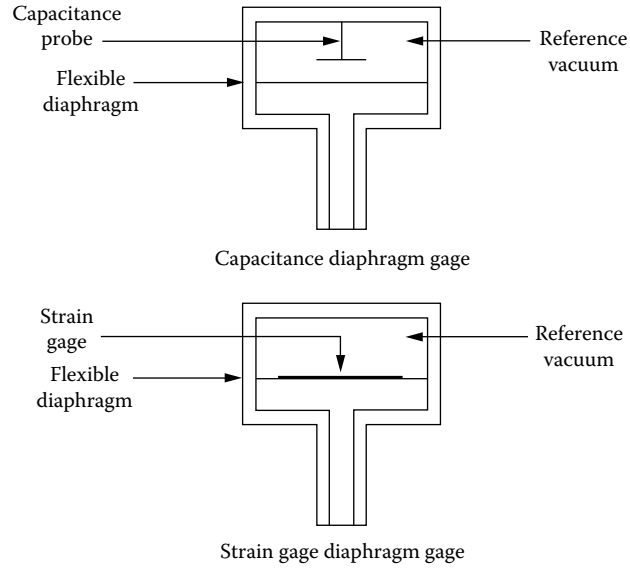


FIGURE 40.6 Diaphragm gage. (From Bayles, W.H. Jr., *Ind. Heat*., 1992. With permission.)

The strain-gage type of diaphragm gage is shown in Figure 40.6. In this case, the deformation of the diaphragm causes a proportional output from the attached strain gage. Sensitivities and dynamic range tend to be less than those of the capacitance diaphragm gage, but the price of the strain-gage-type diaphragm gage is usually lower.

Both of these gages are prone to errors caused by small temperature changes due to the inherent high sensitivity of this gage type. Temperature-controlled heads or correction tables built into the electronics have been used to minimize this problem. Other sources of error in all solid wall gages are hysteresis and metal fatigue.

40.3 Indirect Reading Gages

Indirect reading gages measure some property of the gas that changes with the density of the gas and usually produces an electric output. Electronic devices amplify and compensate this output to provide a pressure reading.

40.3.1 Thermal Conductivity Gages

Thermal conductivity gages utilize the property of gases in which reduced thermal conductivity corresponds to decreasing density (pressure). The thermal conductivity decreases from a nearly constant value above ~1 torr to essentially 0 at pressures below 10^{-2} torr. The gage controllers are designed to work with a specific sensor tube, and substitutions are limited to those that are truly functionally identical. Heat transfer at various pressures is related to the Knudsen number, as is shown in Figure 40.7 for various heat transfer regimes. The Knudsen number can then be related to pressure through the geometry of the sensor, providing a relationship of heat transfer to pressure for a particular design thermal conductivity gage.

40.3.1.1 Pirani Gages

The Pirani gage is perhaps the oldest indirect gage that is still used today. In operation, a sensing filament carrying current and producing heat is surrounded by the gas to be measured. As the pressure changes, the thermal conductivity changes with it too, thus altering the temperature of the sensing filament.

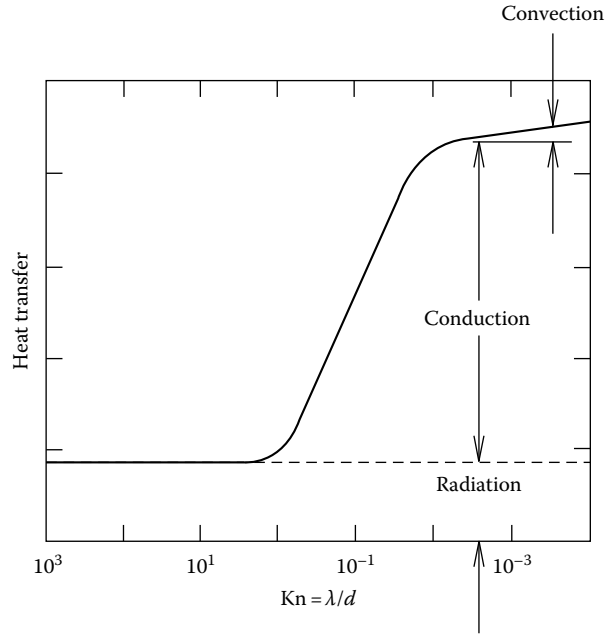


FIGURE 40.7 Heat transfer regimes in a thermal conductivity gage. (From O'Hanlon, J.F., *A User's Guide to Vacuum Technology*, John Wiley & Sons, Hoboken, NJ, 1980.)

The temperature change causes a change in the resistance of the sensing filament. The sensing filament is usually one leg of a Wheatstone bridge. The bridge can be operated so that the voltage is varied to keep the bridge balanced; that is, the resistance of the sensing filament is kept constant.

This method is called the constant-temperature method and is deemed the fastest, most sensitive, and most accurate. To reduce the effect of changing ambient temperature, an identical filament sealed off at very low pressure is placed in the leg adjacent to the sensing filament as a balancing resistor. Because of its high thermal resistance coefficient, the filament material is usually a thin tungsten wire. It has been demonstrated that a 10 W light bulb works quite well [6] (Figure 40.8).

A properly designed, compensated Pirani gage with sensitive circuitry is capable of measuring upto 10^{-4} torr. However, the thermal conductivity of gases varies with the gas being measured, causing a

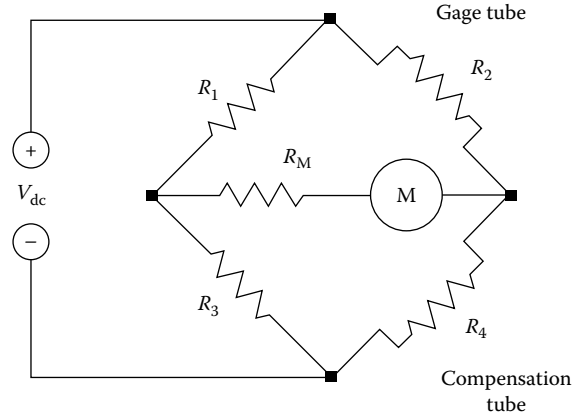


FIGURE 40.8 Pirani gage. (From O'Hanlon, J.F., *A User's Guide to Vacuum Technology*, John Wiley & Sons, Hoboken, NJ, 1980.)

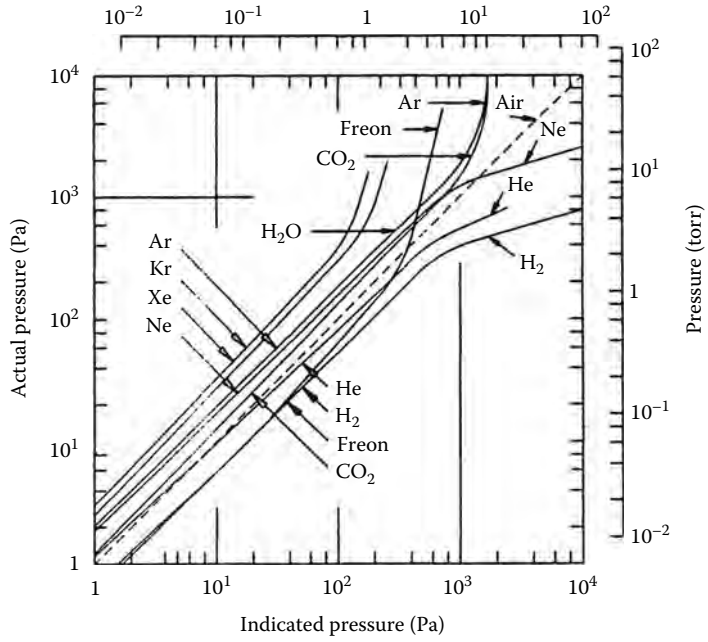


FIGURE 40.9 Calibration curves for the Pirani gage. (Reprinted with permission from Leybold-Herqeus GMblt, Köhn, Germany.)

variation in gage response. These variations can be as large as a factor of 5 at low pressures and as high as 10 at high pressures (see Figure 40.9). The correction for these variations can be made on the calibration curves supplied by the manufacturer if the composition of the gas is known. The operation in the presence of high partial pressures of organic molecules such as oils is not recommended.

40.3.1.2 Thermistor Gages

In the thermistor gage, a thermistor is used as one leg of a bridge circuit. The inverse resistive characteristics of the thermistor element unbalance the bridge as the pressure changes, causing a corresponding change in current. Sensitive electronics measure the current and are calibrated in pressure units. The thermistor gage measures approximately the same pressure range as the thermocouple. The exact calibration depends on the gas measured. In a well-designed bridge circuit, the plot of current vs. pressure is practically linear in the range 10^{-3} to 1 torr [7]. Modern thermistor gages use constant-temperature techniques.

40.3.1.3 Thermocouple Gages

Another example of an indirect reading thermal conductivity gage is the thermocouple gage. This is a relatively inexpensive device with proven reliability and a wide range of applications. In the thermocouple gage, a filament of resistance alloy is heated by the passage of a constant current (see Figure 40.10). A thermocouple is welded to the midpoint of the filament or preferably to a conduction bridge at the center of the heated filament. This provides a means of directly measuring the temperature. With a constant current through the filament, the temperature increases as the pressure decreases as there are fewer molecules surrounding the filament to carry the heat away. The thermocouple output voltage increases as a result of the increased temperature and varies inversely with the pressure. The thermocouple gage can also be operated in the constant-temperature mode.

Gas composition effects apply to all thermal conductivity gages. The calibration curves for a typical thermocouple gage are shown in Figure 40.11. The thermocouple gage can be optimized for operation

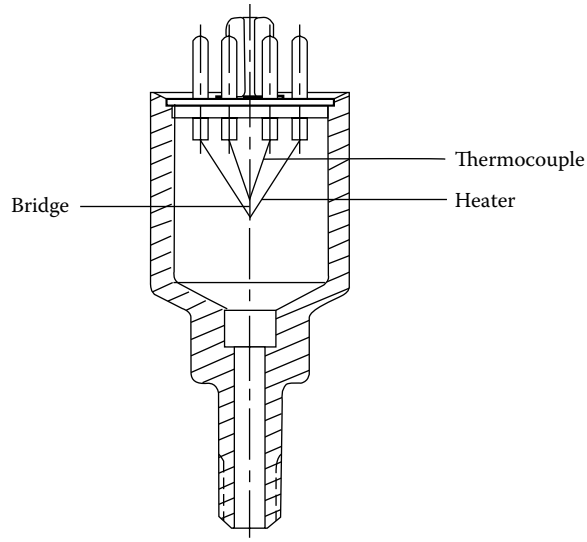


FIGURE 40.10 Thermocouple gage. (Reprinted with permission from Engleman and Televac Division, The Fredericks Co., Huntingdon Valley, PA.)

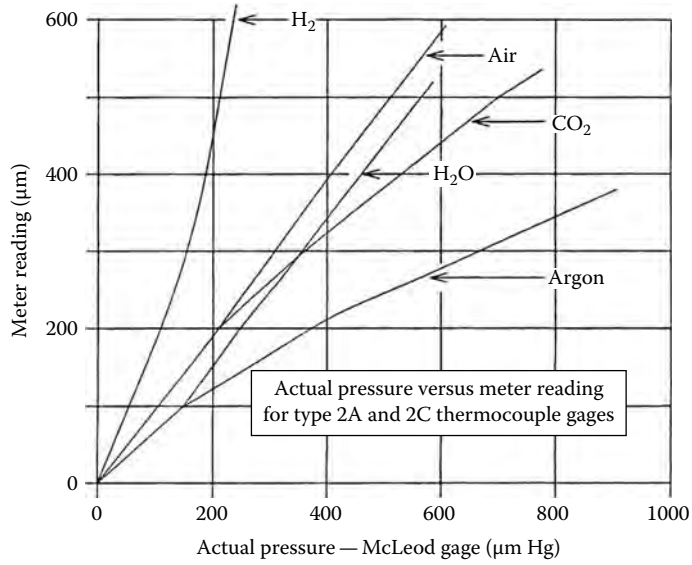


FIGURE 40.11 Calibration curves for the thermocouple gage. (Reprinted with permission from Engleman and Televac Division, The Fredericks Co., Huntingdon Valley, PA.)

in various pressure ranges. The operation of the thermocouple gage in high partial pressures of organic molecules such as oils should be avoided. One manufacturer pre-oxidizes the thermocouple sensor for stability in "dirty" environments and for greater interchangeability in clean environments.

40.3.1.4 Convection Gages

Below 1 torr, a significant change in thermal conductivity occurs as the pressure changes. Thus, the thermal conductivity gage is normally limited to 1 torr.

At pressures above 1 torr, there is, in most gages, a small contribution to heat transfer caused by convection. Manufacturers have developed gages that utilize this convection effect to extend the usable range

to atmospheric pressure and slightly above [8–12]. The orientation of a convection gage is critical because this convection heat transfer is highly dependent on the orientation of the elements within the gage.

The Convectron™ uses the basic structure of the Pirani with special features to enhance convection cooling in the high-pressure region [13]. To utilize the gage above 1 torr (133 Pa), the sensor tube must be mounted with its major axis in a horizontal position. If the only area of interest is below 1 torr, the tube can be mounted in any position. As mentioned earlier, the gage controller is designed to be used with a specific model sensor tube; because extensive use is made of calibration curves and lookup tables stored in the controller, no substitution is recommended.

The Televac convection gage uses the basic structure of the thermocouple gage except that two thermocouples are used [14]. As in any thermocouple gage, the convection gage measures the pressure by determining the heat loss from a fine wire maintained at constant temperature. The response of the sensor depends on the gas type. A pair of thermocouples is mounted a fixed distance from each other (see Figure 40.12). The one mounted lower is heated to a constant temperature by

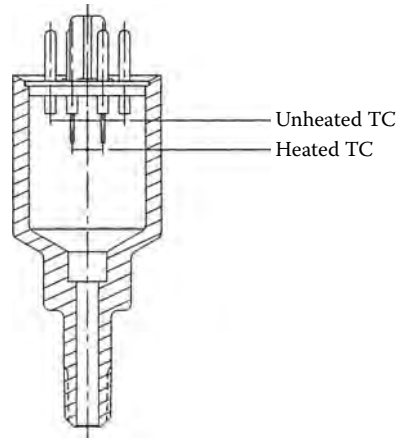


FIGURE 40.12 Convection gage. (Reprinted with permission from Engleman and Televac Division, The Fredericks Co., Huntingdon Valley, PA.)

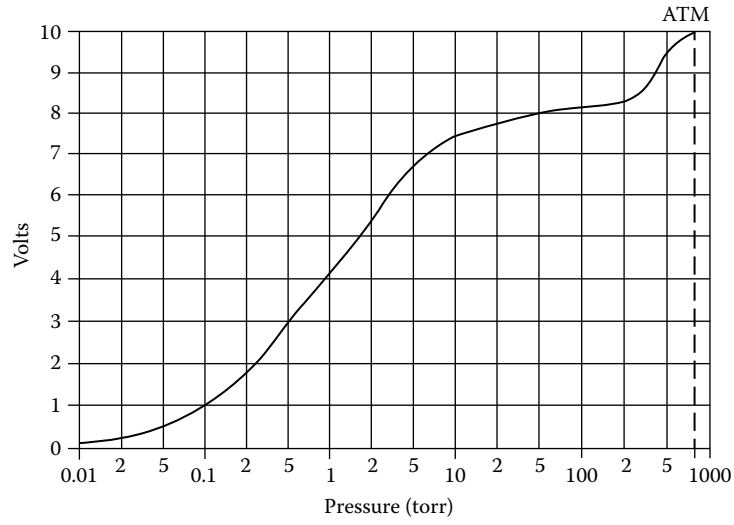


FIGURE 40.13 Output curve for the convection gage. (Reprinted with permission from Engleman and Televac Division, The Fredericks Co., Huntingdon Valley, PA.)

a variable current power supply. Power is pulsed to this lower thermocouple and the temperature is measured between heating pulses. The second (upper) thermocouple measures convection effects and also compensates for ambient temperature. At pressures below ~2 torr (270 Pa), the temperature in the upper thermocouple is negligible. The gage tube operates as a typical thermocouple in the constant-temperature mode. Above 2 torr, convective heat transfer causes heating of the upper thermocouple. The voltage output is subtracted from that of the lower thermocouple, thus requiring more current to maintain the wire temperature. Consequently, the range of pressure that can be measured (via current change) is extended to atmospheric pressure (see Figure 40.13). The orientation of the sensor is with the axis vertical.

The use of convection gages with process control electronics allows for automatic pump down with the assurance that the system will neither open under vacuum nor be subject to overpressure during backfill to atmospheric pressure. These gages, with their controllers, are relatively inexpensive. In oil-free systems, they afford long life and reproducible results.

40.4 Cathode Ionization Gages

40.4.1 Hot Cathode Ionization Gages

Hot cathode ionization gage designs consist of triode gages, Bayard–Alpert gages, and others.

40.4.1.1 Triode Hot Cathode Ionization Gages

For over 80 years, the triode electron tube has been used as an indirect way to measure vacuum [15,16]. A typical triode connection is as an amplifier, as shown in Figure 40.14. A brief description of its operation is given here, but more rigorous treatment of triode performance is given in [17–20]. However, if the triode is connected as in Figure 40.15 so that the grid is positive and the plate is negative with respect to the filament, then the ion current collected by the plate for the same electron current to the grid is greatly increased [21].

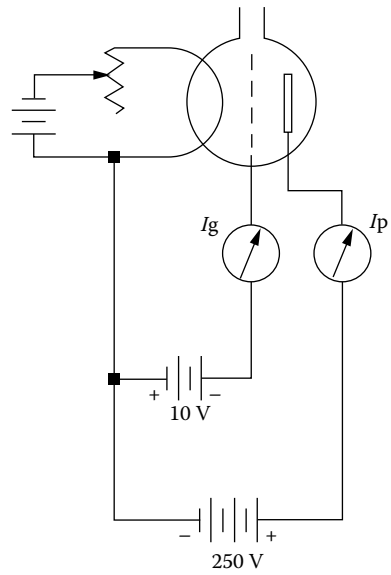


FIGURE 40.14 Typical triode connection.

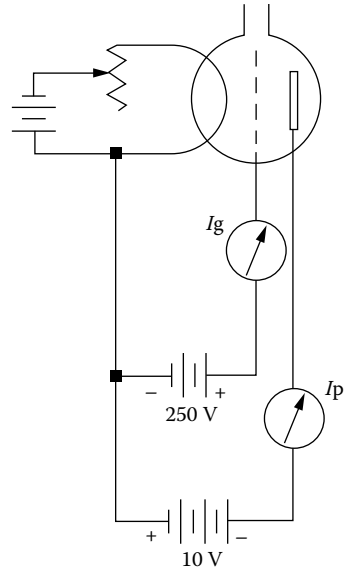


FIGURE 40.15 Alternative triode connection.

Today, the triode gage is used in this higher sensitivity mode. Many investigators have shown that a linear change in molecular density (pressure) results in a linear change in ion current [15,21,22]. This linearity allows a sensitivity factor S to be defined such that

$$I_i = S \times I_e \times P \tag{40.1}$$

where
I_i is the ion current (A)
I_e is the electron current (A)
P is the pressure
S is the sensitivity (in units of reciprocal pressure)

Additional details are found in [23–25]. In nearly all cases, except at relatively high pressures, the triode gage has been replaced by the Bayard–Alpert gage.

40.4.1.2 Bayard–Alpert Hot Cathode Ionization Gages

It became apparent that the pressure barrier observed at 10^{-8} torr was caused by a failure in measurement rather than pumping [26,27]. A solution to this problem was proposed by Bayard and Alpert [28] that is now the most widely used gage for general UHV measurement.

The Bayard–Alpert gage is similar to a triode gage but has been redesigned so that only a small quantity of the internally generated x-rays strike the collector. The primary features of the Bayard–Alpert gage and its associated circuit are shown in Figures 40.16 and 40.17. The cathode has been replaced by a thin collector located at the center of the grid, and the cathode filament is now outside and several millimeters away from the grid. The Bayard–Alpert design utilizes the same controller as the triode gage, with corrections for sensitivity differences between the gage designs. When a hot filament gage is exposed to high pressures, burnout of the tungsten filaments often occurs. To prevent this, platinum metals were coated with refractory oxides to allow the gage to withstand sudden exposure to atmosphere with the filament hot [29,30]. Typical materials include either thoria or yttria coatings on iridium. Bayard–Alpert and triode gages of identical structure and dimensions but with different filaments

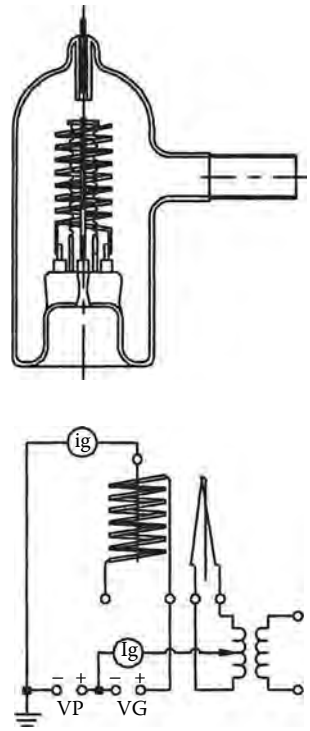


FIGURE 40.16 Bayard–Alpert hot cathode ionization gage.

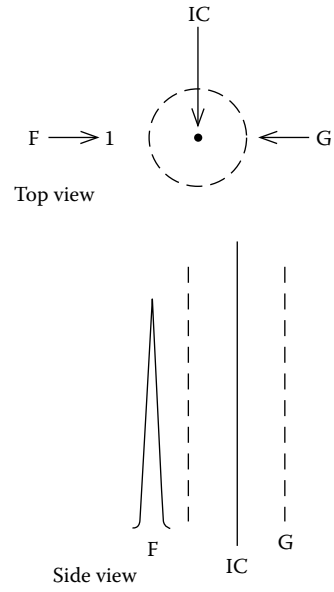


FIGURE 40.17 Bayard–Alpert gage configuration. F, filament; G, grid; IC, ion collector.

(i.e., tungsten vs. thoria iridium) were observed to have different sensitivities, with the tungsten filament versions being 20%–40% more sensitive than the iridium of the same construction.

The lowest pressure that can be measured is limited by low-energy x-rays striking the ion collector and emitting electrons. Several methods to reduce this x-ray limit were developed. Gage designs with

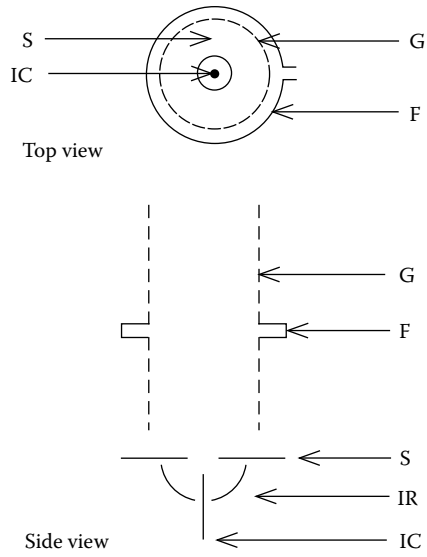


FIGURE 40.18 Extractor ionization gage. F, filament; G, grid; S, shield; IR, iron reflector; and IC, ion collector.

very small diameter collectors have been made that extend the high vacuum range down to 10^{-12} torr, but accuracy was lost at the high pressures [31,32].

The modulated gage was designed by Redhead [33] with an extra electrode near the ion collector. In this configuration, the x-ray current could be subtracted by measuring the ion current at two modulator potentials, thus increasing the range to 5×10^{-12}. Other gages use suppressor electrodes in front of the ion collector [34,35].

The extractor gage (Figure 40.18) is the most widely used UHV hot cathode gage for those who need to measure 10^{-2} torr [36]. In this gage, the ions are extracted out of the ionizing volume and deflected or focused onto a small collector. More recent designs have been developed [37,38]. The use of a channel electron multiplier [39] has reduced the low pressure limit to 10^{-15} torr.

The Bayard–Alpert gage suffers from some problems, however. The ion current is geometry dependent. Investigators have reported on the sensitivity variations, inaccuracy, and instability of Bayard–Alpert gages with widely differing results [40–49]. Investigators have developed ways to reduce or eliminate some of these problems [50,51].

40.4.2 Cold Cathode Ionization Gages

To measure pressures below 10^{-3} torr, Penning [52] developed the cold cathode discharge gage. Below 10^{-3} torr, the mean free path is so high that little ionization takes place. The probability of ionization was increased by placing a magnetic field parallel to the paths of ions and electrons to force these particles into helical trajectory.

This gage consists of two parallel cathodes and an anode, which is placed midway between them (see Figure 40.19). The anode is a circular or rectangular loop of metal wire whose plane is parallel to that of the cathodes. A few kilovolts potential difference is maintained between the anode and the cathodes. Furthermore, a magnetic field is applied between the cathodes by a permanent magnet usually external to the gage body. Electrons emitted from either of the two cathodes must travel in helical paths due to the magnetic field eventually reaching the anode, which carries a high positive charge. During the travel along this long path, many electrons collide with the molecules of the gas, thus creating positive ions

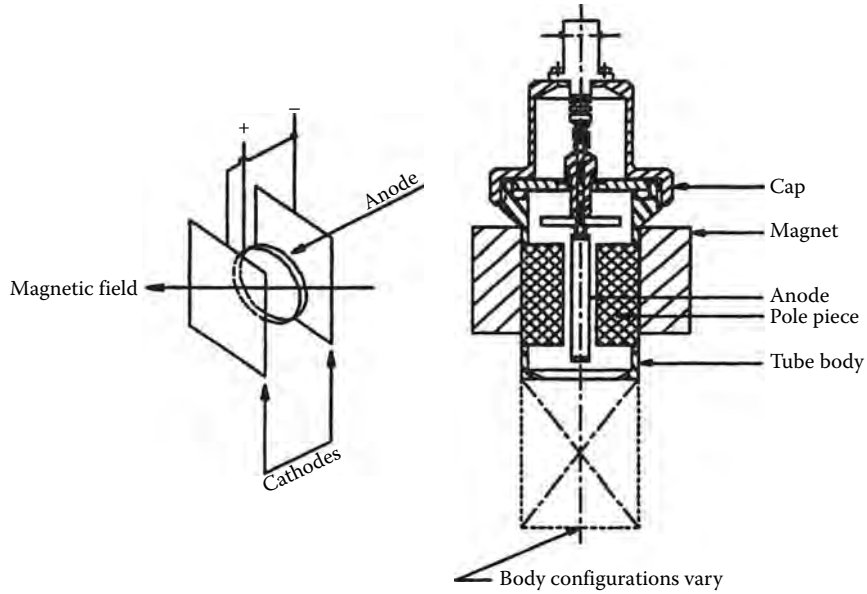


FIGURE 40.19 Penning gages. (a) Basic arrangement. (From O'Hanlon, J.F., *A User's Guide to Vacuum Technology*, John Wiley & Sons, Hoboken, NJ, 1980.) (b) Complex assembly. (Reprinted with permission from Engleman and Televac Division, The Fredericks Co., Huntingdon Valley, PA.)

that travel more directly to the cathodes. The ionization current thus produced is read out on a sensitive current meter as pressure.

This is a rugged gage used for industrial applications such as in leak detectors, vacuum furnaces, electron beam welders, and other industrial processes. The Penning gage is rugged, simple, and inexpensive. Its range is typically 10^{-3}–10^{-6} torr, and some instability and lack of accuracy has been observed [53]. The magnetron design [54] and the inverted magnetron design [55] extended the low pressure range to 10^{-12} torr [56] or better. These improvements produced a better gage, but instability, hysteresis, and starting problems remain [57,58]. Magnetrons currently in use are simpler and do not use an auxiliary cathode.

More recently, a double inverted magnetron was introduced [59]. This gage has greater sensitivity (amp/torr) than the other types (see Figure 40.20). It has been operated successfully at $\sim 1 \times 10^{-11}$ torr. The gage consists of two axially magnetized, annular-shaped magnets (1) placed around a cylinder (2) so that the north pole of one magnet faces the north pole of the other one. A nonmagnetic spacer (3) is placed between the two magnets, and thin shims (4) are used to focus the magnetic fields. This gage has been operated to date at 10^{-11} torr and stays ignited and reignites quickly when power is restored at this pressure. Essentially instantaneous reignition has been demonstrated by the use of radioactive triggering [60].

40.5 Resonance Gages

One example of a resonance-type vacuum gage is the quartz friction vacuum gage [61]. A quartz oscillator can be built to measure pressure by a shift in resonance frequency caused by static pressure of the surrounding gas or by the increased power required to maintain constant amplitude. Its range is from near atmospheric pressure to about 0.1 torr. A second method is to measure the resonant electric impedance of a tuning fork oscillator. Test results for this device show accuracy within ±10% for pressure from 10^{-3} to 10^{3} torr. There is little commercial use to date for these devices.

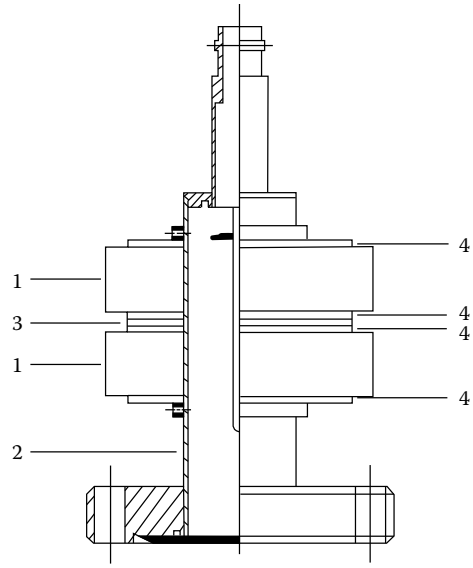


FIGURE 40.20 Double inverted magnetron. (Reprinted with permission from Engleman and Televac Division, The Fredericks Co., Huntingdon Valley, PA.)

40.6 Molecular Drag (Spinning Rotor) Gages

Meyer [62] and Maxwell [63] introduced the idea of measuring pressure by means of the molecular drag of rotating devices in 1875. The rotors of these devices were tethered to a wire or thin filament. The gage was further enhanced by Holmes [64], who introduced the concept of the magnetic rotor suspension, leading to the spinning rotor gage. Nearly 10 years later, Beams et al. [65] disclosed the use of a magnetically levitated, rotating steel ball to measure pressure at high vacuum. Fremerey [66] reported on the historical development of this gage.

The molecular drag gage (MDG), often referred to as the spinning rotor gage, received wider acceptance after its commercial introduction in 1982 [67]. It is claimed to be more stable than other gages at lower pressures [68].

The principle of operation of the modern MDG is based on the fact that the rate of change of the angular velocity of a freely spinning ball is proportional to the gas pressure and inversely proportional to the mean molecular velocity. When the driving force is removed, the angular velocity is determined by measuring the ac voltage induced in the pickup coils by the magnetic moment of the ball (see Figure 40.21).

In current practice, a small rotor (steel ball bearing) about 4.5 mm in diameter is magnetically levitated and spun up to about 400 Hz by induction. The ball, enclosed in a thimble connected to the vacuum system, is allowed to coast by turning off the inductive drive. Then, the time of a revolution of the ball is measured by timing the signal induced in a set of pickup coils by the rotational component of the ball's magnetic moment. Gas molecules will exert a drag on the ball, slowing it at a rate set by the pressure P, its molecular mass m, the temperature T, and the coefficient of momentum transfer σ, between the gas and the ball. A perfectly smooth ball would have a value of unity. There is also a pressure-independent residual drag (RD) caused by eddy current losses in the ball and surrounding structure. There will also be temperature effects that will cause the ball diameter and moment of inertia to change.

The pressure in the region of molecular flow is given by

$$P = \frac{\pi \rho a \bar{c}}{10\sigma_{\text{eff}}}\left(\frac{-\omega' - \text{RD} - 2\alpha T'}{\omega}\right) \tag{40.2}$$

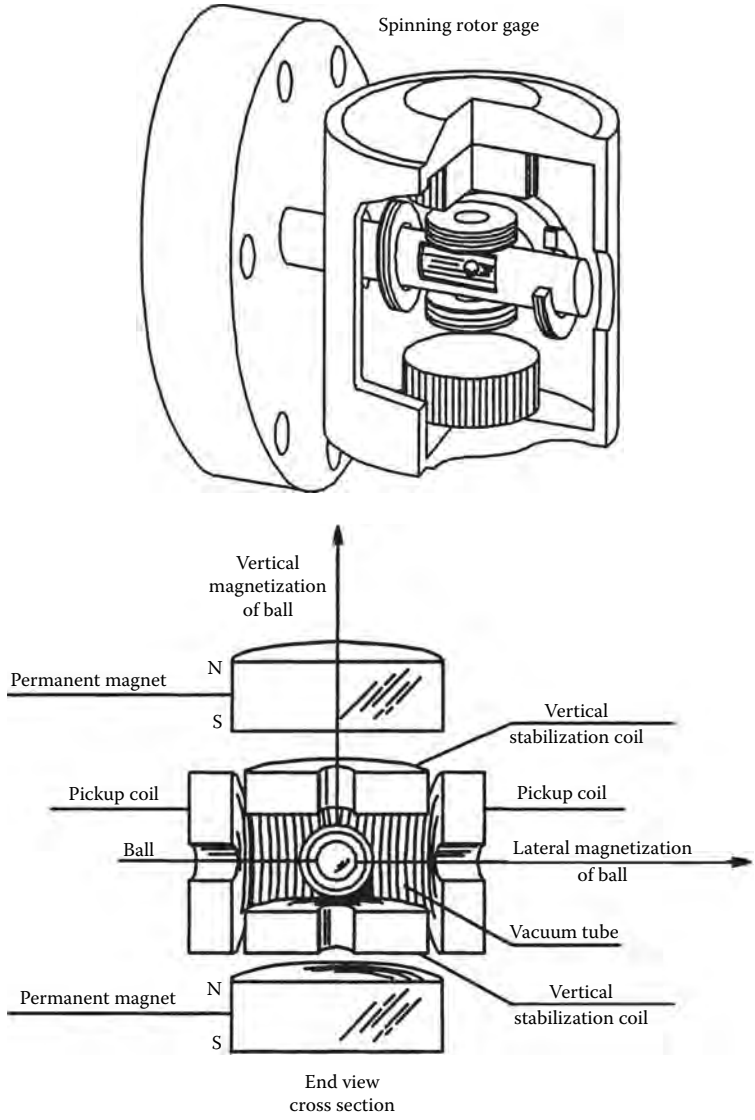


FIGURE 40.21 MDG (spinning rotor).

Note: Some sources include the term

$$\frac{(8kT')}{(\pi m)}$$

where

ρ is the density of the rotor
a is the radius of the rotor
ω'/ω is the fractional rate of slowing of the rotor
$\bar{c}$ is the mean gas molecular velocity
α is the linear coefficient of expansion of the ball
T' is the rate of change of the ball's temperature

All of the terms in the first part of the equation can be readily determined except for the accommodation coefficient σ, which depends on the surface of the ball and the molecular adhesion between the gas and the surface of the ball. The accommodation coefficient σ must be determined by calibration of the MDG against a known pressure standard or if repeatability is more important than the highest accuracy, by assuming a value of 1 for σ. Measurements of σ on many balls over several years have been repeatedly performed by Dittman et al. [68]. The values obtained ranged from 0.97 to 1.06 for 68 visually smooth balls, so using a value of 1 for σ would not introduce a large error and would allow the MDG to be considered a primary standard (Fremerey [66]).

The controller [68] contains the electronics to power and regulate the suspension and drive, detect and amplify the signal from the pickup coils, and then time the rotation of the ball. It also contains a data processor that stores the calibration data and computes the pressure.

The MDG is perhaps the best available transfer standard for the pressure range of 10^{-2}–10^{-7} torr (1–10^{-5} Pa) because it is designed for laboratory use in controlled relatively vibration-free environments [69,70].

40.7 Partial Pressure Measurements and Mass Spectrometers

The theory and the practical applications of partial pressure measurements and mass spectrometers are discussed in detail in the literature [71]; however, an overview is presented herein [72].

A simple device for measuring the partial pressure of nitrogen as well as the total pressure in a vacuum system is the residual nitrogen analyzer (RNA). Operating in high vacuum, it is used to detect leaks in a vacuum system. It is effective because various gases are pumped at different rates and nitrogen is readily pumped, leaving a much lower percentage than is present at atmospheric pressure. Thus, the presence of a significant percentage of nitrogen at high vacuum indicates an air leak. The RNA consists of a cold cathode gage with an optical filter and a photomultiplier tube. Since ionization in the cold cathode tube produces light and the color is determined by the gases present, the RNA filters out all except for that corresponding to nitrogen and is calibrated to give the partial pressure of nitrogen.

A more complex device to measure the partial pressure of many gases in a vacuum chamber is the mass spectrometer–type residual gas analyzer (RGA). This device comes in many forms and several sizes. The quadrupole mass spectrometer is shown in Figure 40.22. The sensing head consists of an ion source, a quadrupole mass filter, and a Faraday cup collector. The quadrupole mass filter consists of two pairs of parallel rods having equal and opposite RF and dc voltages. For each combination of voltages, only ions of a specific mass will pass through the filter. The mass filter is tuned to pass only ions of a specific mass-to-charge ratio at a given time. As the tuning is changed to represent increasing mass numbers, a display such as Figure 40.23 is produced, showing the relative intensity of the signal vs. the mass number. This display can then be compared electronically with similar displays for known gases to determine the composition of the gases in the vacuum chamber.

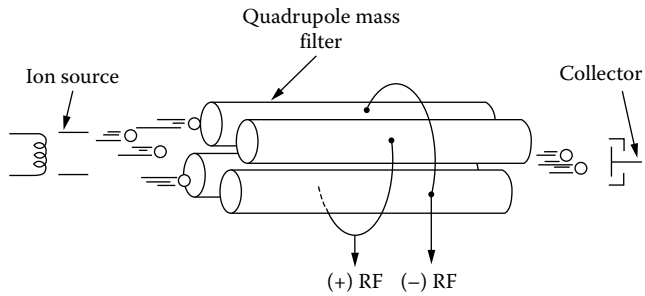


FIGURE 40.22 Quadrupole mass spectrometer. (From Varian Associates, *Basic Vacuum Practice*, Varian Associates, Inc., Lexington, MA, 1992. With permission.)

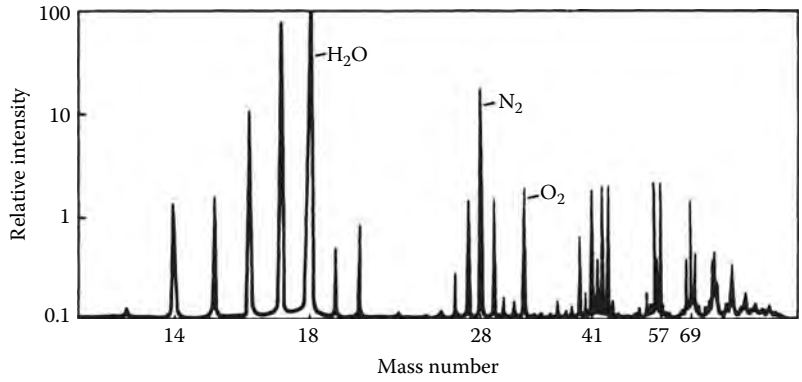


FIGURE 40.23 Relative intensity versus mass number. (From Varian Associates, *Basic Vacuum Practice*, Varian Associates, Inc., Lexington, MA, 1992. With permission.)

The head can operate only at high vacuum. However, by maintaining the head at high vacuum and using a sampling technique, the partial pressures of gases at higher pressures can be determined.

The calibration of mass spectrometers can be accomplished by equating the integral (the total area under all the peaks, taking into account the scale factors) to the overall pressure as measured by another gage (cold cathode, BA, or MDG). Once calibrated, the mass spectrometer can be used as a sensitive monitor of system pressure. It is also important when monitoring system pressure to know what gases are present. Most gages have vastly different sensitivities to different gas species.

Additional references on the details of the MDG and other types of gages available and on calibration are found in the literature [69–71,73,75–77]. The material in this chapter was summarized from an article on the fundamentals of vacuum measurement, calibration, and certification [78] from the authors' contribution to *The Handbook of Vacuum Technology* [71] and from other referenced sources.

References

1. J.F. O'Hanlon, *A User's Guide to Vacuum Technology*, New York: John Wiley & Sons, 1980, p. 47.
2. H. McLeod, *Phil. Mag.*, 47, 110, 1874.
3. C. Engleman and Televac Div. The Fredericks Co., Huntingdon Valley, PA, 19006, private communication.
4. Wallace and Tiernan Div., High Precision Gages, Pennwalt Corp., Bellville, NJ, http://www.wallaceandtiernan-usa.com/ (accessed on March 15, 2013).
5. R.W. Hyland and R.L. Shaffer, *J. Vac. Sci. Tech. A*, 9(6), 2843, 1991.
6. K.R. Spangenberg, *Vacuum Tubes*, New York: McGraw-Hill, 1948, p. 766.
7. S. Dushman, *Scientific Foundations of Vacuum Technique*, 2nd edn., J.M. Lafferty (ed.), New York: John Wiley & Sons, 1962.
8. W. Steckelmacher and B. Fletcher, *J. Physics E.*, 5, 405, 1972.
9. W. Steckelmacher, *Vacuum*, 23, 307, 1973.
10. A. Beiman, *Total Pressure Measurement in Vacuum Technology*, Orlando, FL: Academic Press, 1985.
11. Granville-Phillips, Boulder CO, 80303.
12. Televac Div., The Fredericks Co., Huntingdon Valley, PA, 19006.
13. Granville-Phillips, Boulder CO, Data Sheet 360127, 3/95.
14. Televac U.S. Patent No. 5351551.
15. O.E. Buckley, *Proc. Natl. Acad. Sci.*, 2, 683, 1916.
16. M.D. Sarbey, *Electronics,* 2, 594, 1931.

17. R. Champeix, *Physics and Techniques of Electron Tubes*, Vol. 1, New York: Pergamon Press, 1961, pp. 154–156.
18. N. Morgulis, *Physik Z. Sowjetunion*, 5, 407, 1934.
19. N.B. Reynolds, *Physics*, 1, 182, 1931.
20. J.H. Leck, *Pressure Measurement in Vacuum Systems*, London, U.K.: Chapman & Hall, 1957, pp. 70–74.
21. S. Dushman and C.G. Found, *Phys. Rev*., 17, 7, 1921.
22. E.K. Jaycock and H.W. Weinhart, *Rev. Sci. Instr*., 2, 401, 1931.
23. G.J. Schulz and A.V. Phelps, *Rev. Sci. Instr*., 28, 1051, 1957.
24. Japanese Industrial Standard (JIS-Z-8570), Method of Calibration for Vacuum Gages.
25. J.W. Leck, op. cit., 69.
26. W.B. Nottingham, *Proceedings of 7th Annual Conference on Physical Electronics*, MIT Press, Cambridge, MA, 1947.
27. H.A. Steinhertz and P.A. Redhead, *Sci. Am*., 206, 78, 1962.
28. R.T. Bayard and D. Alpert, *Rev. Sci. Instr*., 21, 571, 1950.
29. O.A. Weinreich, *Phys. Rev*., 82, 573, 1951.
30. O.A. Weinreich and H. Bleecher, *Rev. Sci. Instr*., 23, 56, 1952.
31. H.C. Hseuh and C. Lanni, *J. Vac. Sci. Technol*., A5, 3244, 1987.
32. T.S. Chou and Z.Q. Tang, *J. Vac. Sci. Technol*., A4, 2280, 1986.
33. P.A. Redhead, *Rev. Sci. Instr*., 31, 343, 1960.
34. G.H. Metson, *Br. J. Appl. Phys*., 2, 46, 1951.
35. J.J. Lander, *Rev. Sci. Inst*., 21, 672, 1950.
36. P.A. Redhead, *J. Vac. Sci. Technol*., 3, 173, 1966.
37. J. Groszkowski, *Le Vide*, 136, 240, 1968.
38. L.G. Pittaway, *Philips Res. Rept*., 29, 283, 1974.
39. D. Blechshmidt, *J. Vac. Sci. Technol*., 10, 376, 1973.
40. P.A. Redhead, *J. Vac. Sci. Technol*., 6, 848, 1969.
41. S.D. Wood and C.R. Tilford, *J. Vac. Sci. Technol*., A3, 542, 1985.
42. C.R. Tilford, *J. Vac. Sci. Technol*., A3, 546, 1985.
43. P.C. Arnold and D.G. Bills, *J. Vac. Sci. Technol*., A2, 159, 1984.
44. P.C. Arnold and J. Borichevsky, *J. Vac. Sci. Technol*., A12, 568, 1994.
45. D.G. Bills, *J. Vac. Sci. Technol*., A12, 574, 1994.
46. C.R. Tilford, A.R. Filippelli, et al., *J. Vac. Sci. Technol*., A13, 485, 1995.
47. T.A. Flaim and P.D. Owenby, *J. Vac. Sci. Technol*., 8, 661, 1971.
48. J.F. O'Hanlon, op. cit., 65.
49. N. Ohsako, *J. Vac. Sci. Technol*., 20, 1153, 1982.
50. P.C. Arnold, D.G. Bills et al., *J. Vac. Sci. Technol*., A12, 580, 1994.
51. ETI Division of the Fredericks Co., Huntingdon Valley, PA, Gage Type 8184.
52. F.M. Penning, *Physica*, 4, 71, 1937.
53. F.M. Penning and K. Nienhauis, *Philips Tech. Rev*., 11, 116, 1949.
54. P.A. Redhead, *Can. J. Phys*., 36, 255, 1958.
55. J.P. Hobson and P.A. Readhead, *Can. J. Phys*., 33, 271, 1958.
56. NRC type 552 data sheet.
57. D. Pelz and G. Newton, *J. Vac. Sci. Technol*., 4, 239, 1967.
58. R.N. Peacock, N.T. Peacock, and D.S. Hauschulz, *J. Vac. Sci. Technol*., A9, 1977 1991.
59. E. Drubetsky, D.R. Taylor, and W.H. Bayles, Jr., *Am. Vac. Soc., New Engl. Chapter, 1993 Symp*.
60. B.R. Kendall and E. Drubetsky, *J. Vac. Sci. Technol*., A14, 1292, 1996.
61. M. Ono, K. Hirata, et al., *J. Vac. Sci. Technol*., A4, 1728, 1986.
62. O.E. Meyer, *Pogg. Ann*., 125, 177, 1865.
63. J.C. Maxwell, *Phil. Trans. R. Soc*., 157, 249, 1866.

64. F.T. Holmes, *Rev. Sci. Instru.*, 8, 444, 1937.
65. J.W. Beams, J.L. Young, and J.W. Moore, *J. Appl. Phys.*, 17, 886, 1946.
66. J.K. Fremery, *Vacuum*, 32, 685, 1946.
67. NIST, Vacuum calibrations using the molecular drag gage, Course Notes, April 15–17, 1996.
68. S. Dittman, B.E. Lindenau, and C.R. Tilford, *J. Vac. Sci. Technol.*, A7, 3356, 1989.
69. K.E. McCulloh, S.D. Wood, and C.R. Tilford, *J. Vac. Sci. Technol.*, A3, 1738, 1985.
70. G. Cosma, J.K. Fremerey, B. Lindenau, G. Messer, and P. Rohl, *J. Vac. Sci. Technol.*, 17, 642, 1980.
71. D.M. Hoffman, B. Singh, and J.H. Thomas, III (eds.), *The Handbook of Vacuum Science and Technology*, Orlando, FL: Academic Press, 1998.
72. Varian Associates, *Basic Vacuum Practice*, Lexington, MA: Varian Associates, Inc., p. 02173, 1992.
73. C.R. Tilford, S. Dittman, and K.E. McCulloh, *J. Vac. Sci. Technol.*, A6, 2855, 1988.
74. M. Hirata, M. Ono, H. Hojo, and K. Nakayama, Calibration of secondary standard ionization gauges, *J. Vac. Sci. Technol.*, 20(4), 1159, 1982.
75. S. Dittman, NIST Special Publication 250–34, 1989.
76. National Conference of Standards Laboratories, Boulder, CO.
77. H. Gantsch, J. Tewes, and G. Messer, *Vacuum*, 35(3), 137, 1985.
78. W.H. Bayles, Jr., Fundamentals of vacuum measurement, calibration and certification, *Ind. Heat.*, October 1992.

Further Readings

Barbour, E., How Vacuum Tubes Work, http://www.vacuumtubes.net/How_Vacuum_Tubes_Work.htm (accessed on July 11, 2012).

O'Hanlon, J. F., *A User Guide to Vacuum Technology*, 3rd edn., Hoboken, NJ: John Wiley & Sons, 2003.

Hoffman, D. M., Bawa S., and Thomas, J. H., *Handbook of Vacuum Science and Technology*, San Diego, CA: Academic Press, 1998.

Partial List of Vendors and Suppliers

9scale Vacuum Products, Inc., 192 Anawan St., Box 1B, Fall River, MA 02721, Phone: 508-646-9352, 800-565-6911 (toll free), Fax: 508-646-9597, www.systemsfab.com

Airserco, 7555 Tyler Blvd., Mentor, OH 44060, Phone: 440-946-2700, 800-473-1754 (toll free), Fax: 440-946-8188, www.airserco.com

McGrath, E., Inc., 35 Osborne St., Salem, MA 01970-2529, Phone: 978-744-3546, 877-744-3546 (toll free), Fax: 978-741-4020, www.emcgrath.com

Precision Metrology, 7350 N. Teutonia Ave., Milwaukee, WI 53209, Phone: 414-351-7420, 888-330-3303 (toll free), Fax: 414-351-7429, www.precisionmetrology.com

TM Limited, P.O. Box 1438, Middleboro, MA 02346, Phone: 508-946-4533, 800-910-2352 (toll free), Fax: 508-947-4686, www.tmlimited.com

Vacuubrand, Inc., 11 Bokum Rd., Essex, CT 06426-1506, Phone: 860-767-5341, 888-882-6730 (toll free), Fax: 860-767-2563, www.vacuubrand.com



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

41

Force Measurement

41.1 General Considerations **41**-1
41.2 Hooke's Law **41**-2
Basic Methods of Force Measurement
41.3 Force Sensors **41**-3
41.4 Strain-Gage Load Cell **41**-4
Strain Gages • Beam-Type Load Cell • Ring-Type Load Cell
41.5 Piezoelectric Methods **41**-7
Resistive Method • Inductive Method • Piezotransistor Method • Multicomponent Dynamometers Using Quartz Crystals as Sensing Elements
41.6 Capacitive Force Transducer **41**-12
41.7 Force-Sensing Resistors (Conductive Polymers) **41**-13
Magnetoresistive Force Sensors
41.8 Magnetoelastic Force Sensors **41**-14
41.9 Torsional Balances **41**-14
41.10 Tactile Sensors **41**-14
Tactile Sensor Requirements • Tactile Array Sensor
References **41**-16
Further Information **41**-17
Partial List of Vendors and Suppliers **41**-17

M.A. Elbestawi
McMaster University

Force, which is a vector quantity, can be defined as an action that will cause acceleration or a certain reaction of a body. This chapter will outline the methods that can be employed to determine the magnitude of these forces.

41.1 General Considerations

The determination or measurement of forces must yield to the following considerations: If the forces acting on a body do not produce any acceleration, they must form a *system of forces in equilibrium*. The system is then considered to be in static equilibrium. The forces experienced by a body can be classified into two categories: internal, where the individual particles of a body act on each other, and external, otherwise. If a body is supported by other bodies while subject to the action of forces, deformations and/or displacements will be produced at the points of support or contact. The internal forces will be distributed throughout the body until equilibrium is established, and then the body is said to be in a state of tension, compression, or shear. In considering a body at a definite section, it is evident that all the internal forces act in pairs, the two forces being equal and opposite, whereas the external forces act singly.

41.2 Hooke's Law

The basis for force measurement results from the physical behavior of a body under external forces. Therefore, it is useful to review briefly the mechanical behavior of materials. When a metal is loaded in uniaxial tension, uniaxial compression, or simple shear (Figure 41.1), it will behave elastically until a critical value of normal stress (S) or shear stress (τ) is reached, and then it will deform plastically [1]. In the elastic region, the atoms are temporarily displaced but return to their equilibrium positions when the load is removed. Stress (S or τ) and strain (e or γ) in the elastic region are defined as indicated in Figure 41.2.

$$\nu = -\frac{e_2}{e_1} \tag{41.1}$$

Poisson's ratio (ν) is the ratio of transverse (e_2) to direct (e_1) strain in tension or compression. In the elastic region, ν is between 1/4 and 1/3 for metals. The relation between stress and strain in the elastic region is given by Hooke's law:

$$S = Ee \text{ (tension or compression)} \tag{41.2}$$

$$\tau = G\gamma \text{ (simple shear)} \tag{41.3}$$

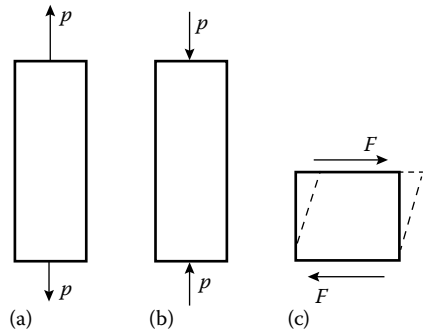


FIGURE 41.1 When a metal is loaded in uniaxial tension (a), uniaxial compression (b), or simple shear (c), it will behave elastically until a critical value of normal stress or shear stress is reached.

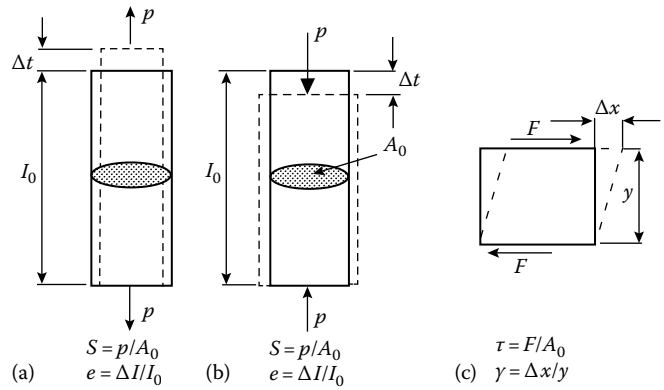


FIGURE 41.2 Elastic stress and strain for (a) uniaxial tension, (b) uniaxial compression, and (c) simple shear. (From Shaw, M.C., *Metal Cutting Principles*, Oxford Science Publications: Clarendon Press, Oxford, U.K., 1989.)

where E and G are the Young's and shear moduli of elasticity, respectively. A small change in specific volume (ΔVol/Vol) can be related to the elastic deformation, which is shown to be as follows for an isotropic material (same properties in all directions):

$$\frac{\Delta \text{Vol}}{\text{Vol}} = e_1(1-2\nu) \tag{41.4}$$

The bulk modulus (K = reciprocal of compressibility) is defined as follows:

$$K = \frac{\Delta p}{\Delta \text{Vol}/\text{Vol}} \tag{41.5}$$

where Δp is the pressure acting at a particular point. For an elastic solid loaded in uniaxial compression (S),

$$K = \frac{S}{\Delta \text{Vol}/\text{Vol}} = \frac{S}{e_1(1-2\nu)} = \frac{E}{1-2\nu} \tag{41.6}$$

Thus, an elastic solid is compressible as long as ν is less than 1/2, which is normally the case for metals. Hooke's law (Equation 41.2) for uniaxial tension can be generalized for a 3D elastic condition.

The theory of elasticity is well established and is used as a basis for force-measuring techniques. Note that the measurement of forces in separate engineering applications is very application specific and care must be taken in the selection of the measuring techniques outlined as follows.

41.2.1 Basic Methods of Force Measurement

An unknown force may be measured by the following means:

1. Balancing the unknown force against a standard mass through a system of levers
2. Measuring the acceleration of a known mass
3. Equalizing it to a magnetic force generated by the interaction of a current-carrying coil and a magnet
4. Distributing the force on a specific area to generate pressure and then measuring the pressure
5. Converting the applied force into the deformation of an elastic element

The aforementioned methods used for measuring forces yield a variety of designs of measuring equipment. The challenge involved with the task of measuring force resides primarily in sensor design. The basics of sensor design can be resolved into two problems:

1. Primary geometric or physical constraints, governed by the application of the force sensor device
2. The means by which the force can be converted into a workable signal form (such as electronic signals or graduated displacements)

The remaining sections will discuss the types of devices used for force to signal conversion and finally illustrate some examples of applications of these devices for measuring forces.

41.3 Force Sensors

Force sensors are required for a basic understanding of the response of a system. For example, cutting forces generated by a machining process can be monitored to detect a tool failure or to diagnose the causes of this failure in controlling the process parameters and in evaluating the quality of the surface produced. Force sensors are used to monitor impact forces in the automotive industry. Robotic handling and assembly tasks are controlled by detecting the forces generated at the end effector. Direct measurement of forces is useful in controlling many mechanical systems.

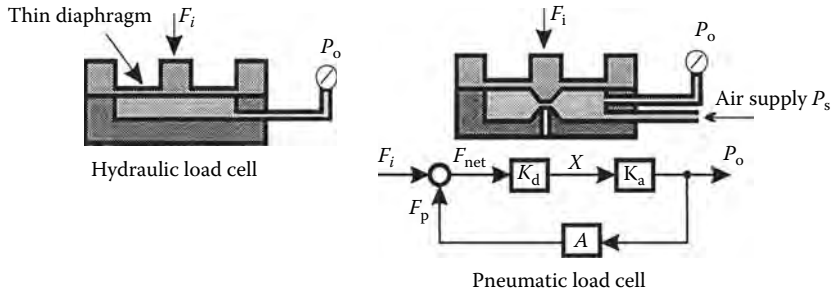


FIGURE 41.3 Different types of load cells. (From Doebelin, E.O., *Measurement Systems, Application and Design*, 4th edn., McGraw-Hill, New York, 1990.)

Some types of force sensors are based on measuring a deflection caused by the force. Relatively high deflections (typically, several micrometers) would be necessary for this technique to be feasible. The excellent elastic properties of helical springs make it possible to apply them successfully as force sensors that transform the load to be measured into a deflection. The relation between force and deflection in the elastic region is demonstrated by Hooke's law. Force sensors that employ strain-gage elements or piezoelectric (quartz) crystals with built-in microelectronics are common. Both impulsive forces and slowly varying forces can be monitored using these sensors.

Of the available force-measuring techniques, a general subgroup can be defined as that of load cells. Load cells are comprised generally of a rigid outer structure, some medium that is used for measuring the applied force, and the measuring gage. Load cells are used for sensing large, static, or slowly varying forces with little deflection and are a relatively accurate means of sensing forces. Typical accuracies are of the order of 0.1% of the full-scale readings. Various strategies can be employed for measuring forces that are strongly dependent on the design of the load cell. For example, Figure 41.3 illustrates different types of load cells that can be employed in sensing large forces for relatively little cost. The hydraulic load cell employs a very stiff outer structure with an internal cavity filled with a fluid. The application of a load increases the oil pressure, which can be read off an accurate gage.

Other sensing techniques can be utilized to monitor forces, such as piezoelectric transducers for quicker response of varying loads, pneumatic methods, and strain gages. The proper sensing technique needs special consideration based on the conditions required for monitoring.

41.4 Strain-Gage Load Cell

The strain-gage load cell consists of a structure that elastically deforms when subjected to a force and a strain-gage network that produces an electric signal proportional to this deformation. Examples of this are beam and ring types of load cells.

41.4.1 Strain Gages

Strain gages use a length of gage wire to produce the desired resistance (which is usually about 120 Ω in the form of a flat coil). This coil is then cemented (bonded) between two thin insulating sheets of paper or plastic. Such a gage cannot be used directly to measure deflection. It has to be first fixed properly to a member to be strained. After bonding the gage to the member, they are baked at about 195 °F (90 °C) to remove moisture. Coating the unit with wax or resin will provide some mechanical protection. The resistance between the member under test and the gage itself must be at least 50 MΩ. The total area of all conductors must remain small so that the cement can easily transmit the force necessary to deform the wire. As the member is stressed, the resulting strain deforms the strain gage and the cross-sectional area diminishes. This causes an increase in resistivity of the gage that is easily determined. In order to

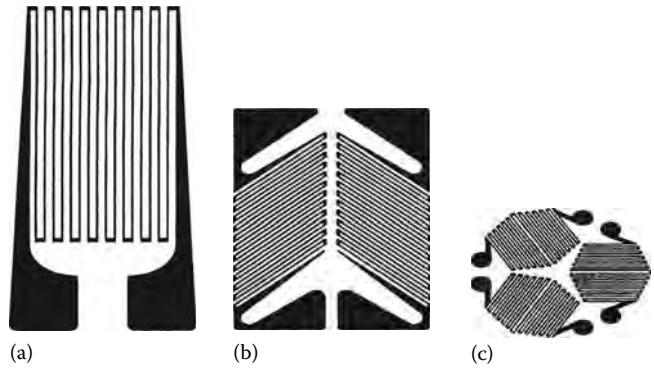


FIGURE 41.4 Configuration of metal-foil resistance strain gages: (a) single element, (b) two elements, and (c) three elements.

measure very small strains, it is necessary to measure small changes of the resistance per unit resistance ($\Delta R/R$). The change in the resistance of a bonded strain gage is usually less than 0.5%. A wide variety of gage sizes and grid shapes are available, and typical examples are shown in Figure 41.4.

The use of strain gages to measure force requires careful consideration with respect to rigidity and environment. By virtue of their design, strain gages of shorter length generally possess higher response frequencies (examples: 660 kHz for a gage of 0.2 mm and 20 kHz for a gage of 60 mm in length). The environmental considerations focus mainly on the temperature of the gage. It is well known that resistance is a function of temperature and, thus, strain gages are susceptible to variations in temperature. Thus, if it is known that the temperature of the gage will vary due to any influence, temperature compensation is required in order to ensure that the force measurement is accurate.

A Wheatstone bridge (Figure 41.5) is usually used to measure this small order of magnitude. In Figure 41.5, no current will flow through the galvanometer (G) if the four resistances satisfy a certain condition. In order to demonstrate how a Wheatstone bridge operates [3], a voltage scale has been drawn at points C and D of Figure 41.5. Assume that R_1 is a bonded gage represented by R_g and that initially Equation 41.7 is satisfied. If R_1 is now stretched so that its resistance increases by one unit ($+\Delta R$), the voltage at point D will be increased from zero to plus one unit of voltage ($+\Delta V$), and there will be a voltage difference of one unit between C and D that will give rise to a current through C. If R_4 is also a bonded gage and at the same time, R_1 changes by $+\Delta R$ and R_4 changes by $-\Delta R$, the voltage at D will move to $+2\Delta V$. Also, if at the same time, R_2 changes by $-\Delta R$ and R_3 changes by $+\Delta R$, then the voltage of point C will move to $-2\Delta V$ and the voltage difference between C and D will now be $4\Delta V$. It is then apparent that although a single gage can be used, the sensitivity can be increased fourfold if two gages are used in tension while two others are used in compression. Similarly, the bridge becomes unbalanced in proportion to the difference in strain when gages are located on

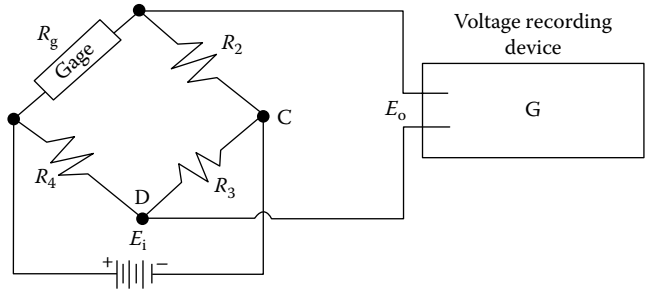


FIGURE 41.5 Wheatstone bridge.

the adjacent leg. The strain gage arrangement determines the relationship between the output and the type of strain being measured for axial, shear, bending, and torsional strains.

$$\frac{R_1}{R_4} = \frac{R_2}{R_3} \tag{41.7}$$

The grid configuration of the metal-foil resistance strain gages is formed by a photo-etching process. The shortest gage available is 0.20 mm; the longest is 102 mm. Standard gage resistance is 120 and 350 Ω. A strain gage exhibits a resistance change $\Delta R/R$ that is related to the strain in the direction of the grid lines by the expression in Equation 41.8 (where S_g is the gage factor or calibration constant for the gage).

$$\frac{\Delta R}{R} = S_g \varepsilon \tag{41.8}$$

41.4.2 Beam-Type Load Cell

Beam-type load cells are commonly employed for measuring low-level loads [3]. A simple cantilever beam (see Figure 41.6a) with four strain gages, two on the top surface and two on the bottom surface (all oriented along the axis of the beam), is used as the elastic member (sensor) for the load cell. The gages are wired into a Wheatstone bridge as shown in Figure 41.6b. The load P produces a moment $M = Px$ at the gage location (x) that results in the following strains:

$$\varepsilon_1 = -\varepsilon_2 = \varepsilon_3 = -\varepsilon_4 = \frac{6M}{Ebh^2} = \frac{6Px}{Ebh^2} \tag{41.9}$$

where

b is the width of the cross section of the beam
h is the height of the cross section of the beam

Thus, the response of the strain gages is obtained from Equation 41.10:

$$\frac{\Delta R_1}{R_1} = -\frac{\Delta R_2}{R_2} = \frac{\Delta R_3}{R_3} = -\frac{\Delta R_4}{R_4} = \frac{6S_g PxE_1}{Ebh^2} \tag{41.10}$$

The output voltage E_o from the Wheatstone bridge, resulting from application of the load P, is obtained from Equation 41.11. If the four strain gages on the beam are assumed to be identical, then Equation 41.11 holds:

$$E_o = \frac{6S_g PxE_1}{Ebh^2} \tag{41.11}$$

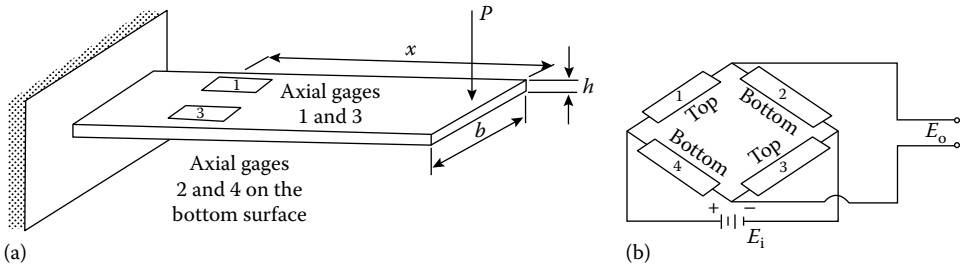


FIGURE 41.6 Beam-type load cells: (a) a selection of beam-type load cells (elastic element with strain gages) and (b) gage positions in the Wheatstone bridge. (From Dally, J.W. et al., *Instrumentation for Engineering Measurements*, John Wiley & Sons, New York, 1984.)

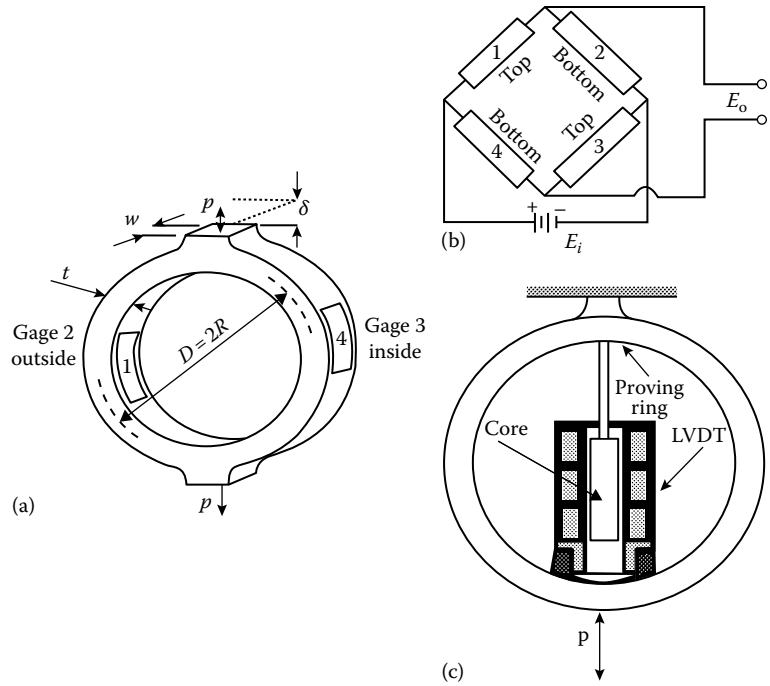


FIGURE 41.7 Ring-type load cells: (a) elastic element with strain-gage sensors, (b) gage positions in the Wheatstone bridge, and (c) elastic element with an LVDT sensor. (From Dally, J.W. et al., *Instrumentation for Engineering Measurements*, John Wiley & Sons, New York, 1984.)

The range and sensitivity of a beam-type load cell depends on the shape of the cross section of the beam, the location of the point of application of the load, and the fatigue strength of the material from which the beam is fabricated.

41.4.3 Ring-Type Load Cell

Ring-type load cells incorporate a proving ring (see Figure 41.7) as the elastic element. The ring element can be designed to cover a very wide range of loads by varying the diameter D, the thickness t, or the depth w of the ring. Either strain gages or linear variable differential transformer (LVDT) can be used as the sensor.

The load P is linearly proportional to the output voltage E_o. The sensitivity of the ring-type load cell with an LVDT sensor depends on the geometry of the ring (R, t, and w), the material from which the ring is fabricated (E), and the characteristics of the LVDT (S and E_i). The range of a ring-type load cell is controlled by the strength of the material used in fabricating the ring.

41.5 Piezoelectric Methods

A piezoelectric material exhibits a phenomenon known as the *piezoelectric effect*. This effect states that when asymmetrical, elastic crystals are deformed by a force, an electric potential will be developed within the distorted crystal lattice. This effect is reversible. That is, if a potential is applied between the surfaces of the crystal, it will change its physical dimensions [4]. Elements exhibiting piezoelectric qualities are sometimes known as electrorestrictive elements.

The magnitude and the polarity of the induced surface charges are proportional to the magnitude and direction of the applied force [4]:

$$Q = dF \tag{41.12}$$

where d is the charge sensitivity (a constant for a given crystal) of the crystal in C/N. The force F causes a thickness variation Δt meters of the crystal:

$$F = \frac{aY}{t}\Delta t \tag{41.13}$$

where

a is area of crystal
t is thickness of crystal
Y is Young's modulus

$$Y = \frac{\text{stress}}{\text{strain}} = \frac{Ft}{a\Delta t} \tag{41.14}$$

The charge at the electrodes gives rise to a voltage $E_o = Q/C$, where C is capacitance in farads between the electrodes and $C = \varepsilon a/t$ where ε is the absolute permittivity:

$$E_o = \frac{dF}{C} = \frac{d}{\varepsilon}\frac{tF}{a} \tag{41.15}$$

The voltage sensitivity = $g = d/\varepsilon$ in volt m/N can be obtained as

$$E_o = g\frac{t}{a}F = gtP \tag{41.16}$$

The piezoelectric materials used are quartz, tourmaline, Rochelle salt, ammonium dihydrogen phosphate (ADP), lithium sulfate, barium titanate, and lead zirconate titanate (PZT) [4]. Quartz and other earthly piezoelectric crystals are naturally polarized. However, synthetic piezoelectric materials, such as barium titanate ceramic, are made by baking small crystallites under pressure and then placing the resultant material in a strong dc electric field [4]. After that, the crystal is polarized, along the axis on which the force will be applied, to exhibit piezoelectric properties. Artificial piezoelectric elements are free from the limitations imposed by the crystal structure and can be molded into any size and shape. The direction of polarization is designated during their production process.

The different modes of operation of a piezoelectric device for a simple plate are shown in Figure 41.8 [4]. By adhering two crystals together so that their electric axes are perpendicular,

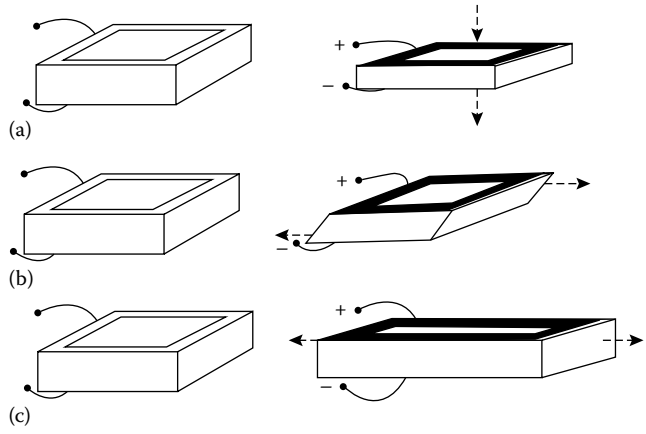


FIGURE 41.8 Modes of operation for a simple plate as a piezoelectric device: (a) Perpendicular force, (b) shear force, and (c) horizontal force. (From Mansfield, P.H., *Electrical Transducers for Industrial Measurement*, The Butterworth Group, London, U.K., 1973.)

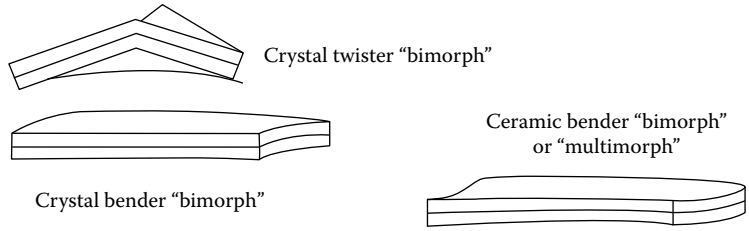


FIGURE 41.9 Curvature of "twister" and "bender" piezoelectric transducers when voltage applied. (From Mansfield, P.H., *Electrical Transducers for Industrial Measurement*, The Butterworth Group, London, U.K., 1973.)

bending moments or torque can be applied to the piezoelectric transducer and a voltage output can be produced (Figure 41.9) [4]. The ranges of forces that can be measured using piezoelectric transducers are from 1 to 200 kN and at a ratio of 2×10^5.

Piezoelectric crystals can also be used in measuring an instantaneous change in the force (dynamic forces). A thin plate of quartz can be used as an electronic oscillator. The frequency of these oscillations will be dominated by the natural frequency of the thin plate. Any distortion in the shape of the plate caused by an external force alters the oscillation frequency. Hence, a dynamic force can be measured by the change in frequency of the oscillator.

41.5.1 Resistive Method

The resistive method employs the fact that when the multiple contact area between semiconducting particles (usually carbon) and the distance between the particles are changed, the total resistance is altered. The design of such transducers yields a very small displacement when a force is applied. A transducer might consist of 2–60 thin carbon disks mounted between a fixed and a movable electrode. When a force is applied to the movable electrode and the carbon disks move together by 5–250 μm per interface, the transfer function of their resistance against the applied force is approximately hyperbolic, that is, highly nonlinear. The device is also subject to large hysteresis and drift together with a high transverse sensitivity.

In order to reduce hysteresis and drift, rings are used instead of disks. The rings are mounted on an insulated rigid core and prestressed. This almost completely eliminates any transverse sensitivity error. The core's resonant frequency is high and can occur at a frequency as high as 10 kHz. The possible measuring range of such a transducer is from 0.1 to 10 kg. The accuracy and linear sensitivity of this transducer is very poor.

41.5.2 Inductive Method

The inductive method utilizes the fact that a change in mechanical stress of a ferromagnetic material causes its permeability to alter. The changes in magnetic flux are converted into induced voltages in the pickup coils as the movement takes place. This phenomenon is known as the *Villari effect* or *magnetostriction*. It is known to be particularly strong in nickel–iron alloys.

Transducers utilizing the Villari effect consist of a coil wound on a core of magnetostrictive material. The force to be measured is applied on this core, stressing it and causing a change in its permeability and inductance. This change can be monitored and used for determining the force.

The applicable range for this type of transducer is a function of the cross-sectional area of the core. The accuracy of the device is determined by a calibration process. This transducer has poor linearity and is subject to hysteresis. The permeability of a magnetostrictive material increases when it is subjected to pure torsion, regardless of direction. A flat frequency response is obtained over a wide range from 150 to 15,000 Hz.

41.5.3 Piezotransistor Method

Devices that utilize *anisotropic stress effects* are described as piezotransistors. In this effect, if the upper surface of a *p–n* diode is subjected to a localized stress, a significant reversible change occurs in the current across the junction. These transistors are usually silicon nonplanar type, with an emitter-base junction. This junction is mechanically connected to a diaphragm positioned on the upper surface of a typical TO-type can [4]. When a pressure or a force is applied to the diaphragm, an electronic charge is produced. It is advisable to use these force-measuring devices at a constant temperature by virtue of the fact that semiconducting materials also change their electric properties with temperature variations. The attractive characteristic of piezotransistors is that they can withstand a 500% overload.

41.5.4 Multicomponent Dynamometers Using Quartz Crystals as Sensing Elements

41.5.4.1 Piezoelectric Effects in Quartz

For force measurements, the *direct piezoelectric effect* is utilized. The direct longitudinal effect measures compressive force; the direct shear effect measures shear force in one direction. For example, if a disk of crystalline quartz (SiO_2) cut normally to the crystallographic *x*-axis is loaded by a compression force, it will yield an electric charge, nominally 2.26 pC/N. If a disk of crystalline quartz is cut normally to the crystallographic *y*-axis, it will yield an electric charge (4.52 pC/N) if loaded by a shear force in one specific direction. Forces applied in the other directions will not generate any output [5].

A charge amplifier is used to convert the charge yielded by a quartz crystal element into a proportional voltage. The range of a charge amplifier with respect to its conversion factor is determined by a feedback capacitor. Adjustment to mechanical units is obtained by additional operational amplifiers with variable gain.

41.5.4.2 Design of Quartz Multicomponent Dynamometers

The main element for designing multicomponent dynamometers is the three-component force transducer (Figure 41.10). It contains a pair of *X*-cut quartz disks for the normal force component and a pair of *Y*-cut quartz disks (shear sensitive) for each shear force component.

Three-component dynamometers can be used for measuring cutting forces during machining. Four three-component force transducers sandwiched between a baseplate and a top plate are shown in Figure 41.10. The force transducer is subjected to a preload as shear forces are transmitted by friction. The four force transducers experience a drastic change in their load, depending on the type and position of force application. An overhanging introduction of the force develops a tensile force for some

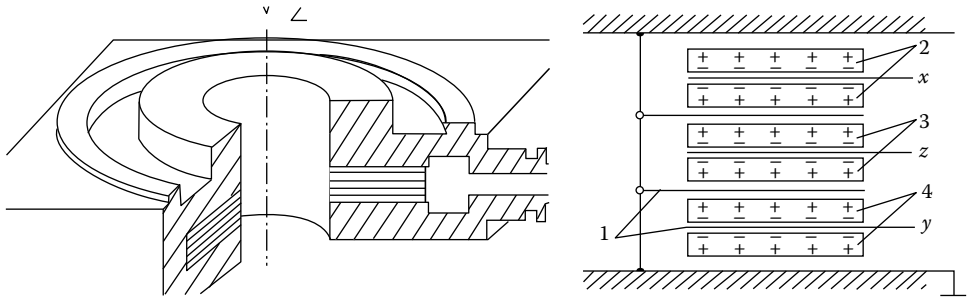


FIGURE 41.10 Three-component force transducer.

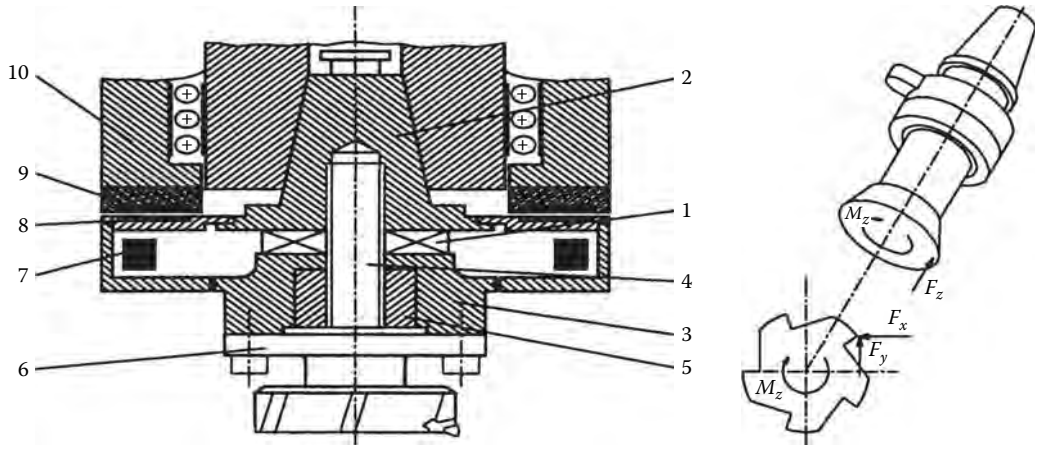


FIGURE 41.11 Force-measuring system to determine the tool-related cutting forces in five-axis milling. (From Spur, G. et al., Measuring the cutting force in five-axis milling, Translated paper "Zerpankraftmessung bei der funfachsigen Frasbearbeitung," Zeitschrift fur wirtschaftliche Fertigung und Automatisierung 9/93 Carl Hanser, Munchen, Kistler Piezo-Instrumentation, 20.162e 9.94.)

transducers, thus reducing the preload. The bending of the top plate of the dynamometer causes bending and shearing stresses. The measuring ranges of a dynamometer depend not only on the individual forces but also on the individual bending stresses.

41.5.4.3 Measuring Signals Transmitted by Telemetry

Figure 41.11 shows the newly designed force-measuring system rotating cutting force dynamometer (RCD). A ring-shaped sensor (1) is fitted in a steep-angle taper socket (2) and a base ring (3) allowing sensing of the three force components F_x, F_y, and F_z at the cutting edge as well as the moment M_z. The physical operating principle of this measuring cell is based on the piezoelectric effect in quartz plates. The quartz plates incorporated in the sensor are aligned so that the maximum cross sensitivity between the force components is 1%. As a result of the rigid design of the sensor, the resonant frequencies of the force-measuring system range from 1200 to 3000 Hz, and the measuring ranges cover a maximum of 10 kN [6].

Force-proportional charges produced at the surfaces of the quartz plates are converted into voltages by four miniature charge amplifiers (7) in hybrid construction. These signals are then filtered by specific electric circuitry to prevent aliasing effects and digitized with 8-bit resolution using a high sampling rate (pulse-code modulation). The digitized signals are transmitted by a telemetric unit consisting of a receiver and transmitter module, an antenna at the top of the rotating force-measuring system (8), as well as a fixed antenna (9) on the splash cover of the two-axis milling head (10). The electric components, the charge amplifier, and the transmitter module are mounted on the circumference of the force-measuring system [6].

The cutting forces and the moment measured are digitized with the force-measuring system described earlier. They are modulated on an FM carrier and transmitted by the rotating transmitter to the stationary receiver. The signals transmitted are fed to an external measured-variable conditioning unit.

41.5.4.4 Measuring Dynamic Forces

Any mechanical system can be considered in the first approximation as a weakly damped oscillator consisting of a spring and a mass. If a mechanical system has more than one resonant frequency, the lowest one must be taken into consideration. As long as the test frequency remains below 10% of the resonant

frequency of the reference transducer (used for calibration), the difference between the dynamic sensitivity obtained and the static calibration will be less than 1%. The aforementioned considerations assume a sinusoidal force signal. The static calibration of a reference transducer is also valid for dynamic calibration purposes if the test frequency is much lower (at least 10 times lower) than the resonant frequency of the system.

41.6 Capacitive Force Transducer

A transducer that uses capacitance variation can be used to measure force. The force is directed onto a membrane whose elastic deflection is detected by a capacitance variation. A highly sensitive force transducer can be constructed because the capacitive transducer senses very small deflections accurately. An electronic circuit converts the capacitance variations into dc-voltage variations [7].

The capacitance sensor illustrated in Figure 41.12 consists of two metal plates separated by an air gap. The capacitance C between terminals is given by the expression

$$C = \varepsilon_0 \varepsilon_r \frac{A}{h} \tag{41.17}$$

where

C is the capacitance in farads (F)
ε_0 is the dielectric constant of free space
ε_r is the relative dielectric constant of the insulator
A is the overlapping area for the two plates
h is the thickness of the gap between the two plates

The sensitivity of capacitance-type sensors is inherently low. Theoretically, decreasing the gap h should increase the sensitivity; however, there are practical electric and mechanical conditions that preclude high sensitivities. One of the main advantages of the capacitive transducer is that moving of one of its plates relative to the other requires an extremely small force to be applied. A second advantage is stability, and the sensitivity of the sensor is not influenced by pressure or temperature of the environment.



FIGURE 41.12 Capacitive force transducer. (From Nachtigal, C.L., *Instrumentation and Control, Fundamentals and Applications*, John Wiley & Sons, New York, 1990.)

41.7 Force-Sensing Resistors (Conductive Polymers)

Force-sensing resistors (FSRs) utilize the fact that certain polymer thick-film devices exhibit decreasing resistance with the increase of an applied force. An FSR is made up of two parts. The first is a resistive material applied to a film. The second is a set of digitizing contacts applied to another film. Figure 41.13 shows this configuration. The resistive material completes the electric circuit between the two sets of conductors on the other film. When a force is applied to this sensor, a better connection is made between the contacts; hence, the conductivity is increased. Over a wide range of forces, it turns out that the conductivity is approximately a linear function of force. Figure 41.14 shows the resistance of the sensor as a function of force. It is important to note that there are three possible regions for the sensor to operate. The first abrupt transition occurs somewhere in the vicinity of 10g of force. In this region, the resistance changes very rapidly. This behavior is useful when one is designing switches using FSRs.

FSRs should not be used for accurate measurements of force because sensor parts may exhibit 15%–25% variation in resistance between each other. However, FSRs exhibit little hysteresis and are considered far less costly than other sensing devices. Compared to piezofilm, the FSR is far less sensitive to vibration and heat.

41.7.1 Magnetoresistive Force Sensors

The principle of *magnetoresistive force sensors* is based on the fact that metals, when cooled to low temperatures, show a change of resistivity when subjected to an applied magnetic field. Bismuth, in particular, is quite sensitive in this respect. In practice, these devices are severely limited because of their high sensitivity to ambient temperature changes.

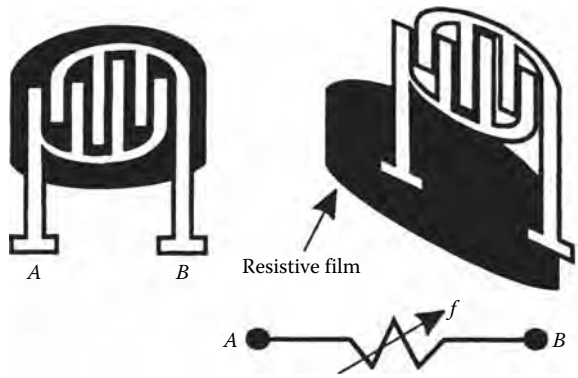


FIGURE 41.13 Diagram of a typical FSR.

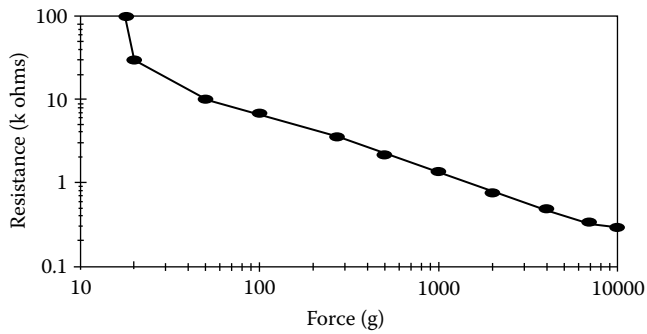


FIGURE 41.14 Resistance as a function of force for a typical FSR.

41.8 Magnetoelastic Force Sensors

Magnetoelastic transducer devices operate based on the Joule effect; that is, a ferromagnetic material is dimensionally altered when subjected to a magnetic field. The principle of operation is as follows: Initially, a current pulse is applied to the conductor within the waveguide. This sets up a magnetic field circumference wise around the waveguide over its entire length. There is another magnetic field generated by the permanent magnet that exists only where the magnet is located. This field has a longitudinal component. These two fields join vectorally to form a helical field near the magnet that, in turn, causes the waveguide to experience a minute torsional strain or twist only at the location of the magnet. This twist effect is known as the *Wiedemann effect* [8].

Magnetoelastic force transducers have a high frequency response (on the order of 20 kHz). Some of the materials that exhibit magnetoelastic include Monel metal, Permalloy, Cekas, Alfer, and a number of nickel–iron alloys. The disadvantages of these transducers include (1) the fact that excessive stress and aging may cause permanent changes, (2) zero drift and sensitivity changes due to temperature sensitivity, and (3) hysteresis errors.

41.9 Torsional Balances

Balancing devices that utilize the deflection of a spring may also be used to determine forces. *Torsional balances* are equal arm scale force-measuring devices. They are comprised of horizontal steel bands instead of pivots and bearings. The principle of operation is based on force application on one of the arms that will deflect the torsional spring (within its design limits) in proportion to the applied force. This type of instrument is susceptible to hysteresis and temperature errors and therefore is not used for precise measurements.

41.10 Tactile Sensors

Tactile sensors are usually interpreted as a touch sensing technique. Tactile sensors cannot be considered as simple touch sensors, where very few discrete force measurements are made. In tactile sensing, a force "distribution" is measured using a closely spaced array of force sensors.

Tactile sensing is important in both grasping and objects identification operations. Grasping an object must be done in a stable manner so that the object is not allowed to slip or damaged. Object identification includes recognizing the shape, location, and orientation of a product, as well as identifying surface properties and defects. Ideally, these tasks would require two types of sensing [9]:

1. Continuous sensing of contact forces
2. Sensing of the surface deformation profile

These two types of data are generally related through stress–strain relations of the tactile sensor. As a result, almost continuous variable sensing of tactile forces (the sensing of the tactile deflection profile) is achieved.

41.10.1 Tactile Sensor Requirements

Significant advances in tactile sensing are taking place in the robotics area. Applications include automated inspection of surface profiles, material handling or parts transfer, parts assembly, and parts identification and gaging in manufacturing applications and fine manipulation tasks. Some of these applications may need only simple touch (force–torque) sensing if the parts being grasped are properly oriented and if adequate information about the process is already available.

Naturally, the main design objective for tactile sensing devices has been to mimic the capabilities of human fingers [9]. Typical specifications for an industrial tactile sensor include

1. Spatial resolution of about 2 mm
2. Force resolution (sensitivity) of about 2 g
3. Maximum touch force of about 1 kg
4. Low response time of 5 ms
5. Low hysteresis
6. Durability under extremely difficult working conditions
7. Insensitivity to change in environmental conditions (temperature, dust, humidity, vibration, etc.)
8. Ability to monitor slip

41.10.2 Tactile Array Sensor

Tactile array sensors (Figure 41.15) consist of a regular pattern of sensing elements to measure the distribution of pressure across the fingertip of a robot. The 8 × 8 array of elements at 2 mm spacing in each direction provides 64 force-sensitive elements. Table 41.1 outlines some of the characteristics of early tactile array sensors. The sensor is composed of two crossed layers of copper strips separated by strips of thin silicone rubber. The sensor forms a thin, compliant layer that can be easily attached to a variety of fingertip shapes and sizes. The entire array is sampled by computer.

A typical tactile sensor array can consist of several sensing elements. Each element or taxel (Figure 41.16) is used to sense the forces present. Since tactile sensors are implemented in applications where sensitivity providing semblance to human touch is desired, an elastomer is utilized to mimic the human skin. The elastomer is generally a conductive material whose electric conductivity changes locally when pressure is applied. The sensor itself consists of three layers: a protective covering, a sheet of conductive elastomer, and a printed circuit board. The printed circuit board consists of two rows of two "bull's-eyes," each with conductive inner and outer rings that compromise the taxels of the sensor.

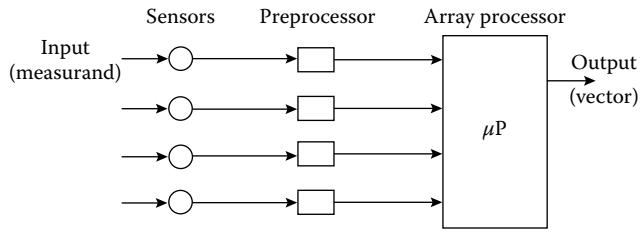


FIGURE 41.15 Tactile array sensor.

TABLE 41.1 Summary of Some of the Characteristics of Early Tactile Array Sensors

	Size of Array		
Device Parameter	(4 × 4)	(8 × 8)	(16 × 6)
Cell spacing (mm)	4.00	2.00	1.00
Zero-pressure capacitance (fF)	6.48	1.62	0.40
Rupture force (N)	18.90	1.88	0.19
Max. linear capacitance (fF)	4.80	1.20	0.30
Max. output voltage (V)	1.20	0.60	0.30
Max. resolution (bit)	9.00	8.00	8.00
Readout (access) time (μs)	—	<20	—

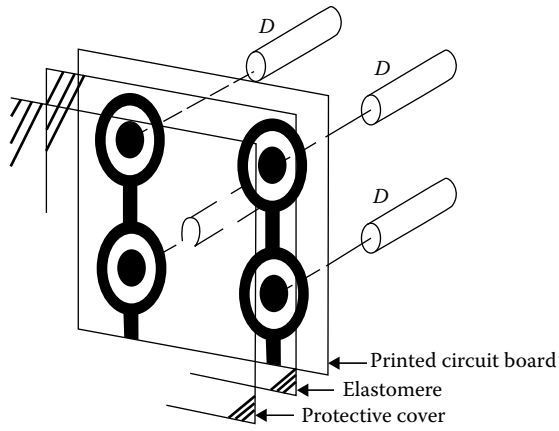


FIGURE 41.16 Typical taxel sensor array.

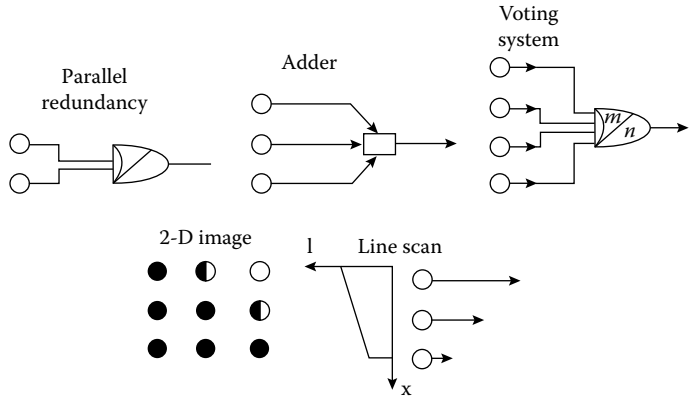


FIGURE 41.17 General arrangement of an intelligent sensor array system. (From Gardner, J.W., *Microsensors Principles and Applications*, John Wiley & Sons, New York, 1995.)

The outer rings are connected together and to a column-select transistor. The inner rings are connected to diodes (D) in Figure 41.16. Once the column in the array is selected, the current flows through the diodes, through the elastomer, and thence through a transistor to ground. As such, it is generally not possible to excite just one taxel because the pressure applied causes a local deformation in neighboring taxels. This situation is called *cross talk* and is eliminated by the diodes [10].

Tactile array sensor signals are used to provide information about the contact kinematics. Several feature parameters, such as contact location, object shape, and pressure distribution, can be obtained. The general layout of a sensor array system can be seen in Figure 41.17. An example of this is a contact and force-sensing finger. This tactile finger has four contact sensors made of piezoelectric polymer strips on the surface of the fingertip that provide dynamic contact information. A strain-gage force sensor provides static grasp force information.

References

1. M. C. Shaw, *Metal Cutting Principles*, Oxford, U.K.: Oxford Science Publications, Clarendon Press, 1989.
2. E. O. Doebelin, *Measurement Systems, Application and Design*, 4th edn., New York: McGraw-Hill, 1990.
3. J. W. Dally, W. F. Riley, and K. G. McConnel, *Instrumentation for Engineering Measurements*, New York: John Wiley & Sons, 1984.

4. P. H. Mansfield, *Electrical Transducers for Industrial Measurement*, London, U.K.: The Butterworth Group, 1973.
5. K. H. Martini, Multicomponent dynamometers using quartz crystals as sensing elements, *ISA Trans.*, 22(1), 35–46, 1983.
6. G. Spur, S. J. Al-Badrawy, and J. Stirnimann, Measuring the cutting force in five-axis milling, Translated paper "Zerpankraftmessung bei der funfachsigen Frasbearbeitung," Zeitschrift fur wirtschaftliche Fertigung und Automatisierung 9/93 Carl Hanser, Munchen, Kistler Piezo-Instrumentation, 20.162e 9.94.
7. C. L. Nachtigal, *Instrumentation and Control, Fundamentals and Applications*, Wiley Series in Mechanical Engineering Practice, New York: Wiley Interscience, John Wiley & Sons, 1990.
8. C. W. DeSilva, *Control Sensors and Actuators*, Englewood Cliffs, NJ: Prentice Hall, 1989.
9. J. W. Gardner, *Microsensors Principles and Applications*, New York: John Wiley & Sons, 1995.
10. W. Stadler, *Analytical Robotics and Mechatronics*, New York: McGraw-Hill, 1995.

Further Information

Lindberg, B. and B. Lindstrom, Measurements of the segmentation frequency in the chip formation process, *Ann. CIRP*, 32(1), 17–20, 1983.

Mechanical measurements and Metrology VIDEO LECTURES, Ebookee, Online on 17-02-2011, http://ebookee.org/Mechanical-Measurements-and-Metrology-VIDEO-LECTURES_1056871.html (accessed on July 25, 2012).

Norton, H. N. *Sensor and Analyzer Handbook*, Englewood Cliffs, NJ: Prentice Hall, 1982.

Dan Mihai Stefanescu, *Handbook of Force Transducers: Principles and Components*, Springer, Berlin, Germany, 2011

Sze, S. M. *Semiconductor Sensors*, New York: John Wiley & Sons, 1994.

Tlusty, J. and G. C Andrews, A critical review of sensors for unmanned machining, *Ann. CIRP*, 32(2), 563–572, 1983.

Wright, C. P. *Applied Measurement Engineering: How to Design Effective Mechanical Measurement Systems*, Englewood Cliffs, NJ: Prentice Hall, 1995.

Partial List of Vendors and Suppliers

Kistler North America, 75 John Glenn Dr., Amherst, NY 14228-2119, Phone: 716-691-5100, 888-KISTLER, 888-547-8537 (toll free), Fax: 716-691-5226, www.kistler.com

Mark-10 Corp, 11 Dixon Ave., Copiague, NY 11726, Phone: 631-842-9200, 888-627-5836 (toll free), 888-MARKTEN (toll free), Fax: 631-842-9201, www.Mark-10.com

PCB Piezotronics, Inc., 3425 Walden Ave., Depew, NY 14043-2417, Phone: 716-684-0001, 800-828-8840 (toll free), Fax: 716-684-0987, www.pcb.com

Strainsert Co., 12 Union Hill Rd., West Conshohocken, PA 19428, Phone: 610-825-3310, 800-529-1502 (toll free), Fax: 610-825-1734, www.strainsert.com

Transducer Techniques, Inc., 42480 Rio Nedo, Temecula, CA 92590, Phone: 951-719-3965, 800-344-3965 (toll free), Fax: 951-719-3900, www.transducertechniques.com

Tekscan, 307 W. First St., South Boston, MA 02127, Phone: 617-464-4500, 800-248-3669 (toll free), Fax: 617-464-4266, www.tekscan.com



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

42

Angle Measurement

42.1 Angle Gage Blocks **42**-1
42.2 Clinometers **42**-5
42.3 Optical Comparator **42**-5
42.4 Protractor **42**-5
42.5 Sine Bar **42**-5
42.6 Sine Plate **42**-6
42.7 Taper **42**-7
Further Information **42**-7

Robert J. Sandberg
University of Wisconsin–Madison

An *angle* is defined as the figure formed by two lines or planes that intersect one another. Such lines or planes may be real, such as the edges and surfaces of an object, or they may be defined, such as lines from an observer to two distant stars.

The units of measurement of angle are degrees (°) (1° is defined as 1/360 of a circle) and radians (rad) (1 rad is defined as $1/(2\pi)$ of a circle). One radian is equal to 57.29578°, and small angles may be expressed in the unit of milliradians (1×10^{-3} rad). A degree of angle is further divided into 60′ (min), and 1′ of angle is divided into 60″ (s). One second of angle, being 1/1,296,000 of a circle, is a very small unit of measurement when related to manufactured parts but is a significant unit when related to much larger dimensions such as the Earth (1″ of angle equals approximately 30 m of a great circle) or in space travel (an included angle of 1″ represents about 9 km on the surface of the moon when it is observed from the Earth during its closest approach to Earth.)

Many terms are used to describe angles in many different fields of expertise. Table 42.1 lists some of these terms, along with very basic definitions.

Many devices and instruments are used to measure or set angles. The following paragraphs describe the variety of equipment involved. See Table 42.2 for a partial list of manufacturers and suppliers of this equipment. See Table 42.3 for a partial list of specific models and approximate prices of angle measurement devices and systems.

42.1 Angle Gage Blocks

Angle gage blocks are right triangle-shaped, hardened, and ground steel; flat; and about 7 mm thick and 60 mm long. They are supplied in sets that include blocks with one of the acute angles equal to 1°, 2°, 3°, 4°, 5°, 6°, 7°, 8°, 9°, 10°, 15°, 20°, or 30°. These blocks can be used in combination to set work pieces or measure objects in 1° increments. Other sets include blocks with as small as 1″ steps. Special angle gage blocks can be made to any acute angle with the aid of a sine bar and thickness gage blocks.

Angle gage blocks provide a durable, simple, and inexpensive method for measuring and setting angles, for example, positioning work pieces in a vice or fixture prior to a machining operation such as milling, drilling, or grinding.

Because they are not adjustable and made of hardened steel, their accuracy can be assumed by simple observation of their physical condition. Look for indications of wear, nicks, or dents before using.

TABLE 42.1 Defining Terms Relating to Angles

Term	Definition
Angle	A figure formed by two lines or planes that intersect one another
Acute angle	An angle less than 90°
Azimuth	The horizontal angle measured along the Earth's horizon, between a fixed reference (usually due south) and an object
Bank	A lateral inclination
Circle	A closed plane curve where all of its points are the same distance from its center point
Declination = declivity	A negative slope
Degree	Equal to 1/360 of a circle
Goniometer	An instrument for measuring angles (from the Greek word *gonio*)
Incline = slope = bias = slant = gradient = grade	The deviation, plus or minus, from horizontal as defined by gravity
Latitude	An angle measured north or south from the equator on a meridian to a point on the earth
Lean = List = Tilt	The deviation from vertical as defined by gravity
Longitude	The angle between the prime meridian (through Greenwich, England) and the meridian of a given place on Earth. This angle is defined as positive moving west
Milliradian	An angle equal to 1/1000 rad
Minute	An angle equal to 1/60°
Oblique angle	An obtuse or acute angle
Obtuse angle	An angle greater than 90°
Quadrant	One quarter of a circle (90°)
Radian	The angle subtended by an arc of a circle equal to the radius of that circle. One radian is equal to 57.29578°.
Rake	Equals the deviation in degrees from being perpendicular (90°) to a line or a plane
Right angle	An angle of 90°
Rise	A positive incline
Second	An angle equal to 1/60′ (1/3600°)
Straight	An angle equal to 180°
Taper	The change in diameter or thickness per unit length of axis
Twist	The angle of turn per unit length of axis, as in a gun barrel or a screw thread

TABLE 42.2 A Partial List of Manufacturers and Supplies of Angle Measurement Equipment

Company	Address
Flexbar Machine Corporation (Representative for Erich Preissr & Co., West Germany)	250 Gibbs Road, Islandia, NY 11722–2697, Tel: (800) 879–7575
Fred V. Fowler Co., Inc.	66 Rowe Street, P.O. Box 299, Newton, MA 02166, Tel: (617) 332–7004
L. S. Starrett Company	121 Crescent Street, Athol, MA 01331, Tel: See local distributor
Brown & Sharpe Mfg. Co.	931 Oakton Street, Elk Grove Village, IL 60007, Tel: (312) 593–5950
Swiss Precision Instruments, Inc., SPI	2206 Lively Blvd, Elk Grove Village, IL 60007, Tel: (708) 981–1300
Edmund Scientific Co.	101 East Gloucester Pike, Barrington, NJ 08007–1380, Tel: (609) 573–6250

TABLE 42.3 Instruments and Devices Used to Measure or Indicate Angles

Type	Manufacturer	Model	Description	Approx. Price
Sine bar	Flexbar	16292	5 in. × 15/16 in. wide	$130.00
		16293	10 in. × 1 in. wide	
		16294	5 in. × 2 in. wide	
		12202	5 in. × 1 in. wide, economy	$30.00
	Fowler	52-455-010	5 in. center to center, 15/16 in. wide	
		52-455-015	10 in. C. to C., 1 in. wide	
		52-455-030	2.5 in. C. to C., 1 in. wide	
	SPI	30-712-4	10 in. C. to C., universal bench center	$3048.00
		98-379-1	5 in. C. to C., 1 in. wide, accuracy between rolls = 0.0003 in.	$31.00
		30-091-3	10 in. C. to C., 1 in. wide, accuracy between rolls = 0.0001 in.	$203.00
	Brown & Sharpe	598-291-121-1	5 in. C. to C., 1 in. wide	
		598-293-121-1	10 in. C. to C., 1 1/8 in. wide	
Sine plate	Flexbar	14612	5 in. C. to C., 6 in. × 3 in. × 2 in.	$320.00
		14615	10 in. C. to C., 12 in. × 6 in. × 2 5/8 in.	$1000.00
	Fowler	57-374-001	5 in. C. to C., 6 in. × 3 in. × 2 in.	
		57-374-004	10 in. C. to C., 12 in. × 6 in. × 2 5/8 in.	
	SPI	77-026-3	10 in. C. to C., 12 in. × 6 in. × 2 5/8 in.	$872.00
	Brown & Sharpe	599-925-10	10 in. C. to C., 12 in. × 6 in. × 2 3/8 in.	
Compound sine plate	Flexbar	14616	5 in. C. to C., 6 in. × 6 in. × 3 1/8 in.	$1100.00
	Fowler	57-375-001	5 in. C. to C., 6 in. × 6 in. × 3 1/8 in.	
	SPI	7-072-7	5 in. C. to C., 6 in. × 6 in. × 3 1/8 in.	$926.00
	Brown & Sharpe	599-926-5	5 in. C. to C., 6 in. × 6 in. × 3 1/2 in.	
Angle computer	Flexbar	19860	3-axis with vernier protractors	$3750.00
Protractor—direct	Flexbar	16337	Rectangular head, 0°–180°	$25.00
	Starrett	RP1224W	Head only. To fit 12, 18, and 24 in. blades	
		C183	Rectangular head, 0°–180° 6 in. blade	
	SPI	30-393-3	Rectangular head, 0°–180°	$23.00
		31-804-8	Head only. To fit 12 in., 18 in., and 24 in. blades	$39.00
Protractor—vernier	Flexbar	16339	360° range, 1′ reading with magnifier, 12 in. and 6 in. blades incl.	$400.00
		16338	360° range, 5′ reading	$75.00
	Starrett	C364DZ	12 in. blade, 0°–90° range through 360°, 5′ graduations	
	SPI	30-395-8	6 in. blade, 0°–90° range through 360°, 5′ graduations	$65.00
		30-390-9	6 in. and 12 in. blades, 0°–90° range through 360°, 1′ graduations with magnifier	$540.00
	Brown & Sharpe	599-490-8	8 in. blade, 0°–90° range through 360°, magnifier optional	

(*continued*)

TABLE 42.3 (continued) Instruments and Devices Used to Measure or Indicate Angles

Type	Manufacturer	Model	Description	Approx. Price
Protractor, digital (inclinometer)	Flexbar	17557	±45° range, ±0.1° resolution	$260.00
		17556	±60° range, ±0.01° resolution, SPC output	$450.00
	Fowler	54-635-600	±45° range, ±0.01° resolution, RS232 output available	
	SPI	31-040-9	±45° range, resolution: ±0.01° (0 to ±10°), 0.1° (10° to 90°)	$329.00
Protractor, dial bevel		30-150-7	8 in. blade, 1 3/8 in. diameter dial, geared to direct read to 5′	$527.30
	Brown & Sharpe	599-4977-8	8 in. blade, dial read degrees and 5′	
Square—reference	SPI	30-392-5	90° fixed angle	$55.00
Optical comparator (projector)	Fowler	53-912-000	12 in. screen diameter, 10× , 20× , 25× lens available, horizontal beam, with separate light source for surface illumination	
	Starrett	HB350	14 in. screen diameter, 10× , 20× , 25×, 31.25× , 50× , 100× lens available, horizontal beam	
		VB300	12 in. screen diameter, 10× through 100× lens available, vertical beam	
		HS1000	40 in. screen diameter, 10× through 100× lens available, horizontal beam	
	SPI	40-350-1	14 in. screen diameter, 10, 20× , 50× lens available, horizontal beam	$2995.00
Optical comparator (handheld)	SPI	40-145-3	10× magnification, pocket style	$57.50
			Additional reticles	$11.00
		40-140-6	7× magnification, pocket style with illuminator	$62.50
			Additional reticles	$10.50
	Edmund Scientific	A2046	6× magnification, pocket style, 360° protractor reticle, 1° increments	$58.75
Angle plate	Fowler	52-456-000	Set of 2, 9/32 in. thick, steel, 30° × 60° × 90°, 45° × 45° × 90°	
	SPI	98-463	Set of 2, 5/16 in. thick, steel, 30° × 60° × 90°, 45° × 45° × 90°	$32.00
Angle positioning Block	SPI	70-997-2	0° to 60°, 10′ vernier (for setting work piece in a vice)	$122.00
Angle gage	Fowler	52-470-180	18 leaves, spring steel, 1°–10°, 14°, 14.5°, 15°, 20°, 25°, 30°, 35°, and 45°	
	SPI	31-375-9	18 gage set, 5°–90° in 5° steps, 5′ accuracy	$49.60
Angle gage blocks	Starrett	Ag18.TR	18 block set, use in combination for steps of l″, 1″ accuracy	
	Starrett AG16.LM		16 block set, use in combination for steps of l″, 1/4″ accuracy	
	SPI	30-140-8	10 block set, 1°, 2°, 3°, 4°, 5°, 10°, 15°, 20°, 25°, 30°, accuracy = ±0.000l″/in.	$170.00
			1/4° and 1/2° blocks optional (each)	$18.00

42.2 Clinometers

A *clinometer* is an electronic device that measures vertical angle with respect to gravitational level. It is rectangular, with each side being 90° to its adjacent sides. With a range of readings of at least ±45°, this shape allows measurements up to a full 360°. Floating zero can be set anywhere and resolutions of ±0.01° are obtainable. Some models will convert readings to inches per foot, percentage of grade, and millimeters per meter (mm m^{-1}).

A clinometer can be used wherever the angle of a surface with respect to gravity or another surface needs to be measured. High accuracy and resolution are obtainable, but calibration should be checked periodically with respect to a known level surface and a known angle. Surfaces that are remote to one another or have an intervening obstruction pose no problem for a clinometer.

42.3 Optical Comparator

An *optical comparator* measures angles, along with other dimensions or profiles, by referencing a visual image (usually magnified) of an object to a reticule that is calibrated in the measurement units desired. A handheld optical comparator is placed directly over the object to be measured, and the operator's eye is moved to the proper distance above the comparator for good focus of the magnified image. Some models contain a battery- or line-powered light source. Reticules for these handheld devices are generally graduated in 1° increments.

Projection-type optical comparators are available as bench or floor models and are made for either horizontal or vertical beam viewing. They use a high-intensity light source and magnifying optics to display an image of an object onto a rear projection, frosted glass screen that is inscribed with angular as well as linear markings. The image displayed is the result of light being projected past the object, referred to as a shadow graph, or of light being reflected off the surface of the object. The method used is determined by the shape of the viewed object and its surface quality.

Magnification of the optical system in these devices can range from 10× by 100×, with screen diameters ranging from 0.3 to 1 m.

These instruments are useful for measuring profiles of parts after final matching for the purpose of quality control or duplication.

As the name implies, the image that is projected can be superimposed on a mask or outline drawing placed directly on the view screen so that any deviations from the required shape can easily be determined. Optical comparators are heavy, nonportable devices that require a fairly high amount of maintenance and are best used in a fairly dark room.

42.4 Protractor

A *protractor* is an instrument used for measuring and constructing angles. A direct-reading protractor usually is graduated in 1° increments and can be semicircular or circular in shape. The simplest models are of one-piece construction and made from metal or plastic. Other models include a blade or pointer pivoted in the center of the graduated circle.

More precise protractors are equipped with a vernier scale that allows an angle to be indicated to 5′ of arc. See Figure 42.1 for an explanation of how to read such a vernier scale.

42.5 Sine Bar

A *sine bar* is a device used to accurately measure angles or to position work pieces prior to grinding and other machining procedures. It is constructed from precisely hardened and ground steel rectangular bar to which two hardened and ground steel cylindrical rods of the same diameter are attached. The axis of each rod is very accurately positioned parallel to the other and to the top surface of the bar.

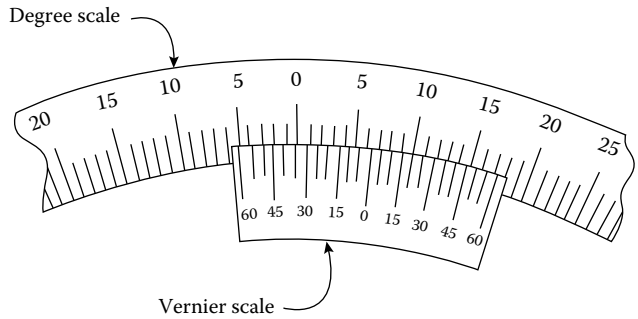


FIGURE 42.1 Vernier protractor. If the zero mark on the vernier scale is to the right of the zero mark on the main scale, as shown in this drawing, then the right side of the vernier scale must be used. Look for the mark on the vernier scale that best aligns with one of the marks of the protractor degree scale. Count the number of marks on the vernier scale from the zero mark to this mark. Each mark thus counted is, in this example, equal to 5′ of arc, and, therefore, the number of minutes to be added to the number of degrees indicated is 5 times the vernier marks counted. (In this example, the fourth mark aligns the best with the main scale indicating 20′). The number of degree indicated is the degree mark just to the left of the zero mark on the vernier scale. The left side of the vernier is similarly used when the indicated angle is to the left of the zero mark on the degree scale.

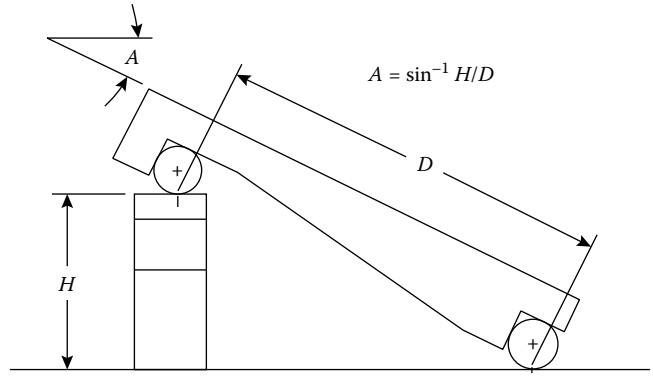


FIGURE 42.2 Sine bar. A sine bar is used in conjunction with precision gage blocks that are placed under one of the cylindrical rods, raising that end a distance, *H*, equal to the sine of the desired angle times the distance, *D*, between the two rods.

A sine bar is used in conjunction with precision gage blocks that are placed under one of the two cylindrical rods to raise that rod above the other rod a distance *H* (see Figure 42.2) equal to the sine of the angle desired, times the distance *D* between the two rods. The standard distance between the rods is 250 mm (5 in.) or 500 mm (10 in.). The governing equation in using a sine bar is $\sin A = H/D$.

A work piece positioned using a sine bar is usually secured with the use of a precision vice. The vice may clamp directly to the work piece or, when using a sine bar that has tapped holes on its top surface, to the sine bar sides with the work piece bolted to the top of the sine bar.

42.6 Sine Plate

A variation of the sine bar is the *sine plate*. A sine plate consists of the three elements of a sine bar plus a bottom plate and side straps used to lock the plate in the desired position. In addition, one of the ground steel rods is arranged to form a hinge between the top and bottom plates. When using a sine plate, a work piece is secured to the top plate using bolts or clamps, and the bottom plate is secured to a machine tool table using clamps or a magnetic chuck.

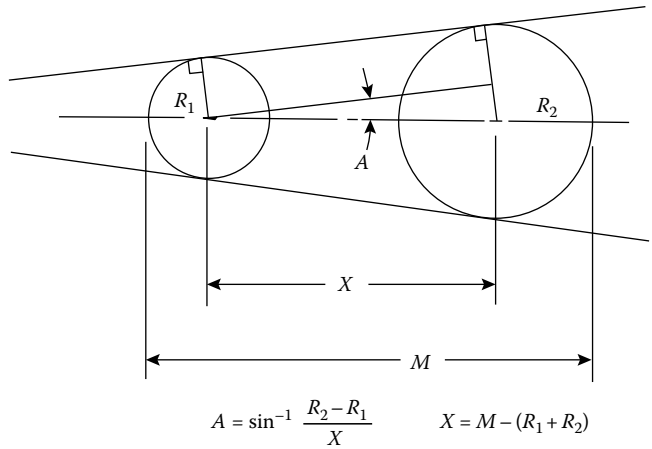


FIGURE 42.3 Measuring tapers. Two balls (for holes) or gage pins (for slots) should be selected to fit the hole size to be measured. The distance between the balls or pins should be as large as possible to allow for the best accuracy of measurement. The position of the balls can be determined and the equations shown in this figure applied to be determined the angle of the taper.

Compound sine plates are bidirectional, allowing angles to be measured and set in each of two orthogonal planes (true compound angles).

42.7 Taper

An accurate method that can be used to measure a tapered hole is described in Figure 42.3. In this method, one uses a pair of steel balls of proper diameters to match the hole size and angle of taper. This method can also be used to measure the angle between two adjacent planes, using either balls or gage pins.

Further Information

Further information on this subject can be found in company catalogs in both print and Internet formats. An additional reference is *Machinery's Handbook*, 25th edn., Industrial Press Inc., 200 Madison Avenue, New York, 10016–4078, 1996.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

43

Mass, Weights, and Instrumentation

43.1 Introduction **43**-1
43.2 Etymology **43**-2
43.3 Definitions of Terms **43**-2
43.4 Theory **43**-3
Quantities Derived From Mass
43.5 Test Instrumentation **43**-5
Classification of Weighing Devices (Scales, Balances, and Mass Comparators) • Mass Measuring Technique • Equal Arm Balance • Scales • Balances and Mass Comparators
43.6 Weighing Methods **43**-8
Direct Reading on an Equal Arm Balance • Direct Reading on Single-Pan Optical-Mechanical and Electronic Balances • Substitution Method • Transposition Method
43.7 Calibration Standards **43**-12
Materials • Design • Maximum Permissible Error • Handling, Storage, Packing, and Shipping
43.8 Range and Accuracy **43**-14
Range • Stability • Inertia • Replaceability • Accuracy (Also Trueness or Justness) • Linearity of a Scale • Sensitivity of a Scale • Mass Traceability • Safety Precautions in Mass Measurement
43.9 Units of Measurement **43**-17
Definition: The Unit of Mass (Kilogram) • Conversion Factors
References **43**-18

Emil Hazarian
California State University

43.1 Introduction

The determination of mass is one of the oldest and most engaged activities of the mankind. Ancient primitive forms of trade as well as today's complex global trade exchanges are employing mass measurements to a high extent. Nevertheless, the physical notion of mass is still not completely clear when considering its triple use in atomic models (microstructures), regular size objects and astronomic models (macrostructures).

Weighing technology today permits detection of differences between 1 kg weights, smaller than 1 μg (microgram) with an uncertainty of 10^{-16} kg.

The unit of mass is materialized in weights. The instruments used to determine the mass are called scales, balances, and mass comparators.

43.2 Etymology

The word *mass* comes from the Latin word *massa* meaning bulge of shapeless material that in turn comes from the Greek *massein* or *maza*. Other older and different meanings may be assumed including the religious meanings of the word; also new meanings are used today such as mass media or mass production.

43.3 Definitions of Terms

Measurement accuracy (accuracy of measurement, accuracy): Closeness of agreement between a measured quantity value and a true quantity value of a measurand [4].

Apparent mass: See conventional mass.

Balance: A measuring device designed to evaluate the mass of the objects; in everyday vocabulary, a balance is used to determine the weight of an object. A balance is usually of higher precision.

Buoyancy: A force that acts on any body located in a liquid or gas by operating in a direction opposed to the force of gravity and thus causes an apparent reduction of weight. According to the Archimedean principle, the buoyancy of a body is equal to the weight force of the amount of gas or liquid displaced.

Air buoyancy varies in the United States from 0.98 mg/cm^3 to 1.21 g/cm^3; air is less dense at higher altitude.

Calibration: The set of operations that establish, under specified conditions, the relationship between values indicated by measuring instrument or measuring system, or values represented by a material measure, and the corresponding known value of measurand. Sometimes implies the adjustment of the measurand.

Conventional mass: Sometimes called mass in air, or apparent mass; this is the value of the mass required to balance the object in conventionally chosen conditions of air density of 1.2 kg/m^3, temperature of 20 °C, and atmospheric pressure of 760 mmHg [7].

If the mass required to balance the object is made out of brass with a conventional density of 8390.94 kg/m^3 at 20 °C, we obtain the conventional mass (or apparent mass) versus brass of that object.

If the mass required to balance the object is made out of stainless steel with a conventional density of 8000 kg/m^3 at 20 °C, we obtain the conventional mass (or apparent mass) versus stainless steel of that object.

(See also true mass)

Correction: The value that added algebraically to the uncorrected results of a measurement compensates for an assumed systematic error. **Note**: The correction is equal to the assumed systematic error but has opposite sign.

Drift: The slow variation with time of a metrological characteristic of a measuring instrument. In the case of a balance or scale, it is the change of the indications while the load remains constant.

Mass: The state scalar physical quantity expressing both the inertial property of the matter and the property of the matter to generate a field of attraction forces (gravitational field). This is also called true mass or mass in vacuum.

Numerically the mass is expressed as the ratio between a force and a material object and its resulting acceleration.

Mass comparator: A weighing device used only as an intermediary between weights with a known value (standards) and weights with unknown value with the purpose of determining the unknown value. **Note**: any scale or balance may be used as a mass comparator.

Scale division: The part of a scale between any two successive scale marks.

Scale interval: The difference between the scale values corresponding to two successive scale marks.

Scale spacing: The distance between any two successive scale marks measured along the same line as the scale length (expressed in units of length).

Sensitivity: The change in the response of a measuring instrument divided by the corresponding change in the stimulus.

Tolerance: The maximum permissible variation or limits allowed a quantity or parameter from a specified value.

True mass: The mass of an object determined in vacuum. The true mass of an object can be calculated from the conventional mass, by applying the buoyancy correction. True mass value is usually determined for special applications such as in nuclear field and biotechnology; it implies weights of ASTM class 1.

True value (*of a quantity*): The value that characterizes a quantity perfectly defined, in the conditions that exist when the quantity is considered. **Note**: The true value is an ideal concept and in general cannot be known exactly.

Uncertainty of measurement: An estimate characterizing the range of values within which the true value of a measurand lies. **Note**: Uncertainty of measurement comprises, in general, many components; some of these components may be estimated on the basis of the statistical distribution of the results of series of measurements and can be characterized by experimental standard deviations; estimates of other components can only be based on scientific judgment, experience, or other information.

Universality (replaceability): The quality of a measuring instrument (e.g., scale) is replaceable by another measuring instrument having the same characteristics.

43.4 Theory

Mass is one of the base quantities in classical mechanics. The concept of mass constitutes a universal characteristic of bodies. As a physical quantity, the mass could be determined in relationship with other physical quantities. Such relationships are given by Newton's second law of motion and the law of universal gravitation.

The mathematical expression of Newton's second law of motion is

$$\Sigma F = ma \tag{43.1}$$

where

ΣF is the summation of all the external forces applied to the particle
a is the acceleration of the particle
m is the mass of the particle

From the preceding equation we can write

$$m = \frac{F}{a} \tag{43.2}$$

The mass is numerically determined by the ratio between force and acceleration.

Newton's law of gravitation, published in 1686, may be stated: Every particle of matter in the universe attracts every other particle with a force that is directly proportional to the product of

the masses of the particles and inversely proportional to the square of the distance between them. Or in a mathematical form

$$F = k\frac{m_1 m_2}{r^2} \tag{43.3}$$

where

F is the attraction force between the two bodies m_1 and m_2
m_1, m_2 is the masses of the two bodies
r is the distance between the centers of the two bodies
k is the gravitational constant ($k = 6.670 \times 10^{-11}$ N m^2 kg^{-2})

An immediate application of the universal attraction to be observed on earth is the weight of the bodies. The force with which an object is attracted to the earth is called its weight. Weight is different from mass, which is a measure of the inertia an object displays. Although they are different physical quantities, mass and weight are closely related.

The weight of an object is the force that causes it to be accelerated when it is dropped. All objects in free fall near the earth surface have a downward acceleration of g, the acceleration of gravity. The value of g varies in different locations of the earth surface, due to the distance to the center of the earth. One value used in common calculations is 9.8 ms^{-2}. If the object's mass is m, then the downward force on it, which is the weight w, can be found from the second law of motion, where $F = W$ and $a = g$.

Evidently we have

$$W = mg \tag{43.4}$$

The weight of any object is equal to its mass multiplied by the acceleration of gravity. Since, for practical purposes, g is a constant near the earth's surface, the weight w of an object is always directly proportional to its mass m: a large mass is heavier than a small one. For more accurate applications, the determination of the gravitational acceleration at the very location of the experiment is required.

The mass of an object is a more fundamental property than its weight, because its mass when at rest is constant and is the same everywhere in the universe whereas the gravitational force on it depends upon its position relative to the earth or to some other astronomical body (at the earth's poles $g = 9.831$ ms^{-2}, which is the maximum value, and at the equator $g = 9.781$ ms^{-2}, which is the minimum value).

Mass remains a constant quantity only in classical mechanics. The modern physics, particularly the theory of relativity, shows that the mass of a body increases with the increase of its velocity. The variation of mass with velocity follows the formula:

$$m = \frac{m_0}{\sqrt{1-(v^2/c^2)}} \tag{43.5}$$

The symbols have the following meanings:

m_0 is the mass measured when object is at rest ("rest mass")
m is the mass measured when the object is in relative motion
v is the velocity of relative motion
c is the velocity of light

Relativistic mass increases only at velocities approaching that of light. For velocities smaller than 10^6 m/s, the difference between m and m_0 is insignificant.

43.4.1 Quantities Derived From Mass

Here are some selected quantities derived from mass:

- Density
- Force
- Pressure
- Electric current
- Voltage
- Electric resistance

43.5 Test Instrumentation

43.5.1 Classification of Weighing Devices (Scales, Balances, and Mass Comparators)

Weighing devices are the measuring devices exclusively designed to determine the mass of an object.

The classification of weighing devices could be accomplished according to the following criteria [2]:

- The mass measuring technique
- The extent of operator participation in the weighing process
- Installation
- The method of obtaining the equilibrium position
- The type of the load-receiving element: pan, platform, bucket, rail, hook, conveyer belt, etc.

43.5.2 Mass Measuring Technique

According to this criterion, we have direct weighing devices and indirect weighing devices.

The devices determining directly the mass of an object are those devices employing levers in their construction. The mass is determined by the equality between the moments created by the measurand and the standard weight. This type of weighing devices is called *balances* if they have a single lever, with equal or unequal arms. If they employ multiple levers, they are called *scales* and are used for weighing voluminous and heavy objects (i.e., 50,000 lb).

The weighing devices determining indirectly the mass are based on one the following principles:

- Elastic element (flexure, torsion, compression, traction, etc.)
- Hydraulic element (variation of a liquid pressure)
- Electronic

43.5.2.1 Operator Participation in the Weighing Process

According to this criterion, we have manual, semiautomated, and automated weighing devices.

A manual weighing device is that type where the operator does all three main operations, loading, weighing, and unloading. Examples are equal arm balances, platform scales, and truck scales.

A semiautomated weighing device is loaded and unloaded manually, but the weighing operation is automated. The most common are those scales with dial indicator.

An automated weighing device has all three operations loading, weighing, and unloading completed without the operator intervention.

Examples are the belt conveyer scale and new generation of electronic balances.

43.5.2.2 Installation

According to this criterion, weighing devices are classified into

- Portable
- Stationary (stable)

43.5.2.3 Method of Obtaining the Equilibrium Position

Based on this criterion, we have two types of weighing devices:

- With constant equilibrium position (i.e., two indicators, one fix and one mobile, and two mobile indicators)
- With variable equilibrium position (i.e., the indicator moves in front of a graduated scale, semi-automated weighing devices)

43.5.3 Equal Arm Balance

The equal arm balance is frequently used for precision mass measurements or weighing. To visualize its operation, let us consider for a moment the beam of Figure 43.1, supported at its center by the fulcrum (A). If the system is to be in equilibrium, the sum of all the forces must equal zero.

If the weight of the beam and its load (w_1 and w_2) are considered as the force acting downward, they must be supported by an equal lifting force provided by the fulcrum (A). Thus, there is no rotation.

If the length of the beam on each side of the fulcrum is assumed equal, then

$$x_1 = x_2 \tag{43.6}$$

Furthermore, if the system is in equilibrium, then the torques (forces momentum) about point A must be equal; therefore,

$$w_1 x_1 = w_2 x_2 \tag{43.7}$$

Since the lengths of the arms are assumed equal, then at equilibrium,

$$w_1 = w_2 \tag{43.8}$$

If one of the weights is known (standard), value may be assigned to the other (unknown) weight. This is the basic operating principle of the equal arm balance. Even though the arms of the balance are not exactly equal, precision weighing procedures are available that account for any small inequality.

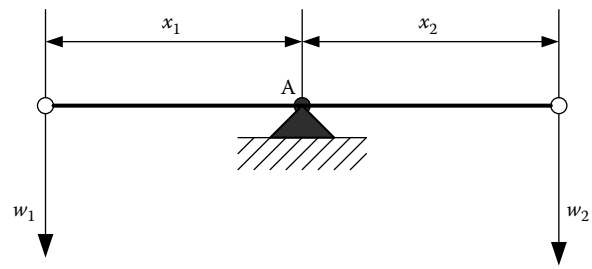


FIGURE 43.1 Beam in equilibrium state.

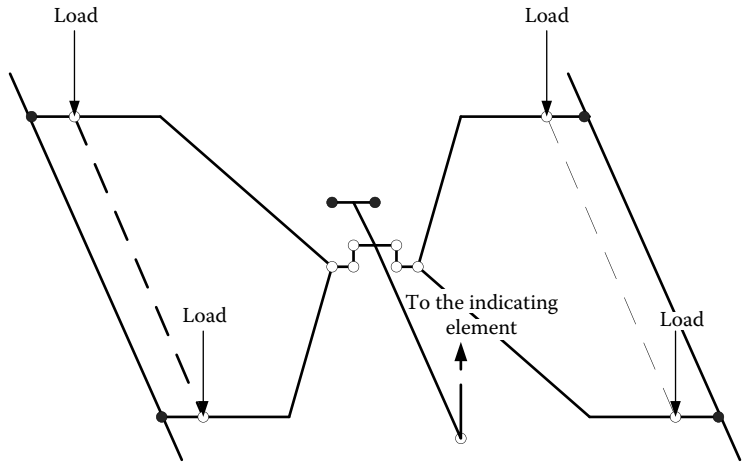


FIGURE 43.2 Typical platform scale lever system. (From Hazarian, E., Some characteristics of weighing devices, Technical paper at *Measurement Science Conference*, 1992, Anaheim, CA.)

43.5.4 Scales

Weighing devices employing multiple levers in their design are called scales. There are a large variety of scales, from which we choose to mention the following categories: platform scales, either portable or built on the floor; vehicle scales; and railroad scales.

The design of scales changed very much from levers and knife-edge joints to levers with flexure joints and lately levers with load cells or just loads cells [1, p. 54]. Nowadays, a scale has three general components: the load-receiving element, the load cell, and an indicating device. Large-capacity scales with load cell elements are achieving higher accuracy than traditional scales employing beams. A typical platform scale lever system is shown in Figure 43.2.

43.5.5 Balances and Mass Comparators

For over three thousand years, the design of balances did not change significantly, mainly because the requirements were unchanged. The last half of the twentieth century changed the situation, and new balances and mass comparators based on electromagnetic force compensation emerged to satisfy new technologies, such as computers, biotechnology and drug manufacturing, and sophisticated accurate weaponry.

43.5.5.1 Top-Loading and Analytical Balances [8, p. 6]

These balances have a single-pan or top-pan load-receiving element [10].

43.5.5.2 Electromagnetic Force Compensation Mass Comparators [10]

These weighing devices are called mass comparators as they perform a comparison between a known weight piece (standard weight) and an unknown weight piece (unknown) in order to determine the mass of the latter. Such devices are designated for high-accuracy mass determinations in the national and primary calibration laboratories.

These weighing devices employ flexure joints in their design, replacing the knife-edge joints. There are no moving parts making these devices practically maintenance-free. Although the electromagnetic force compensation principle was known well before, it is the last generation of mass comparators, created in the last decade, that achieved the accuracy of former equal arm balances combined with electro-optical systems. Another significant advantage of the new mass comparators is the possibility of computer interfacing. Using the computer to record data eliminates one of the writing sources of errors and gives the possibility of instantaneous statistical analysis of real-time data, so critical for making the right decisions.

43.6 Weighing Methods

Selecting the weighing method for your application depends on a series of factors such as

- Required level of accuracy and precision
- Available equipment
- Available trained operators
- Cost-effectiveness

Sometimes a "lower"-level procedure may be sufficient for a specific application, rather than a laborious state-of-the-art intercomparison. Having handy the entire palette of weighing methods will offer a better opportunity for selecting the most appropriate one.

43.6.1 Direct Reading on an Equal Arm Balance

Using an equal arm balance for simple weighing could be accomplished by loading one pan with the measurand (unit to be measured or unknown weight) and the other pan with the necessary weights, until the balance is brought in a normal equilibrium position. This is a lower-level weighing method, for applications that do not require special care with regard to accuracy. In this case, the *mass of the measurand is equal with the total mass of the added weights*.

Same method may proffer a higher accuracy and could be used in industrial laboratories in which case the balance equilibrium position is not approximated but rather determined by reading the balance indicator elongations and by calculation. Bringing the balance in the equilibrium position is a meticulous task, if not impossible. Moreover, it could be considered useless, as soon as we can determine the mass of the measurand even in an intermediate position, close to what is called the normal equilibrium position.

Let us consider the case of an equal arm balance (Figure 43.3) with its beam in horizontal position; therefore, the balance is in a normal equilibrium position (*a*). The second case is when mass Q is heavier than Q_1; therefore, the balance will stop in an inclined position (*b*). The equilibrium equations applied for both situations will give

$$Q_1 \times l = P \times l' \tag{43.9}$$

or

$$m_{Q_1} \times g \times l = m_P \times g \times l' \tag{43.10}$$

We also may write

$$Q \times l \cos\alpha = P \times l' \cos\alpha + G \times a \sin\alpha \tag{43.11}$$

$$m_Q \times g \times l \cos\alpha = m_P \times g \times l' \cos\alpha + m_g \times g \times a \sin\alpha \tag{43.12}$$

If we divide the two equations by g and $g\cos\alpha$, respectively, and consider

$$m_Q - m_{Q_1} = e$$

$$tg\alpha = \frac{d \times n}{l}$$

where

d is the one division on the graduated scale
n is the number of divisions the indicator moved due to the addition to mass m_Q not balanced by m_P
l is the length of the indicator

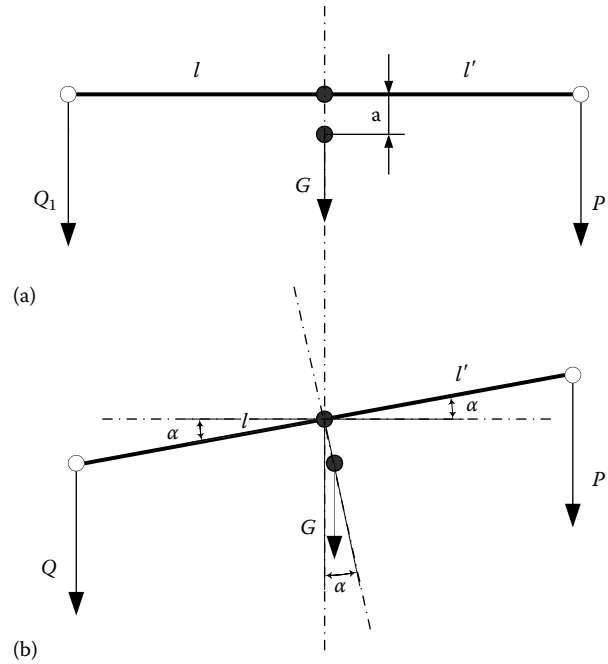


FIGURE 43.3 Direct reading on an equal arm balance: (a) Balance in equilibrium, (b) balance out of equilibrium.

we have

$$m_Q = m_P + S_P \times n \tag{43.13}$$

where

m_Q is the mass of the measurand (unknown)
m_P is the mass of the weights
n is the number of divisions (elongation) of the indicator

$$S_P = m_G \frac{a \times d}{l \times i}, \text{sensitivity}$$

From the preceding equation, we can state that the *mass of the measurand (m_Q) is equal to the algebraic sum of the mass of the weights (m_p) and the product between sensitivity (S_P) and the number of divisions (n), the indicator moved.*

Practically we determine n for every weighing, whereas S_P is determined once for several weighing.

In order to calculate n and S_P, we have to calculate first the equilibrium position L, for which we read the turning points e_i (i = 1, 2, 3) of the balance indicator:

$$L = \frac{e_1 + 2e_2 + e_3}{4} \tag{43.14}$$

The turning point readings (Figure 43.4) are all positive, if the zero graduation is placed at the left end of the graduated scale (unilateral scale). If the zero graduation is placed in the middle of the graduated scale (bilateral scale), the readings will be both positive and negative.

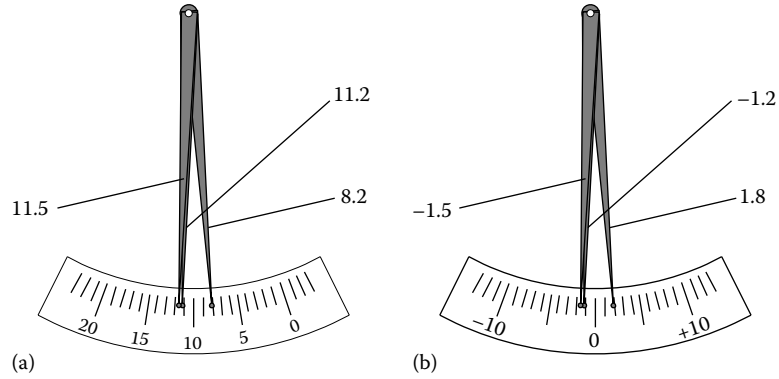


FIGURE 43.4 Example of turning point readings: (a) Unilateral scale, (b) bilateral scale.

Example

The following readings were made on a balance having unilateral scale, the most usual case:

e_1 = 11.5 div.
e_2 = 8.2 div.
e_3 = 11.2 div.

Therefore, $L = (11.5 + 2 \times 8.2 + 11.2)/4 = 9.78$ div.

If the same balance has a bilateral scale, the readings will be

e_1 = −1.5 div.
e_2 = +1.8 div.
e_3 = −1.2 div.

Thus,

$$L = \frac{[(-1.5) + 2(+1.8) + (-1.2)]}{4} = +0.22 \text{ div.}$$

Notice that the two values obtained, 9.78 div. and 0.22 div., are representing the same equilibrium position placed at 0.2 div. to the right of the middle graduation.

Direct reading on an equal arm balance is generally adequate at the industrial laboratories level, a lower level than standard laboratories; however, we still need to have confidence in our measurements. It also requires accuracy and traceability.

$$L = \frac{e_1 + 3e_2 + 3e_3 + e_4}{8} \tag{43.15}$$

43.6.2 Direct Reading on Single-Pan Optical-Mechanical and Electronic Balances

The value of the measurand, in our case the unknown mass, is determined by placing it on the pan of the balance, releasing the pan and read the value of the mass on the balance display. Certainly, we have to verify if the display shows "zero" when the balance is unloaded.

Single-pan optical-mechanical scales provide apparent mass value, depending on balance adjustment. Both single-pan optical-mechanical and electronic scales rely on internal weights of the balance. For that reason, the internal weights should be tested periodically. Also the optical scale should be tested for the same reason.

43.6.3 Substitution Method

The substitution method (Figure 43.5), known also as the Borda method, eliminates the systematic error induced in the weighing process by the unequal arms of the balance beam.

The substitution method consists in the following:

The measurand is placed on the left pan of the balance, while on the right pan, we place tare weights until the balance remains in the normal equilibrium position or close to this position. We replace then the measurand with known weights, adding them until obtaining the same equilibrium position as previously. The mass m_X of the measurand is equal with m_P, the mass of the known weights. In order to demonstrate this, we have to write the equality of the two moments of forces created in the two situations:

$$X\,l = T\,l'$$

$$P\,l = T'\,l$$

also

$$m_X l = m_T l'$$

$$m_P = m_T l'$$

resulting in

$$m_X = m_P \tag{43.16}$$

In practice, the balance may be in equilibrium in any intermediate position, the aforementioned particular case being chosen for convenience purpose only.

The substitution method is largely employed in most of the laboratories performing mass measurements.

1. Single substitution
   In the single substitution method the sequence is as follows:
   a. Standard, S, observation O1
   b. Unknown, X, observation O2
   c. Unknown plus sensitivity, X + sw, observation O3
2. Double substitution
   The measurement sequence in double substitution is as follows:
   a. Standard, S, observation O1
   b. Unknown, X, observation O2
   c. Unknown plus sensitivity, X + sw, observation O3
   d. Standard plus sensitivity, S + sw, observation O4

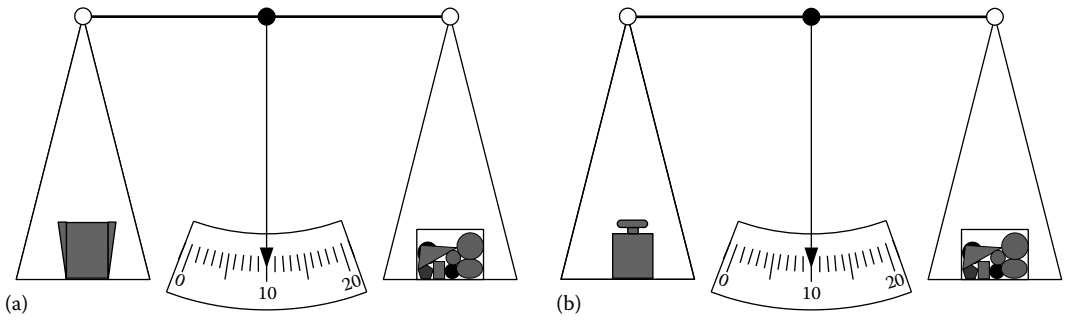


FIGURE 43.5 Substitution method: (a) With measurand, (b) with standard weights.

43.6.4 Transposition Method

The transposition method or Gauss method, after the German mathematician and astronomer Carl Friedrich Gauss (1777–1855), consists of the following:

- The measurand is placed on the left pan of the balance, while on the right pan, we place known weights (standard) until the balance remains in the normal equilibrium position or close to this position (observation O1).
- We permute the measurand and the standard, adding known weights (standards) to the left pan, while the measurand is placed on the right pan (observation O2).
- A sensitivity weight is added to the left pan (observation O3).

43.7 Calibration Standards

Calibration standards for mass measures, called (improperly) weights, represent in the International System of Units (SI), the materialization of the mass unit, the kilogram, its decimal multiples and submultiples, and also their half and double values. The pound unit, in the pound–inch system, derives its value from the kilogram, since there is no international pound mass standard.

Both the SI and pound–inch systems are using the following denominations for weights: 5321, 5221, and 52111. Denominations of weights are selected in order to cover the most possible nominal values with a minimum number of weights. This is particularly important when someone needs to calibrate customary weights using metric standards; in this case because of the conversion factors, the equivalent number of weights composing the standard is larger than initially designed; therefore, the uncertainty of the calibration will be larger versus a one-to-one calibration.

There are three major classifications of weights in use in the United States concentrated in documents generated by three organizations:

1. National Institute of Standards and Technology (NIST), formerly National Bureau of Standards, with the following documents:
   a. Circular 547, Section 1, 1954
      "Precision Laboratory Standards of Mass and Laboratory Weights"
      Note: This classification was withdrawn.
   b. Circular 3, 1918
      "Design and Test of Standard of Mass"
      Note: This classification was withdrawn.
   c. NIST Handbook 105-1, Revised 1990
      "Specifications and Tolerances for Reference Standards and Field Standards Weights and Measures"
   d. Specifications and Tolerances for Filed Standards Weights (NIST Class F),
      Note: The NIST, Weights and Measures Division currently use this classification, for the States Laboratory Program, in Legal Metrology.
2. American Society for Testing Materials, with the following document.
   ASTM E617-97(2008) "Standard Specification for Laboratory Weights and Precision Mass Standards"
3. International Organization for Legal Metrology
   a. OIML R111: 2004
      "Weights of classes E1, E2, F1, F2, M1, M2, M3"

43.7.1 Materials

The following materials are mostly used in the manufacturing of weights:

- Platinum–Iridium (90% Pt, 10% Ir) used for manufacturing the international prototype and national standards. Density: 21.5 g/cm^3
- Stainless steel, density 8.0 g/cm^3
- Tantalum, density 16.0 g/cm^3
- Aluminum, density 2.7 g/cm^3
- Brass, density 8.4 g/cm^3
- Cast iron, density 7.8 g/cm^3

43.7.2 Design

Weights design varies with the nominal value and application. Small weights, from 1 to 500 mg, are leaf shaped with a vertical lip to accommodate handling using fine-tip nonmetallic tweezers.

Fractional pounds (i.e., 0.005 lb) may be shaped as a dish.

Larger weights, up to 50 kg and 100 lb, are usually of cylindrical or rectangular shape with or without adjusting cavity. The adjusting cavity is sealed either with a threaded screw and aluminum seal or a threaded knob.

43.7.3 Maximum Permissible Error

Each weight within a class is assigned a maximum permissible error (MPE) (or tolerance). Unless it is specified otherwise, the maximum permissible error is assigned both plus and minus sign.

Example: ASTM class 6, 1 kg weight has 6 100 mg MPE, meaning the weight value may be anywhere between 1 kg – 100 mg and 1 kg + 100 mg values.

43.7.4 Handling, Storage, Packing, and Shipping

- Weights of ASTM classes 5 or smaller should not be handled with bare hands, due to possible contamination that will change the mass. There are special nonmetallic, nonmagnetic tweezers to handle small and big weights up cca. 100 g. Special lifters are used for bigger weights with cushion material at the point of contact. Chains with soft material sleeves or high-density nylon belts are used for lifting large weights.
- Weights are usually offered as a set. Precision weights are provided in cases made of wood or plastic material with seats for each weight. All weights need to be stored in places free of dust or any other contamination with temperature and humidity not to exceed the prescribed limits. In general, a temperature of 20 °C (68 °F) is considered acceptable for storage.
- Weights of small denomination are packed in their cases in such a way that they are not loosened and lost in transportation. Precision weights are wrapped individually with dry, wax-free paper and placed in their case seats. Cases are wrapped individually with one inch thick foam or other similar wrapping material and placed in cardboard boxes. Weights up to 10 kg or 25 lb are wrapped individually and should not touch each other. Large weights are packed in specially made wooden boxes with 1 in. thick cushion material inside.
- It is preferable that precision weights are hand carried. In general, boxes containing weights should be treated similar with fragile materials and sensitive equipment, for transportation.

43.8 Range and Accuracy

All measuring devices, including weighing devices, have to meet certain requirements and specifications. These requirements are built in what are commonly known as technical characteristics such as range, stability, inertia, replaceability, and metrological characteristics such as accuracy, repeatability, reproducibility, sensitivity, and traceability.

43.8.1 Range

Mass measures or weights have each a nominal value. Usually the weights are presented in sets. The range of a weight set is the difference between the largest and smallest weights in the set.

Example: A set of 21 weights from 1 mg to 100 g has a range of

$$100\ \text{g} - 1\ \text{mg} = 100\ \text{g} - 0.001\ \text{g} = 99.999\ \text{g}$$

Weight value may vary from

- 1 mg (milligram) to 5000 kg (kilogram)
- 0.00001–10,000 lb and 0.0001–12 oz

Scales range is defined as the difference between the highest and smallest indicated mass values on the scale display.

Example: A floor scale with digital indicator has the value of division (smallest indication) of 1 lb and the highest 1000 lb. The scale range will be

$$1000 - 1\ \text{lb} = 999\ \text{lb}$$

Scales ranges may vary from few grams to 50,000 kg or up to 100,000 lb or higher.

43.8.2 Stability

For weights, stability is defined as the capability of the weight piece to retain its mass value unchanged for a longer period of time.

Note: The considered mass value has an uncertainty associated with it. As long as the mass value and its uncertainty are still in statistical control, the weight piece may be considered stable.

For a weighing device, stability is defined as the device capability to restore its equilibrium state. If we consider an unloaded equal arm balance and we initiate an oscillation, the balance beam will oscillate until the state of stable equilibrium is restored. In the case of an electronic scale, the damping process and equilibrium restoration are very fast, almost instantaneous.

43.8.3 Inertia

The inertia describes how the weighing devices are responding to different variations occurring during their operation. It is quantifying the reaction time between changing the mass on the receiving element and the indication on the reading element.

43.8.4 Replaceability

Replaceability applies to both weights and weighing devices; it is the technical characteristic that permits replacing a weight or a weighing device with an identical weight or weighing device, for similar measurements.

43.8.5 Accuracy (Also Trueness or Justness)

The accuracy or justness of a weight is the closeness between the weight calibrated value and the nominal value indicated on the weight.

Accuracy of a scale—the capability of a scale to indicate values closer to the corrected value of a mass (weight) placed on the scale-receiving element. Sometimes, instead of the corrected value, the nominal value of the mass (weight) may be considered. The closer the scale indication, the more accurate it is.

43.8.6 Linearity of a Scale

A balance design and manufacturing target is to obtain a direct proportional relation, graphically represented by a straight line, between the load and the scale reading. That is, if we represent graphically the different readings on scale when we apply increasing different masses (weights), we should obtain a straight line for a perfect linear scale. Again we may consider either the corrected masses (weights) or their nominal values.

Discussion: A scale must be accurate to the same degree in every point of its range, in order to obtain a valid representation of its linearity. If the balance is not accurate in a particular point, the graphic will show a deviation from the straight line. In that point, we have a nonlinearity caused by inaccuracy.

43.8.7 Sensitivity of a Scale

The magnitude of the variation of the scale reading is caused by the variation in the scale load. The smaller the load that creates a perceptible variation of the scale reading, the more sensitive the scale is considered.

The sensitivity of a scale may have different magnitudes at different levels of load.

Here is a discussion of the three parameters: accuracy, linearity, and sensitivity.

Scale is accurate, linear, and sensitive: Good. This is a scale showing values close to the masses' (weights') values we put on, is doing that along the entire range, and is very receptive when we add a small mass (weight) anywhere in the range, that is. It detects better the imperfections (errors) of the masses placed on it.

Scale is not accurate, but is linear and sensitive: Not possible. Inaccuracy generates nonlinearity regardless if the scale is sensitive.

Scale is not accurate, not linear but sensitive: Not good. Who needs a scale not reading close to the value of mass (weight) we put on it, doesn't do that along the entire range, but shows changes in readings when we add a small mass (weight) to the load anywhere in the scale range.

Scale is not accurate, not linear, and not sensitive: Not good.

Scale is accurate, not linear, and sensitive: Not good. It may have segments of linearity along its range, but we don't know where, and it is less probable we would know the law of variation.

Scale is accurate, not linear, and not sensitive: Not good and worst than the aforementioned.

Scale is accurate, linear, but not sensitive: Partially good, for the scale is showing values close to the mass values we put on and is doing that along the entire range, but is not receptive when we add a small mass anywhere in the range, that is, is not detecting the imperfections (errors) of the masses placed on it.

An example will be shipping scales were tolerances are large.

Discussion of the two concepts

- Scale is both accurate and linear: good
- Scale is not accurate but it is linear: not possible
- Scale is accurate, but is not linear: not good (possibly good if we can find the law of variation)
- Scale is not accurate and is not linear: not good, buy a new scale

Measurement repeatability: The measurement precision under a set of repeatability conditions of measurement. VIM 2.21 (3.6)

It is the closeness in the agreement of the results of successive weighing of the same weights that carried out subject to all of the following conditions:

- The same method of weighing
- The same observer
- The same weighing device
- The same location
- The same condition of use
- Repetition over a short period of time

Note:

1. Repeatability may be expressed quantitatively in terms of the dispersion of the results. *Measurement reproducibility*: The measurement precision under reproducibility conditions of measurement. VIM 2.25 (3.7)

Reproducibility is the closeness of the agreement between the results of weighing of the same weights, where individual measurements are carried out changing conditions such as

- Method of weighing
- Observer
- Weighing device
- Location
- Condition of use
- Time

Notes:

- A valid statement of reproducibility requires specification of the conditions changed.
- Repeatability may be expressed quantitatively in terms of the dispersion of the results.

43.8.8 Mass Traceability

Metrological traceability: The property of a measurement result whereby the result can be related to a reference through a documented unbroken chain of calibrations each contributing to the measurement uncertainty. VIM 2.41 (6.10)

Mass traceability—the property of a result of a mass measurement whereby it can be related to appropriate standard weights, generally international or national standards, through an unbroken chain of comparisons each contributing to the measurement uncertainty.

43.8.9 Safety Precautions in Mass Measurement

Safety measures and precautions in mass measurement are related to the following activities:

- Usage of cleaning solvents such as ethylic alcohol and acetone. These substances will be monitored and kept under control. Personnel using these substances need to have appropriate training for storage, usage, disposal, and first aid in case of an accident.
- Usage of paint removal substances, paint thinner, and paint. Wearing adequate protective equipment including mask respirators is mandatory.
- Manipulating metal such as lead and brass. Wearing gloves to avoid direct contact with metallic surface and metallic oxides.
- Manipulating weights to prevent dropping them.
- Manipulating large and very large weights up to 2500 kg and 5000 lb using electrical or compressed air–powered lifting devices (i.e., hoist, forklift) needs to be done by authorized personnel only, wearing adequate protective equipment including hard shoes.

43.9 Units of Measurement

The metric unit of mass is the kilogram. Mass is one of the seven independent base quantities of the International System and is the only base quantity having its measuring unit, the kilogram, defined as an artifact and not based on a natural or physical law.

In the customary (British) system, the unit of mass is the *slug* and the unit of force is the *pound* (lb).

Note: The abbreviation "lb" for pound arrives from the Latin *libra*, which was used in the ancient Roman Empire for balance, scale, or pound.

A body with a mass of 1 slug experiences an acceleration of 1 ft/s^2 when a force of 1 lb works on it. Accordingly, 1 lb = 1 slug · ft/s^2.

The slug is not a familiar unit because in the everyday life, weights rather than masses are specified in the customary system. We are asking for 8 lb of oranges, not for 1/4 slug of oranges.

In the metric system, we normally indicate masses in kilograms not in newtons. To avoid the general misconception, the following stipulations should be taken into consideration:

1. In the metric system
   a. The unit for mass is called kilogram.
   b. The unit for force is called newton.
2. In the pound–inch system
   a. The unit for mass is called slug.
   b. The unit for force is called pound.

Customary and erroneously, when pound–inch system is employed, people are using the unit pound to designate the mass of an object.

Customary and erroneously, the term weight is used to designate the mass of an object (i.e., we say "the weight of the watermelon was…" or "the person's weight is…").

43.9.1 Definition: The Unit of Mass (Kilogram)

The first CGPM (1889) sanctioned the international prototype of the kilogram and declared: this prototype shall henceforth be considered to be the unit of mass.

The third CGPM (1901), in a declaration intended to end the ambiguity that existed as to the meaning of the word "weight" in popular usage, confirmed that the kilogram is the unit of mass; it is equal to the mass of the international prototype of the kilogram.

This international prototype is kept at the International Bureau of Weights and Measures (BIPM) at the Pavillon de Breteuil, Sevres, near Paris. It is made of platinum–iridium alloy.

The same conference adopted the conventional value for the standard acceleration due to gravity that is g_n = 9.806 65 m/s^2.

The United States is in possession of two national prototype kilograms identified as K20 and K4, made of the same alloy of 90% platinum and 10% iridium, as the international prototype. Also the international prototypes as well as the national prototypes are right circular cylinders having the height equal to the diameter and equal to 39 mm (1.535 ft). The material density is 21.5 g/cm^3.

The national prototype kilograms are kept at NIST, located in Gaithersburg, Maryland, near Washington DC.

43.9.2 Conversion Factors

The terms avoirdupois, apothecary, and troy describe units of mass within customary system. The avoirdupois units are extensively used in commerce and trade in the United States.

The following are some conversion factors [6]. Bold face figures are exact.

1 pound avoirdupois (lb avdp) = 0.453 592 37 kilograms (kg)
= 7 000 grains
= 1.215 troy or apothecary pounds

1 ounce avoirdupois (oz avdp) = 28.350 grams (g)
= 437.5 grains
= 0.911 troy or apothecaries ounces

1 pound troy or apothecary (lb t or lb ape) = 373.242 grams (g)
= 5 760 grains
= 0.823 avoirdupois pound

1 ounce troy or apothecaries (oz t or oz ap) = 31.103 grams (g)
= 480 grains
= 1.097 avoirdupois ounces

1 kilogram (kg) = 2.205 pounds

1 carat (c) = 200 milligrams (mg)
= 3.086 grains

1 gram (g) = 0.035 avoirdupois ounce
= 15.432 grains

1 grain = 64.798 91 milligrams (mg)

1 metric ton (t) = 2 204.623 pounds
= 1.102 net tons

1 ton (net or short) = 2000 pounds
= 0.907 metric ton

1 slug = 14.6 kilograms (kg)

References

1. Bucher, J., *The Metrology Handbook*, 2004, ASQ Quality Press, Milwaukee, WI.
2. Hazarian, E., *Techniques of Mass Measurement Workbook*, 1994, Los Angeles, CA.
3. Hazarian, E., Some characteristics of weighing devices, Technical paper at *Measurement Science Conference*, 1992, Anaheim, CA.
4. International Bureau of Weights and Measures, Joint Committee for Guides in Metrology (JCGM) 200:2012, *International Vocabulary of Metrology—Basic and General Concepts and Associated Terms* (VIM), 3rd edition, 2008 version with minor corrections.
5. Jones, F. E., Schoonover, R. M., *Handbook of Mass Measurement*, 2002, CRC Press, Washington, DC.
6. Kochsiek, M., Gläser, M. (eds.), *Comprehensive Mass Metrology*, 2000, Wiley-VCH, Berlin, Germany.
7. Mettler, T., *Glossary of Weighing Terms, A Practical Guide to the Terminology of Weighing*, With the assistance of Prof. Dr. M. Kochsiek, Physikalisch-Technische Bundesanstalt, PTB, English Translation by J. Penton, Zurich, 1997, Berlin, Germany.
8. NBS Metric Information, Letter Circular LC 1071, Factors for High-Precision Conversion, 1976, Gaithersburg, MD.
9. Hazarian, E., Mass measurement techniques IV: Scales uncertainty, *Measurement Science Conference*, Tutorial, 2002, Anaheim, CA.
10. Hazarian, E., *Elements of Measurement Techniques*, California State University Dominguez Hills, MSQA Program, QAS 516 Measurement and Testing Techniques, Supplement, 2000.

44

Torque and Power Measurement

44.1 Fundamental Concepts **44**-1
Angular Displacement, Velocity, and Acceleration • Force, Torque, and Equilibrium • Stress, Rigidity, and Strain • Work, Energy, and Power
44.2 Arrangement of Apparatus for Torque and Power Measurement **44**-4
44.3 Torque Transducer Technologies **44**-5
Surface Strain • Twist Angle • Stress
44.4 Torque Transducer Constructions, Operation, and Application **44**-7
Mechanical Considerations • Electric Considerations • Costs and Options
44.5 Apparatus for Power Measurement **44**-10
Absorption Dynamometers • Driving and Universal Dynamometers • Foursquare Dynamometers • Measurement Accuracy • Costs
References **44**-15

Ivan J. Garshelis
Magnova, Inc.

Torque, speed, and power are the defining parameters of rotating machinery. The ability to accurately measure these quantities is essential for determining a machine's efficiency and to establish operating regimes that are safe and conducive to long and reliable service. On-line measurements of these quantities enable real-time control, help to ensure consistency in product quality, and can provide early indications of impending problems. Torque and power measurements are used in testing new designs of engines, motors, pumps, and other rotating machines and in the development of new machine components. Torque measurements are used routinely for controlling and verifying the tightness of threaded fasteners. Starting from basic concepts, this chapter describes the various methods and devices currently employed for the measurement of torque and power; the measurement of speed, or more precisely, angular velocity, is discussed elsewhere in this handbook [1].

44.1 Fundamental Concepts

44.1.1 Angular Displacement, Velocity, and Acceleration

The concept of *rotational* motion is readily formalized: All points within a rotating rigid body move in parallel or coincident planes while remaining at fixed distances from a line called the *axis*. In a perfectly rigid body, all points also remain at fixed distances from each other. Rotation is perceived as a change in the angular position of a reference point on the body, that is, as its *angular displacement*, $\Delta\theta$, over some time interval, Δt. The motion of that point, and therefore of the whole body, is characterized by its

clockwise (CW) or counterclockwise (CCW) *direction* and by its *angular velocity*, $\omega = \Delta\theta/\Delta t$. If during a time interval Δt, the velocity changes by $\Delta\omega$, the body is undergoing an *angular acceleration*, $\alpha = \Delta\omega/\Delta t$. With angles measured in radians and time in seconds, units of ω become radians per second (rad/s) and of α, radians per second per second (rad/s^2). Angular velocity is often referred to as *rotational speed* and measured in numbers of complete revolutions per minute (rpm) or revolutions per second (rps).

44.1.2 Force, Torque, and Equilibrium

Rotational motion, as with motion in general, is controlled by *forces* in accordance with Newton's laws. Since a force directly affects only that component of motion in its line of action, forces or components of forces acting in any plane that includes the axis produce no tendency for rotation about that axis. Rotation can be initiated, altered in velocity, or terminated only by a *tangential force* F_t acting at a finite radial distance l from the axis. The effectiveness of such forces increases with both F_t and l; hence, their product, called a "moment," is the activating quantity for rotational motion. A moment about the rotational axis constitutes a *torque*. Figure 44.1a shows a force F acting at an angle β to the tangent at a point P, distant l (called the "moment arm") from the axis. The torque T is found from the *tangential component* of F as

$$T = F_t l = (F\cos\beta)l. \tag{44.1}$$

The combined effect, known as the *resultant*, of any number of torques acting on a body is found from their *algebraic* sum, wherein opposite signs are assigned to torques tending to cause rotation in CW and CCW directions. Forces, hence torques, arise from physical contact with other solid bodies or motional interaction with fluids or via gravitational (including inertial), electric, or magnetic force fields. The *source* of each such torque is subjected to an equal, but oppositely directed, *reaction* torque. With force measured in newtons and distance in meters, Equation 44.1 shows the unit of torque to be a "newton meter" (Nm, sometimes N · m).

A nonzero resultant torque will cause the body to undergo a proportional angular acceleration, found, by application of Newton's second law, from

$$T_r = I\alpha, \tag{44.2}$$

where I, having units of kilogram meter^2 (kg m^2), is the moment of inertia of the body around the axis (i.e., its *polar* moment of inertia). Equation 44.2 is applicable to anybody regardless of its state of motion. When $\alpha = 0$, Equation 44.2 shows that T_r is also zero; the body is said to be in *equilibrium*. For a body to be in equilibrium, there must be either more than one *applied* torque or none at all.

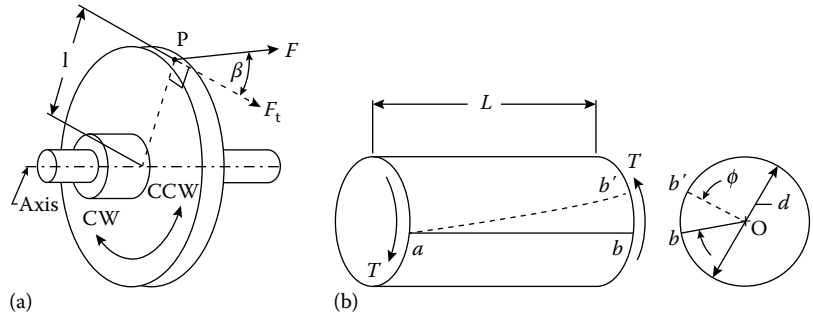


FIGURE 44.1 (a) Off-axis force F at P produces a torque $T = (F\cos\beta)l$ tending to rotate the body in the CW direction. (b) Transmitting torque T over length L twists the shaft through angle ϕ.

44.1.3 Stress, Rigidity, and Strain

Any portion of a rigid body in equilibrium is also in equilibrium; hence, as a condition for equilibrium of the portion, any torques applied thereto from *external* sources must be balanced by equal and directionally opposite *internal* torques from adjoining portions of the body. Torque is *transmitted* internally between adjoining portions by the collective action of *stresses* over their common cross sections. In a solid homogeneous body having a round cross section, for example, a typical shaft, and within a material-dependent upper limit, the *shear stress* τ varies linearly from zero at the axis to a maximum value, τ_m, at the surface. For a shaft of diameter d, transmitting a torque T, τ_m, is found from

$$\tau_m = \frac{16T}{\pi d^3}. \tag{44.3}$$

Real materials are not *perfectly* rigid but have instead a *modulus of rigidity*, G, which expresses the finite ratio between τ and *shear strain*, γ. The maximum strain therefore also exists at the shaft surface and can be found from

$$\gamma_m = \frac{\tau_m}{G} = \frac{16T}{\pi d^3 G}. \tag{44.4}$$

Figure 44.1b shows the manifestation of shear strain as an angular displacement between axially separated cross sections. Over the length L, the solid round shaft shown will be *twisted* by the torque through an angle φ (for small angles, $\varphi = \tan\varphi$) found from

$$\phi = \frac{2L\gamma_m}{d} = \frac{32LT}{\pi d^4 G}. \tag{44.5}$$

44.1.4 Work, Energy, and Power

If during the time of application of a torque T, the body rotates through some angle, θ, mechanical work

$$W = T\theta \tag{44.6}$$

is performed. If the torque acts in the same CW or CCW sense as the displacement, the work is said to be done *on* the body; else, it is done by the body. If the work being done on the body exceeds that being done by the body, the body accelerates; the net work appears as an increase in *kinetic energy* (KE = $I\omega^2/2$). The net work done by the body causes deceleration, with a corresponding decrease in kinetic energy. If the body is not accelerating, any work done on it at one location must be done by it at another location. Work and energy are each measured in units called a joule (J). Equation 44.6 shows that 1 J is equivalent to 1 Nm rad, which is, since a radian is a dimensionless ratio, ≡ 1 Nm. To avoid confusion with torque, it is preferable to quantify mechanical work in units of mN or, better still, in J.

The *rate* at which work is performed is termed *power*, P. If a torque T acts over a small interval of time Δt, during which there is an angular displacement $\Delta\theta$, work equal to $T\Delta\theta$ is performed at the rate $T\Delta\theta/\Delta t$. Replacing $\Delta\theta/\Delta t$ by ω, power is found simply as

$$P = T\omega. \tag{44.7}$$

The unit of power follows from its definition and is given the special name watt (W). 1 W = 1 J/s = 1 m · N/s. Historically, power has also been measured in horsepower (Hp), where 1 Hp = 746 W. Rotating bodies effectively transmit power between locations where torques from external sources are applied.

44.2 Arrangement of Apparatus for Torque and Power Measurement

Equations 44.1 through 44.7 express the physical bases for torque and power measurement. Figure 44.2 illustrates a generalized measurement arrangement, the actual apparatus being selected to fulfill specific measurement purposes. In general, a driving torque originating within a device at one location (B in Figure 44.2) is resisted by an opposing torque developed by a different device at another location (F). The driving torque (from, e.g., an electric motor, a gasoline engine, a steam turbine, a muscular effort) is coupled through connecting members C, transmitting region D, and additional couplings E, to the driven device (an electric generator, a pump, a machine tool, a mated threaded fastener, etc.) within which the resisting torque is met at F. The torque at B or F is the quantity to be measured. They may be *indirectly* determined from a correlated physical quantity, for example, an electric current or fluid pressure associated with the operation of the driving or driven device, or more directly determined by measuring either the *reaction* torque at A or G or the torque *transmitted* through D.

To the extent that the frames of the driving and driven devices and their mountings to the "earth" are *perfectly* rigid, the reaction at A will *at every instant* equal the torque at B, as will the reaction at G equal the torque at F. Under equilibrium conditions, these equalities are independent of the compliance of any member. Also under equilibrium conditions, and except for usually minor *parasitic* torques (due, e.g., to bearing friction and air drag over rapidly moving surfaces), the driving torque at B will equal the resisting torque at F.

Torque is often determined from measurements of reaction forces at dimensionally known locations, for example, mounting feet at A or G, using Equation 44.1. Transmitted torque is determined from measurements, on a suitable member within region D, of τ_m, γ_m, or φ and applying Equations 44.3 through 44.5 (or analogous expressions for members having other than solid round cross sections [2]). *Calibration*, that is, measurement of the stress, strain, or twist angle resulting from the application of *known* torques (within the range of interest), makes knowledge of details about the member within D unnecessary. When $\alpha \neq 0$, and is measurable with suitable accuracy, T may be determined from Equation 44.2. Requiring only noninvasive, observational measurements, this method is especially useful for determining transitory torques, for example, those associated with firing events in multicylinder internal combustion engines [3].

Equations 44.6 and 44.7 are applicable *only* during rotation since in the absence of motion, no work is done and power transfer is zero. Equation 44.6 may be used to determine *average* torque from calorimetric measurements of the heat generated (equal to the mechanical work W) during a totalized number of revolutions ($\equiv\theta/2\pi$). Equation 44.7 is routinely applied in power measurement, wherein T is determined by methods based on Equations 44.1, 44.3, 44.4, or 44.5 and ω is measured by any suitable means [4].

F, T, and φ are sometimes measured by simple mechanical methods. For example, a "torque wrench" is often used for the controlled tightening of threaded fasteners. In one such device, torque is indicated

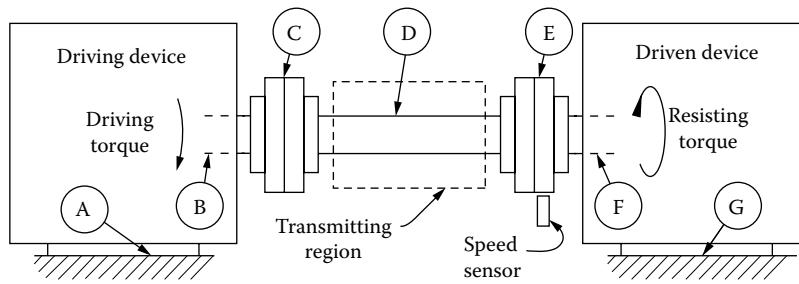


FIGURE 44.2 Schematic arrangement of devices used for the measurement of torque and power.

by the position of a "needle" moving over a calibrated curved scale in response to the deflection of an elastic member, in the simplest case, the bending (as a beam) of the wrench handle [5]. More generally, instruments, variously called *sensors* or *transducers*, are used to convert the torque being measured into a linearly proportional electric signal, typically a voltage or a frequency. The determination of P usually involves multiplication of the two signals from separate T and ω sensors. A transducer wherein the power transmitted by a rotating shaft is indicated by the amplitude of a *single* signal has also been described [6].

44.3 Torque Transducer Technologies

Various physical interactions serve to convert F, τ, γ, or φ into proportional electric signals. Each requires that some axial portion of the shaft be dedicated to the torque sensing function. Figure 44.3 shows typical configurations of the functional parts needed to apply the technologies most often utilized in present-day torque transducers.

44.3.1 Surface Strain

Figure 44.3a illustrates a sensing region configured to convert surface shear strain (γ_m) into an electric signal proportional to the transmitted torque. Surface strain became the key basis for measuring both force and torque following the invention of bonded wire strain gages by E. E. Simmons Jr. and Arthur C. Ruge in 1938 [7]. A modern strain gage consists of an elongated electric conductor (or semiconductor), generally in the form of a thin foil or film in a serpentine pattern, enclosed within a thin insulating carrier. The carrier is securely attached, usually with a thin adhesive layer, to the surface of the torque-carrying member. Strain is sensed as a change in gage resistance. Since these changes are generally too small (e.g., ~0.1%) to be accurately measured directly, it is common practice to employ multiple gages arranged in "half" (two gages) or "full" (four gages) Wheatstone bridge circuits. A four-gage bridge, comprised of two diametrically opposite pairs of matched strain gages, each aligned along a *principal strain* direction, provides independence from axial and bending loads as well as some temperature compensation. Limiting strains, as determined from Equation 44.4 (with τ_m equal to the shear proportional limit of the shaft material), rarely exceed a few parts in 10^3. Measurement sensitivity is maximized by choosing the shaft diameter in the sensing region to attain the limiting strain at the highest value of torque to be measured.

Torque transducers utilizing strain gages based on surface acoustic wave (SAW) technology [8] have been more recently introduced. In one configuration, illustrated in Figure 44.3b, piezoelectric "resonators," aligned with each of the principal strain directions, are attached to the transducer shaft. Under torque, the spacing between those "Interdigital Electrodes" is subjected to compressive strain decrease. The resonant frequencies of SAWs generated on the two substrates by electromagnetically coupled voltages applied to the electrodes of each resonator thereby decrease and increase, respectively. The difference between these two frequencies is thus a measure of the strain, hence (from Equation 44.4) of torque.

44.3.2 Twist Angle

If the shaft is *slender* enough, for example, $L > 5d$, φ, at limiting values of τ_m for typical shaft materials, can exceed 1°, enough to be resolved with sufficient accuracy for practical torque measurements. (φ at τ_m can be found by manipulating Equations 44.3 through 44.5.) Figure 44.3c shows a common arrangement wherein torque is determined from the difference in tooth-space phasing between two identical (and initially aligned for $T = 0$) "toothed" wheels attached at opposite ends of a compliant "torsion bar." The phase displacement of the periodic electric signals from the two "pickups" is proportional to the peripheral displacement of "teeth" or other salient features on the two wheels, hence to the twist angle

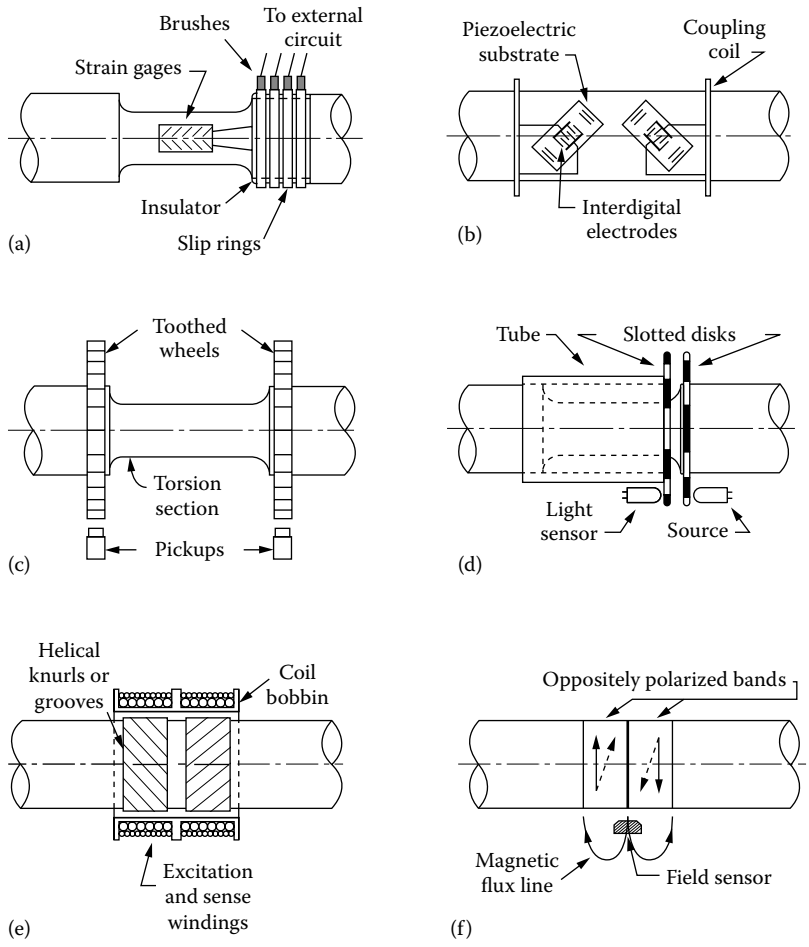


FIGURE 44.3 Six techniques in present use for measuring transmitted torque. (a) Torsional strain in the shaft upsets the balance of the Wheatstone bridge (comprised of four strain gages, two not seen). In the embodiment shown, external connections to the bridge are made through slip rings and brushes. (b) Torsional strain alters the frequencies of the two SAW resonators, increasing one and decreasing the other. Power is inductively coupled to the resonators. (c) Twist of the torsion section causes angular displacement of the surface features on the toothed wheels. This creates a phase difference in the signals from the two pickups. (d) Twist of the shaft section inside the tube causes angular displacement of the two slotted disks, thereby allowing more or less light from the source to reach the sensor. (e) The permeability of the two grooved regions of the shaft changes oppositely with torsional stress. This is sensed as a difference in the output voltages of the two sense windings. (f) Torsional stress causes the initially circumferential magnetizations in the bands (solid arrows) to tilt (dashed arrows). Magnetic poles thereby appear at the band ends and at their interface. The resulting magnetic field is sensed by the field sensor.

of the torsion bar and thus to the torque. These features are designed to enable sensing by one or other noncontacting magnetic, optical, or capacitive techniques. A rigid tube is sometimes used to bring the angular position of an axially remote part of the shaft close to the axial location of one of the sensors. This facilitates the alignment of the two pickups and simplifies associated wiring. It also allows for the use of a wider range of position measurement devices, especially those based on optical techniques as illustrated in Figure 44.3d and those in which the relative angular position of the two wheels (or disks) appears as the amplitude of a *single* electric signal, thus providing for the measurement of torque even on a stationary shaft.

44.3.3 Stress

In addition to elastic strain, the stresses by which torque is transmitted also affect the magnetization orientation in ferromagnetic shaft materials. Unlike strain, this "magnetoelastic interaction" [9] is detectable without contacting the shaft. Two types of magnetoelastic (sometimes called "magnetostrictive") torque transducers are in present use: Type 1 derives output signals from torque-induced variations in permeability; type 2 "measures" magnetic fields created in response to the torque. Type 1 transducers typically employ "branch," "cross," or "solenoidal" constructions [10]. In branch and cross designs, torque is detected as an imbalance in the permeabilities of the shaft surface material along the principal stress directions (orthogonal 45° helical paths). In solenoidal constructions, torque is detected by differences in the *axial* permeabilities of two adjacent surface regions, pre-endowed with symmetrical magnetic "easy" axes (typically in the principal stress directions). A unique variation of this basic construction utilizing copper-plated "V"-shaped stripes has more recently been developed [11]. While branch- and cross-type sensors are readily miniaturized [12], local variations in magnetic properties of typical shaft surfaces compromise their attainable accuracy. Solenoidal designs, illustrated in Figure 44.3e, avoid this pitfall by effectively averaging these local variations.

As indicated by the solid arrows in Figure 44.3f, the sensing region in type 2 transducers consists of (usually 2) remanently magnetized (in, respectively, opposite circumferential directions) adjacent bands. These "polarized" bands are instilled directly either into the near-surface regions of the shaft itself or in a ring of magnetoelastically active material affixed to the shaft [13]. When torque is applied, the magnetizations tilt into helical directions (dashed arrows) causing magnetic poles to develop at the band interface and, with opposite polarity, at their outer ends. Torque is determined from the output signal of one or more magnetic field sensors (e.g., Hall effect, magnetoresistive, or fluxgate devices) located to sense the intensity and polarity of the magnetic field that thereby arises in the space near the ring or band.

44.4 Torque Transducer Constructions, Operation, and Application

Although the application (automotive, aerospace) often requires that a torque sensing region should be created directly on an existing shaft, it is more usual to install a preassembled and calibrated, *modular* torque transducer into the driveline. Standard models with capacities from 0.001 to 300,000 Nm are obtainable from more than 50 manufacturers, worldwide [14]. At least one unit having a 2 MNm measuring range for wind turbine application has been made.

The descriptions of the measuring technology as well as installation and operating instructions can be found in the literature or on the websites of many manufacturers. The patent literature [15] often reveals key construction features in a product line of any one manufacturer.

Trade names generally identify families of transducers, from a single manufacturer, based on a specific operating principle, or possessing a collection of unique features; for example, *Torductor®-S* [11] identifies a unique and proprietary type of type 1 magnetoelastic transducer, *Torquetronic* [16] is a class of transducers using wrap-around twist angle sensors, *TorqSense* identifies a family of SAW-based transducers [17], and *smarttorque®* refers to a bearingless, flange-type, strain-gage-based "intelligent measuring system" [18]. A sampling of other well-known manufacturers is listed in [19–25]. Many of these devices show generic similarities that transcend both their specific sensing technology and torque range. Figure 44.4 identifies these common features in a classical torque transducer construction wherein a shaft having a sensing region is supported on bearings within a housing that encloses ancillary components. In recent times, these often include various types of signal conditioning and signal transmitting (e.g., wireless) electronics. Figure 44.5 illustrates the generic features of a relatively newer, bearingless construction, commonly called a "torque flange." All presently available torque flanges

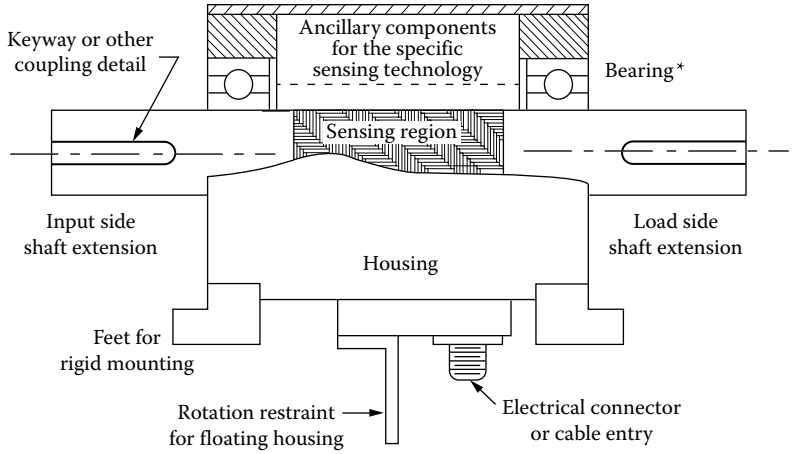


FIGURE 44.4 Generic features of a modular torque transducer. Bearings* are used only if the shaft is to rotate. One, but not both shaft extensions, can be rigidly connected to the driving or driven machine; the coupling at the other end must possess sufficient angular, radial, and axial flexibilities to suit the application. Rigid mounting of the housing requires flexible couplings at both shaft ends. Axially short, dual flange constructions have no bearings (see Figure 44.5).

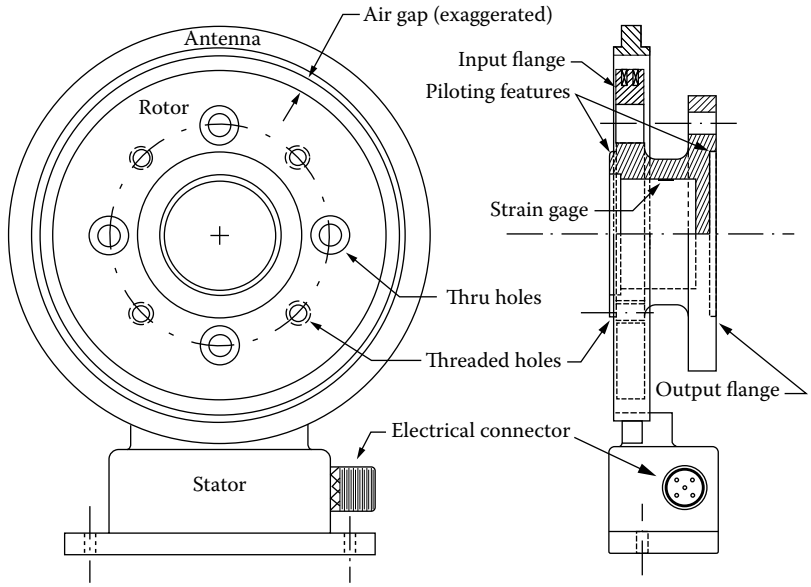


FIGURE 44.5 "Torque-flange" type of torque transducer. "Antennas" through which power and signal are coupled between the rotor and the stator vary greatly in configuration and angular extent between manufacturers and among models of any one manufacturer. Rotors contain a significant amount of electronic circuitry. Some rotors include encoder details for speed measurement. Rotor temperature measurement is sometimes available. Digital and analog telemetries often coexist. A common feature is the use of threaded holes on the input flange and thru holes on the output flange. This allows fastening bolts to both flanges to be installed from the same side.

utilize resistive strain-gage technology, either in torsional strain configurations or, for minimizing the axial width, to sense bending stresses in spokes that connect the concentric input and output flanges.

It is sometimes the case that the torque to be measured has a relatively large constant or slowly varying component, on top of which there are small rapidly varying components. This occurs with many types of motors, pumps, and geared machines. Transducers that respond to the rate of change of

torque (dT/dt) rather than T can more sensitively measure this type of dynamic component. Devices for this purpose typically employ either piezoelectric [26] or magnetoelastic [27] sensing elements that, by their nature, are well suited for such use.

44.4.1 Mechanical Considerations

Maximum operating speeds vary widely; upper limits depend upon the size, operating principle, type of bearings, lubrication, and dynamic balance of the rotating assembly. Ball bearings, lubricated by grease, oil, or oil mist, are typical. Parasitic torques associated with bearing lubricants and seals limit the accuracy of low-end torque measurements. (Minute capacity units have no bearings [23].) Forced lubrication allows operation up to 80,000 rpm [16]. Ceramic races and/or balls allow even higher speeds (120,000 rpm [16]). For any given torque range, flange-type transducers have higher maximum speed ratings than those with bearings. High-speed operation requires consideration of the effects of centrifugal stresses on the sensed quantity as well as of critical (vibration inducing) speed ranges. Torsional oscillations associated with resonances of the shaft elasticity with the rotational inertia of coupled masses can corrupt the measurement, damage the transducer by dynamic excursions above its rated overload torque, and *even be physically dangerous*.

Housings either *float* on the shaft bearings or are *rigidly mounted*. Free-floating housings are restrained from rotating by such "soft" means as a cable, spring, or compliant bracket or by an eccentric external feature simply resting against a fixed surface. In free-floating installations, the axes of the driving and driven shafts must be carefully aligned. Torsionally rigid "flexible" couplings at each shaft end are used to accommodate small angular and/or radial misalignments. Alternatively, the use of dual flexible couplings at one end will allow direct coupling of the other end. Rigidly mounted housings are equipped with mounting feet or lugs similar to those found on the frames of electric motors. Free-floating models are sometimes rigidly mounted using adapter plates fastened to the housing. Rigid mountings are preferred when it is difficult or impractical to align the driving and driven shafts, as, for example, when driving or driven machines are changed often. Rigidly mounted housings *require* the use of dual flexible couplings at *both* shaft ends.

Having neither bearings nor slip rings, torque flanges are completely wear-free. When mounted into an existing drive train, they require no more consideration than if mounting a spacer. One flange is directly connected (usually) to the loaded machine and the other, through universal joints, multiple disk, or similar couplings, is connected to the driving machine.

Transducers designed for zero or limited rotation applications have no need for bearings or noncontact constructions. To ensure that *all* of the torque applied at the ends is sensed, it is important in such "reaction"-type torque transducers that the housing be attached to the "shaft" at only one side of the sensing region. Whether rotating or stationary, the external shaft ends generally include such torque coupling details as flats, keyways, splines, tapers, flanges, and male/female square drives.

44.4.2 Electric Considerations

By their nature, transducers require some electric input power or *excitation*. The "raw" output signal of the actual sensing device also generally requires "conditioning" into a level and format appropriate for display on a digital or analog meter or to meet the requirements of data acquisition equipment. Excitation and signal conditioning are supplied by electronic circuits designed to match the characteristics of the specific sensing technology. For example, strain-gage bridges are typically powered with 10–20 V (dc or ac) and have rated load outputs in the range of 1.5–3.0 mV/V of excitation. Raising these millivolt signals to more usable levels requires amplifiers having gains of 100 or more. With ac excitation, oscillators and demodulators (or rectifiers) are also needed. Circuit elements of these types are normal when inductive elements are used, either as a vital part of the sensor or simply to implement a contactless construction.

Strain gages, differential transformers, and related sensing technologies obviously require that electric components be mounted *on* the torqued member. Various methods are employed to bring electric power to, and signals from, these components on rotating shafts. The most direct and common approach is to use conductive means wherein brushes (typically of silver graphite) bear against (silver) slip rings. Useful life is extended by providing means to lift the brushes off the rotating rings when measurements are not being made. Several "noncontacting" methods are also used. For example, power may be supplied via inductive coupling between stationary and rotating transformer windings [18–25], by the illumination of shaft-mounted photovoltaic cells or even by batteries strapped to the shaft [28] (limited by centrifugal force to relatively low speeds). Output signals are coupled off the shaft through rotary transformers, by frequency-modulated (infrared) LEDs [21] or by radio-frequency (FM) telemetry [28]. Inductive and telemetric components are incorporated into the associated torque flanges. Where shaft rotation is limited to no more than a few full rotations, as in steering gear, valve actuators, or oscillating mechanisms, hard wiring both power and signal circuits is often suitable. Flexible cabling minimizes incidental torques and makes for a long and reliable service life. All such wiring considerations are avoided when noncontact technologies or constructions are used.

44.4.3 Costs and Options

Prices of torque transducers reflect the wide range of available capacities, performance ratings, types, styles, optional features, and accessories. In general, prices of any one type increase with increasing capacity. Reaction types cost about half of similarly rated rotating units. A typical foot mounted, 500 N·m capacity, strain-gage transducer with either slip rings or rotary transformers and integral speed sensor, specified nonlinearity and hysteresis each within ±0.1%, costs about $4000. Compatible accessories providing transducer excitation signal conditioning and analog output with digital display of torque and speed add about $2000. A comparable magnetoelastic transducer with ±0.5% accuracy costs about $1300. Low-range, 1% accuracy magnetoelastic units cost under $500. High-capacity transducers for extreme speed service with appropriate lubrication options can cost well over $50,000. Price ranges for standard units are often posted on manufacturer's web pages. Cost estimates for mass produced torque transducers for use in automotive power steering applications are in the range of $20.

44.5 Apparatus for Power Measurement

Rotating machinery exists in specific types without limit and may operate at power levels from fractions of a watt to some tens of megawatts, a range spanning more than 10^8. Apparatus for power measurement exists in a similarly wide range of types and sizes. Mechanical power flows from a *driver* to a *load*. This power can be determined *directly* by application of Equation 44.7, simply by measuring, in addition to ω, the output torque of the driver or the input torque to the load, whichever is the device under test (DUT). When the DUT is a driver, measurements are usually required over its full service range of speed and torque. The test apparatus therefore must act as a controllable load and be able to *absorb* the delivered power. Similarly, when the DUT is a pump or fan or other type of load, or one whose function is simply to alter speed and torque, for example, a gearbox, the test apparatus must include a *driver* capable of supplying power over the DUT's full rated range of torque and speed. Mechanical power can also be determined *indirectly* by conversion into (or from) another form of energy, for example, heat or electricity, and measuring the relevant calorimetric or electric quantities. In view of the wide range of readily available methods and apparatus for accurately measuring both torque and speed, indirect methods need only be considered when special circumstances make direct methods difficult.

Dynamometer is the special name given to power measuring apparatus that includes absorbing or/and driving means and wherein torque is determined by the reaction forces on a stationary part (the *stator*). An effective dynamometer is conveniently assembled by mounting the DUT in such a manner as to allow measurement of the reaction torque on its frame. Figure 44.6 shows a device

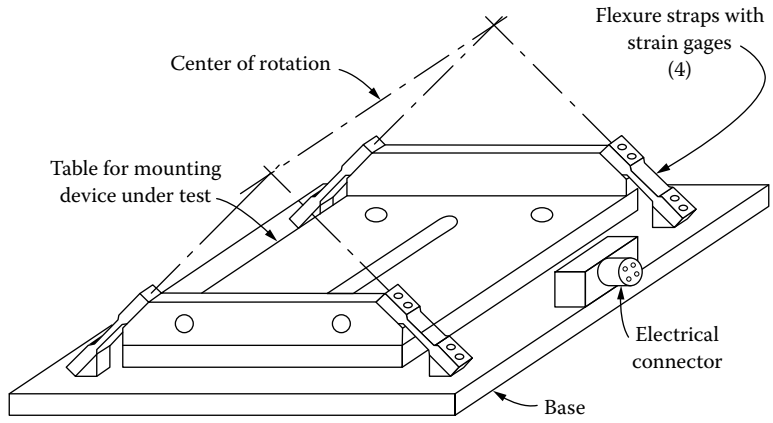


FIGURE 44.6 Support system for measuring the reaction torque of a rotating machine. The axis of the machine must be accurately set on the "center of rotation." The holes and the keyway in the table facilitate machine mounting and alignment. Holes in the front upright provide for attaching a lever arm from which calibrating weights may be hung. (From Supplement to ASME performance test codes, *Measurement of Shaft Power*, ANSI/ASME PTC 19.7-1980 (Reaffirmed 1988); Lebow Products (Honeywell Sensors), 1728 Maplelawn Road, Troy, MI 48099, https://measurementsensors.honeywell.com/pages/ProductInfo.aspx.)

designed to facilitate such measurements. Commercial models (Torque Table® [20]) rated to support DUTs weighing 222–4900 N are available with torque capacities from 1.3 to 226 N·m. "Torque tubes" [4] or other DUT mounting arrangements are also used. Other than for possible rotational/elastic resonances, these systems have no speed limitations. More generally, and especially for large machinery, dynamometers include a specialized driving or absorbing machine. Such dynamometers are classified according to their function as *absorbing* or *driving* (sometimes *motoring*). A *universal* dynamometer can function as either a driver or an absorber.

44.5.1 Absorption Dynamometers

Absorption dynamometers, oftentimes called *brakes* since their operation depends on the creation of a controllable *drag* torque, convert mechanical work into heat. A drag torque, as distinguished from an active torque, can act only to restrain and not to initiate rotational motion. Temperature rise within a dynamometer is controlled by carrying away the heat energy, usually by transfer to a moving fluid, typically air or water. Drag torque is created by inherently dissipative processes such as friction between rubbing surfaces, shear or turbulence of viscous liquids, flow of electric current, or magnetic hysteresis. Gaspard Riche de Prony (1755–1839), in 1821 [29], invented a highly useful form of a friction brake to meet the needs for testing the steam engines that were then becoming prevalent. Brakes of this type are rarely used with modern prime movers, their main use being for instructional purposes, since they embody all of the underlying principles and major operating considerations for all types of absorption dynamometers. Figure 44.7 shows the basic form and constructional features of a *prony brake*. The power that would normally be delivered by the shaft of the driving engine to the driven load is (for measurement purposes) converted instead into heat via the work done by the frictional forces between the friction blocks and the flywheel rim. Adjusting the tightness of the clamping bolts varies the frictional drag torque as required. Heat is removed from the inside surface of the rim by arrangements (not shown) utilizing either a continuous flow or evaporation of water. There is no need to know the magnitude of the frictional forces nor even the radius of the flywheel (facts recognized by Prony), since, while the drag torque tends to rotate the clamped-on apparatus, it is held stationary by the equal but opposite reaction torque Fr. F at the end of the torque arm of radius r (a fixed dimension of the apparatus) is monitored by a scale or load cell. The power is found from Equations 44.1 and 44.7 as $P = Fr\omega = Fr2\pi N/60$ where N is in rpm.

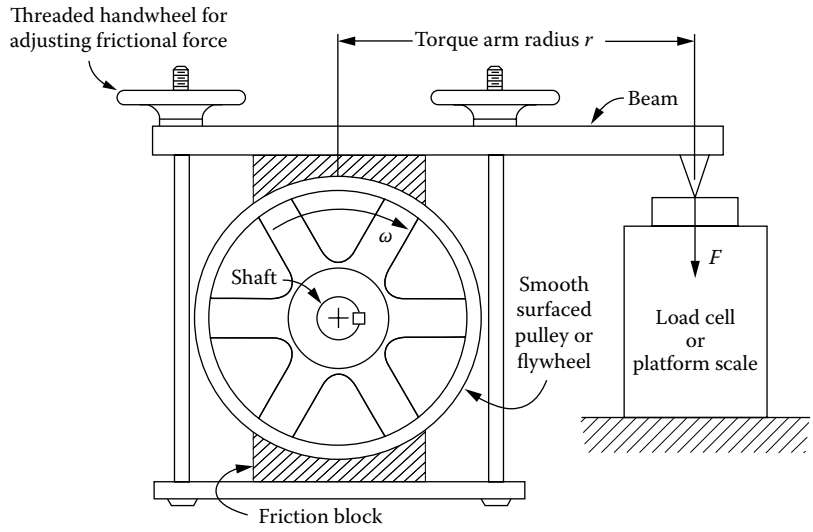


FIGURE 44.7 Classical prony brake. This brake embodies the defining features of all absorbing dynamometers: conversion of mechanical work into heat and determination of power from measured values of reaction torque and rotational velocity.

Uneven retarding forces associated with fluctuating coefficients of friction generally make rubbing friction a poor way to generate drag torque. Nevertheless, because they may be easily constructed, ad hoc variations of prony brakes, often using only bare ropes or wooden cleats connected by ropes or straps, find use in the laboratory or wherever undemanding or infrequent power measurements are to be made. More sophisticated prony brake constructions are used in stand-alone dynamometers with self-contained cooling water tanks in sizes up to 746 kW (1000 Hp) for operation up to 3600 rpm with torques to 5400 N·m [30]. Available in stationary and mobile models, they find use in testing large electric motors as well as engines and transmissions on agricultural vehicles. Prony brakes allow full drag torque to be imposed down to zero speed.

William Froude (1810–1879) [31] invented a *water brake* (1877) that does not depend on rubbing friction. Drag torque within a *Froude brake* is developed between the rotor and the stator by the momentum imparted by the rotor to water contained within the brake casing. Rotor rotation forces the water to circulate between cuplike pockets cast into facing surfaces of both rotor and stator. The rotor is supported in the stator by bearings that also fix its axial position. Labyrinth-type seals prevent water leakage while minimizing frictional drag and wear. The stator casing is supported in the dynamometer frame in cradle fashion by *trunnion* bearings. The torque that prevents rotation of the stator is measured by reaction forces in much the same manner as with the prony brake. Drag torque is adjusted by a valve, controlling either the back pressure in the water outlet piping [32] or the inlet flow rate or sometimes (to allow very rapid torque changes) with two valves controlling both [34]. In any case, the absorbed energy is carried away by the continuous water flow. Other types of cradle-mounted water brakes, while externally similar, have substantially different internal constructions and depend on other principles for developing the drag torque; for example, smooth rotors develop viscous drag by shearing and turbulence. Nevertheless, all *hydraulic dynamometers* purposefully function as *inefficient* centrifugal pumps. Regardless of internal design and valve settings, maximum drag torque is low at low speeds (zero at standstill) but can rise rapidly, typically varying with the square of rotational speed. The irreducible presence of some water, as well as windage, places a speed-dependent lower limit on the *controllable* drag torque. In any one design, wear and vibration caused by cavitation place upper limits on the speed and power level. Hydraulic dynamometers are available in a wide range of capacities between 300 and 25,000 kW with some portable units having capacities as low as 75 kW [33]. The largest ever built [34], absorbing up to about 75,000 kW (100,000 Hp), has been used to test propulsion systems

for nuclear submarines. Maximum speeds match the operating speeds of the prime movers that they are built to test and therefore generally decrease with increasing capacity. High-speed gas turbine and aerospace engine test equipment may operate as high as 30,000 rpm [32].

In 1855, Jean B. L. Foucault (1819–1868) [29] demonstrated the conversion of mechanical work into heat by rotating a copper disk between the poles of an electromagnet. This simple means of developing drag torque, based on *eddy currents*, has, since *ca*. 1935, been widely exploited in dynamometers. Figure 44.8 shows the essential features of this type of brake. Rotation of a toothed or spoked steel rotor through a spatially uniform magnetic field, created by direct current through coils in the stator, induces locally circulating (eddy) currents in electrically conductive (copper) portions of the stator. Electromagnetic forces between the rotor, which is magnetized by the uniform field, and the field arising from the eddy currents create the drag torque. This torque and hence the mechanical input power are controlled by adjusting the *excitation* current in the stator coils. Electric input power is less than 1% of the rated capacity. The dynamometer is effectively an internally short-circuited generator since the power associated with the resistive losses from the generated eddy currents is dissipated *within* the machine. Being heated by the flow of these currents, the stator must be cooled, sometimes (in smaller-capacity machines) by air supplied by blowers [30], more often by the continuous flow of water [32,34,35]. In *dry-gap* eddy current brakes (the type shown in Figure 44.7), water flow is limited to passages within the stator. Larger machines are often of the *water in gap* type wherein water also circulates around the rotor [35]. Water in contact with the moving rotor effectively acts as in a water brake adding a non-electromagnetic component to the total drag torque, thereby placing a lower limit to the controllable torque. Windage limits the minimum value of controllable torque in dry-gap types. Since drag torque is developed by the motion of the rotor, it is zero at standstill for any value of excitation current. Initially rising rapidly, approximately linearly, with speed, torque eventually approaches a current limited saturation value. As in other cradled machines, the torque required to prevent rotation of the stator is measured by the reaction force acting at a fixed known distance from the rotation axis. Standard model eddy current brakes have capacities from less than 1 kW [30,36] to more than 2,000 kW [34,35] with maximum speeds from 12,000 rpm in the smaller-capacity units to 3,600 rpm in the largest units. Special units with capacities of 3,000 Hp (2,238 kW) at speeds to 25,000 rpm have been built [35].

Hysteresis brakes [36] develop drag torque via magnetic attractive/repulsive forces between the magnetic poles established in a reticulated stator structure by a current through the field coil and those created in a "drag-cup" rotor by the stator field gradients. The rotation of the special steel rotor, through the spatial field pattern established by the stator, results in a cyclical reversal of the polarity

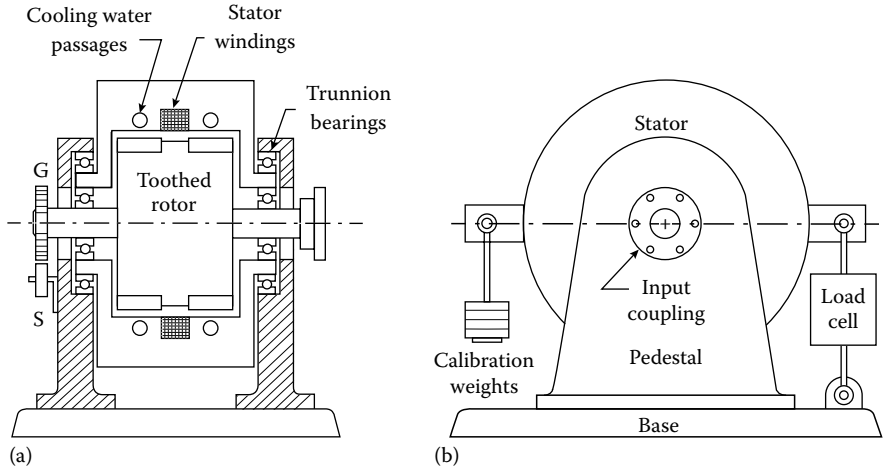


FIGURE 44.8 (a) Cross section and (b) front view of an eddy current dynamometer. G is a gear wheel and S is a speed sensor. Hoses carrying cooling water and cable carrying electric power to the stator are not shown.

of its local magnetizations. The energy associated with these reversals (proportional to the area of the hysteresis loop of the rotor material) is converted into heat within the drag cup. Temperature rise is controlled by forced air cooling from a blower or compressed air source. As with eddy current brakes, the drag torque of these devices is controlled by the excitation current. In contrast with eddy current brakes, rated drag torque is available down to zero speed. (Eddy current effects typically add only 1% to the drag torque for each 1000 rpm.) As a result of their smooth-surfaced rotating parts, hysteresis brakes exhibit low parasitic torques and hence cover a dynamic range as high as 200–1. Standard models are available having continuous power capacities up to 6 kW (12 kW with two brakes in tandem cooled by two blowers). Intermittent capacities per unit (for 5 min or less) are 7 kW. Some low-capacity units are convection cooled; the smallest has a continuous rating of just 7 W (35 W for 5 min). Maximum speeds range from 30,000 rpm for the smallest to 10,000 rpm for the largest units. Torque is measured by a strain-gage bridge on a moment arm supporting the machine stator.

44.5.2 Driving and Universal Dynamometers

Electric generators, both ac and dc, offer another means for developing a controllable drag torque, and they are readily adapted for dynamometer service by cradle mounting their stator structures. Moreover, electric machines of these types can also operate in a motoring mode wherein they can deliver controllable *active* torque. When configured to operate selectively in either driving or absorbing modes, the machine serves as a universal dynamometer. With dc machines in the absorbing mode, the generated power is typically dissipated in a convection-cooled resistor bank. Air cooling the machine with blowers is usually adequate, since *most* of the mechanical power input is dissipated externally. Nevertheless, *all* of the mechanical input power is accounted for by the product of the reaction torque and the rotational speed. In the motoring mode, torque and speed are controlled by adjustment of both field and armature currents. Modern ac machines utilize regenerative input power converters to allow braking power to be returned to the utility power line. In the motoring mode, speed is controlled by high-power, solid-state, adjustable-frequency inverters. Internal construction is that of a simple three-phase induction motor, having neither brushes, slip rings, nor commutators. The absence of rotor windings allows for higher-speed operation than dc machines. Universal dynamometers are "four-quadrant" machines, a term denoting their ability to produce torque in the same or opposite direction as their rotational velocity. This unique ability allows the effective drag torque to be reduced to zero at any speed. Universal dynamometers [32,35] are available in a relatively limited range of capacities (56–450 kW), with commensurate torque (110–1900 N · m) and speed (4,500–13,500 rpm) ranges, reflecting their principal application in automotive engine development. Special dynamometers for testing transmissions and other vehicular drive train components insert the DUT between a diesel engine or electric motor prime mover and a hydraulic or eddy current brake [37].

44.5.3 Foursquare Dynamometers

It is often desirable to perform endurance tests of the various load-carrying parts (shafts, gears, bearings, lubricants, etc.) of the rotating machinery or of the entire assemblies such as automotive or helicopter transmissions or even the gearboxes used in modern wind turbines. This requires that the particular DUT be operated under full torque (or full range of torque variation) at full rated speed for hundreds, sometimes for thousands of hours. To do this with any of the previously described dynamometers could obviously require the expenditure of enormous quantities of energy. Apparatus such as a "foursquare dynamometer" provides a simple means of circulating large (and adjustable) amounts of power while utilizing only the amount of external power required to overcome the static and dynamic friction associated with its operation. Due to the wide range of torques, speeds, and devices to be tested, there are few suppliers of "standard" models of foursquare dynamometers [38], although the principle is often applied in the construction of specialized units for specific applications [39].

44.5.4 Measurement Accuracy

Accuracy of power measurement (see discussion in [4]) is generally limited by the torque measurement (±0.25% to ±1%) since rotational speed can be measured with almost any desired accuracy. Torque errors can arise from the application of extraneous (i.e., not indicated) torques from hose and cable connections, from windage of external parts, and from miscalibration of the load cell. Undetected friction in the trunnion bearings of cradled dynamometers may compromise the torque measurement accuracy. Ideally, well-lubricated antifriction bearings should make no significant contribution to the restraining torque. In practice, however, the unchanging contact region of the balls or other rolling elements on the bearing races makes them prone to brinelling (a form of denting) from forces arising from vibration and unsupported weight of attached devices or even inadvertently during the alignment of connected machinery. The problem can be alleviated by periodic rotation of the (primarily outer) bearing races. In some bearing-in-bearing constructions, the central races are continuously rotated at low speeds by an electric motor, while still others avoid the problem by supporting the stator on hydrostatic oil lift bearings [35].

44.5.5 Costs

The wide range of torque, speed, and power levels, together with the variation in sophistication of associated instrumentation, is reflected in the very wide range of dynamometer prices. Suspension systems of the type illustrated in Figure 44.6 (for which the user must supply the rotating machine) cost $4000–$6000, increasing with capacity [20]. A 100 Hp (74.6 kW) *portable* water brake equipped with a strain-gage load cell and a digital readout instrument for torque, speed, and power cost $5000–$15,000 depending on the sophistication of the data acquisition equipment [33]. Stationary (and some *transportable* [30]) hydraulic dynamometers cost from $125/kW in the smaller sizes [32] down to $40 kW for the very largest [34]. Transportation, installation, and instrumentation can add significantly to these costs. Eddy current dynamometers cost from as little as $65/kW to nearly $800/kW depending on the rated capacity, type of control system, and instrumentation [32,34,35]. Hysteresis brakes with integral speed sensors cost from $4,000 to $20,000 according to capacity [36]. Compatible controllers, from manual to fully programmable for PC test control and data acquisition via an IEEE-488 or other standard interface, vary in price from $700 to $5000. The flexibility and high performance of ac universal dynamometers is reflected in their comparatively high prices of $670–$2200/kW [32,35]. Prices for a wide range of dynamometers, from $13,000 to $1,200,000, are quoted on at least one manufacturer's website [37].

References

1. C. P. Pinney and W. E. Baker, Velocity measurement, In *The Measurement, Instrumentation and Sensors Handbook*, Webster, J. G., ed., Boca Raton, FL: CRC Press, 1999.
2. S. Timoshenko, *Strength of Materials*, 3rd edn., New York: Robert E. Kreiger, 1956, Part I, pp. 281–290, Part II, pp. 235–250.
3. S. J. Citron, On-line engine torque and torque fluctuation measurement for engine control utilizing crankshaft speed fluctuations, U.S. Patent No. 4,697,561, 1987.
4. Supplement to ASME performance test codes, Measurement of shaft power, ANSI/ASME PTC 19.7-1980 (Reaffirmed 1988).
5. See for example the catalog of torque wrench products of Consolidated Devices, Inc., 19220 San Jose Ave., City of Industry, CA 91748, http://www.cditorque.com (accessed on March 16, 2013).
6. I. J. Garshelis, C. R. Conto, and W. S. Fiegel, A single transducer for non-contact measurement of the power, torque and speed of a rotating shaft, SAE Paper No. 950536, 1995.
7. C. C. Perry and H. R. Lissner, *The Strain Gage Primer*, 2nd edn., New York: McGraw-Hill, 1962, p. 9. (This book covers all phases of strain gage technology.)
8. A. Lonsdale, Dynamic rotary torque measurement using surface acoustic waves, *Sensors*, 18, 51–55, 2001.

9. B. D. Cullity, *Introduction to Magnetic Materials*, Reading, MA: Addison-Wesley, 1972, Section 8.5, pp. 266–274.
10. W. J. Fleming, Magnetostrictive torque sensors—Comparison of branch, cross and solenoidal designs, SAE Paper No. 900264, 1990.
11. Torductor®-S, ABB Torque Measurement http://www05.abb.com/global/scot/scot264.nsf/veritydisplay/eef8846525950f25c1257380003623c6/$file/3bse022227r0101_-001.pdf (accessed on March 16, 2013).
12. Y. Nonomura, J. Sugiyama, K. Tsukada, M. Takeuchi, K. Itoh, and T. Konomi, Measurements of engine torque with the intra-bearing torque sensor, SAE Paper No. 870472, 1987.
13. I. J. Garshelis, Circularly magnetized non-contact torque sensor and method for measuring torque using same, U.S. Patent 5,351,555, 1994 and 5,520,059, 1996.
14. Directories of torque transducer (torque sensor) manufacturers can be found at: http://www.thomasnet.com/prodsearch.html?cov=NA&which=prod&what=torque+measurement, http://www.globalspec.com/SpecSearch/Suppliers?Comp=11&ShowAll=true (accessed on March 16, 2013).
15. Search existing U.S. Patents at: http://www.uspto.gov/patents/process/search/index.jsp (accessed on March 16, 2013).
16. Torquemeters Ltd., Ravensthorpe, Northampton NN6 8EH, England/(In USA: Torquetronics, Inc., P.O. Box 100, Allegany, NY 14707), http://www.torquemeters.com (accessed on March 16, 2013).
17. Sensor Technology Ltd., P.O. Box 36, Banbury, Oxon OX15 6JB, UK, http://www.sensors.co.uk/torqsense/ (accessed on March 16, 2013).
18. HBM, Inc., 19 Bartlett St., Marlborough, MA 01752, http://www.hbm.com (accessed on March 16, 2013).
19. S. Himmelstein & Co., 2490 Pembroke, Hoffman Estates, IL 60195, http://www.himmelstein.com (accessed on March 16, 2013).
20. Lebow Products (Honeywell Sensors), 1728 Maplelawn Road, Troy, MI 48099, https://measurementsensors.honeywell.com/pages/ProductInfo.aspx (accessed on March 16, 2013).
21. Sensor Developments, Inc., P.O. Box 290, Lake Orion, MI 48361-0290, http://www.sendev.com (accessed on March 16, 2013).
22. Futek Advanced Sensor Technology, Inc., 10 Thomas, Irvine, CA 92618, http://www.futek.com (accessed on March 16, 2013).
23. Schlenker Enterprises Ltd., P.O. Box 9277, Lombard, IL 60148-9277, http://www.schlenkent.com/torque_transducers.htm (accessed on March 16, 2013).
24. Vibrac Corporation, 16 Columbia Drive, Amherst, NH 03031, http://www.vibrac.com (accessed on March 16, 2013).
25. Kistler North America, 30 Hudson Drive, Novi, MI 48377, http://www.kistler.com (accessed on March 16, 2013).
26. J. Neely et al., Torque ripple sensor and mitigation mechanism, U.S. Patent No. 7,117,754, October 10, 2006.
27. I. J. Garshelis, R. J. Kari, and S. P. L. Tollens, Devices and methods for detecting rates of change of torque, U.S. Patent No. 7,832,289 B2, November 10, 2010.
28. Binsfield Engineering, Inc., 4571 W. MacFarlane, Maple City, MI 49664, http://www.binsfield.com (accessed on March 16, 2013).
29. C. C. Gillispie (ed.), *Dictionary of Scientific Biography*, Vol. XI, New York: Charles Scribner's Sons, 1975.
30. AW Dynamometer, Inc., 10001 W. North Pontiac, IL 61764, http://www.awdynamometer.com (accessed on March 16, 2013).
31. R. Porter (ed.), *The Biographical Dictionary of Scientists*, 2nd edn., New York: Oxford University Press, 1994.
32. Froude Hoffman, Inc., 45225 Polaris Court, Plymouth, MI 48170, http://www.froudehoffman.com (accessed on March 16, 2013).

33. Dyne Systems, Inc. W209 N17391 Industrial Drive, Jackson WI 53037, http://www.dynesystems.com (accessed on March 16, 2013).
34. Taylor Dynamometer, 3602 West Wheelhouse Road, Milwaukee, WI 53208, http://www.taylordyno.com (accessed on March 16, 2013).
35. Drive Source International, Inc., 7900 Durand Ave., Sturtevant, WI, http://www.drivesourceusa.com (accessed on March 16, 2013).
36. Magtrol, Inc., 70 Gardenville Parkway, Buffalo, NY 14224-1322. http://www.magtrol.com (accessed on March 16, 2013).
37. Land and Sea, Inc., 25 Henniker Street, Concord, NH 03301-8528, http://www.land-and-sea.com (accessed on March 16, 2013).
38. MTS, 1400 Technology Drive, Eden Prairie, MN 55344, http://www.mts.com/en/about/news/DEV_002528 (accessed on March 16, 2013).
39. J. Schmitt, Performance and analysis of a dynamic loading four square machine for crawler tractor power train development, SAE Technical Paper 821028, 1982.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

45

Density Measurement

45.1 Introduction **45**-1
45.2 Solid Density **45**-2
45.3 Fluid Density **45**-2
Pycnometric Densitometers • Buoyancy-Type Densitometers • Hydrometers • Hydrostatic Weighing Densitometers • Balance-Type Densitometers • Balanced-Flow Vessel • Chain-Balanced Float Densitometers • Gas Specific Gravity Balance • Buoyancy Gas Balance • Column-Type Densitometers • Vibrating Element Densitometers • Magnetic Drives • Piezoelectric Drives • Vibrating Tube Densitometers • Vibrating Cylinder Densitometers • Tuning Fork Densitometers • Radioactive Densitometers • Refractometer and Index of Refraction Densitometers • Coriolis Densitometers • Absorption-Type Densitometers
Bibliography **45**-16
Partial List of Manufacturers and Suppliers **45**-17

Halit Eren
Curtin University

45.1 Introduction

Density measurements are significant part of instrumentation. Density measurements are made for at least two important reasons: (1) for the determination of the mass and the volume of products and (2) for the quality of the products. In many industrial applications, density measurement ascertains the quality and the value of the products.

Density is defined as the mass of a given volume of a substance under fixed physical conditions. However, ultimate care must be exercised in the measurements since the density varies with pressure and temperature. This variation is much greater and significant in gases.

In many modern applications, the densities of the products are obtained by the sampling techniques. In industrial operations, there are two basic concepts: *static density measurements* and *dynamic (online) density measurements*. Within each concept, there are many different methods that can be employed. These methods are based on different physical principles. In many cases, the application itself and the characteristics of the process determine the best suitable method to be used. Generally, static methods are well developed, are lower in cost, and tend to be more accurate. Dynamic density measurements can be expensive, may require high degree of automation, and use microprocessor and computer-based signal-processing devices. Nevertheless, nowadays, many static methods are also computerized, offering easy-to-use, flexible, and self-calibrating features.

There is no single universally applicable density measurement technique that can be applied to all types of measurements due to different characteristics of materials.

In many cases, density is normalized under reference conditions and the density of a substance is determined by dividing the density of that substance by the density of a standard substance obtained

under the same conditions. This dimensionless ratio is called the *specific gravity* (SG), also termed the *relative density*. The specific gravities of liquid and gases under reference conditions are given by

$$\text{Liquid SG} = \frac{\text{Density of liquid}}{\text{Density of water}} \tag{45.1}$$

$$\text{Gas SG} = \frac{\text{Density of gas}}{\text{Density of air}} \tag{45.2}$$

Commonly accepted sets of conditions are *normal temperature and pressure* (NTP) and *standard temperature and pressure* (STP). NTP is usually taken as the temperature of 0.00 °C and a pressure of 760 mmHg. The NTP is accepted as 15.56 °C at a pressure of 101.325 kPa.

45.2 Solid Density

If the uniformity is maintained, the determination of density of solids is a simple task. Once the volume of the solid and its mass are known, the density can be found using the basic ratio: density = mass/volume (kg m^{-3}).

However, in many applications, solids have different constituents and materials having different ratios of mixture. Their volumes can also change often. In these cases, dynamic methods are employed, such as radioactive absorption types, ultrasonic, or other techniques. Some of these methods are described in detail later in this chapter.

45.3 Fluid Density

The measurement of fluid density is much more complex than the solids and there are many different available techniques. This complexity is mainly due to the nature in processes, the variations in fluid densities, and the diverse characteristics of processes as well as characteristics of fluids themselves. Some of these methods are custom designed and applicable to only situations they are designed for. Others are very similar in principles and technology, and they may be applicable on different types of fluids. At present, apart from conventional methods, there are many novel and unusual techniques that are undertaken in research and development efforts. For example, densitometers based on electromagnetic principles can be only small part of an intelligent instrumentation system.

Fluids can be divided to liquids and gases. Extra care and further considerations are necessary in gas density measurements. For example, perfect gases contain an equal number of molecules under the same conditions and volumes. Therefore, molecular weights can be used in the density measurements.

Depending on the applications, fluid densities can be measured in either *static* or *dynamic* forms. In general, static density measurements of fluids are well developed, precise, and have greater resolution and accuracy than most dynamic techniques. Pycnometers and buoyancy are examples of static techniques that can be adapted to cover small density ranges with a high resolution and precision. Nowadays, many manufacturers offer dynamic instruments previously known to be static. Also, many static density measurement devices are computerized and come with appropriate hardware and software. In general, static-type measurements are employed in laboratory conditions, and dynamic methods are employed in real-time measurements where the properties of a fluid can vary from time to time.

In this chapter, the discussion will concentrate on the commonly applied, modern density measuring devices, which include

1. Pycnometric densitometers
2. Buoyancy-type densitometers
3. Hydrometers

4. Hydrostatic weighing densitometers
5. Balance-type densitometers
6. Column-type densitometers
7. Vibrating element densitometers
8. Radioactive densitometers
9. Refractometer and index of reflection densitometers
10. Coriolis densitometers
11. Absorption-type densitometers

45.3.1 Pycnometric Densitometers

Pycnometers are static devices. They are manufactured as fixed-volume vessels that can be filled with the sample liquid. The density of the fluid is measured by weighing the sample. The simplest version consists of a vessel in the shape of a bottle with a long stopper containing a capillary hole, as shown in Figure 45.1. The capillary is used to determine the exact volume of the liquid, thus giving high resolution when filling the pycnometer. The bottle is first weighed empty and then with distilled–aerated water to determine the volume of the bottle. The bottle is then filled with the process fluid and weighed again. The density is determined by dividing the mass by the volume. The SG of the liquid is found by the ratio of the fluid mass to water mass. When pycnometers are used, for good precision, ultimate care must be exercised during the measurements; that is, the bottle must be cleaned after each measurement, the temperature must be kept constant, and precision balances must be used. In some cases, to ensure filling of the pycnometer, twin capillary tubes are used. The two capillaries, made of glass, are positioned such that the fluid can be driven into the vessel under vacuum conditions. Accurate filling to graduation marks on the capillary is then made.

The pycnometers have to be lightweight, strong enough to contain samples, and they need to be nonmagnetic for accurate weighing to eliminate possible ambient magnetic effects. Very high-resolution balances must be used to detect small differences in weights of gases and liquids. Although many

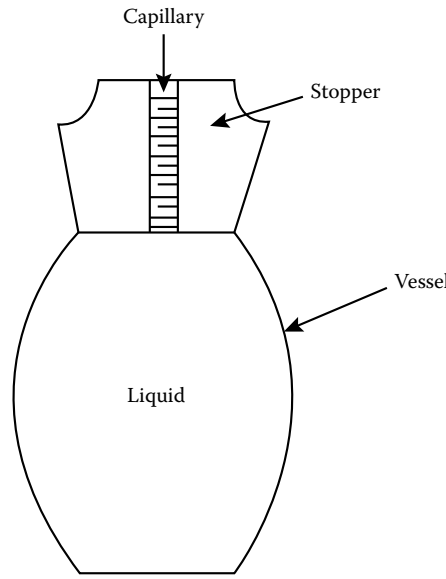


FIGURE 45.1 A pycnometer. A fixed-volume container is filled with liquid and weighed accurately; the capillary is used to determine the exact volume of the liquid.

pycnometers are made of glass, they can also be made from metals to give enough strength for the density measurements of gases and liquids at extreme high pressures. In many cases, metal pycnometers are necessary for taking samples from the line of some rugged processes.

Pycnometers have advantages and disadvantages. Advantages are that if used correctly, they are accurate, and they can be used for both density and SG measurements. The disadvantages include

1. Great care must be exercised for accurate results.
2. The sample must be taken off-line, with consequent time lag in results. This creates problems of relating samples to the materials that exist in the actual process.
3. High-precision pycnometers are expensive. They require precision weighing scales and controlled laboratory conditions. Specialized techniques must be employed to take samples in high-pressure processes and hostile conditions, such as offshore installations.
4. Their good performances might depend on the skill of the operator.

45.3.2 Buoyancy-Type Densitometers

The buoyancy method basically uses Archimedes principle. A suspended sinker, with a known mass and volume attached to a fine wire, is totally immersed in the sample liquid. A precision force balance is used to measure the force to support the sinker. Once the mass, volume, and supporting weight of the sinker are known, the density of the liquid can be calculated. However, some corrections need to be made for surface tension on the suspension wire, the cubicle expansion coefficient of the sinker, and the temperature of process. Buoyancy-type densitometers give accurate results and are used for the calibration of the other liquid density transducers.

One advanced version of the buoyancy technique is the magnetic suspension system. The sinker is fully enclosed in a pressure vessel, thus eliminating surface tension errors. Their uses can also be extended to applications such as the SG measurements under low vapor pressures and density measurements of hazardous fluids.

45.3.3 Hydrometers

Hydrometers are the most commonly used devices for measurement of the density of liquids. They are so commonly used that their specifications and procedure of use are described by national and international standards, such as ISO 387. The buoyancy principle is used as the main technique of operation. The volume of fixed mass is converted to a linear distance by a sealed bulb-shaped glass tube containing a long-stem measurement scale, shown in Figure 45.2. The bulb is ballasted with a lead shot and pitch, the mass of which depends on the density range of the liquid to be measured. The bulb is simply placed into the liquid and the density is read from the scale. The scale is graduated in density units such as kg m^{-3}. However, many alternative scales are offered by manufacturers, such as SG, American Petroleum Institute (API) gravity, Brix, and Brine. Hydrometers can be calibrated for different ranges for surface tensions and temperatures. Temperature corrections can be made for set temperature such as 15°C, 20°C, or 25°C. ISO 387 covers a density range of 600–2000 kg m^{-3}. Hydrometers have a number of advantages and disadvantages. The advantages include

1. Relatively low cost and easy to use
2. Good resolution for small range
3. Traceable to national and international standards

The disadvantages include

1. They have small span; therefore, a number of meters are required to cover a significant range.
2. They are made from glass and fragile. Metal and plastic versions are not as accurate.

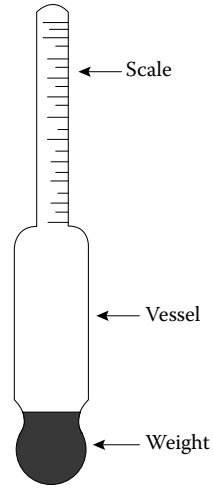


FIGURE 45.2 A hydrometer. A fixed-weight and fixed-volume bulb is placed into the liquid. The bulb sinks in the liquid, depending on its density. The density is read directly from the scale. Temperature correction is necessary.

3. The fluid needs to be an off-line sample, not representing the exact conditions of the process. There are pressure hydrometers for low vapor pressure hydrocarbons, but this adds a need for accurately determining the pressure too.
4. If good precision is required, they are difficult to use, needing surface tension and temperature corrections. Further corrections could be required for opaque fluids.

45.3.4 Hydrostatic Weighing Densitometers

The most common device using a hydrostatic weighing method consists of a U-tube that is pivoted on flexible end couplings. A typical example is shown in Figure 45.3. The total weight of the tube changes, depending on the density of fluid flowing through it. The change in the weight needs to be measured accurately, and there are a number of methods employed to do this. The most common commercial

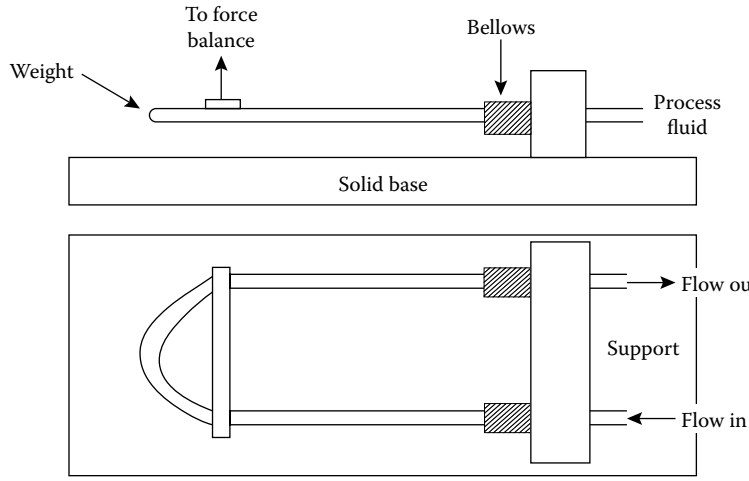


FIGURE 45.3 Hydrostatic weighing. The total weight of a fixed-volume tube containing liquid is determined accurately. The density is calculated using mass-to-volume ratio.

meters use a force balance system. The connectors are stainless steel bellows. In some cases, rubber or PTFE are used, depending on the process fluid characteristics and the accuracy required. There are temperature and pressure limitations due to bellows and the structure of the system may lead to a reading offset. The meter must be securely mounted on a horizontal plane for optimal accuracy.

The advantages of this method include

1. They give continuous reading and can be calibrated accurately.
2. They are rugged and can be used for two-phase liquids such as slurries, sugar solutions, and powders.

The disadvantages of these meters include

1. They must be installed horizontally on a solid base. These meters are not flexible enough to adapt to any process; thus, the process must be designed around it.
2. They are bulky and cumbersome to use.
3. They are unsuitable for gas density measurements.

45.3.5 Balance-Type Densitometers

Balance-type densitometers are suitable for liquid and gas density measurements. There are many different types; four of the most commonly used ones are discussed in the following.

45.3.6 Balanced-Flow Vessel

A fixed-volume vessel, shown in Figure 45.4, is employed for the density measurements. While the liquid is flowing continuously through the vessel, it is weighed automatically by a sensitive scale by means of a spring balance system or a pneumatic force balance transmitter. Because the volume and the weight of the liquid are known, the density or SG can easily be calculated and scaled in respective units. In the design process, extra care must be exercised for the flexible end connections.

45.3.7 Chain-Balanced Float Densitometers

In chain-balanced float densitometers, a self-centering, fixed-volume, submerged plummet is used, as illustrated in Figure 45.5. The plummet is located entirely under the liquid surface. At balance, the plummet operates without friction and is not affected by surface contamination. Under steady-state

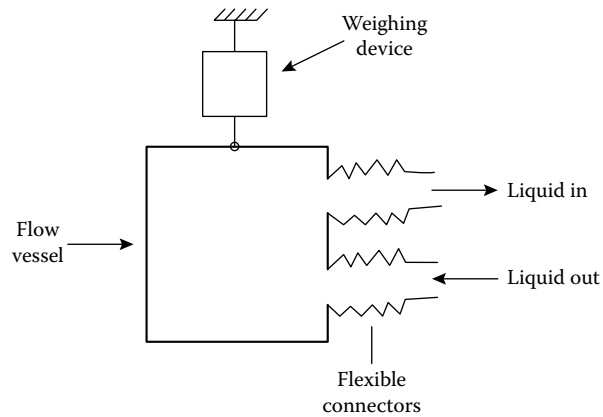


FIGURE 45.4 A balanced-flow vessel. An accurate spring balance or force balance system is used to weigh the vessel as the liquid flows through it.

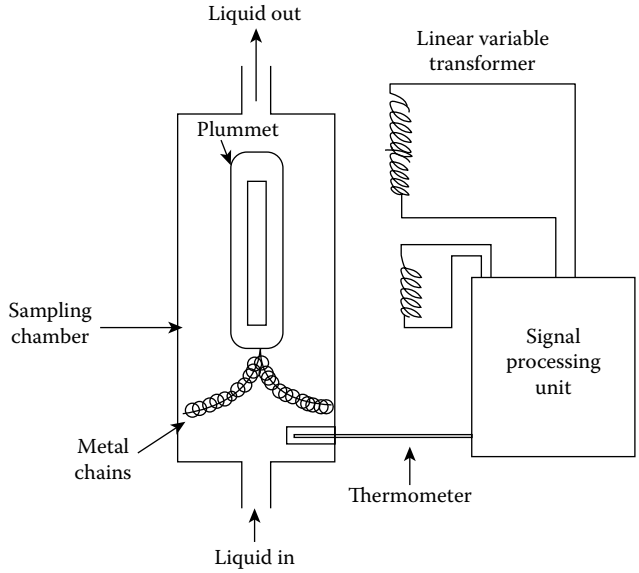


FIGURE 45.5 Chain-balanced float. The fixed-volume and fixed-weight plummet totally suspended in the liquid assumes equilibrium position, depending on the density. The force exerted by the chains on the plummet is a function of plummet position; hence, the measured force is proportional to the density of the liquid.

conditions, the plummet assumes a stable position. The effective weight of the chain on the plummet varies, depending on the position of the plummet, which in turn is a function of the density of the liquid. The plummet contains a metallic transformer core that transmits changes in the position to be measured by a pickup coil. The voltage differential, a function of plummet displacement, is calibrated as a measure of variations in SG. A resistance thermometer bridge is used for the compensation of the effects of temperature on the density.

45.3.8 Gas Specific Gravity Balance

A tall column of gas is weighed by the floating bottom of the vessel. This weight is translated into the motion of an indicating pointer, which moves over a scale graduated in units of density or SG. This method can be employed for the density measurement of gases.

45.3.9 Buoyancy Gas Balance

In this instrument, a displacer is mounted on a balance beam in a vessel, as shown in Figure 45.6. The displacer is balanced for air, and the manometer reading is noted at the exact balance pressure. The air is then displaced by gas, and the pressure is adjusted until the same balance is restored. The ratio of the pressure of air to the pressure of gas is then the density of the gas relative to air. This method is commonly applied under laboratory conditions and is not suitable for continuous measurements.

45.3.10 Column-Type Densitometers

There are a number of different versions of column methods. As a typical example, a reference column method is illustrated in Figure 45.7. A known head of sample liquid and water from the respective bubbler pipes is used. A differential pressure-measuring device compares the pressure differences, proportional to relative densities of the liquid and the water. By varying the depth of immersion of the pipes,

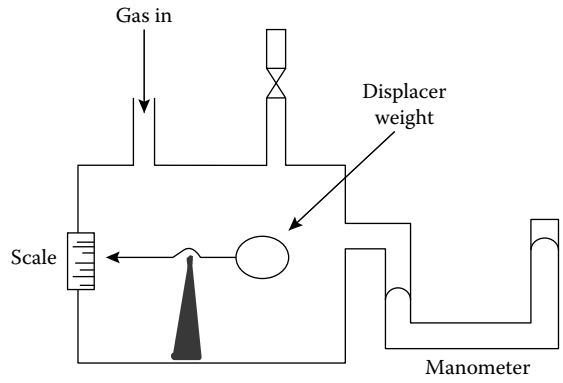


FIGURE 45.6 Buoyancy gas balance. The position of the balance beam is adjusted by a set pressure air. The air is then displaced by gas of unknown density having the same pressure. The difference in the reading of the balance beam gives the SG of the gas. The pressures are read on the manometer.

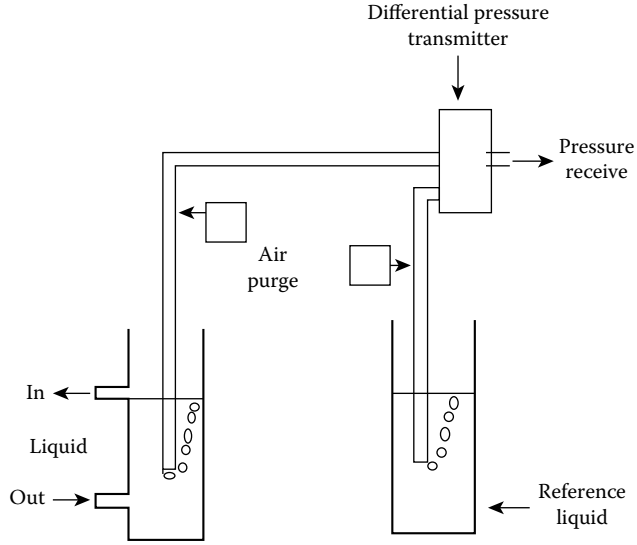


FIGURE 45.7 A reference column densitometer. Two identical tubes, having the same distance from the surface, are placed in water and liquid. Water with known density characteristics is used as the reference. The pressure necessary to displace the fluid inside the tubes is proportional to the density of the fluid. The pressure difference at the differential pressure transmitter is translated into relative densities.

a wide range of measurements can be obtained. Both columns must be maintained at the same temperature to avoid the necessity for corrections due to possible temperature effects.

A simpler and more widely used method of density measurement is achieved by the installation of two bubbler tubes as illustrated in Figure 45.8. The tubes are located in the sample fluid such that the end of one tube is higher than that of the other. The pressure required to bubble air into the fluid from both tubes is equal to the pressure of the fluid at the end of the bubbler tubes. The outlet of one of the tubes is higher than the other one. The distances of the openings of the tubes are fixed; hence, the difference in the pressure is the same as the weight of a column of liquid between the ends. Therefore, the differential pressure measurement is equivalent to the weight of the constant volume of the liquid, and calibrations can be made such that they have a direct relationship to the density of the liquid. This method is accurate to within 0.1%–1% SG. It must be used with liquids that do not crystallize or settle in the measuring chamber during measurements.

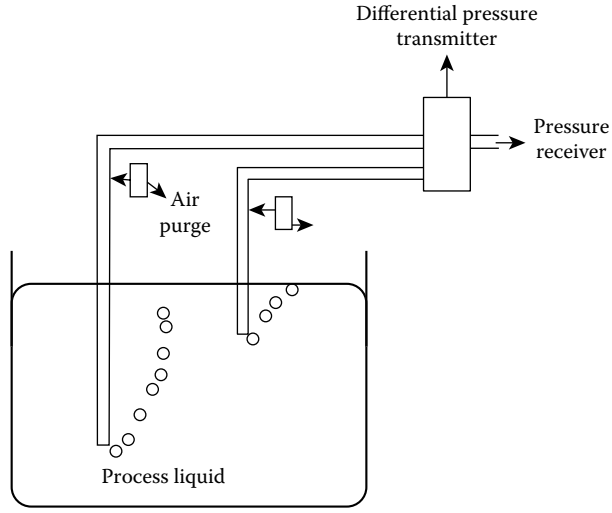


FIGURE 45.8 A two-tube column densitometer. The pressure difference at the differential pressure transmitter depends on the relative positions of the openings of the pipes and the density of liquid. Once the relative positions are fixed, the pressure difference can be related to the equivalent weight of the liquid column at the openings of the pipes, hence the density of the liquid.

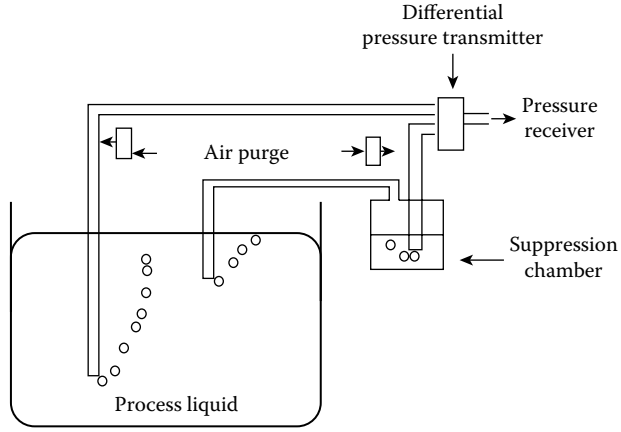


FIGURE 45.9 Suppression-type, two-tube column densitometer. Operation principle is the same as in densitometers in Figures 45.7 and 45.8. In this case, the suppression chamber gives greater accuracy in readings.

Another version is the range suppression type, which has an additional constant pressure drop chamber as shown in Figure 45.9. This chamber is in series with the low-pressure side to give advantages in scaling and accurate readings of densities.

45.3.11 Vibrating Element Densitometers

If a body containing or surrounded by a fluid is set to resonance at its natural frequency, then the frequency of oscillation of the body will vary with the properties of the fluid. The natural frequency is directly proportional to the stiffness of the body and inversely proportional to the combined mass of the body and the fluid. Vibrations are also dependent on the shape, size, and elasticity of the material,

induced stress, mass, and mass distribution of the body. Basically, the vibration of the body can be equated to motion of a mass attached to a mechanical spring. Hence, an expression for the frequency can be written as

$$\text{Resonant frequency} = \text{SQRT}\frac{K}{M + k\rho} \tag{45.3}$$

where
K is the system stiffness
M is the transducer mass
k is the system constant
ρ is the fluid density

A factor common to all types of vibrating element densitometers is the problem of setting the element in vibration and maintaining its natural resonance. There are two drives for the purpose.

45.3.12 Magnetic Drives

Magnetic drives of the vibrating element and the pickup sensors of vibrations are usually achieved using small coil assemblies. Signals picked up by the sensors are amplified and fed back as a drive to maintain the disturbing forces on the vibrating body of the meter.

In order to achieve steady drives, the vibrating element sensor can be made from nonmagnetic materials. In this case, small magnetic armatures are attached.

The main advantage of magnetic drive and pickup systems is that they are noncontact methods. They use conventional copper windings and they are reliable within a wide temperature range of –200 °C to +200 °C.

45.3.13 Piezoelectric Drives

A selective range of piezoelectric materials are available to meet the requirements of driving vibrating elements. These materials demonstrate good temperature characteristics as do magnetic drive types. They also have the advantage of being low in cost. They have high impedances thus making the signal conditioning circuitry relatively easy. They do not require magnetic sensors.

The piezoelectric drives are mechanically fixed on the vibrating element by adhesives. Therefore, attention must be paid to the careful placement of the mount in order to reduce the strain experienced by the piezo element due to thermal and pressure stresses while the instrument is in service.

A number of different types of densitometers have been developed that utilize this phenomenon. The three main commercial types are introduced here.

45.3.14 Vibrating Tube Densitometers

Vibrating tube densitometers are suitable for applications in highly viscous liquids and slurries. The mode of operation of vibration tube meters is based on the transverse vibration of tubes as shown in Figure 45.10. The tube and the driving mechanisms are constrained to vibrate in a single plane. As the liquid moves inside the tube, the density of the entire liquid is measured. The tube length is set to approximately 20 times longer than the tube diameter.

A major design problem with the vibrating tube method is the conflict of limiting the vibrating element to a finite length and fixing the nodes accurately. Special attention must be paid to avoid any exchange of vibration energy outside the sensory tube. In one type, a single tube is used and it has one important disadvantage of presenting obstruction to the flow, thus experiencing some pressure losses. In another type, a twin tube is used, which offers a very small blockage (Figure 45.11), and it can easily

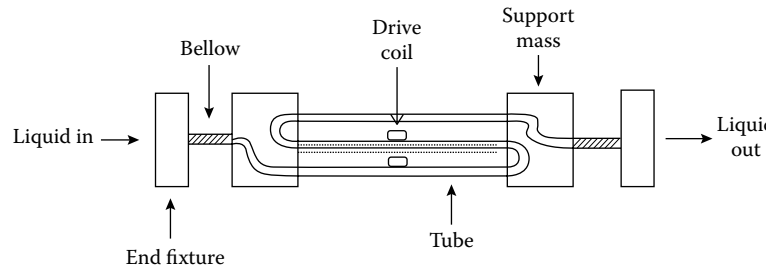


FIGURE 45.10 A vibrating tube densitometer. Tube containing fluid is vibrated at resonant frequency by an electromagnetic vibrator. The resonant frequency, which is a function of the density of the fluid, can be measured accurately. The tube is isolated from the fixtures by carefully designed bellows.

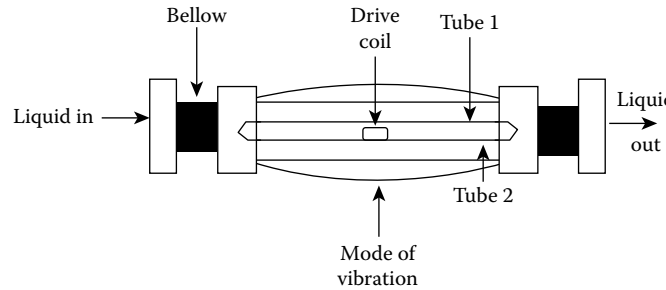


FIGURE 45.11 Two-tube vibrating densitometer. The two tubes are vibrated in antiphase for greater accuracy. Twin-tube densitometers are compact in size and easy to use.

be inspected and cleaned. The twin-tube types are compact size, hence offering another distinct advantage. Often, the twin tubes are designed to achieve a good dynamic balance, with the tubes vibrating in antiphase. Their nodes are fixed at the ends, hence demonstrating maximum sensitivity to installation defects, clamping, and mass loading.

The main design problems of the vibrating tube sensors are in minimizing the influence of end padding and overcoming the effects of pressure and temperature. Bellows are used at both ends of the sensor tubes to isolate the sensors from external vibrations. Bellows also minimize the end loadings due to differential expansions and installation stresses.

The fluid runs through the tubes; therefore, no pressure balance is required. Nevertheless, in some applications, the pressure stresses the tubes, resulting in changes in stiffness. Some manufacturers modify the tubes to minimize the pressure effects. In these cases, corrections are necessary only when high accuracy is mandatory. The changes in the Young's modulus with temperature can be reduced to near zero by using materials such as Ni-span-C compounds whenever corrosive properties of fluids permit. Usually, manufacturers provide pressure and temperature correction coefficients for their products.

It is customary to calibrate each vibration element densitometer against others as a transfer of standards. Often, the buoyancy method is used for calibration purposes. The temperature and pressure coefficients are normally found by exercising the transducer over a range of temperatures and pressures on a liquid with well-known properties. Prior to calibration, the vibration tube densitometers are subjected to a programmed burn-in cycle to stabilize them against temperature and pressure fluctuations.

45.3.15 Vibrating Cylinder Densitometers

A thin-walled cylinder, with a 3:1 length to diameter ratio, is fixed with stiff ends. The thickness of the cylinder wall varies from 25 to 300 μm, depending on the density range and type of fluid used. The cylinder can be excited to vibrate in a hoop mode by magnetic drives mounted either in or outside the cylinder.

For good magnetic properties, the cylinder is made of corrosion-resistant magnetic materials. Steel such as FV520 is often used for this purpose. Such materials have good corrosion-resistance characteristics; unfortunately, due to their poor thermoelastic properties, they need extensive corrections for temperature. Nickel–iron alloys such as Ni-span-C are often used to minimize the effects of temperature. Once correctly treated, the Ni-span-C alloy has near-zero Young's modulus properties. Because the cylinder is completely immersed in the fluid, there are no pressure coefficients.

The change in the resonant frequency is determined by the local mass loading of the fluid in contact with the cylinder. The curve of frequency against density is nonlinear and has a parabolic shape, thus requiring linearization to obtain practical outputs. The resonant frequency varies from 2 to 5 kHz, depending on the density range of the instrument. The cylinders need precision manufacturing and thus are very expensive to construct. Each meter needs to be calibrated individually for different temperatures and densities to suit specific applications. In the case of gas density applications, gases with well-known properties (e.g., pure argon or nitrogen) are used for calibrations. In this case, the meters are subjected to a gas environment with controlled temperature and pressure. The calibration curves are achieved by repetitions to suit the requirements of individual customer for a particular application. In the case of liquids, the meters are calibrated with liquids of known densities, or they are calibrated against another standard (e.g., pycnometer or buoyancy-type densitometers).

Vibration cylinder-type densitometers have zero pressure coefficients and they are ideal for liquefied gas products or refined liquids. Due to relatively small clearances between cylinder and housing, they require regular cleaning. They are not suitable for liquids or slurries with high viscous properties.

45.3.16 Tuning Fork Densitometers

These densitometers make use of the natural frequency of low-mass tuning forks, shown in Figure 45.12. In some cases, the fluid is taken into a small chamber in which the electromechanically driven forks are situated. In other cases, the fork is inserted directly into the liquid. Calibration is necessary in each application. The advantages of vibrating element meters include

1. They are suitable for both liquids and gases with reasonable accuracy.
2. They can be designed for real-time measurements.
3. They can easily be interfaced because they operate on frequencies and they are inherently digital.
4. They are relatively robust and easy to install.
5. Programmable and computerized versions are commonly available. Programmed versions make all the corrections automatically. They provide the output of live density, density at reference conditions, relative density, SG, concentration, solid contents, etc.

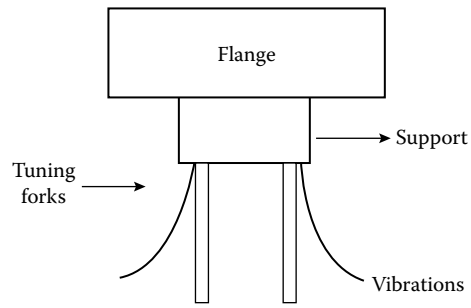


FIGURE 45.12 The tuning fork densitometer. Twin forks are inserted into the fluid, and the natural frequencies are measured accurately. The natural frequency of the forks is a function of the density of the fluid.

The disadvantages include

1. They do not relate directly to primary measurements; therefore, they must be calibrated.
2. They all have problems in measuring multiphase liquids.

45.3.17 Radioactive Densitometers

As radioactive isotopes decay, they emit radiation in the form of particles or waves. This physicalphenomenon can be used for the purposes of density measurement. For example, γ rays can be passed through the samples and their rate of arrivals, which is dependent on the density of the materials in between, can be determined using ion- or scintillation-based detectors. Generally, γ-ray mass absorption rate is independent of material composition; hence, they can be programmed for wide-range applications. Densitometers based on radiation methods can provide accuracy up to +0.0001 g mL^{-1}. Many of these devices have self-diagnostic capabilities and are able to compensate for drifts caused by the decay of the source, thus pinpointing any signaling problems.

If γ rays of intensity J_0 penetrate a material of a density ρ and thickness d, then the intensity of the radiation after passing through the material can be expressed by

$$J = J_0 \exp(n\rho d) \tag{45.4}$$

where n is the mass absorption coefficient.

The accuracy of the density measurement depends on the accuracy of the measurement of the intensity of the radiation and the path length d. A longer path length through the material gives a stronger signal to be detected.

For accurate operations, there are many arrangements for relative locations of transmitters and detectors, some of which are illustrated in Figures 45.13 and 45.14. Generally, the source is mounted in a lead container clamped onto the pipe or the container wall. In many applications, the detector is also clamped onto the wall.

The advantages of using radioactive methods include

1. The sensor does not touch the sample; hence, there is no blockage to the path of the liquid.
2. Multiphase liquids can be measured.
3. They come in programmable forms and are easy to interface.
4. They are most suitable in difficult applications, such as mining, chemical, metallurgical, and other heavy process industries.

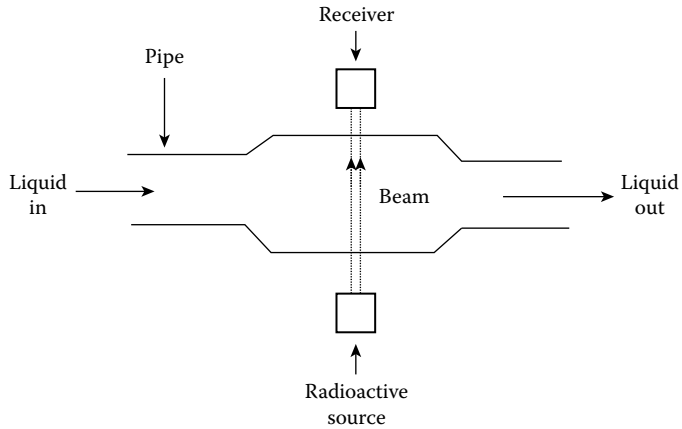


FIGURE 45.13 Fixing radioactive densitometer on an enlarged pipe. The pipe is enlarged to give longer beam length through the liquid and hence better attenuation of the radioactive energy.

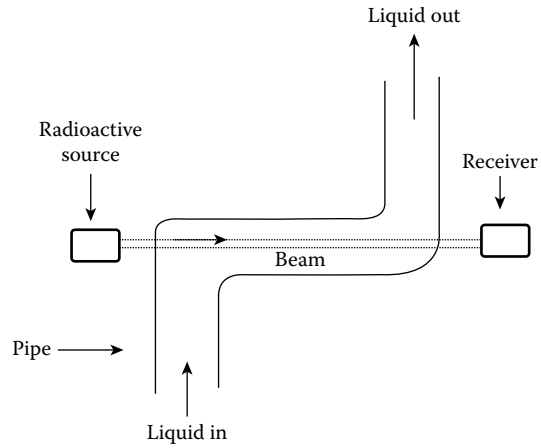


FIGURE 45.14 Fixing radioactive densitometer on an elongated pipe. Elongated path yields a longer path length of the radioactive energy through the liquid, hence a stronger attenuation.

The disadvantages include

1. A radioactive source is needed; hence, there is difficulty in handling.
2. For reasonable accuracy, a minimum path length is required.
3. There could be long time constants, making them unsuitable in some applications.
4. They are suitable only for solid and liquid density measurements.

45.3.18 Refractometer and Index of Refraction Densitometers

Refractometers are essentially optical instruments operating on the principles of refraction of light traveling in liquid media. Depending on the characteristics of the samples, measurement of refractive index can be made in a variety of ways (e.g., critical angle, collimation, and displacement techniques). Usually, an in-line sensing head is employed, whereby a sensing window (commonly known as a prism) is wetted by the product to be measured. In some versions, the sensing probes must be installed inside the pipelines or in tanks and vessels. They are most effective in reaction-type process applications where blending and mixing of liquids take place. For example, refractometers can be used to measure the contents of the dissolved soluble solids accurately.

Infrared diodes, lasers, and other lights may be used as the light sources. However, this measurement technique is not recommended in processes containing suspended solids, high turbidity, entrained air, heavy colors, and poor transparency or opacity, or substances having extremely high flow rates. The readings are automatically corrected for variations in process temperature. The processing circuitry can include signal outputs adjustable in both the frequency and the duration.

Another version of a refractometer is the index of refraction-type densitometer. For example, in the case of position-sensitive detectors, the index of refraction of liquid under test is determined by measuring the lateral displacement of a laser beam. When the laser beam impinges on the cell at an angle of incidence, as in Figure 45.15, the axis of the emerging beam is displaced by the cell wall and by the inner liquid. The lateral displacement can accurately be determined by position-sensitive detectors. For maximum sensitivity, the devices need to be calibrated with the help of interferometers.

Refractometers are often used for the control of adulteration of liquids of common use (e.g., edible oils, wines, and gasoline). They also find application in pulp and paper, food and beverage, sugar, dairy, and other chemical industries.

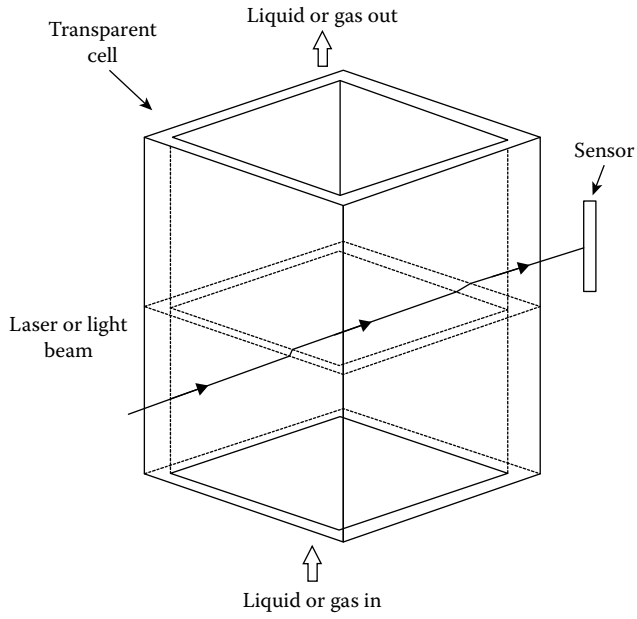


FIGURE 45.15 The index of refraction-type densitometer. The angle of refraction of the beam depends on the shape, size, and thickness of the container and the density of fluid in the container. Because the container has the fixed characteristics, the position of the beam can be related to density of the fluid. Accurate measurement of the position of the beam is necessary.

45.3.19 Coriolis Densitometers

The Coriolis density metering systems employ similar principles as vibrating tube densitometers with slight variations in the design. Often, they contain a number of sensors and signal-processing transmitters. Each sensor consists of one or two flow tubes enclosed in the sensor housing. They are manufactured in various shapes and sizes. The sensor tubes are securely anchored at the fluid inlet and outlet points and force is vibrated at the free end, as shown in Figure 45.16. The sensor operates by applying Newton's second law of motion ($F = ma$).

Inside the housing, the tubes are vibrated in their natural frequencies using drive coils and related feedback circuits. This resonant frequency of the assembly is a function of the geometry of the vibrating element, the material of construction, and the mass of the tube assembly. The tube mass comprises two parts: the mass of the tube itself and the mass of the fluid inside the tube. The mass of the tube is fixed for a given sensor. The mass of fluid in the tube is equal to the fluid density multiplied by the volume.

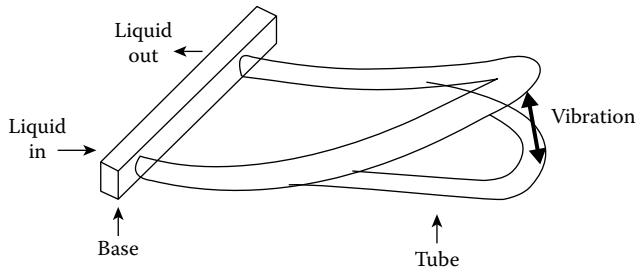


FIGURE 45.16 Coriolis densitometer. Vibration of the tube is detected and related to the mass and flow rate of the fluid. Further calibrations and calculations must be made to determine the densities of fluids in the pipes.

Because the tube volume is constant, the frequency of oscillation can be related directly to the fluid density. Therefore, for a given geometry of tube and the material of the construction, the density of the fluid can be determined by measuring the resonant frequency of the vibration. Temperature sensors are used for overcoming the effects of changes in modulus of elasticity of the tube. The fluid density is calculated using a linear relationship between the density and the tube period and the calibration constants.

Special peripherals, based on microprocessors, are offered by various manufacturers for a variety of measurements. However, all density peripherals employ the natural frequency of the sensor coupled with the sensor temperature to calculate online density of the process fluid. Optional communication, interfacing facilities, and appropriate software are also additional features of these densitometers.

45.3.20 Absorption-Type Densitometers

Absorption techniques are also used for density measurements in some custom-designed and custom-specific applications. X-rays, visible light, UV light, and sonic absorptions are typical examples of this method. Essentially, attenuation and phase shift of a generated beam going through the sample is sensed to be related to the density of the sample. Most absorption-type densitometers are carefully designed for applications having particular characteristics. Two typical examples are (1) UV absorption or x-ray absorptions that are used for determining the local densities of mercury deposits in arc discharge lamps and (2) ultrasonic density sensors that are used in difficult density measurements (e.g., density measurement of slurries). The lime slurry, for example, is a very difficult material to handle. It has a strong tendency to settle out and coat all equipment with which it comes in contact with. An ultrasonic density control sensor can fully be emerged into the agitated slurry, thus avoiding the problems of coating and clogging. Inasmuch as the attenuation of the ultrasonic beam is proportional to the suspended solids, the resultant electronic signal is proportional to the SG of the slurry. Such devices can give accuracy up to 0.01%. The ultrasonic device measures the percentage of the suspended solids in the slurry by providing a close approximation of the SG.

Bibliography

Boiten, W., *Hydrometry: A Comprehensive Introduction to the Measurement of Flow in Open Channels*, CRC Press, Boca Raton, FL, 2008.

Davis, R. S., *Measurement of Mass and Density*, Encyclopedia of Applied Physics, Wiley On-Line Library. http://onlinelibrary.wiley.com/doi/10.1002/3527600434.eap220.pub2/abstract (accessed on July 2, 2012).

Density, New world encyclopedia. http://www.newworldencyclopedia.org/entry/Density (accessed on October 2, 2011).

Density Measurement, Technical data sheet, Emerson process. http://www.emersonprocess.com/rosemount/document/tds/3208_00q.pdf (accessed on July 3, 2012).

Fujii, K., Precision density measurements of solid materials and hydrostatic weighing, *Measurement and Science Technology*, 17, 2551–2559, 2006.

Grimes, C. A., Kouzoudis, D., and Mungle, C., Simultaneous measurement of liquid density and viscosity using remote query magnetoelastic sensors, *Review of Scientific Instruments*, 71(10), 3822–3824, 2000.

Gupta, S. V., *Practical Density Measurements and Hydrometer*, Taylor & Francis Group, Boca Raton, FL, 2002.

Mass and Related Quantities Capabilities, National Measurement Institute, Australia http://www.measurement.gov.au/Services/calibrationtesting/Pages/Massandrelatedquantities.aspx (accessed on August 3, 2012).

Pendrill, L. R., Refractometry and gas density, *Metrologia*, 41(2), l, 2004.

Wiederseiner, S., Andreini, N., Epely-Chauvin, G., and Ancey, C., Refractive-index and density matching in concentrated particle suspensions: A review, *Experiments in Fluids*, 50(5), 1183–1206, 2011.

Partial List of Manufacturers and Suppliers

ABB Automation, Inc.—TOTALFLOW® Products Group, 3010 Briarpark, Houston, TX 77042, Phone: 713-821-8128, 866-365-8299 (toll free), Fax: 713-821-3530, www.totalflow.com

AMETEK Power Instruments, 255 N. Union St., Rochester, NY 14605, Phone: 585-263-7700, 800-881-4156 (toll free), Fax: 585-262-4777, www.ametekpower.com

Analtech, Inc., 75 Blue Hen Dr., Newark, DE 19713, Phone: 302-737-6960, 800-441-7540 (toll free), Fax: 302-737-7115, www.ichromatography.com/

Analytical Process, Inc., 3311 Richmond Ave., Suite 110, Houston, TX 77098, Phone: 713-526-6552, Fax: 713-522-1563, www.processanalytical.com

Automation Products, Inc., 3030 Max Roy St., Houston, TX 77008, Phone: 713-869-0361, 800-231-2062 (toll free), Fax: 713-869-7332, www.dynatrolusa.com

CCSI, 221 Beaver St., Akron, OH 44301, Phone: 330-376-3600, Fax: 330-376-8500, www.ccsi-inc.com

Central States Group, 8720 S. 137th Circle, Omaha, NE 68138-6198, Phone: 402-894-1003, 800-383-2747 (toll free), Fax: 402-894-1339, www.centralstatesgroup.com

Density Analytics LLC, 12 Grand Banks Circle, Marlton, NJ 08053, Phone: 856-797-1449, www.densityanalytics.com/

Dynatrol® Div., Automation Products, Inc., 3030 Maxroy St., Houston, TX 77008, Phone: 713-869-0361, 800-231-2062 (toll free), Fax: 713-869-7332, www.dynatrolusa.com

Eagle Eye Power Solutions LLC, 4230 N. Oakland Ave. #176, Milwaukee, WI 53211, Phone: 414-962-3377, 877-805-3377 (toll free), Fax: 414-962-3660, eepowersolutions.com/

Flexial Corp., Edge Welded Metal Bellows, P.O. Box 3105, Cookeville, TN 38502-3105, Phone: 931-432-1853, Fax: 931-432-1889, www.flexial.com

FMC Technologies-Material Handling Solutions, P.O. Box 1370, Tupelo, MS 38802, Phone: 662-869-5711, Fax: 662-869-7401, www.fmctechnologies.com

Graham Optical Systems, 9530 Topanga Canyon Blvd., Chatsworth, CA 91311, Phone: 818-700-1263, Fax: 818-700-1627, www.grahamoptical.com

Grainger Industrial Supply, 100 Grainger Pkwy., Lake Forest, IL 60045, Phone: 847-535-1000, 888-361-8649 (toll free), Fax: 847-535-9123, www.grainger.com/thomasnet

Honeywell Sensing and Control, 1985 Douglas Dr. North, MN10-192B, Golden Valley, MN 55422, Phone: 763-954-4818, 800-446-6555 (toll free), sensing.honeywell.com/index.cfm/ci_id/15…

ITT Barton, 800 S. Turnbull Canyon Rd., City of Industry, CA 91745, Phone: 626-961-2547, 800-291-3550 (toll free), Fax: 800-408-2168

Ives Equipment Corp., 601 Croton Rd., King of Prussia, PA 19406-3111, Phone: 610-265-3071, 877-768-1600 (toll free), Fax: 610-768-1605, www.ivesequipment.com

KROHNE, Inc., 7 Dearborn Rd., Peabody, MA 01960, Phone: 978-535-6060, 800-356-9464 (toll free), Fax: 978-535-1720, www.krohne.com/index.php?id=12076

LND, Inc., 3230 Lawson Blvd., Oceanside, NY 11572, Phone: 516-678-6141, Fax: 516-678-6704, www.lndinc.com

Mettler-Toledo, Inc., 5 Barr Rd., Ithaca, NY 14850-9117, Phone: 607-257-6000, 800-836-0836 (toll free), Fax: 607-257-5232, us.mt.com/us/en/home.html

Micromeritics Instruments Corp., 4356 Communications Dr., Norcross, GA 30093, Phone: 770-662-3633, Fax: 770-662-3696, www.micromeritics.com

Micro Motion, a Division of Emerson Process Management, 7070 Winchester Circle, Boulder, CO 80301, Phone: 303-530-8400, 800-760-8119 (toll free), Fax: 503-530-8459, www2.emersonprocess.com/en-us/brands/mic…

MPE-Modern Process Equipment, Inc., 3125 S. Kolin Ave., Chicago, IL 60623-4831, Phone: 773-254-3929, Fax: 773-254-3935, www.mpechicago.com

MC Schroeder Co., 3625 Denver Dr., Denver, NC 28037, Phone: 704-489-6171, 800-333-3830 (toll free), Fax: 704-489-6175, www.mcschroeder.com/

NDS Products, Inc., 111 Anderson St., Pasadena, TX 77506, Phone: 713-475-2986, 800-413-4750 (toll free), Fax: 713-477-6741

PICSINC, 81 Lancaster Ave., Suite 202, Malvern, PA 19355, Phone: 610-644-7701, Fax: 610-644-9297, www.picscontrols.com

Princo Instruments, Inc., 1020 Industrial Blvd., Southampton, PA 18966-4095, Phone: 800-221-9237 (toll free), Fax: 215-355-1500, www.princolevelcontrols.com/

Sargent-Welch Scientific Co., Industrial-Collegiate Div., 911 Commerce Court, Buffalo Grove, IL 60089, Phone: 847-677-0600, Fax: 847-677-4869

SC Controls, Inc., 1366 Heritage Ct. Escondido, CA 92027, Phone: 760-480-0473, Fax: 760-480-8309

Schneider Electric-Hyde Park, 1875 Founders Dr., Dayton, OH 45420, Phone: 937-514-7207, Fax: 937-258-5830, www.sesensors.com

Sentinel, 6765 Langley Dr., Baton Rouge, LA 70809, Phone: 225-751-5893, Fax: 225-756-0365, www.sentinelndt.com/

Sierra Monitor Corp., 1991 Tarob Court, Milpitas, CA 95035, Phone: 408-262-6611, 800-727-4377 (toll free), Fax: 408-262-9042, www.sierramonitor.com

Solartron Mobrey, 19408-T Park Row, Suite 320, Houston, TX 77084, Phone: 281-398-7890, 888-696-0706 (toll free), Fax: 281-398-7891, www.solartronmobrey.com

Terriss Consolidated Industries, Inc., 807 Summerfield Ave., Asbury Park, NJ 07712, Phone: 732-988-0909, 800-342-1611 (toll free), Fax: 732-502-0526

Thermo Fisher Scientific, 81 Wyman St., Waltham, MA 02451-1105, Phone: 781-622-1000, 800-678-5599 (toll free), Fax: 781-622-1207, www.thermofisher.com/global/en/home.asp

Toshiba International Corp., 13131 W. Little York Rd., Houston, TX 77041, Phone: 713-466-0277, Fax: 713-466-8773, www.toshiba.com/ind/

Warren-Knight Instrument Co., 2045 Bennett Dr., Philadelphia, PA 19116-3019, Phone: 215-464-9300, Fax: 215-464-9303, www.warrenind.com

46

Fluid Viscosity Measurement

46.1 Shear Viscosity **46**-1
46.2 Newtonian and Non-Newtonian Fluids **46**-3
Dimensions and Units of Viscosity
46.3 Viscometer Types **46**-5
Rheometer • Rotational Methods/Concentric Cylinders • Cone-and-Plate Viscometers • Parallel Disks
46.4 Pressure and Gravity Flow Methods **46**-11
Capillary Viscometers • Glass Capillary Viscometers • Orifice/Cup, Short Capillary: Saybolt Viscometer
46.5 Falling Body Methods **46**-13
Falling Sphere • Falling Cylinder • Falling Methods in Opaque Liquids • Rising Bubble/Droplet
46.6 Oscillation Methods **46**-17
46.7 Acoustic Methods **46**-20
46.8 Microrheology **46**-22
Particle-Tracking Microrheology • Dynamic Light Scattering
46.9 High-Pressure Rheology **46**-25
High-Pressure Rheometry
References **46**-29
Further Information **46**-31

R.A. Secco
University of Western Ontario

M. Kostic
Northern Illinois University

J.R. deBruyn
University of Western Ontario

46.1 Shear Viscosity

An important mechanical property of fluids is *viscosity*. Physical systems and applications as diverse as fluid flow in pipes, the flow of blood, lubrication of engine parts, the dynamics of raindrops, volcanic eruptions, and planetary and stellar magnetic field generation, to name just a few, all involve fluid flow and are controlled to some degree by fluid viscosity. *Viscosity* is the tendency of a fluid to resist flow and can be thought of as the internal friction of a fluid. Microscopically, viscosity is related to molecular diffusion and depends on the interactions between molecules or, in complex fluids, larger-scale flow units. Diffusion tends to transfer momentum from regions of high momentum to regions of low momentum, thus smoothing out variations in flow velocity. In this sense, the internal friction of a fluid is analogous to the macroscopic mechanical friction, which causes an object sliding across a planar surface to slow down. In the mechanical system, energy must be supplied to sustain the motion of the object over the plane, while in a fluid, energy must be supplied to sustain a flow. Since viscosity is related to the diffusive transport of momentum, the viscous response of a fluid is called a *momentum transport process*. The flow velocity within a fluid will vary, depending on location. Consider a viscous fluid at constant pressure between two closely spaced parallel plates as shown in Figure 46.1. A force, *F*, applied to the

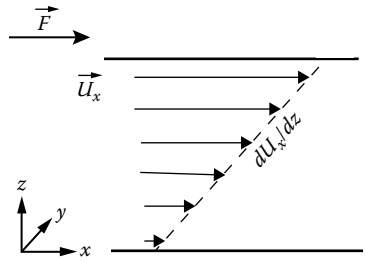


FIGURE 46.1 System for defining Newtonian viscosity. When the upper plate is subjected to a force, the fluid between the plates is dragged in the direction of the force with a velocity of each layer that diminishes away from the upper plate. The reducing velocity eventually reaches zero at the lower plate boundary.

top plate causes the fluid adjacent to the upper plate to move in the direction of $\vec{F}$. The fluid adjacent to the top plate is constrained by the no-slip boundary condition to move at the same speed as the plate. Similarly, the fluid next to the stationary bottom plate must be stationary. The motion of the top plate thus causes the fluid to flow with a velocity profile across the liquid that decreases linearly from the upper to the lower plate, as shown in Figure 46.1. This arrangement is referred to as *simple shear*. The applied force is called a *shear*, and the force per unit area a *shear stress*. The resulting deformation rate of the fluid, or equivalently the *velocity gradient* dU_x/dz, is called the *shear strain rate*, $\dot{\gamma}_{zx}$. The mathematical expression describing the viscous response of the system to the shear stress is simply

$$\tau_{zx} = \frac{\eta dU_x}{dz} = \eta \dot{\gamma}_{zx} \tag{46.1}$$

where

τ_{zx}, the shear stress, is the force per unit area exerted on the upper plate in the x-direction (and hence is equal to the force per unit area exerted by the fluid on the upper plate in the negative x-direction)

dU_x/dz is the gradient of the x-velocity in the z-direction in the fluid, that is, the shear strain rate

η is the *coefficient of viscosity*

Note that in general, the shear strain rate is a more complex function of the fluid velocity-gradient tensor. In this case, because one is concerned with a shear force that produces the fluid motion, η is more specifically called the *shear dynamic viscosity*. In fluid mechanics, where the motion of a fluid is considered without reference to force, it is common to define the *kinematic viscosity*, ν, which is simply given by

$$\nu = \frac{\eta}{\rho} \tag{46.2}$$

where ρ is the mass density of the fluid.

The viscosity defined by Equation 46.1 is relevant only for *laminar* (i.e., layered or sheetlike) or streamline flow as depicted in Figure 46.1, and it refers to the molecular viscosity or *intrinsic viscosity*. The molecular viscosity is a property of the material that depends microscopically on interactions between individual molecules and is manifested macroscopically as the fluid's resistance to flow. When the flow is turbulent, small-scale turbulent vortices can contribute to the overall diffusion of momentum, resulting in an effective viscosity, sometimes called the *eddy viscosity*, that, depending on the Reynolds number, can be as much as 10^6 times larger than the intrinsic viscosity.

Molecular viscosity is separated into *shear viscosity* and bulk or *volume viscosity*, η_v, depending on the type of strain involved. Shear viscosity is a measure of resistance to isochoric flow in a shear field, whereas volume viscosity is a measure of resistance to volumetric flow in a 3D stress field.

For most liquids, including hydrogen bonded weakly associated or unassociated, and polymeric liquids as well as liquid metals, $\eta/\eta_v \approx 1$, suggesting that shear and structural viscous mechanisms are closely related [1].

The shear viscosity of most liquids decreases with temperature, which is opposite to what occurs in gases. Under most conditions, including those relevant in engineering applications, viscosity can be considered to be independent of pressure, as temperature effects are very much larger. In the context of planetary interiors, however, the effects of high pressure cannot be ignored as viscosity normally increases as pressure increases. However, depending on composition, pressure can cause fundamental changes in the molecular structure of the fluid that can result in an anomalous viscosity decrease with increasing pressure [2].

46.2 Newtonian and Non-Newtonian Fluids

Equation 46.1 is known as Newton's law of viscosity. If the viscosity throughout the fluid is independent of strain rate, then the fluid is said to be a *Newtonian fluid*. The constant of proportionality is called the coefficient of viscosity, and a plot of stress versus strain rate for Newtonian fluids yields a straight line with a slope of η, as shown by the solid line flow curve in Figure 46.2. Examples of Newtonian fluids are pure, single-phase, unassociated gases; liquids; and solutions of low molecular weight such as water. There is, however, a large group of fluids for which the viscosity is dependent on the strain rate. Such fluids are said to be non-Newtonian and their study is called *rheology*. In differentiating between Newtonian and non-Newtonian behavior, it is helpful to consider the time scale (as well as the normal stress differences and phase shift in dynamic testing) involved in the process of a liquid responding to a shear perturbation. In general, the shear strain rate $\dot{\gamma}$ is a more complex function of the second invariant of the fluid velocity-gradient tensor; however, in a simple shear flow, like on Figure 46.1, it reduces to the velocity gradient, dU_x/dz. Therefore, the time scale related to the applied shear perturbation about the equilibrium state is t_s, where $t_s = \dot{\gamma}^{-1}$. A second time scale, t_r, called the *relaxation time*, characterizes the rate at which the relaxation of the strain in the fluid can be accomplished and is related to the time it takes for a typical flow unit to move a distance equivalent to its mean diameter. For Newtonian water, $t_r \sim 10^{-12}$ s, and, because shear rates greater than 10^6 s^{-1} are rare in practice, the time required for the fluid to respond to the strain perturbation in water is much less than the time scale of the perturbation itself (i.e., $t_r \ll t_s$). However, for non-Newtonian macromolecular liquids like polymeric liquids, for colloidal and fiber suspensions, and for pastes and emulsions, the long response times of large viscous flow units (e.g., polymer molecules or aggregates of interacting particles) can easily make $t_r > t_s$. An example

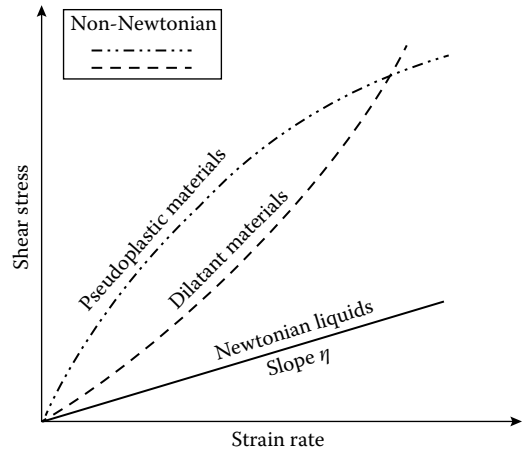


FIGURE 46.2 Flow curves illustrating Newtonian and non-Newtonian fluid behavior.

of a non-Newtonian fluid is liquid elemental sulfur, in which long chains (polymers) of up to 100,000 sulfur atoms form flow units that are easily entangled, which bind the liquid in a "rigid-like" network. Another example of a well-known non-Newtonian fluid is tomato ketchup.

With reference to Figure 46.2, Equation 46.1 can be written in a more general form:

$$\tau_{xy}(\dot{\gamma}) = \tau_{xy}(0) + \frac{\dot{\gamma}\partial\tau_{xy}(0)}{\partial\dot{\gamma}} + O(\dot{\gamma}^2) \tag{46.3}$$

Here $\tau_{xy}(0)$ is a yield stress required for flow to start, the term proportional to $\dot{\gamma}$ is the usual linear Newtonian term in the limit of small strain rates, and the nonlinear term $O(\dot{\gamma}^2)$ accounts for the nonlinear viscous response of some materials.

For a Newtonian fluid, the initial stress at zero shear rate and the nonlinear function $O(\dot{\gamma}^2)$ are both negligible (zero), so Equation 46.3 reduces to Equation 46.1, since $\partial\tau_{xy}(0)/\partial\dot{\gamma}$ then equals η. For a non-Newtonian fluid, $\tau_{xy}(0)$ may be zero but the nonlinear term $O(\dot{\gamma}^2)$ is nonzero. This characterizes fluids in which shear stress varies nonlinearly with strain rate, as shown by the dashed-dotted and dashed flow curves in Figure 46.2. The former type of fluid behavior is known as *shear thickening* or dilatancy, and an example is a concentrated suspension of cornstarch in water. The latter, much more common type of fluid behavior is known as *shear thinning* or pseudoplasticity; cream, blood, most polymers, and liquid cement are all examples. Both behaviors result from particle or molecular reorientations or rearrangements in the fluid that increase or decrease, respectively, the internal friction to shear. Non-Newtonian behavior can also arise in fluids whose viscosity changes with time of applied shear stress. Fluids whose viscosity increases over time when a shear stress is applied are called *rheopectic*. Conversely, liquids whose viscosity decreases with time are called *thixotropic fluids*. Nondrip paint, which will not flow until a shear stress is applied by the paint brush for a sufficiently long time, is a common example of a thixotropic fluid.

Fluid deformation that is not recoverable after removal of the stress is typical of the purely viscous response. The other extreme response to an external stress is purely elastic and is characterized by an equilibrium deformation that is fully recovered on removal of the stress. There are an infinite number of intermediate or combined viscous/elastic responses to external stress, which are grouped under the behavior known as *viscoelasticity*. Fluids that behave elastically in some stress range require a limiting or yield stress before they will flow as a viscous fluid. A simple, empirical, constitutive equation often used for this type of rheological behavior is of the form

$$\tau_{yx} = \tau_y + \dot{\gamma}^n \eta_p \tag{46.4}$$

where τ_y is the *yield stress*, η_p is an *apparent viscosity* called the plastic viscosity, and the exponent n allows for a range of non-Newtonian responses: $n = 1$ describes pseudo-Newtonian behavior and is called a *Bingham fluid*; $n < 1$ describes shear thinning behavior; and $n > 1$ describes shear thickening behavior. Interested readers should consult [3–9] for further information on applied rheology.

46.2.1 Dimensions and Units of Viscosity

From Equation 43.1, the dimensions of dynamic viscosity are $M\,L^{-1}\,T^{-1}$, and the basic SI unit is the Pascal second (Pa · s), where 1 Pa · s = 1 N s m^{-2}. The c.g.s. unit of dyn s cm^{-2} is the poise (P). The dimensions of kinematic viscosity, from Equation 46.2, are $L^2\,T^{-1}$, and the SI unit is $m^2\,s^{-1}$. For most practical situations, this is usually too large and so the c.g.s. unit of $cm^2\,s^{-1}$, or the stoke (St), is preferred. Table 46.1 lists some common fluids and their shear dynamic viscosities at atmospheric pressure and 20 °C.

TABLE 46.1 Shear Dynamic Viscosity of Some Common Fluids at 20 °C and 1 atm

Fluid	Shear Dynamic Viscosity (Pa·s)
Air	1.8×10^{-4}
Water	1.0×10^{-3}
Mercury	1.6×10^{-3}
Automotive engine oil (SAE 10W30)	1.3×10^{-1}
Dish soap	4.0×10^{-1}
Corn syrup	6.0

46.3 Viscometer Types

46.3.1 Rheometer

A viscometer (also called viscosimeter) is an instrument used to measure the viscosity of a fluid. For liquids with viscosities that vary with flow conditions, an instrument called a rheometer is used. Viscometers only measure under one flow condition. The instruments for viscosity measurements are designed to determine "a fluid's resistance to flow," a fluid property defined earlier as viscosity. The fluid flow in a given instrument geometry defines the strain rates, and the corresponding stresses are the measure of resistance to flow. If strain rate or stress is set and controlled, then the other one will, everything else being the same, depend on the fluid viscosity. If the flow is simple (one dimensional, if possible) such that the strain rate and stress can be determined accurately from the measured quantities, the absolute dynamic viscosity can be determined; otherwise, the relative viscosity will be established. For example, the fluid flow can be set by dragging the fluid with a sliding or rotating surface, or a body falling through the fluid, or by forcing the fluid (by external pressure or gravity) to flow through a fixed geometry, such as a capillary tube, annulus, a slit (between two parallel plates), or orifice. The corresponding resistance to flow is measured as the boundary force or torque or pressure drop. The flow rate or efflux time represents the fluid flow for a set flow resistance, like pressure drop or gravity force. The viscometers are classified, depending on how the flow is initiated or maintained, as in Table 46.2.

The basic principle of all viscometers is to provide as simple flow kinematics as possible, preferably 1D (*isometric*) *flow*, in order to determine the shear strain rate accurately, easily, and independent of fluid type. The resistance to such flow is measured, and thereby the shearing stress is determined. The shear viscosity is then easily found as the ratio between the shearing stress and the corresponding shear strain rate. Practically, it is never possible to achieve desired 1D flow or ideal geometry, and a number of errors, listed in Table 46.3, can occur and need to be accounted for [4–8]. A sample list of manufacturers/distributors of commercial viscometers/rheometers is given in Table 46.4. There is no implied endorsement of any manufacturer on this list.

46.3.2 Rotational Methods/Concentric Cylinders

The main advantage of the rotational viscometer as compared to many other viscometers is its ability to operate continuously at a given shear rate, so that other steady-state measurements can be conveniently performed. That way, time dependency, if any, can be detected and determined. Also, subsequent measurements can be made with the same instrument and sampled at different shear rates, temperature, etc. For these and other reasons, *rotational viscometers* are among the most widely used class of instruments for rheological measurements.

TABLE 46.2 Viscometer/Rheometer Classification and Basic Characteristics

Type/Geometry	Basic Characteristics/Comments
Drag flow types: Flow set by motion of instrument boundary/surface using external or gravity force	
Rotating concentric cylinders (Couette)	Good for low viscosity, high shear rates; for $R_2/R_1 \cong 1$, see Figure 46.3; difficult to clean thick fluids
Rotating cone and plate	Homogeneous shear, best for non-Newtonian fluids and normal stresses; need good alignment, problems with loading and evaporation
Rotating parallel disks	Similar to cone and plate but inhomogeneous shear; shear varies with gap height, easy sample loading
Sliding parallel plates	Homogeneous shear, simple design, good for high viscosity; difficult loading and gap control
Falling body (ball, cylinder)	Very simple, good for high temperature and pressure; need density and special sensors for opaque fluids, not good for viscoelastic fluids
Rising bubble oscillating body	Similar to falling body viscometer; for transparent fluids
Oscillating body	Needs instrument constant, good for low-viscous liquid metals
Pressure flow types: Fluid set in motion in fixed instrument geometry by external or gravity pressure	
Long capillary (Poiseuille flow)	Simple, very high shears and range but very inhomogeneous shear, variable time dependency, and is time consuming
Orifice/cup (short capillary)	Very simple, reliable but not for absolute viscosity and non-Newtonian fluids
Slit (parallel plates) pressure flow	Similar to capillary but difficult to clean
Axial annulus pressure flow	Similar to capillary, better shear uniformity but more complex, eccentricity problem and difficult to clean
Others/miscellaneous	
Ultrasonic	Good for high-viscosity fluids, small sample volume, gives shear and volume viscosity, and elastic property data; problems with surface finish and alignment, complicated data reduction

Source: Adapted from Macosko, C. W., *Rheology: Principles*, *Measurements*, *and Applications*, VCH, New York, 1994.

TABLE 46.3 Typical Causes of Viscometer/Rheometer Errors

Error/Effect	Cause/Comment
End/edge effect	Energy losses at the fluid entrance and exit of main test geometry
Kinetic energy losses	Loss of pressure to kinetic energy
Secondary flow	Energy loss due to unwanted secondary flow, vortices, etc.; increases with Reynolds number
Nonideal geometry	Deviations from ideal shape, alignment, and finish
Shear rate nonuniformity	Important for non-Newtonian fluids
Temperature variation and viscous heating	Variation in temperature, in time, and in space, influences the measured viscosity
Turbulence	Partial and/or local turbulence often develops even at low Reynolds numbers
Surface tension	Difference in interfacial tensions
Elastic effects	Structural and fluid elastic effects
Miscellaneous effects	Depend on test specimen, melt fracture, thixotropy, rheopexy

TABLE 46.4 Viscometer/Rheometer Manufacturers and Distributors (for measurement at 1 atm)

A&D Engineering Co. Ltd	Hydramotion	Perten
Ametek	Kinematica AG	Physica Messtechnik GmbH
Ankersmid Lab	Kittiwake	REL
Anton Parr	Koehler Inst. Co., Inc.	Reologica Instr.
ATS	Lamy Rheology	Rheometric Scientific, Inc.
Avenisense	Lauda	Rheotec
Bartec	Lemis	Santam Eng. Energy Co. Ltd
Brookfield Eng. Labs, Inc.	Malvern	SI Analytics
BYK Instr.	Marimex	So Fraser
Cannon Inst. Co.	Micro Motion	Stanhope-Seta
Ceramic Instr.	Nametre Co.	T.A. Instruments, Inc.
Cole-Palmer Inst. Co.	Norcross	Tanaka
Dong Jin Inst. Corp.	Orion	Thermo Scientific
Fungilab	Paar Physica USA, Inc.	TMI
Galvanic Applied Sciences, Inc.	PAC	Toyo Seiki Seisaku-Sho Ltd.
Gottfert Werkstorf-Prufmaschinen GmbH	Parker	Weschler Instr.

Note: All the above-mentioned manufacturers/distributors can be found on the Internet, along with their most recent contact information, product description, and, in some cases, pricing. A variety of products using different viscosity measuring methods are available from this list. There is no implied endorsement of any manufacturer on this list.

Concentric cylinder-type viscometers/rheometers are usually employed when absolute viscosity is required, which in turn requires knowledge of well-defined shear rate and shear stress. Such instruments are available in different configurations and can be used for almost any fluid. There are models for low and high shear rates. More complete discussion on concentric cylinder viscometers/rheometers is given elsewhere [4–9]. In the *Couette-type viscometer*, the rotation of the outer cylinder, or cup, minimizes centrifugal forces, which cause Taylor vortices. The latter can be present in the *Searle-type viscometer* when the inner cylinder, or bob, rotates.

Usually, the torque on the stationary cylinder and rotational velocity of the other cylinder are measured for determination of the shear stress and shear rate, respectively, which is needed for viscosity calculation. Once the torque, T, is measured, it is simple to describe the fluid shear stress $\tau_{r\theta}(r)$ at any point with radius r between the two cylinders, as shown in Figure 46.3:

$$\tau_{r\theta}(r) = \frac{T}{2\pi r^2 L_e} \tag{46.5}$$

In Equation 46.5, $L_e = (L + L_c)$ is the effective length of the cylinder at which the torque is measured. In addition to the cylinder's length L, it takes into account the end-effect correction L_c [4–8].

For a narrow gap between the cylinders ($\beta = R_2/R_1 \cong 1$), regardless of the fluid type, the velocity profile can be approximated as linear, and the shear rate, $\dot{\gamma}(r)$, within the gap will be uniform:

$$\dot{\gamma}(r) \cong \frac{\Omega \bar{R}}{(R_2 - R_1)} \tag{46.6}$$

where

$\Omega = (\omega_2 - \omega_1)$ is the relative rotational speed

$\bar{R} = (R_1 + R_2)/2$ is the mean radius of the inner (1) and outer (2) cylinders

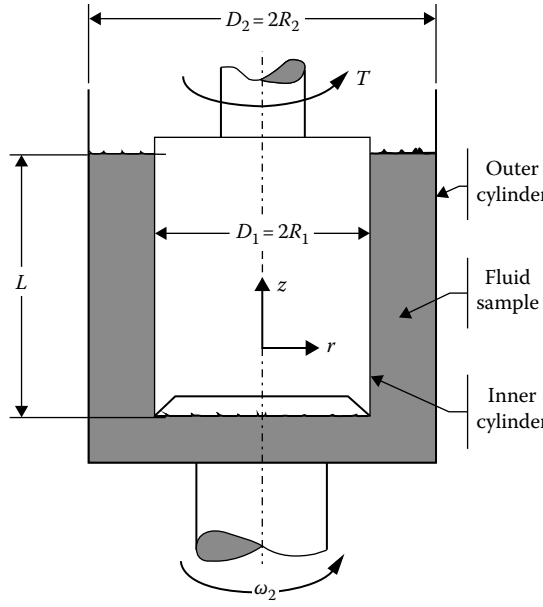


FIGURE 46.3 Concentric cylinder viscometer geometry.

Actually, the shear rate profile across the gap between the cylinders depends on the relative rotational speed, radii, and the unknown fluid properties, which seems an "open-ended" enigma. The solution of this complex problem is given elsewhere [4–8] in the form of an infinite series and requires the slope of a logarithmic plot of T as a function of Ω in the range of interest. Note that for a stationary inner cylinder ($\omega_1 = 0$), which is the usual case in practice, Ω becomes equal to ω_2. However, there is a simpler procedure [10] that has also been established by German standards [11]. For any fluid, including non-Newtonian fluids, there is a radius at which the shear rate is virtually independent of the fluid type for a given Ω. This radius, being a function of geometry only, is called the *representative radius*, R_R, and is determined as the location corresponding to the so-called representative shear stress, $\tau_R = (\tau_1 + \tau_2)/2$, the average of the stresses at the outer and inner cylinder interfaces with the fluid, that is,

$$R_R = R_1\left\{\frac{[2\beta^2]}{[1+\beta^2]}\right\}^{1/2} = R_2\left\{\frac{2}{[1+\beta^2]}\right\}^{1/2} \tag{46.7}$$

Since the shear rate at the representative radius is virtually independent on the fluid type (whether Newtonian or non-Newtonian), the representative shear rate is simply calculated for Newtonian fluid ($n = 1$) and $r = R_R$, according to [10]:

$$\dot{\gamma}_R = \dot{\gamma}_{r=R_R} = \omega_2\left\{\frac{[\beta^2+1]}{[\beta^2-1]}\right\} \tag{46.8}$$

The accuracy of the representative parameters depends on the geometry of the cylinders (β) and fluid type (n).

It is shown in [10] that, for an unrealistically wide range of fluid types ($0.35 < n < 3.5$) and cylinder geometries ($\beta = 1$–1.2), the maximum errors are less than 1%. Therefore, the error associated with the representative parameter concept is virtually negligible for practical measurements.

Finally, the (apparent) fluid viscosity is determined as the ratio between the shear stress and corresponding shear rate using Equations 46.5 through 46.8, as

$$\eta = \eta_R = \frac{\tau_R}{\dot{\gamma}_R} = \left\{\frac{[\beta^2 - 1]}{[4\pi\beta^2 R_1^2 L_e]}\right\}\frac{T}{\omega_2} = \left\{\frac{[\beta^2 - 1]}{[4\pi R_2^2 L_e]}\right\}\frac{T}{\omega_2} \tag{46.9}$$

For a given cylinder geometry (β, R_2, and L_e), the viscosity can be determined from Equation 46.9 by measuring T and ω_2.

As already mentioned, in Couette-type viscometers, the Taylor vortices within the gap are virtually eliminated. However, vortices at the bottom can be present, and their influence becomes important when the Reynolds number reaches the value of unity [10,11]. Furthermore, flow instability and turbulence will develop when the Reynolds number reaches values of 10^3–10^4. The Reynolds number, Re, for the flow between concentric cylinders is defined [11] as

$$\mathrm{Re} = \left\{\frac{[\rho\omega_2 R_1^2]}{2\eta}\right\}[\beta^2 - 1] \tag{46.10}$$

46.3.3 Cone-and-Plate Viscometers

The simple cone-and-plate viscometer geometry provides a uniform rate of shear and direct measurements of the first normal stress difference. It is the most popular instrument for measurement of non-Newtonian fluid properties. The working shear stress and shear strain rate equations can be easily derived in spherical coordinates, as indicated by the geometry in Figure 46.4, and are, respectively,

$$\tau_{\theta\phi} = \frac{3T}{[2\pi R^3]} \tag{46.11}$$

and

$$\dot{\gamma} = \frac{\Omega}{\theta_0} \tag{46.12}$$

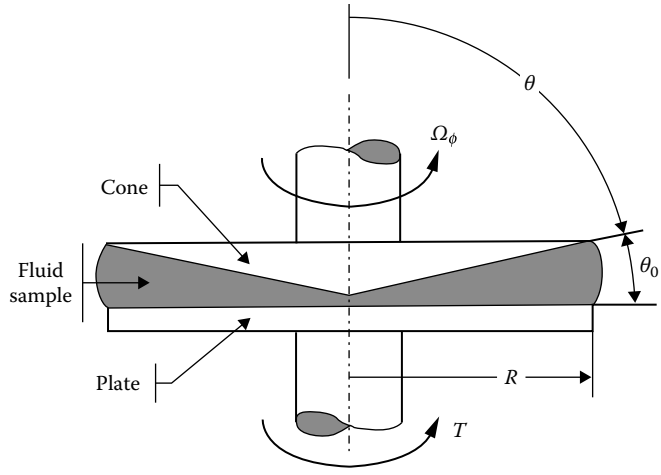


FIGURE 46.4 Cone-and-plate viscometer geometry.

where R and $\theta_0 < 0.1$ rad (≈6°) are the cone radius and angle, respectively. The viscosity is then easily calculated as

$$\eta = \frac{\tau_{\theta\phi}}{\dot{\gamma}} = \frac{[3T\theta_0]}{2\pi\Omega R^3} \tag{46.13}$$

Inertia and secondary flow increase, while shear heating decreases the measured torque (T_m). For more details, see [4,5]. The torque correction is given as

$$\frac{T_m}{T} = 1 + 6 \cdot 10^{-4}\,\mathrm{Re}^2 \tag{46.14}$$

where

$$\mathrm{Re} = \frac{\left\{\rho[\Omega\theta_0 R]^2\right\}}{\eta} \tag{46.15}$$

46.3.4 Parallel Disks

This geometry (Figure 46.5), which consists of a disk rotating in a cylindrical cavity, is similar to the cone-and-plate geometry, and many instruments permit the use of either one. However, the shear rate is no longer uniform but depends on radial distance from the axis of rotation and on the gap h, that is,

$$\dot{\gamma}(r) = \frac{r\Omega}{h} \tag{46.16}$$

For Newtonian fluids, after integration over the disk area, the torque can be expressed as a function of viscosity, so that the latter can be determined as

$$\eta = \frac{2Th}{[\pi\Omega R^4]} \tag{46.17}$$

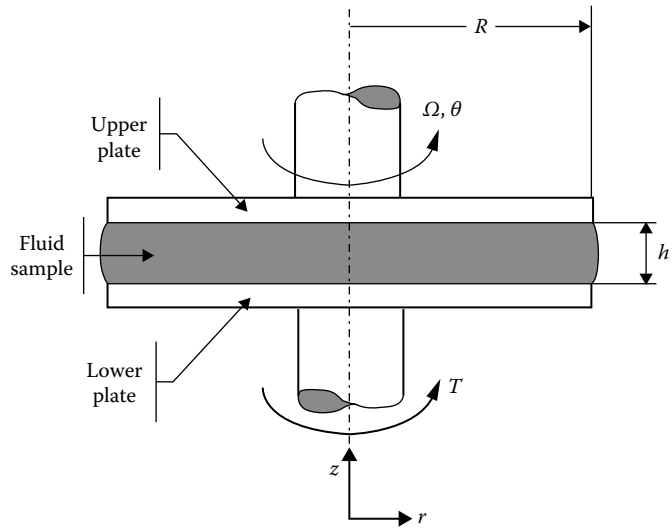


FIGURE 46.5 Parallel disks viscometer geometry.

46.4 Pressure and Gravity Flow Methods

46.4.1 Capillary Viscometers

The *capillary viscometer* is based on the fully developed laminar tube flow theory (*Hagen–Poiseuille flow*) and is shown in Figure 46.6. The capillary tube length is many times larger than its small diameter, so that entrance flow is neglected or accounted for in more accurate measurement or for shorter tubes. The expression for the shear stress at the wall is

$$\tau_w = \left[\frac{\Delta P}{L}\right] \cdot \left[\frac{D}{4}\right] \tag{46.18}$$

and

$$\Delta P = (P_1 - P_2) + (z_1 - z_2) - \frac{[C\rho V^2]}{2} \tag{46.19}$$

where $C \cong 1.1$, P, z, $V = 4Q/[\pi D^2]$, and Q are correction factor, pressure, elevation, the mean flow velocity, and the fluid volume flow rate, respectively. The subscripts 1 and 2 refer to the inlet and outlet, respectively.

The expression for the shear rate at the wall is

$$\dot{\gamma} = \left\{\frac{[3n+1]}{4n}\right\} \cdot \left\{\frac{8V}{D}\right\} \tag{46.20}$$

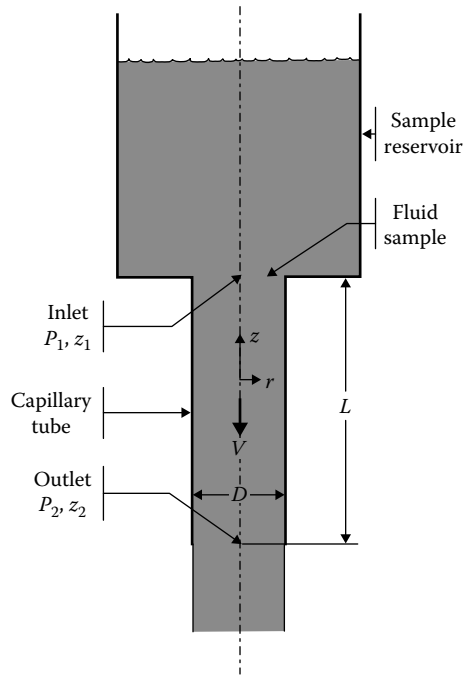


FIGURE 46.6 Capillary viscometer geometry.

where $n = d[\log\tau_w]/d[\log(8V/D)]$ is the slope of the measured $\log(\tau_w) - \log(8V/D)$ curve. Then, the viscosity is simply calculated as

$$\eta = \frac{\tau_w}{\dot{\gamma}} = \left\{\frac{4n}{[3n+1]}\right\} \cdot \left\{\frac{\Delta PD^2}{[32LV]}\right\} = \left\{\frac{4n}{[3n+1]}\right\} \cdot \left\{\frac{[\Delta PD^4\pi]}{[128QL]}\right\} \tag{46.21}$$

Note that $n = 1$ for a Newtonian fluid, so the first term $[4n/(3n + 1)]$ becomes unity and disappears from the earlier equations. The advantages of capillary over rotational viscometers are low cost, high accuracy (particularly with longer tubes), and the ability to achieve very high shear rates, even with high-viscosity samples. The main disadvantages are high residence time and variation of shear across the flow, which can change the structure of complex test fluids, as well as shear heating with high-viscosity samples.

46.4.2 Glass Capillary Viscometers

Glass capillary viscometers are very simple and inexpensive. Their geometry resembles a U-tube with at least two reservoir bulbs connected to a capillary tube passage with inner diameter D. The fluid is drawn up into one bulb reservoir of known volume, V_0, between etched marks. The efflux time, Δt, is measured for that volume to flow through the capillary under gravity.

From Equation 46.21 and taking into account that $V_0 = (\Delta t)VD^2\pi/4$ and $\Delta P = \rho g(z_1 - z_2)$, the kinematic viscosity can be expressed as a function of the efflux time only, with the last term, $K/\Delta t$, added to account for error correction, where K is a constant [7], that is,

$$\nu = \frac{\eta}{\rho} = \left\{\left[\frac{4n}{(3n+1)}\right] \cdot \frac{[\pi g(z_1 - z_2)D^4]}{128LV_o}\right\}(\Delta t - K\Delta t) \tag{46.22}$$

Note that for a given capillary viscometer and $n \cong 1$, the curl-bracketed term is a constant. The last correction term is negligible for a large capillary tube ratio, L/D, where kinematic viscosity becomes linearly proportional to measured efflux time. Various kinds of commercial glass capillary viscometers, like Cannon–Fenske type or similar, can be purchased from scientific and/or supply stores. They are the modified original *Ostwald viscometer* design in order to minimize certain undesirable effects, to increase the viscosity range, or to meet specific requirements of the tested fluids, like opacity. Glass capillary viscometers are often used for low-viscosity fluids.

46.4.3 Orifice/Cup, Short Capillary: Saybolt Viscometer

The principle of these viscometers is similar to that of glass capillary viscometers, except that the flow through a short capillary ($L/D \ll 10$) does not satisfy or even approximate the Hagen–Poiseuille, fully developed, pipe flow. The influences of entrance end effect and changing hydrostatic heads are considerable. The efflux time reading, Δt, represents relative viscosity for comparison purposes and is expressed as "viscometer seconds," like the *Saybolt seconds* or *Engler seconds* or degrees. Although the conversion formula, similar to glass capillary viscometers, is used, the constants k and K in Equation 46.23 are purely empirical and dependent on fluid types:

$$\nu = \frac{\eta}{\rho} = k\Delta t - \frac{K}{\Delta t} \tag{46.23}$$

where k = 0.00226, 0.0216, 0.073 and K = 1.95, 0.60, 0.0631, for Saybolt universal ($\Delta t < 100$ s), Saybolt Furol ($\Delta t > 40$ s), and Engler viscometers, respectively [12,13]. Due to their simplicity, reliability, and low cost, these viscometers are widely used for Newtonian fluids, in the oil and other industries, where simple correlations between the relative properties and desired results are needed. However, these viscometers are not suitable for absolute viscosity measurement nor for non-Newtonian fluids.

46.5 Falling Body Methods

46.5.1 Falling Sphere

The falling sphere viscometer is one of the earliest and simplest methods to determine the absolute shear viscosity of a Newtonian fluid. In this method, a sphere is allowed to fall freely a measured distance through a viscous liquid medium, and its velocity is determined. The viscous drag of the falling sphere results in the creation of a restraining force, *F*, described by Stokes' law:

$$F = 6\pi\eta r_s U_t \tag{46.24}$$

where

r_s is the radius of the sphere
U_t is the *terminal velocity* of the falling body

If a sphere of density ρ_2 is falling through a fluid of density ρ_1 in a container of infinite extent, then by balancing Equation 46.24 with the net force of gravity and buoyancy exerted on a solid sphere, the resulting equation of absolute viscosity is

$$\eta = 2gr_s^2 \frac{(\rho_2 - \rho_1)}{9U_t} \tag{46.25}$$

Equation 46.25 shows the relation between the viscosity of a fluid and the terminal velocity of a sphere falling within it. Having a finite container volume necessitates the modification of Equation 46.25 to correct for effects on the velocity of the sphere due to its interaction with container walls (*W*) and ends (*E*). Considering a cylindrical container of radius *r* and height *H*, the corrected form of Equation 46.25 can be written as

$$\eta = 2gr_s^2 \frac{(\rho_2 - \rho_1)W}{(9U_t E)} \tag{46.26}$$

where

$$W = 1 - 2.104\left(\frac{r_s}{r}\right) + 2.09\left(\frac{r_s}{r}\right)^3 - 0.95\left(\frac{r_s}{r}\right)^5 \tag{46.27}$$

$$E = 1 + 3.3\left(\frac{r_s}{H}\right) \tag{46.28}$$

The wall correction was empirically derived [14,15] and is valid for $0.16 \leq r_s/r \leq 0.32$. Beyond this range, the effects of container walls significantly impair the terminal velocity of the sphere, thus giving rise to a false high-viscosity value.

Figure 46.7 is a schematic diagram of the falling sphere method and demonstrates the attraction of this method—its simplicity of design. The simplest and most cost-effective approach in applying this method to transparent liquids would be to use a sufficiently large graduated cylinder filled with the liquid. With a distance marked on the cylinder near the axial and radial center (the region least influenced by the container walls and ends), a sphere (such as a ball bearing or a material that is nonreactive with the liquid) with a known density and sized to within the bounds of the container correction free falls the