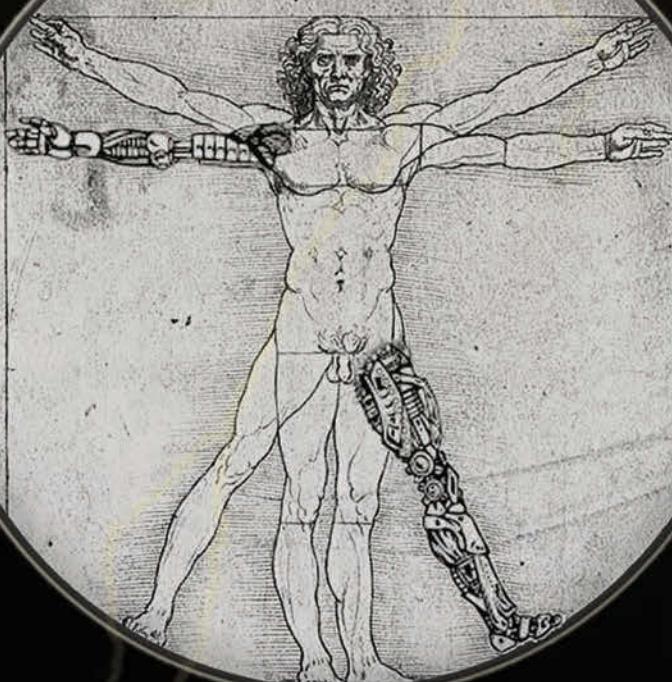


# INTRODUCTION



TO

# BIOMECHATRONICS

Graham M. Brooker

# **Introduction to Biomechatronics**

Graham Brooker  
University of Sydney, Australia



Raleigh, NC  
[scitechpub.com](http://scitechpub.com)



Published by SciTech Publishing, Inc.  
911 Paverstone Drive, Suite B  
Raleigh, NC 27615  
(919) 847-2434, fax (919) 847-2568  
[scitechpublishing.com](http://scitechpublishing.com)

Copyright © 2012 by SciTech Publishing, Raleigh, NC. All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 646-8600, or on the web at [copyright.com](http://copyright.com). Requests to the Publisher for permission should be addressed to the Publisher, SciTech Publishing, Inc., 911 Paverstone Drive, Suite B, Raleigh, NC 27615, (919) 847-2434, fax (919) 847-2568, or email [editor@scitechpub.com](mailto:editor@scitechpub.com).

The publisher and the author make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation warranties of fitness for a particular purpose.

Editor: Dudley R. Kay  
Editorial Assistant: Katie Janelle  
Production Manager: Robert Lawless  
Typesetting: MPS Limited, a Macmillan Company  
Cover Design: Brent Beckley  
Printer: Sheridan Books, Inc., Chelsea, MI

This book is available at special quantity discounts to use as premiums and sales promotions, or for use in corporate training programs. For more information and quotes, please contact the publisher.

Printed in the United States of America  
10 9 8 7 6 5 4 3 2 1

ISBN: 978-1-891121-27-2

**Library of Congress Cataloging-in-Publication Data**

Brooker, Graham.  
Introduction to biomechatronics / Graham Brooker.  
p.; cm.  
Includes bibliographical references and index.  
ISBN 978-1-891121-27-2 (hardcover : alk. paper) – ISBN 978-1-936059-84-3  
(ebook)  
I. Title.  
[DNLM: 1. Biomedical Engineering--instrumentation. 2. Robotics--instrumentation.  
3. Electronics, Medical--instrumentation. 4. Prostheses and Implants.  
5. Signal Processing, Computer-Assisted. QT 26]  
LC classification not assigned  
610.285–dc23

2011053250

# Contents

Preface xi

Acknowledgments xiii

---

## 1 Introduction to Biomechatronics 1

1.1 Introduction 1

1.2 Biomechatronic Systems 2

1.2.1 *The Human Subject* 2

1.2.2 *Stimulus or Actuation* 3

1.2.3 *Transducers and Sensors* 3

1.2.4 *Signal Processing Elements* 3

1.2.5 *Recording and Display* 3

1.2.6 *Feedback Elements* 4

1.3 Physiological Systems 4

1.3.1 *Biochemical System* 4

1.3.2 *Nervous System* 5

1.3.3 *Cardiovascular System* 5

1.3.4 *Respiratory System* 5

1.3.5 *Musculoskeletal System* 5

1.4 Summary of Contents 6

1.5 The Future of Biomechatronic Systems 6

1.6 References 7

---

## 2 Sensors and Transducers 9

2.1 Introduction 9

2.2 Switches 9

2.2.1 *Toggle Switches* 10

2.2.2 *Push-Button Switches* 10

2.2.3 *Limit Switches* 10

2.2.4 *Rotary Switches* 11

2.2.5 *Optical Switches* 11

2.2.6 *Other Switches* 11

2.2.7 *Relays* 12

2.3 Power Supplies 13

2.3.1 *Linear Power Supplies* 13

2.3.2 *Switch-Mode Power Supplies* 15

2.3.3 *Batteries* 16

2.3.4 *Energy Scavenging* 19

<b>2.4</b>	Sensors and Transducers	22
2.4.1	<i>Resistive Displacement Sensors</i>	22
2.4.2	<i>Inductive Displacement Sensors</i>	28
2.4.3	<i>Magnetic Displacement Sensors</i>	31
2.4.4	<i>Capacitive Displacement Sensors</i>	34
2.4.5	<i>Optical Displacement Sensors</i>	34
2.4.6	<i>Ranging Sensors</i>	38
2.4.7	<i>Time-of-Flight Ranging</i>	41
2.4.8	<i>Measuring Rate and Angular Rate</i>	43
2.4.9	<i>Accelerometers</i>	50
2.4.10	<i>Tilt Sensors</i>	55
2.4.11	<i>Pressure Measurement</i>	56
2.4.12	<i>Sound Pressure</i>	60
2.4.13	<i>Flow</i>	61
2.4.14	<i>Temperature Sensors</i>	67
2.4.15	<i>Tactile Sensing</i>	73
2.4.16	<i>Chemical Sensors</i>	77
2.4.17	<i>Optical Chemical Sensors</i>	80
<b>2.5</b>	Electrodes	83
2.5.1	<i>Body-Surface Biopotential Electrodes</i>	84
<b>2.6</b>	References	88

---

<b>3</b>	Actuators	91
<b>3.1</b>	Introduction	91
<b>3.2</b>	Electromechanical Actuators	91
3.2.1	<i>Solenoids and Voice Coils</i>	94
3.2.2	<i>Direct Current Motors</i>	99
3.2.3	<i>Brushless DC Motors</i>	113
3.2.4	<i>Stepper Motors</i>	117
3.2.5	<i>Linear Actuators</i>	124
3.2.6	<i>Servo Motors</i>	130
3.2.7	<i>AC Motors</i>	134
<b>3.3</b>	Hydraulic Actuators	137
<b>3.4</b>	Pneumatic Actuators	139
3.4.1	<i>Pneumatic Muscles</i>	140
<b>3.5</b>	Shape Memory Alloy	142
3.5.1	<i>Principle of Operation</i>	142
3.5.2	<i>Biomechatronic Applications</i>	145
<b>3.6</b>	Mechanical Amplification	145
3.6.1	<i>Linkages and Levers</i>	145
3.6.2	<i>Cams</i>	148
3.6.3	<i>Gears and Belt Drives</i>	149
3.6.4	<i>Translation Screw Devices</i>	153
<b>3.7</b>	Prosthetic Hand Actuation	154
3.7.1	<i>Shape Memory Alloys</i>	155
3.7.2	<i>Electric Motors</i>	155

3.7.3 *Pneumatic Artificial Muscles* 156

**3.8** References 157

---

**4** Feedback and Control Systems 159

**4.1** Introduction 159

**4.2** Biological Feedback Mechanisms 160

**4.3** Biomechatronic Feedback Mechanisms 160

4.3.1 *Limit Switches* 161

4.3.2 *Proportional and Higher-Order Controllers* 161

**4.4** System Representation 162

**4.5** System Models 164

4.5.1 *Mechanical Elements* 164

4.5.2 *Mechanical Model* 166

4.5.3 *Electrical Elements* 168

4.5.4 *Electrical Model* 169

4.5.5 *Similarities of the Two Models* 171

4.5.6 *Fluid Flow Elements* 171

**4.6** System Response 174

4.6.1 *Partial Fraction Expansion* 178

4.6.2 *Analyzing Complex Models* 179

**4.7** System Stability 181

4.7.1 *Root Locus* 184

4.7.2 *Steady-State Error* 188

**4.8** Controllers 188

4.8.1 *Proportional Controller* 188

4.8.2 *Integral Controller* 198

4.8.3 *Proportional Plus Integral Controller* 198

4.8.4 *Proportional–Integral–Derivative Controller* 200

**4.9** Controller Implementation 201

4.9.1 *Selection of Controller Gains* 201

4.9.2 *Controller Hardware* 202

**4.10** References 205

---

**5** Signal Processing 207

**5.1** Introduction 207

**5.2** Biomedical Signals 207

5.2.1 *Bioelectric Signals* 208

5.2.2 *Signals Characterized by Source* 210

5.2.3 *Signals Characterized by Type* 210

**5.3** Signal Acquisition 211

5.3.1 *Noise* 212

5.3.2 *Amplifiers* 216

5.3.3 *Practical Considerations* 222

5.3.4 *Op Amp Specifications* 223

<b>5.4</b>	Analog Signal Processing	224
5.4.1	<i>Frequency Content of a Signal</i>	224
5.4.2	<i>Analog Filters</i>	225
5.4.3	<i>Other Analog Circuits</i>	234
<b>5.5</b>	Digital Signal Processing	241
5.5.1	<i>The Comparator</i>	241
5.5.2	<i>Signal Acquisition and Processing Overview</i>	241
5.5.3	<i>ADCs and DACs</i>	243
5.5.4	<i>Signal Aliasing</i>	245
5.5.5	<i>Digital Filters</i>	248
5.5.6	<i>Filter Time-Domain Response</i>	258
5.5.7	<i>Envelope Detection</i>	259
5.5.8	<i>Spectral Estimation</i>	260
<b>5.6</b>	Statistical Techniques and Machine Learning	264
5.6.1	<i>Statistical Techniques</i>	264
5.6.2	<i>Data Mining</i>	267
5.6.3	<i>Machine Learning</i>	267
<b>5.7</b>	Isolation Barriers	270
5.7.1	<i>Implant Systems</i>	270
5.7.2	<i>Isolation Amplifiers</i>	272
<b>5.8</b>	References	274

---

<b>6</b>	Hearing Aids and Implants	277
<b>6.1</b>	Introduction	277
<b>6.2</b>	What Is Sound?	278
6.2.1	<i>Characteristic Impedance (Z) and Sound Pressure</i>	278
6.2.2	<i>Sound Intensity (I)</i>	279
<b>6.3</b>	How Hearing Works	281
6.3.1	<i>The Outer Ear</i>	281
6.3.2	<i>The Middle Ear</i>	281
6.3.3	<i>The Inner Ear</i>	283
6.3.4	<i>Hearing Statistics</i>	283
<b>6.4</b>	Hearing Loss	285
6.4.1	<i>Causes</i>	285
6.4.2	<i>Diagnosis</i>	286
6.4.3	<i>Treatment</i>	288
<b>6.5</b>	Hearing Aids	289
6.5.1	<i>History</i>	289
6.5.2	<i>Hearing Aid Operation</i>	292
<b>6.6</b>	Bone Conduction Devices	300
<b>6.7</b>	Middle Ear Implants	302
6.7.1	<i>Piezoelectric Devices</i>	303
6.7.2	<i>Electromagnetic Hearing Devices</i>	307
6.7.3	<i>Issues with Implantable Middle Ear Devices</i>	311
<b>6.8</b>	Direct Acoustic Cochlear Stimulatory Devices	312
6.8.1	<i>Actuator Design</i>	312

- 6.9** Cochlear Implants 314  
    6.9.1 *Historical Background* 314  
    6.9.2 *How Cochlear Implants Work* 315  
    6.9.3 *Installation of the Electrode* 320  
    6.9.4 *Signal Processing and Cochlear Stimulation* 320  
    6.9.5 *Spectral Maxima Strategies* 326  
    6.9.6 *Strategies to Enhance Vocal Pitch* 326

- 6.10** Auditory Brainstem Implants 328  
    6.10.1 *Electrodes* 329  
    6.10.2 *Stimulus Mapping* 330

- 6.11** References 330
- 

**7** Sensory Substitution and Visual Prostheses 333

- 7.1** Introduction 334  
**7.2** Anatomy and Physiology of the Visual Pathway 335  
**7.3** Main Causes of Blindness 339  
**7.4** Optical Prosthetics—Glasses, Thermal Imagers, Night Vision 339  
**7.5** Sonar-Based Systems 341  
    7.5.1 *Some Existing Systems* 344  
    7.5.2 *Issues with Sonar-Based Systems* 350  
**7.6** Laser-Based Systems 350  
**7.7** Sensory Substitution 350  
    7.7.1 *Auditory Substitution* 351  
    7.7.2 *Electrotactile and Vibrotactile Transducers* 356  
**7.8** GPS-Based Systems 370  
**7.9** Visual Neuroprostheses 371  
    7.9.1 *Historical Perspective* 371  
    7.9.2 *Potential Sites for Visual Neuroprostheses* 371  
    7.9.3 *Components* 372  
    7.9.4 *Worldwide Research Activity* 375  
    7.9.5 *Subretinal Implants* 375  
    7.9.6 *Epiretinal Implants* 381  
    7.9.7 *Alternative Implants* 386  
    7.9.8 *Optic Nerve Stimulation* 387  
    7.9.9 *Visual Cortex Implants* 388  
**7.10** The Future 391  
**7.11** References 392
- 

**8** Heart Replacement 395

- 8.1** Introduction 396  
**8.2** The Heart as a Pump 397  
    8.2.1 *Heart Valves* 398  
    8.2.2 *The Pump Cycle* 399  
    8.2.3 *The Cardiac Output* 401

8.2.4	<i>Pressure Regulation</i>	401
8.2.5	<i>Heart Disease</i>	402
8.2.6	<i>Biomechatronic Perspective</i>	402
<b>8.3</b>	Heart–Lung Machines	403
8.3.1	<i>History</i>	403
8.3.2	<i>Modern Heart–Lung Machines</i>	404
<b>8.4</b>	Artificial Hearts	408
8.4.1	<i>History</i>	409
8.4.2	<i>Implanting an Artificial Heart</i>	417
<b>8.5</b>	Ventricular Assist Devices	417
8.5.1	<i>History</i>	419
8.5.2	<i>Extracorporeal Ventricular Assist Devices</i>	420
8.5.3	<i>Intracorporeal Left Ventricular Assist Devices</i>	421
8.5.4	<i>Generation 1 LVADs</i>	421
8.5.5	<i>Pulsatile Pump Technology</i>	429
8.5.6	<i>Generation 2 VADs</i>	433
8.5.7	<i>Generation 3 VADs</i>	435
8.5.8	<i>Generation 4 VADs</i>	439
8.5.9	<i>Toward an Ideal Replacement Heart</i>	443
8.5.10	<i>Other Pump Types</i>	443
<b>8.6</b>	Engineering in Heart Assist Devices	446
8.6.1	<i>Fluid Dynamics in Pulsatile LVADs</i>	446
8.6.2	<i>Fluid Dynamics in Centrifugal and Axial LVADs</i>	448
8.6.3	<i>Estimation and Control of Blood Flow</i>	450
8.6.4	<i>Transcutaneous Energy Transfer</i>	452
<b>8.7</b>	Pump Types	455
8.7.1	<i>Centrifugal and Axial Pump Characteristics</i>	456
8.7.2	<i>Rotary Pump Characteristics</i>	460
8.7.3	<i>Reciprocating Pump Characteristics</i>	462
8.7.4	<i>Bearings</i>	466
<b>8.8</b>	References	466

---

<b>9</b>	Respiratory Aids	471
<b>9.1</b>	Introduction	472
<b>9.2</b>	Construction	473
<b>9.3</b>	The Mechanics of Respiration	476
9.3.1	<i>Physical Properties</i>	477
9.3.2	<i>Lung Elasticity</i>	480
9.3.3	<i>Frictional Forces</i>	481
9.3.4	<i>Inertia</i>	485
<b>9.4</b>	Energy Required for Breathing	485
<b>9.5</b>	Measuring Lung Characteristics	488
9.5.1	<i>Spirometry</i>	488
9.5.2	<i>Pneumotachography</i>	492

<b>9.6</b>	Mechanical Ventilation	494
9.6.1	<i>Early History</i>	494
9.6.2	<i>Polio</i>	495
9.6.3	<i>External Negative-Pressure Ventilators</i>	497
9.6.4	<i>The Drinker Respirator</i>	499
9.6.5	<i>The Both Respirator</i>	501
9.6.6	<i>Homemade Iron Lungs</i>	501
9.6.7	<i>The Emerson Respirator</i>	504
9.6.8	<i>The Alligator Cabinet Respirator</i>	505
9.6.9	<i>Portable Respirators</i>	506
9.6.10	<i>Other Uses for Negative-Pressure Ventilation</i>	507
<b>9.7</b>	The Physics of External Negative-Pressure Ventilation	508
<b>9.8</b>	Positive-Pressure Ventilators	511
9.8.1	<i>Historical Background</i>	511
9.8.2	<i>The Need for Positive-Pressure Ventilation</i>	512
9.8.3	<i>Ventilation Modes</i>	513
9.8.4	<i>Controlled Mandatory Ventilation</i>	514
9.8.5	<i>Volume-Controlled Mandatory Ventilation</i>	514
9.8.6	<i>Pressure-Controlled Mandatory Ventilation</i>	516
9.8.7	<i>Spontaneous Ventilation</i>	517
9.8.8	<i>Continuous Positive Airway Pressure</i>	517
9.8.9	<i>Portable Ventilators</i>	517
9.8.10	<i>Sleep Apnea</i>	519
<b>9.9</b>	References	520

---

<b>10</b>	Active and Passive Prosthetic Limbs	523
<b>10.1</b>	Introduction	524
10.1.1	<i>A Brief History of Prosthetics</i>	524
<b>10.2</b>	Structure of the Arm	529
10.2.1	<i>Wrist</i>	529
10.2.2	<i>Elbow</i>	530
10.2.3	<i>Shoulder</i>	530
<b>10.3</b>	Kinematic Model of the Arm	531
<b>10.4</b>	Structure of the Leg	532
10.4.1	<i>The Hip Joint</i>	532
10.4.2	<i>The Knee Joint</i>	533
10.4.3	<i>The Ankle Joint and the Foot</i>	533
<b>10.5</b>	Kinematic Model of the Leg	534
10.5.1	<i>Walking</i>	534
10.5.2	<i>Normal Walking Dynamics</i>	535
<b>10.6</b>	Kinematics of Limb Movement	536
10.6.1	<i>Center of Mass and Moment of Inertia of a Limb Segment</i>	536
10.6.2	<i>Angular Acceleration</i>	538
10.6.3	<i>Center of Mass and Moment of Inertia of a Complete Limb</i>	538
<b>10.7</b>	Sensing	538

- 10.8 Passive Prosthetics 538**
  - 10.8.1 Actuation and Control of Upper Limb Prostheses 539*
  - 10.8.2 Walking Dynamics Using a Passive Prosthesis 541*
  - 10.8.3 Knee Prosthetics 542*
  - 10.8.4 Foot Prosthetics 545*
- 10.9 Active Prosthetics 547**
  - 10.9.1 Arm Mechanisms 548*
  - 10.9.2 Hand Mechanisms 554*
  - 10.9.3 Hand Research and Applications 562*
  - 10.9.4 Control of Prosthetic Arms and Hands 563*
  - 10.9.5 Leg Mechanisms 574*
  - 10.9.6 Ankle–Foot Mechanisms 580*
- 10.10 Prosthesis Suspension 582**
  - 10.10.1 Conventional Suspension Methods 583*
  - 10.10.2 Osseointegration 583*
- 10.11 References 584**
- Index 587**

# Preface

At some point in their lives almost everyone you know will rely on a mechatronic device to keep them from dying or to maintain or improve the quality of their existence. These devices range from neurally controlled multi degree-of-freedom prosthetic limbs to replace those lost to accident, disease or war, through ultra-reliable ventricular assist devices to support damaged hearts in patients awaiting transplants, to tiny actuators that drive the bones in the middle ear, allowing the profoundly deaf to hear again. This book aims to enlighten engineering and medical students, as well as practitioners, with how these amazing devices work and how they have evolved from their primitive origins to their current status as products in the multibillion dollar medical appliances industry.

## University Courses

Until about twenty years ago, universities typically offered degrees in electronics, software or mechanical engineering with only a small amount of mixing. However, as electromechanical devices became smarter with the inclusion of integrated computers and microprocessors, it became clear that a new breed of engineer was required. Mechatronic engineering curricula aim to fulfil that requirement by amalgamating aspects of mechanical engineering, electronics, control engineering and computer science into a single degree. This book would help fill that need in biomedical subject areas.

Biomedical degrees or specializations have been available for many years, but these tended to focus on biomechanical or bioelectronics aspects of engineering. It was only four or five years ago that biomechatronic courses started to appear in any number, but to a large extent these were based on existing biomechanical courses involving limb prosthetics and exoskeletons. As an electrical engineer with a radar background, I didn't feel constrained by these norms and set out to develop a biomechatronics course that offered a wider focus. I wanted the course to meet the needs of medical and biomedical engineering students interested in mechatronic systems, as well as mechatronic engineering students with a biomedical bent.

Over a number of years, as the course was presented, the notes evolved to include individual chapters focussing on fundamental engineering principles, including sensing, actuation, signal processing and control that aimed to provide background for the biomedical engineering students. In addition, the notes also included chapters in which these principles were applied to medical systems. I felt that it was important to examine these systems from an holistic perspective so that students could understand how and why they developed. To achieve this, each of these chapters includes some basic anatomy and physiology for the mechatronics students, and an historical view of the evolution of the technology. Because the field is so wide, it was not possible to include everything I would have liked to in a single semester course, so I had to limit the focus to biomechatronic aspects of prosthetics and orthotics for vision, hearing, circulation, respiration and locomotion.

## Practitioners and Interested Laypersons

Though there is a considerable amount of piecemeal information available both on the Web and in specialist printed texts, nothing covered the range of topics in which I was interested, and in the degree of detail that I thought appropriate at an introductory level, so I approached SciTech Publishing with a proposal to turn my notes into a book. The good people at SciTech reminded me that a good upper division textbook will attract interest among practitioners as well as students. So here is my book in an exciting, rapidly developing field of interest, for all who would like to learn more.

I have read widely and incorporated much of what I have read in a distilled form into this book, and for that would like to thank the myriad authors referenced. Where I have used photographs or figures and tables unaltered, I have attempted to contact the original source for permission to publish. However, obviously with a book of this nature, a few errors are bound to slip through even the most stringent editing process, so I would appreciate feedback from readers to correct these in future printings.

Graham Brooker  
University of Sydney, Australia  
October 2011

# Acknowledgments

Not many people get to do something completely different and interesting in the autumn of their working career, so I have been really fortunate that my position as a lecturer at the University of Sydney has given me the opportunity to develop a Biomechatronics course and to convert the notes into a book. However, to embark on something of this magnitude requires a great deal of time, time I have once again stolen from my family and particularly from my wife, Mandi. I can't thank you enough for the support you have offered. Surely this will be the last time!

I would also like to thank my mom-in-law, Mary, for proof reading during a very difficult time in her life, and also to Liz Magdas for the entertaining drawings that embellish some of the chapter headings.

This book is dedicated to ends and to new beginnings—to my Mom, Trish, and my father-in-law, Ric, both of whom passed away during its writing, and to my grandchildren Charlotte and Daniel who were born.



# Introduction to Biomechatronics

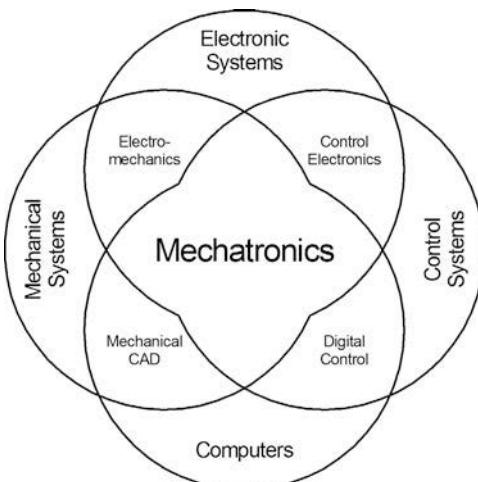
## Chapter Outline

1.1	Introduction .....	1
1.2	Biomechatronic Systems .....	2
1.3	Physiological Systems .....	4
1.4	Summary of Contents .....	6
1.5	The Future of Biomechatronic Systems .....	6
1.6	References .....	7

## 1.1 INTRODUCTION

*Mechatronic engineering* is the synergistic combination of mechanical, electronic, computer, and control systems along with a dash of systems engineering as illustrated in Figure 1-1. This interdisciplinary combination brings together the requisite technology and skills to design new and to improve existing electromechanical systems.

Biomechatronics is the application of mechatronic engineering to human biology, and, as such, it forms an important subset of the overall biomedical engineering discipline. As with mechatronics, which is often synonymous with robotics, biomechatronics is often thought of as restricted to the development of prosthetic limbs. However, in reality,



**FIGURE 1-1** ■  
Mechatronic engineering as a combination of mechanical, electronic, computer, and control systems.

biomechatronics covers a much wider genre than this, and along with prosthetic limbs this book examines some of the more interesting applications including those related to hearing, respiration, vision, and the cardiovascular system.

## 1.2 BIOMECHATRONIC SYSTEMS

Ultimately, biomechatronics can be thought of in a similar manner to any other engineering system with one of its elements, generally the most complex one, being the human being. Unfortunately, the human element is not only the most complex and least understood but also the most difficult to interface to. Attempts to measure and stimulate the human body are not completely deterministic, and repeated application of a set of inputs will not always produce the same response. In fact, even when under conscious control, responses (or actions) are seldom identical. Consider, for example, the best sportsmen in the world: With practice and talent they are able to produce fairly repeatable performances, but subtle changes in initial conditions, within and external to their bodies, results in some variations. This uncertainty is manifest across the complete range of physiological responses, from slight variations in the resting heart rate through the apparently chaotic nature of firing neurons.

In a typical biomechatronic system, a number of components can be identified. These include the following:

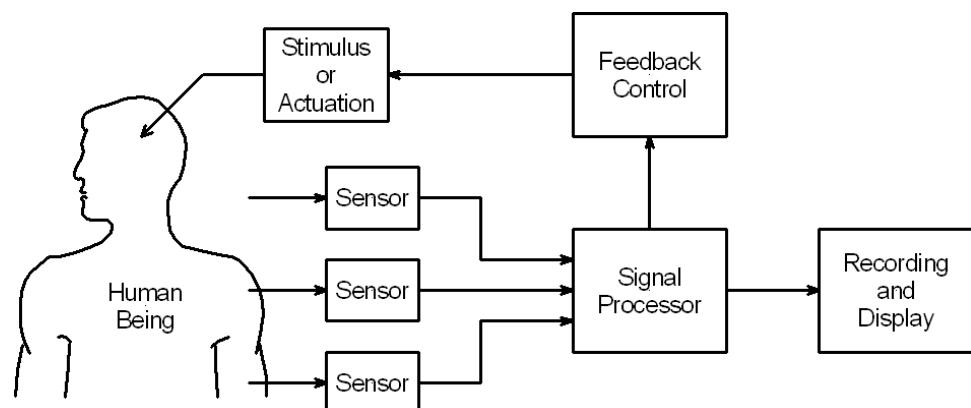
- The human (or animal) subject
- Stimulus or actuation
- Transducers and sensors
- Signal conditioning elements
- Recording and display
- Feedback elements

These are clearly illustrated in Figure 1-2.

### 1.2.1 The Human Subject

The human subject adds the *bio* to this mechatronic control and monitoring process. What makes biomechatronics particularly interesting compared with other mechatronic systems

**FIGURE 1-2** ■  
Block diagram showing the elements of a biomechatronic system.



is the diversity and complexity of human physiology. Unlike the usual engineering systems, the behavior of which can be more or less predicted, each human being is unique and ever changing. This is not a book on human physiology, but to give the reader some essential background its various aspects are considered in more detail in later sections of this book when a specific type of biomechatronic system is introduced.

### 1.2.2 Stimulus or Actuation

The process of stimulation can be introduced as a feedback element, as shown in Figure 1-2, or as a naturally occurring input. Sources of stimuli can encompass any modality that has an effect on the human element. This can include electrical stimuli, an audio tone, control of air or blood flow, a source of light, a tactile stimulus, or even the physical actuation of a limb. This book includes an entire chapter on actuators, their implementation and analysis in general terms, and their consideration as part of a number of individual biomechatronic systems.

### 1.2.3 Transducers and Sensors

Transducers and sensors are the devices that convert physiological outputs into signals that can be used. In most cases, these are sensors that amplify electrical signals or convert them from chemical concentration, temperature, pressure, or flow into electrical signals that can be further processed. Interfacing to the human body is not a trivial task, as embedded sensors must be biocompatible, flexible, and extremely robust to survive in the aggressive internal environment, while surface sensors, particularly electrodes, must be able to form a compatible and relatively stable conductive interface across the skin. As with actuators, sensors and sensing of physiological processes are sufficiently important to devote a full chapter to the physical mechanisms that underpin their operation.

### 1.2.4 Signal Processing Elements

Signal processing involves modification of the electrical signal to some form that is more useful. This generally involves amplification and filtering to extract salient features. However, it often involves the conversion of the analog signal to a digital equivalent that allows for the application of complex algorithms to obtain subtler characteristics. This book includes a chapter that concentrates on the classical aspects of signal processing, before introducing the reader to modern machine learning algorithms. The latter are becoming essential tools in the quest to better understand complex physiological processes.

### 1.2.5 Recording and Display

In many cases, the biomechatronic device functions to monitor a physiological process or response. In these cases it may be important to display the information in a form that is easy to interpret, or to store it for later analysis. Common examples of such devices are the now ubiquitous 12-lead electrocardiograph, pneumotachographs and sphygmomanometers. In the past many of these devices were mechanical and outputs were recorded onto paper tape or photographic film, but with the advent of modern electronics, most have been replaced by their electronic equivalents—random access memory (RAM) and liquid crystal displays.

### 1.2.6 Feedback Elements

In a closed-loop control application, any stimulus or excitation signal is conditioned by the processed outputs of one or a number of sensors monitoring the physiological process. The link that connects the sensing output back to the stimulus includes further processing through control elements. This feedback can be used to close an external loop or one that operates through the human being. An example of the latter is a vibrotactile stimulus that is triggered in response to some changes in the direction of the earth's magnetic field to help a blind person navigate. The chapter on classical control investigates the use of feedback elements in terms of their effect on loop performance.

## 1.3 | PHYSIOLOGICAL SYSTEMS

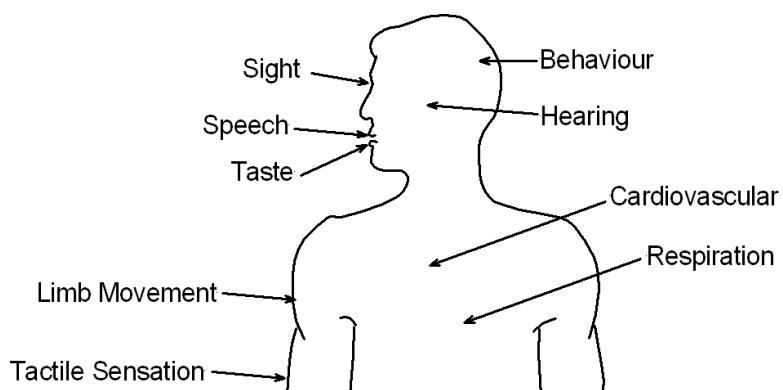
To interact effectively with the human part of the structure shown in Figure 1-2, it is essential to have some understanding of the subject on which the measurements are being made or to which the stimulus is applied. The major functional systems of the body include the cardiovascular, respiratory, and musculoskeletal systems along with those that interpret of taste, sight, hearing, and touch. These are illustrated in Figure 1-3.

Most of the second half of this book considers the application of biomechatronic systems to these physiological components.

### 1.3.1 Biochemical System

The human body is controlled and powered by a complex system of chemical processes. Biochemical processes convert the food we eat into amino acids that are used for building and repair; they break down sugars and fats and store them for later use as sources of energy. In addition, our blood and tissues are awash with hormones and other organic molecules that control, signal, and regulate the amazingly interconnected functions that keep the system alive and healthy. Later sections of this book examine some of the sensors and actuators that interact with these biochemical processes for monitoring and control purposes.

**FIGURE 1-3 ■**  
Major physiological components that can be incorporated into biomechatronic systems. [Adapted from Cromwell, Weibell, et al. (1973).]



### 1.3.2 Nervous System

Whereas the biochemical system is concerned mostly with long-term signaling and control, the nervous system is the high-speed communications network for the body. At its center is the brain, which performs the processing and memory storage tasks, and to it are connected myriad input/output nerve channels. These nerves convey sensory and status information to the brain from specialized sensory organs like the eyes, ears, and skin and from the various internal organs. Nerve outputs from the brain provide feedback to control some of the internal processes and to adjust the tension in the hundreds of muscles in the body that allow us to interact with the environment. Later chapters discuss some of these interactions in more detail, particularly in regard to biomechatronic prostheses and interfaces that can restore function when sensory organs have been damaged.

### 1.3.3 Cardiovascular System

From an engineering perspective, the cardiovascular system is a closed-circuit hydraulic system comprising a dual-action multichambered electrically controlled pump (the heart) interconnected through hundreds of kilometers of flexible tubing (arteries, veins, and capillaries). Pressure and flow regulation are achieved by changing the pump stroke and rate as well as altering the diameters of the arteries. The hydraulic fluid (blood plasma) contains organic molecules for regulation as well as larger particles to aid with puncture repair (blood platelets), defense against intruders (white cells), and the transport of oxygen to the tissues and waste products back to the lungs (red cells). One chapter of this book is dedicated to analyzing the performance of the heart and its mechatronic replacement.

### 1.3.4 Respiratory System

The respiratory system is pneumatic and consists of an air pump (the diaphragm and ribcage) that alternately produces negative and positive pressures in a sealed chamber (the thoracic cavity) drawing air into and then expelling it from a pair of balloon-like organs (the lungs) linked to the atmosphere. The lungs are designed to provide hundreds of square meters of highly vascularized membranes (the alveoli) that allow for the free exchange of gas between the regularly replenished air and the blood to ensure that essential oxygen levels remain high and that waste carbon dioxide is flushed from the system. The pump regulatory mechanism is automatic, but it is provided with a conscious override that can be used to accelerate or inhibit flow as the circumstances dictate. One chapter of this book deals with some of the biomechatronic mechanisms that can augment or replace the regulatory process should its performance become degraded or even if it fails altogether.

### 1.3.5 Musculoskeletal System

The musculoskeletal system performs two major functions. First, it maintains the integrity of the body by providing a firm structure to both support and protect the internal organs. Second, it provides a means for the organism to interact with the outside world by means of locomotion and manipulation. For human beings, locomotion is primarily provided by the legs and feet, and our capability for manipulation by the arms and the hands, and particularly the fingers. A chapter of this book is concerned with replacing lost limbs and the augmentation of limb function in cases of diminished performance.

## 1.4 | SUMMARY OF CONTENTS

---

Most biomechatronic systems operate by sensing some particular aspect of the environment or the human body, processing the information, and then responding in some way. It is therefore important that the student is aware of sensor technology, processing methods, and actuation systems. These are all covered in the first half of this book. Chapter 2 introduces the concept of biometrics, the science of measuring physiological signals, and myriad sensors that can be used to measure these. Such measurements include bioelectric signals in the form of the electrocardiogram, the electroencephalogram, and the electromyogram. Microphones can be used to measure heart sounds, while pressure and flow sensors measure the characteristics of the cardiovascular and respiratory systems. Other sensors measure the position and rates of limb elements, real and prosthetic. Chapter 3 considers various forms of actuation that use electrics, hydraulics, or pneumatics. Chapter 4 closes the loop around the measurement and actuation process. It examines various feedback types and methods of analysis. Chapter 5 is the last of the background chapters, and it introduces concepts of both analog and digital signal processing that are applicable to biomechatronic systems. These include filtering in the analog and digital domains, rectification and detection of signals, and sampling and digitization. It also introduces the reader to machine learning algorithms, which are becoming important in improving our ability to interpret complex physiological processes.

The second half of this book is focused on specific aspects of human physiology with respect to biomechatronic possibilities for their improvement or rehabilitation. Chapter 6 is concerned with hearing, and it considers conventional hearing aids, bone-anchored hearing aid (BAHA) devices, middle ear implantable hearing devices (MEIHDs), as well as cochlear and brainstem implants. Chapter 7 examines one of the less mature fields—sensory substitution and ocular prosthetics based on neural and cortical implants. The following three chapters address the rehabilitation or replacement of organs or systems that move, with Chapter 8 focusing on the cardiovascular system, particularly in regard to the total artificial heart and ventricular assist devices. Chapter 9 examines respiration with a focus on negative pressure ventilators (iron lungs) and positive pressure respiration devices. Finally, Chapter 10 examines powered and passive limb prostheses in detail, as these are some of the most important biomechatronic devices.

## 1.5 | THE FUTURE OF BIOMECHATRONIC SYSTEMS

---

An introductory text cannot hope to cover even a small fraction of the biomechatronic devices and systems that exist or will be available in the foreseeable future. This book provides some insight into the technology and applications that have been developed over the last 60 years or so. But, as with all technology, the rate of advance in the field is accelerating, so it is expected that some amazing new biomechatronic devices will become available within the next decade.

Improvements in our ability to interface directly to the human neural system will provide the greatest advances in a range of applications. Already, neural reinnervation that offers both actuation and feedback is providing better interfaces to prosthetic limbs (MacIsaac and Englehart, 2006; Kuiken, Miller et al., 2007). This is just an interim measure, and direct interface to neurons in the brain will allow seamless integration of

increasingly sophisticated limb prostheses to human amputees (Hochberg, Serruma et al., 2006; Musallam, Corneil et al., 2004; Nicolelis 2011; Serruma and Donoghue, 2003). Research conducted at Brown University trained a rhesus monkey to track visual targets on a computer screen without using a joystick (Serruma, Hatsopoulos et al., 2002). More recently, experiments conducted by the same group in 2006, using a similar brain-computer interface, showed that a monkey could feed itself using a robotic arm controlled directly by its own brain. On this front, progress is being made in our understanding of how the brain operates, based primarily on functional magnetic resonance imaging (fMRI) studies, and our ability to use light to trigger neural activity (Nagel, Brauner et al., 2005). Some futurists believe that within the next few decades such interfaces will no longer be necessary as we will be capable of replacing the human brain with a computer (Kurtzweil, 2005).

Where the body is intact but is dysfunctional, new lightweight materials, improved batteries, and small but powerful actuators will be used to provide full-body powered exoskeletons that will allow the wheelchair bound to walk again (Nicolelis, 2011; Pons, 2008). Already, exoskeletons have been developed as walking aids or assistive limbs for the aged and as mechanisms to provide users with superhuman strength (Stevens, 2010).

Other exciting biomechatronic applications that have been promised for some time but are not yet available are nano-machines that could be injected into the body to perform microsurgery in inaccessible areas. To date, improvements in microelectromechanical system (MEMS) technology have facilitated the development of small devices that are capable of locomotion through liquids (Sanchez, Solovev et al., 2011).

Micro actuators have been developed that are capable of stimulating the ossicles within the human ear to restore hearing to the deaf. This technology is advancing quickly and may replace audio amplification based hearing aids in the near future (Shohet, 2008).

Advances in signal processing and the reliable and safe electrical stimulation of neurons have made the cochlear implant the most common prosthetic in the world. This success has fostered research into other sensory prosthetics focused mostly on restoring sight to the blind. Electrode arrays are now routinely inserted into the visual cortex and onto the retina to provide rudimentary vision. It is envisaged that within a decade these implants will allow blind people to navigate through their environments with confidence and even to read again (Lovell, Hallum et al., 2007). Ultimately, visual implants may offer color vision or even hyperspectral capabilities—the ability to see up into the ultraviolet or down in frequency into the infrared.

Biomechatronic devices are already replacing diseased hearts to prolong and improve the quality of the lives of patients. Improvements in electromechanical devices, materials technology, and computational fluid dynamics continue this trend with patients now able to lead almost normal lives (Deng and Naka, 2007).

In the future it may be possible to provide other artificial organs including lungs (Downs, 2002), improved kidneys, and maybe even livers.

## 1.6 | REFERENCES

- Cromwell, L., F. Weibell, and E. A. Pfeiffer. (1973). *Biomedical Instrumentation and Measurements*. Englewood Cliffs, NJ: Prentice-Hall, Inc.
- Deng, M. and Y. Naka. (2007). *Mechanical Circulatory Support Therapy in Advanced Heart Failure*. London: Imperial College Press.
- Downs, M. (2002). "Artifical Lung Closer to Clinical Trial." *WebMD*. Retrieved March 2011 from <http://www.webmd.com/lung/features/artificial-lung-closer-to-clinical-trial>

- Hochberg, L., M. Serruma, G. M. Friehs, J. A. Mukand, M. Saleh, A. H. Caplan, et al. (2006). “Neuronal Ensemble Control of Prosthetic Devices by a Human with Tetraplegia.” *Nature* 442(7099): 164–171.
- Kuiken, T., L. Miller, R. D. Lipschutz, B. A. Lock, K. Stubblefield, P. D. Marasco, et al. (2007). “Targeted Reinnervation for Enhanced Prosthetic Function in a Woman with a Proximal Amputation: A Case Study.” *Lancet* 369(9559): 371–380.
- Kurtzweil, R. (2005). *The Singularity Is Near: When Humans Transcend Biology*. New York: Viking.
- Lovell, N., L. Hallum, S. C. Chen, S. Dokos, P. Byrnes-Preston, R. Green, et al. (2007). Advances in Retinal Neuroprosthetics. In *Handbook of Neural Engineering*, M. Akay (Ed.). Wiley-IEEE Press.
- MacIsaac, D. and K. Englehart. (2006). “The Science Fiction’s Artificial Men.” *CrossTalk—The Journal of Defense Software Engineering*, October.
- Musallam, S., B. Corneil, B. Greger, H. Scherberger, and R. A. Andersen. (2004). “Cognitive Control Signals for Neural Prosthetics.” *Science* 305(5681): 258.
- Nagel, G., M. Brauner, J. F. Liewald, N. Adeishvili, E. Bamberg, and A. Gottschalk. (2005). “Light Activation in Channelrhodopsin-2 in Excitable Cells of *Caenorhabditis Elegans* Triggers Rapid Behavioral Responses.” *Current Biology* 15(24): 2279–2284.
- Nicolelis, M. (2011). Mind Out of Body. *Scientific American*, February, 61–63.
- Pons, J. (Ed.). (2008). *Wearable Robots—Biomechatronic Exoskeletons*. Chichester, UK: John Wiley & Sons.
- Sanchez, S., A. Solovey, S. M. Harazim, and O. G. Schimdt. (2011). “Microbots Swimming in the Flowing Streams of Microfluidic Channels.” *Journal of the American Chemical Society* 133(4): 701–703.
- Serruma, M. and J. Donoghue. (2003). Design Principles of a Neuromotor Prosthetic Device. In *Neuroprosthetics: Theory and Practice*, K. Horch and G. Dhillon (Eds.). Imperial College Press, Chapter 3.
- Serruma, M., N. Hatsopoulos, L. Paninski, M. R. Fellows, and J. P. Donoghue. (2002). “Instant Neural Control of a Movement Signal.” *Nature* 416: 141–142.
- Shohet, J. (2008). “Implantable Hearing Devices.” Retrieved July 2008 from <http://www.emedicine.com/ent/TOPIC479.HTM>
- Stevens, T. (2010). “HULC Exo-skeleton Ready for Testing, Set to Hit the Ground Running Next Year.” Retrieved March 2010 from <http://www.engadget.com/2010/07/21/hulc-exo-skeleton-ready-for-testing-set-to-hit-the-ground-runni>

# Sensors and Transducers

## Chapter Outline

2.1	Introduction .....	9
2.2	Switches .....	9
2.3	Power Supplies .....	13
2.4	Sensors and Transducers .....	22
2.5	Electrodes .....	83
2.6	References .....	88

## 2.1 | INTRODUCTION

Much of biomechatronic engineering involves the measurement of physical processes stemming either from the biological organism or from an associated prosthesis. These measurements can include voltage, chemical concentration, pressure, position, displacement, and rate. In general, they are processed in some way and then used to apply some form of actuation such as the motion of a prosthetic arm, the contraction of an artificial heart, or the electrical stimulation of the cochlea.

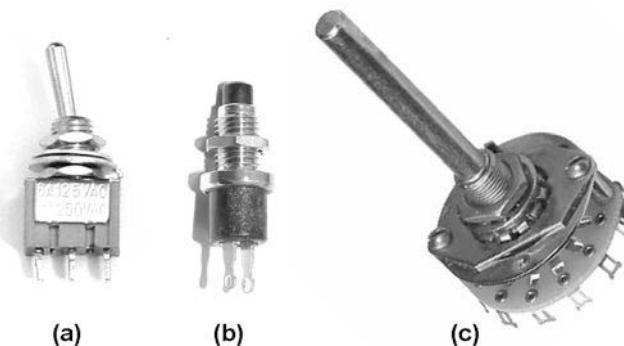
In most cases the measured parameter is converted to an electrical signal for analysis or processing, though a number of devices such as the sphygmomanometer, used to measure blood pressure, often still use pneumatics. However, since the introduction of low-cost lightweight, and reliable electronics in the last 60 years, and particularly since the invention of the microprocessor, far more complex analysis and processing can be undertaken.

This chapter examines some of the more common sensors and transducers that are used in biomechatronic engineering.

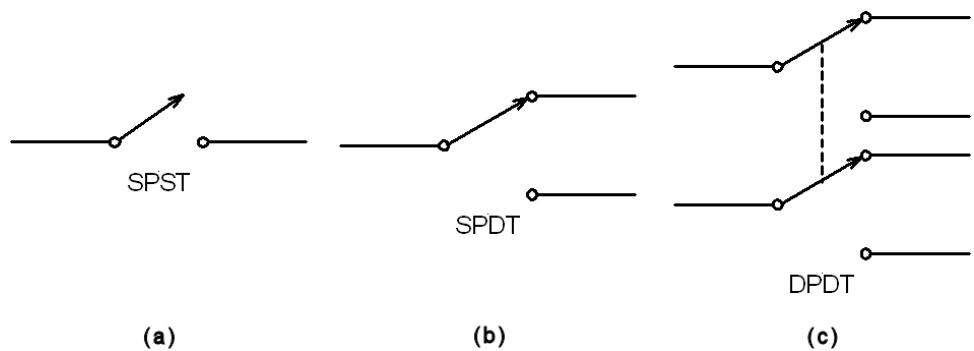
## 2.2 | SWITCHES

The fundamental electrical building block is the switch, of which a number of examples are shown in Figure 2-1. These are devices that control the flow of current in a binary (on–off) manner, using either a mechanical contact or an electronic device such as a transistor. Switches are commonly used to isolate the supply of current to a device, are used as selectors, or, often in mechatronic devices, are used to identify the limits of travel of some mechanical structure.

**FIGURE 2-1** ■  
Examples of some panel switches.  
(a) Toggle. (b) Push button. (c) Rotary.



**FIGURE 2-2** ■  
Fundamental switching configurations.  
(a) Single pole, single throw.  
(b) Single pole, double throw.  
(c) Double pole, double throw.



### 2.2.1 Toggle Switches

Toggle switches are available in a number of different configurations depending on the application and are characterized by the number of switching contacts (or poles) and the number of terminals. For example, a single on-off switch shown in Figure 2-2a is referred to as a single pole single throw (SPST) device. In the case where the switch can connect to a pair of terminals, as shown in Figure 2-2b, it is referred to as a single pole double throw (SPDT) device. Also common are multipole switches, in which individual switching elements are ganged together to form double, triple, and even quadruple devices. The nomenclature follows that of the previous cases, so a double pole double throw switch would be referred to as DPDT.

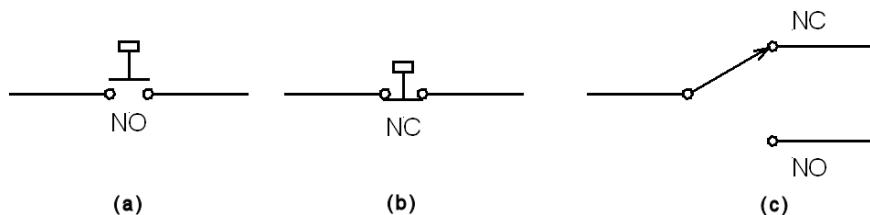
All double throw switches are break-before-make in that the pole never contacts both of the terminals simultaneously. This is important to ensure proper isolation and to avoid short circuits.

### 2.2.2 Push-Button Switches

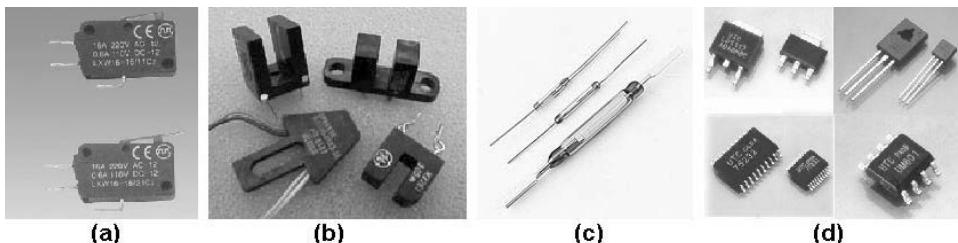
Push-button switches are momentary contact devices because they operate only when force is applied. They come in two varieties, with a normally open (NO) or a normally closed (NC) contact for single throw devices, but are often also configured as double throw devices as shown in Figure 2-3c.

### 2.2.3 Limit Switches

Mechanical limit switches are a variation of the standard push-button configuration fitted with a lever arm. These are often known as microswitches because of their small size.



**FIGURE 2-3** ■ Push-button switch configurations.  
 (a) Normally open.  
 (b) Normally closed.  
 (c) double throw.



**FIGURE 2-4** ■ Examples of limit switches.  
 (a) Micro-switch.  
 (b) Optical. (c) Reed.  
 (d) Hall device.

In the interests of reliability, mechanical limit switches have mostly been superseded by optical or hall switches (discussed later), which have no moving parts and are generally easier to interface. Figure 2-4 shows photos of some examples of microswitches and their electronic equivalents.

## 2.2.4 Rotary Switches

Rotary switches consist of a number of circular wafers containing many poles operated by a spindle that passes through the center of each wafer. There are many types, including both shorting (make-before-break) and nonshorting (break-before-make). However, as with most other mechanical switches, these are not particularly reliable and are being replaced by their electronic counterparts.

### 2.2.5 Optical Switches

Optical switches consist of an infrared light-emitting diode (IR LED) and a phototransistor sensitive to a similar optical wavelength, mounted into the same plastic holder. These are mostly configured so that the IR LED is directed toward the phototransistor across a slot in a transmission configuration, but some are mounted at an angle to operate in a reflective configuration. In the transmission configuration, the phototransistor operates as a switch that is conductive (on) if it is illuminated by the IR LED. In the transmission configuration if an opaque object is introduced into the slot, then the light path is blocked and the phototransistor turns off, whereas in the reflective case the switch is off until a reflective material is introduced to reflect the light from the IR LED back into the phototransistor.

## 2.2.6 Other Switches

Reed switches generally consist of a normally open contact pair mounted on ferromagnetic reeds in a hermetically sealed glass tube so that, in the presence of a strong magnetic field, the contacts are drawn together to close the switch.

The Hall effect generates a potential difference orthogonal to the direction of current flow in the presence of a magnetic field. Hall switches are integrated circuits (ICs) that exploit this effect to produce the electronic equivalent of reed switches.

As well as operating as limit switches, these electronic switches are often also used to detect rotary motion. In the optical case a slotted disk is used, while in the other cases a disk that has been appropriately magnetized can be used. However, because these devices have a large sensitive region they are effective only for low count/revolution applications. Dedicated higher-resolution rotary encoders are discussed later in this chapter.

It is important to note that all switches, whether mechanical or electronic, have a limited current and voltage rating that must be adhered to if reliable operation is required. This specification is particularly important if inductive loads such as motors or relays are switched. The high induced voltage during turnoff can induce arcing across the contacts of mechanical switches and will destroy electronic switches.

Switches of all varieties are common in most mechatronic devices. These range from the conventional toggle on–off switch that controls the power supplied to the device through a range of limit switches ensuring that actuators remain within the required physical limits. One common application for microswitches is in the socket of conventional powered upper limb prostheses as discussed in Chapter 10. These can be controlled by the remaining muscles above the amputation or, in the case of thalidomide deformity, by using the nub digits that sometimes remain.

### 2.2.7 Relays

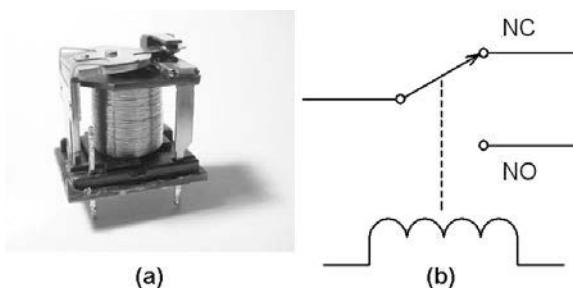
Relays are electrically controlled switches. In their mechanical embodiment they consist of an electromagnet that controls a mechanical lever arm to operate a switching mechanism, as illustrated in Figure 2-5. As with toggle switches, many switching configurations are possible.

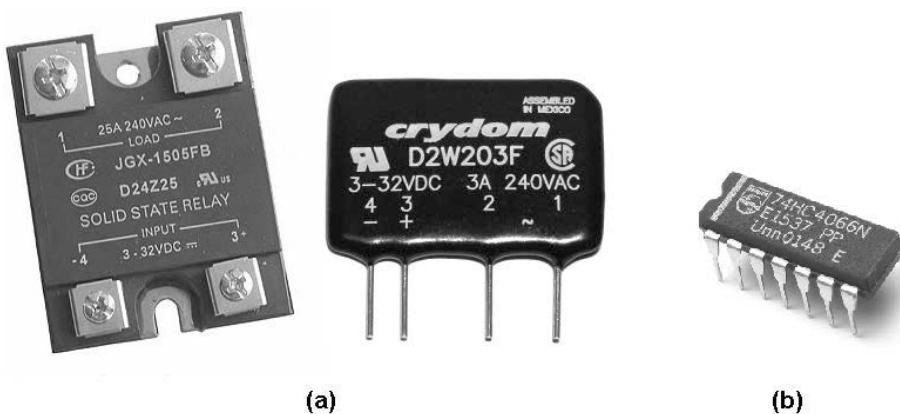
Electronic switches have replaced their mechanical counterparts in many applications both for switching high-voltage mains and low-voltage control signals. Solid-state relays (SSRs) generally have an opto-isolated low-voltage input that controls high current-handling transistors (usually metal–oxide–semiconductor field-effect transistors [MOSFETs]) or thyristors. When designed for lower current applications, these devices are usually called analog switches. Some examples of these devices are shown in Figure 2-6.

In biomechatronic applications, relays are often used to provide the mandatory electrical isolation between the various parts of a system to ensure patient safety. Solid-state relays are also quite common where gross on–off control of electric motors is required in primitive prostheses.

**FIGURE 2-5 ■**

Mechanical relay.  
 (a) Photograph.  
 (b) SPDT switching configuration.





**FIGURE 2-6** ■ Examples of electronic switches.  
 (a) Solid-state relays.  
 (b) Analog switch.

## 2.3 | POWER SUPPLIES

Medical and biomechatronic devices generally run from mains power if they are static and from batteries if they need to be portable. In most cases this supply voltage will need to be converted to one or a number of other voltages to power the various modules within any device. For example, a prosthetic arm may require 12 V to power the motors,  $+/-5$  V for the analog electronics, and 3.3 V for the signal processor.

It is now common to provide power to embedded systems such as cochlear devices and artificial hearts using electromagnetic induction through the unbroken skin, as this eliminates a common source of infection. This technology is discussed in more detail later in the book. An alternative that is becoming feasible as the efficiency of implants improves is to scavenge power from flexing muscles or changes in pressure driven by the heartbeat.

### 2.3.1 Linear Power Supplies

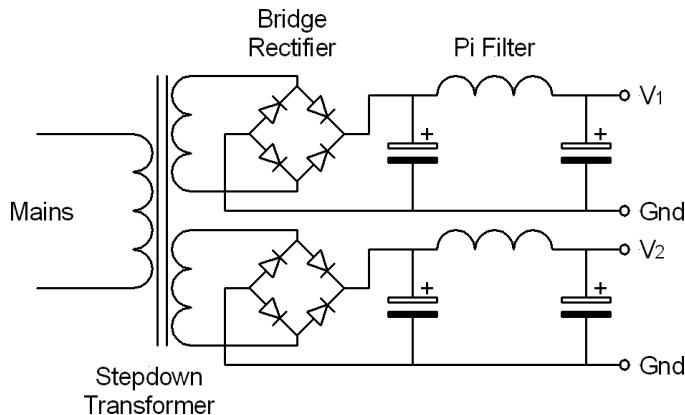
Until recently, most mains powered devices consisted of a step-down transformer followed by a full-wave rectifier to produce rectified alternating current (AC). This was followed by large capacitors and chokes to supply the smoothed direct current (DC) voltage required by the system. If different voltages were required, the transformer would include a number of secondary windings as shown in Figure 2-7. Unfortunately, transformers are heavy and expensive, so modern power supplies often rectify the mains directly and then use switch-mode power supplies to provide the various output voltages.

Irrespective of the method used, good isolation must be maintained between the input and output voltages, and also the raw DC must be further regulated and filtered to remove mains hum or switching noise and to maintain a constant and clean output voltage.

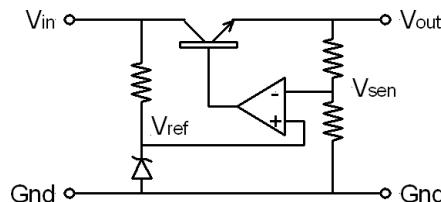
Linear regulators are active control systems that compare the output voltage with a fixed reference and use the error to adjust a series transistor to keep the output constant irrespective of changes in the input voltage or the load, as shown schematically in Figure 2-8. In this example, a Zener diode is used to generate the reference voltage, which is compared with the output sense voltage using an operational amplifier (op amp), the output of which drives the NPN transistor.

It is possible to construct voltage regulators from discrete components, but modern voltage regulator integrated circuits are cost-effective and extremely sophisticated. They

**FIGURE 2-7 ■**  
Schematic diagram for a dual rail mains power supply.



**FIGURE 2-8 ■**  
Schematic diagram of a generic voltage regulator.



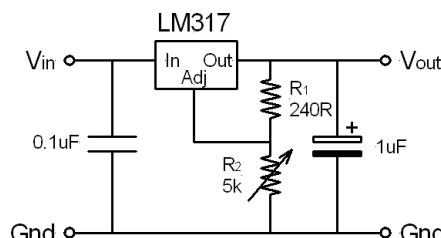
offer excellent regulation, low dropout (LDO) voltage (the difference between  $V_{in}$  and  $V_{out}$ ), and good input noise rejection. Fixed three-terminal regulators like the LM7812 (+12 V) and LM7912 (-12 V) and their relatives operating at other voltages are easy to use, require few external components, and provide good performance for loads of up to about 1 A (depending on the voltage drop). Adjustable three-terminal regulators like the LM317 (for +ve voltages) and LM337 (for -ve voltages) have become the regulators of choice for many designs. They are low cost, require few peripheral components, as can be seen from the schematic in Figure 2-9, and are capable of adjusting the output voltage between 1.2 V and 37 V for a 40 V input voltage.

Because all linear regulators produce output voltages that are lower than the input, some power,  $P_{disp}$ , is dissipated. This is equal to the product of the average current,  $I_{out}$ , and the difference between the input and the output voltage

$$P_{disp} = I_{out} (V_{in} - V_{out}) \quad (2.1)$$

A wide range of linear regulators is available commercially for most applications, as can be seen from Table 2-1. For more information on these and other devices, examine the catalogs of electronics mail-order companies such as Element14 (previously Farnell) or RS-Electronics.

**FIGURE 2-9 ■**  
Schematic diagram of a voltage regulator based on the LM317.



**TABLE 2-1** ■ Specifications of a Range of Linear Voltage Regulators

Component	Feature	Output Voltage (V)	Dropout Voltage	Current (max)	Line Regulation	Load Regulation	Ripple Reject
LM317L	Adj	1.2 to 37		100 mA	0.01%/V	0.1%	80 dB
LM337L	Adj	-1.2 to -37		100 mA	0.01%/V	0.1%	80 dB
LM317	Adj	1.2 to 37		1.5 A	0.01%/V	0.3%	80 dB
LM337	Adj	-1.2 to -37		1.5 A	0.01%/V	0.3%	77 dB
LM2941	Adj LDO	5 to 20	0.5V @ 1 A	1 A			0.005%/V
LM2991	Adj LDO	-3 to -24	0.6V @ 1 A	1 A			60 dB
LM340T-xx	Fixed	5,12,15 (2%)		1 A			70–80 dB
LM2940-xx	Fixed LDO	5,12,15	0.5V @ 1 A	1 A	50 mV	50 mV	54–60 dB
LM2990-xx	Fixed LDO	-5,-12,-15	0.6V @ 1 A	1 A	40 mV	40 mV	58 dB

Notes: Adj, adjusted. LDO, low dropout.

### 2.3.2 Switch-Mode Power Supplies

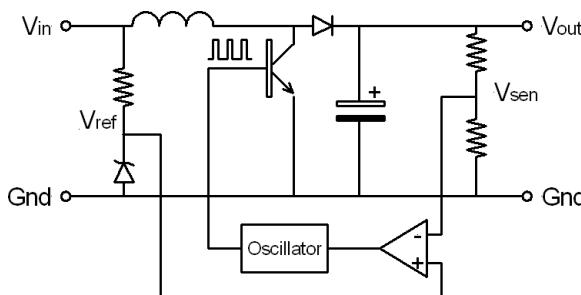
Another method of generating a regulated DC voltage is to use a switching regulator. A transistor operating as a switch periodically applies the unregulated input voltage across an inductor. The inductor's current builds up during the on period, storing energy in its magnetic field, and this energy is transferred to a filter capacitor at the output, where it is smoothed. As with linear regulators, feedback is used to control the output, but in this case it adjusts the oscillator's duty cycle or frequency (Horowitz and Hill, 1989). A generic example of a step-up switching regulator is shown in Figure 2-10.

Switching regulators have some unusual properties that have made them particularly useful in portable applications. Because the control element is either on or off, there is very little power dissipation within the device, which makes the conversion efficiency very good even if there is a large voltage drop between input and output. Switching regulators are also capable of generating output voltages higher than the input as well as outputs with the opposite polarity.

Switching power supplies can be designed with no DC path from the input to the output and are therefore useful for isolation. This allows them to be run directly from the rectified mains with no step-down transformer to produce small, lightweight, and efficient DC power supplies.

The main problems with switch-mode power supplies are residual switching noise on the output and also noise fed back onto the input lines. This requires filtering and results in a decreased overall efficiency and a higher component count.

Root mean square (RMS) output noise of a typical switcher can be 50 mV, which may be acceptable for digital applications but poses a major problem for sensitive analog



**FIGURE 2-10** ■ Schematic diagram of a generic switching regulator.

electronics. Passive LC filtering can reduce this somewhat, but it is often practical to install a linear regulator with good noise rejection at the output. Many DC–DC converters include these considerations and are also housed within a shielded enclosure to minimize noise radiation. In spite of this, circuits including these power supplies must be carefully designed to minimize residual noise problems that may swamp sensitive biological signals.

Switching power supplies are also available with multiple outputs. However, these are generally provided from multiple windings on the same transformer, with feedback regulation applied to only one of the rails. This results in cross-coupling between the sections and poorer regulation on the uncontrolled outputs.

A final consideration is that switching power supplies often have a minimum load current below which their operation can be erratic, or they may not even start. Care should be taken during the design and implementation phases to ensure that this does not occur.

The following advice (Horowitz and Hill, 1989) should be followed when designing power supplies:

- For digital systems, which usually need +5 V at a high current, use a commercial line-powered switcher.
- For analog circuits with low-level signals (small-signal amplifiers with input levels of less than 100  $\mu$ V), use a linear regulator unless efficient operation from a battery supply is required. Switchers will cause problems.
- For high-power applications, use switchers as they are lighter and more efficient than transformer-based power supplies.
- For high-voltage low-power applications (e.g., photomultipliers, defibrillators), use low-power step-up converters.

In general, low-power DC–DC converters are easy to design and require few external components, thanks to the available IC building blocks. However, wherever possible it is preferable to use complete off-the-shelf modules as there are many types ranging from less than 1 W to hundreds of watts, as can be seen from the list in Table 2-2.

### 2.3.3 Batteries

Many of the biomechatronic systems discussed in this book are portable and therefore need to operate using battery power. Where the power consumption is significant and the

**TABLE 2-2** ■ Specifications of a Small Range of Switch-Mode Power Supply Modules

Manufacturer	Model	Power Rating	Input	Output	Efficiency	Noise
C&D Tech.	LME0515SC	250 mW	5 V +/- 10%	15 V	<75%	50 mV pp
C&D Tech.	NME0515SC	1 W	5 V +/- 10%	15 V	<80%	110 mV pp
C&D Tech	NYD2405C	3 W	18–36 V	5 V	<83%	
ASTEC	AA05E-024L-150D	5 W	18–36 V	15 V @ 160 mA –15 V @ 160 mA		
ASTEC	AEE02A24	10 W	18–36 V	5 V @ 2 A	78–85%	
Power One	HAS030YJ-A	25–30 W	18–36 V	15 V @ 2 A		150 mV pp
C&D tech	WPA60R48D0515	60 W	36–75 V	5 V @ 12 A 1.5 V @ 12 A	<90%	90 mV pp
SunPower	SDS-100B05	100 W	18–36 V	5 V @ 20 A	72%	
Mascot	9970 24/12V	276 W	20–30 V	13.8 V @ 20 A	High	

battery pack is integral to the device for reliability and sealing issues or where access to the device is difficult, rechargeable batteries (secondary batteries) are used. Such devices include prosthetic limbs and artificial hearts. In low-power applications where battery life can be months and even years, then it is more practical to use primary batteries.

The oldest form of rechargeable battery still in use is the lead-acid battery. These are *wet cell* devices and mostly need to be kept upright and placed in well-ventilated areas as they generate hydrogen if overcharged. One convenient alternative form is the gel cell, which contains a semisolid electrolyte that prevents spillage.

Most portable rechargeable batteries are *dry cell* types, which are hermetically sealed. These include nickel cadmium, nickel metal hydride, and lithium types. From a practical perspective, the following rules of thumb can be used when selecting batteries for a specific application:

- Nickel cadmium (NiCd): These are good for devices that include motors and other high-discharge requirements. They can accommodate heavy drain rates, but their mAh rating is lower than more modern rechargeables and they also have a strong memory effect.
- Nickel metal hydride (NiMH): These batteries have a high mAh rating and can sustain moderate to high current drain.
- Lithium ion (Li-Ion): These have a very long shelf life and are excellent for moderate to low-power applications.
- Lithium polymer (Li-Po): These have similar chemistry to Li-Ion, but because they are manufactured in flat sheets rather than cylinders they have a higher power density.

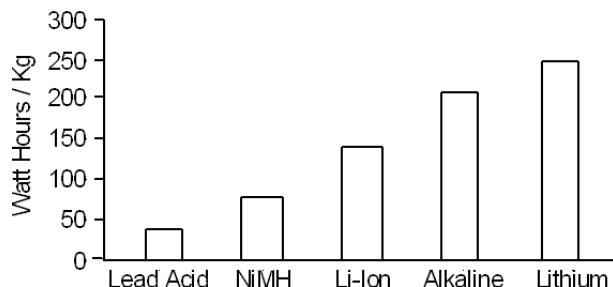
A large number of primary battery types are available for different applications:

- Zinc carbon: This is a low-cost battery good for light current drain applications.
- Zinc chloride: This is similar to the zinc carbon but with a slightly higher power density.
- Alkaline: These are long-life batteries and are suitable for low and high current drain applications. Their energy density is significantly higher than that of zinc carbon types.
- Silver oxide: This type is commonly used for hearing aids and watches where the current drain is low.
- Mercury: This type was formerly used in a wide range of devices but is seldom used today because of toxicity issues.
- Zinc air: These are generally used in hearing aids.

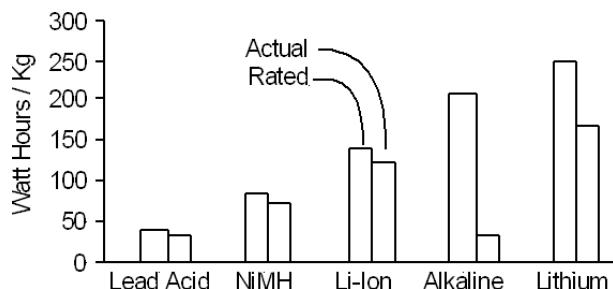
The main advantages of primary batteries are their high energy density, long storage life, and ease of disposal (most primary batteries contain little toxic material). A regular alkaline battery provides 50% more power than a Li-Ion, and a primary lithium battery has three times the energy of a similar sized Li-Ion battery.

Because primary batteries have relatively high internal resistances, the maximum current rating is limited. The energy density ratings shown in Figure 2-11 are determined at the optimum discharge rate for that particular cell type. However, if the comparison is made for a high-current mode application such as a prosthetic limb, then the usable energy for each battery type is as shown in Figure 2-12.

**FIGURE 2-11** ■ Energy density comparison of primary and secondary batteries.  
[Adapted from (Buchmann 2005).]



**FIGURE 2-12** ■ Energy comparison under load.  
[Adapted from (Buchmann 2005).]

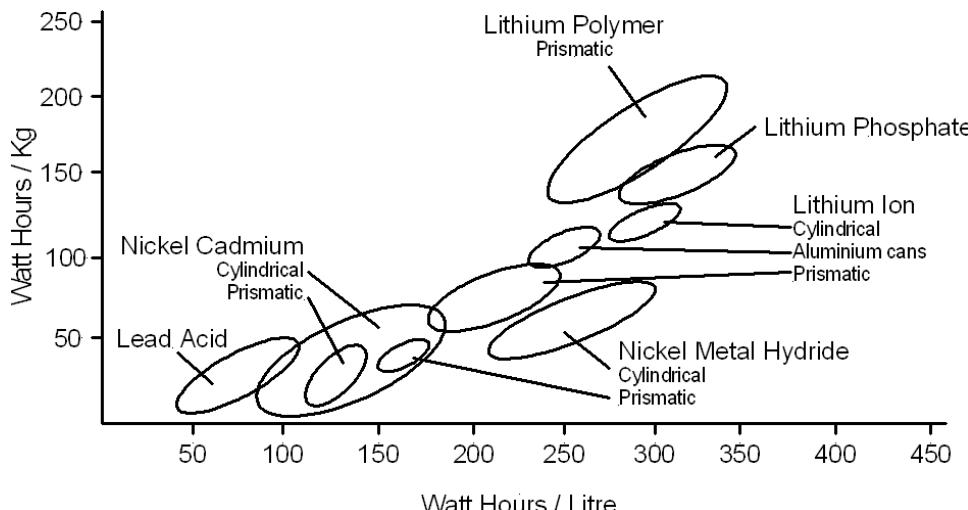


It can be seen that although the alkaline battery is good for low discharge rates its performance in this case is poor. It can also be seen that the performance of the Li-Ion secondary battery is approaching that of the primary lithium type. A comparison of the energy densities of a number of rechargeable batteries is shown in Figure 2-13.

Rechargeable batteries have a very low internal resistance. This allows high current to be drawn on demand, which is essential for many digital devices, and to drive switch-mode power supplies, which have high *inrush* currents.

Their main disadvantages include a limited shelf life and high self-discharge. Whereas a primary battery may have a shelf life of 10 years, lithium-based batteries are good for 2–3 years in normal use, though cool storage at moderate charge prolongs longevity. Nickel-based batteries have a shelf life of up to 5 years but require careful priming to regain performance after long storage.

**FIGURE 2-13** ■ Rechargeable battery energy density.



Nickel-based batteries exhibit a 10–20% self-discharge per month compared with a 5–10% value for lithium- and lead-based batteries. This rate increases at high temperature.

Finally, secondary batteries provide a limited number of charge–discharge cycles. The number that can be achieved is determined by the depth of discharge, the environmental conditions, charge methods, and maintenance. Nickel-based batteries need to be deep discharged periodically to reverse crystal growth (memory effects), while lithium- and lead-based batteries have no memory and can therefore be operated in shallow-cycle applications, with only an occasional deep discharge to verify performance.

### 2.3.4 Energy Scavenging

Energy scavenging from a biomechatronic perspective can include various sources of power including gross body motion, vibration, changes in shape or volume and pressure, temperature gradient, and even fuel cells that oxidize blood glucose. These can range in size from external devices weighing 1 kg or more that provide sufficient energy to power devices like wireless electrocardiography (ECG) monitors to minute implantable devices to monitor medical conditions like hypoglycemia (Campbell, 2010).

#### 2.3.4.1 External Devices

Compared with the amount of power used by the human body, as listed in Table 2-3, the proportion that can be usefully extracted is rather insignificant. A naked human being radiates about 150 W of infrared heat into the surroundings if the ambient temperature is 17 °C (290 K), but with a Carnot efficiency of only 5.5% a maximum of 8 W could be harvested from his complete surface area. Even this is optimistic, as the best thermoelectric generators are only about 1% efficient for temperature gradients of between 5 and 20 °C and it is unlikely that the whole body could be used. In a typical application where the exposed neck comprising 1% of the total surface area of the body is covered by thermoelectric generators with an efficiency of 1%, only about 15 mW of power could be harvested.

**TABLE 2-3 ■ Human Energy Use for a Range of Activities**

Activity	Power (W)
Sleeping	80
Laying quietly	95
Sitting	120
Standing at ease	130
Talking	130
Eating a meal	130
Strolling	165
Driving a car	165
Playing violin or piano	165
Housekeeping	175
Carpentry	270
Hiking (6.5 km/h)	410
Swimming	580
Mountain climbing	700
Long distance running	1050
Sprinting	1630

*Source:* Starner, T. and J. Paradiso, in *Low-Power Electronics*, C. Piguet (Ed.), Boca Raton, FL: CRC Press, pp. 45-1–45-35, 2004.

Capturing all of the work of exhalation based on the differential pressure that can be exerted without interfering with normal breathing is another potential source of power. It is shown in Chapter 9 that the average breathing rate at rest is about 15 breaths per minute for a tidal volume of 400 ml, making the total flow  $Q = 6$  lit/min (0.1 lit/s). For a pressure differential,  $\Delta P$ , of 10 cm H<sub>2</sub>O (980 Pa) the power is

$$\begin{aligned}P &= Q\Delta P \\&= 0.1 \times 10^{-3} \times 980 \\&= 98 \text{ mW}\end{aligned}$$

A face mask housing a small turbine driven generator with a good overall efficiency would be required to capture this energy.

A tight band around the chest can also be used to capture energy from breathing. For normal breathing an increase in circumference,  $\Delta x$ , of about 10 mm occurs during inspiration. For a restraining force,  $F$ , of 100 N the total energy per breath is

$$\begin{aligned}E &= F\Delta x \\&= 100 \times 10 \times 10^{-3} \\&= 1 \text{ J}\end{aligned}$$

For a breathing rate of 15 breaths per minute, or 0.25 breaths per second, the total power output is only 250 mW. Once again, a harvesting efficiency of about 50% for ratchet and flywheel or a piezoelectric material could be expected. The use of internal piezoelectric materials is discussed in the next section.

The power required for arm motion can easily be calculated from its mass and the change in height of the center of gravity per unit time. For example, to lift an arm weighing 3 kg through a height of 0.6 m once per second requires about 18 W. This is obviously unnatural and would be unsustainable over long periods. However, normal arm movement could still provide a reasonable amount of energy without loading the limb or joint significantly.

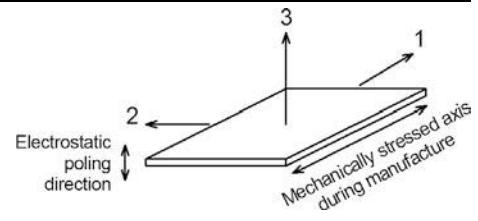
Activities specifically designed to generate electricity such as shake-driven torches or wind-up radios and cell phone chargers can be reasonably efficient. For example, the power supply of the Freeplay radio stores about 500 J of energy from 60 s of winding (Starner and Paradiso, 2004).

Of course, running uses the most energy of any human activity, so capturing even a fraction of that could produce significant amounts of power. A good example of one of the larger devices is a knee brace designed and built by researchers at Simon Fraser University in British Columbia led by Max Donelan. The 1.5 kg device consists of a flexible joint that converts knee flexion into rotary motion that drives a generator. When used at end swing (when the leg is decelerating), it is capable of producing 5 W of electrical energy for a measured expenditure of only 5 W. This should be considered in comparison with the 6 W of energy required to produce each watt of electrical power for a hand-cranked generator.

Unfortunately, just carrying the prototype device, even when it is not generating any energy, uses about 60 W. This is a significant percentage of the 300 W used for walking. The researchers suggest that this overhead could be reduced to about 15 W by moving the device higher up the leg and reducing its mass. When these overheads are taken into account, it is not any more efficient than other techniques such as shoes with power generation mechanisms in the soles or oscillating backpacks (Johnston, 2008).

**TABLE 2-4** ■ Piezoelectric Characteristics of PVDF and PZT

Material Property	PVDF	PZT
Density $\rho$ (g/cm <sup>3</sup> )	1.78	7.6
Relative permittivity $\epsilon$	12	1700
Young's modulus $E$ (N/m <sup>2</sup> )	$3 \times 10^9$	$83 \times 10^9$
Piezoelectric constant (C/N)	$d_{31} = 20 \times 10^{-12}$ $d_{33} = 30 \times 10^{-12}$	$d_{31} = 180 \times 10^{-12}$ $d_{33} = 360 \times 10^{-12}$
Coupling constant (CV/Nm)	0.11	$k_{31} = 0.35$ $k_{33} = 0.69$



Piezoelectric or rotary generators that convert some of the heel-strike energy into electrical power for storage and later use are the commonest energy scavenging devices. Consider a device that can capture the energy over a distance  $x = 30$  mm during heel-strike of a typical 70 kg person. The work per stride per leg is

$$\begin{aligned} W &= mgx \\ &= 70 \times 9.8 \times 30 \times 10^{-3} \\ &= 20.6 \text{ J} \end{aligned}$$

At two strides per second, the total power available is 41 W.

Rotary generators need to spin rapidly to achieve good efficiencies, so pure mechanical coupling involves high gear ratios and a fairly complex mechanism that is prone to failure. Alternatives include miniature hydraulic pump/turbine combinations or pneumatic systems that store power in compressed air.

Piezoelectric materials produce an electrical charge when mechanically stressed. It is interesting to note that human skin and bone both exhibit this property, albeit with very low coupling efficiencies. Common alternatives are polyvinylidene fluoride (PVDF) and lead zirconate titanate (PZT), whose characteristics are shown in Table 2-4.

The coupling constant is the efficiency with which the material converts mechanical energy to electrical, with the subscripts indicating the direction of the interaction in the three axes. For example,  $d_{31}$  is the strain caused to axis-1 by an electrical charge gradient along axis-3.

It is not feasible to extract power by compressing the piezoelectric material as the Young's modulus is too high. However, it is possible to bend the material to take advantage of the 31 mode. In the case of a PZT cantilever beam, the maximum allowed deflection,  $x$  (m), at the tip is determined by the yield stress,  $S = 50$  MPa (Fett, Munz et al., 1999) using the following relationship

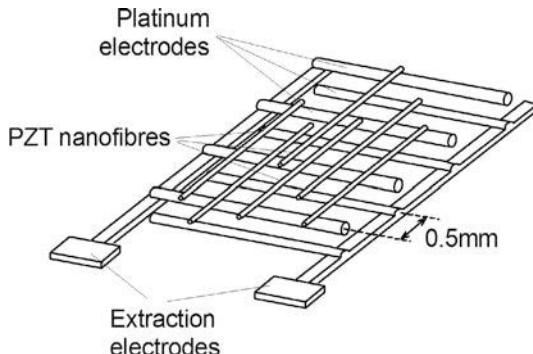
$$x = \frac{2l^2}{3Et}S \quad (2.2)$$

where  $l$  (m) is the length of the beam,  $t$  (m) is the thickness of the material, and  $E$  (N/m<sup>2</sup>) is the Young's modulus for PZT.

Consider mounting a 150 mm cantilever beam made of a 0.5 mm thick slab of PZT within the sole of a shoe. The maximum deflection before the material fractures will be

$$\begin{aligned} x &= \frac{2l^2S}{3Et} \\ &= \frac{2 \times (250 \times 10^{-3})^2 \times 50 \times 10^6}{3 \times 83 \times 10^9 \times 0.5 \times 10^{-3}} \\ &= 50 \text{ mm} \end{aligned}$$

**FIGURE 2-14 ■**  
Nanofiber-based energy scavenging module. [Adapted from (Chen, Shiuou et al., 2010).]



It is therefore feasible to develop the mechanics to deflect the beam by 30 mm while maintaining a reasonable safety margin. This consideration notwithstanding, the superior flexibility of PVDF film makes it a better candidate for power extraction.

Because the total energy output is proportional to the volume of the film stressed, it would appear that thicker films should be used. However, because they generate higher voltages but form smaller capacitors, it is often preferable to use laminates made of many layers of thinner films separated by a compatible material such as polyester.

A 16-layer bimorph PVDF insole developed at the MIT Media Lab produced peak powers of about 15 mW at heel-up into a matched resistive load, with an average of about 1.3 mW. Improved efficiency can be achieved by “tuning” the device so that its electrical resonance is matched to the mechanical excitation frequency (Starner and Paradiso, 2004).

In the last few years, research has focused on using PZT or PVDF nanofibers to produce smaller devices, as shown in Figure 2-14. The ultimate goal is to weave these fibers into normal fabrics so that they can be used to produce power directly from a vest or shirt (Chen, Shiuou et al., 2010).

### 2.3.4.2 Internal Devices

Zhong Lin Wang and his colleagues from the Georgia Institute of Technology have developed a nanogenerator that is able to scavenge power from involuntary movements such as breathing or a beating heart. Zinc oxide nanofibers, which exhibit piezoelectric properties, are bonded onto a  $2 \times 5$  mm flexible polymer substrate before the complete unit is coated in silicone to protect it from corrosion. When bonded to a rat’s diaphragm using a tissue adhesive, the device generated a current of 4 pA at 2 mV. The output increased to 30 pA and 3 mV when a similar device was attached to a rat’s heart.

Other researchers have implanted larger piezoceramic devices into muscles to obtain power from normal body movement (Campbell, 2010).

## 2.4 | SENSORS AND TRANSDUCERS

### 2.4.1 Resistive Displacement Sensors

#### 2.4.1.1 Strain Gauges

Small displacements are generally measured using a strain gauge bridge configuration. A strain gauge consists of a long narrow metal conductor such as a piece of metal foil (usually constantan) mounted on a polyimide film or fine gauge wire stretched over a frame. If it

is stretched within its elastic limits, it will increase in length,  $L$  (m), and decrease in cross sectional area,  $A$  ( $\text{m}^2$ ). Because the resistance between the two ends of this foil or wire can be given by

$$R = \rho \frac{L}{A} \quad (2.3)$$

where  $\rho$  is the electrical resistivity of the material, this stretching will result in an increase in resistance. Because the change in length can be only small for the foil or wire to remain within its elastic limit, the resultant change in resistance will also be small.

To determine how this resistance changes under deformation, the derivative of equation (2.3) must be obtained (Alciatore and Histand, 2003). First take the natural log of both sides

$$\ln R = \ln \rho + \ln L - \ln A. \quad (2.4)$$

Taking the derivative yields the following expression for the change in resistance as a function of the changes in the geometry of the conductor and the material property

$$\frac{dR}{R} = \frac{d\rho}{\rho} + \frac{dL}{L} - \frac{dA}{A} \quad (2.5)$$

Since the cross sectional area of the conductor is the product of the width,  $w$  (m), and the height,  $h$  (m), then the area derivative is

$$\frac{dA}{A} = \frac{w \cdot dh + h \cdot dw}{wh} = \frac{dh}{h} + \frac{dw}{w} \quad (2.6)$$

Poisson's ratio,  $v$ , is defined as the ratio of the transverse  $\varepsilon_{transverse}$ , and axial  $\varepsilon_{axial}$ , strain for a cylinder with diameter,  $D$  (m), and length,  $L$  (m)

$$v = \frac{\varepsilon_{transverse}}{\varepsilon_{axial}} = -\frac{\Delta D/D}{\Delta L/L} \quad (2.7)$$

Therefore

$$\frac{dh}{h} = -v \frac{dL}{L} \quad \text{and} \quad \frac{dw}{w} = -v \frac{dL}{L} \quad (2.8)$$

which makes

$$\frac{dA}{A} = -2v \frac{dL}{L} = -2v \varepsilon_{axial} \quad (2.9)$$

Substituting equation (2.9) back into equation (2.5)

$$\frac{dR}{R} = \varepsilon_{axial}(1 + 2v) + \frac{d\rho}{\rho} \quad (2.10)$$

Dividing through by  $\varepsilon_{axial}$  gives

$$\frac{dR}{R} \frac{1}{\varepsilon_{axial}} = 1 + 2v + \frac{d\rho}{\rho} \frac{1}{\varepsilon_{axial}} \quad (2.11)$$

Note that the first two terms on the right-hand side (RHS), 1 and  $2v$ , represent the change in resistance due to the increased length and decreased diameter of the conductor.

The last term represents the piezoresistive effect of the material, that is, the change in resistivity of the material with strain.

The relative sensitivity of such a device is given by the gauge factor,  $\gamma$ , which is defined as

$$\gamma = \frac{\Delta R/R}{\varepsilon_{axial}} \quad (2.12)$$

where  $\Delta R$  is the change in resistance when the structure is stretched by an amount  $\Delta L$ . This factor is usually specified on the data sheet of the strain gauge, and it can be used to determine the axial strain directly from the change in resistance.

$$\varepsilon_{axial} = \frac{\Delta R/R}{\gamma} \quad (2.13)$$

Typical strain gauges have  $R \approx 120 \Omega$  and  $\gamma \approx 2$ .

### WORKED EXAMPLE

---

#### Strain Gauge

If a  $120 \Omega$  strain gauge with a gauge factor of 2.0 measures a strain of  $50 \mu\varepsilon (50 \times 10^{-6})$  (read as “50 microstrain”), what is the change in resistance from the unloaded to the loaded state?

From equation (2.13)

$$\begin{aligned}\Delta R &= R\varepsilon_{axial}\gamma \\ &= 120 \times 50 \times 10^{-6} \times 2 \\ &= 0.012 \Omega\end{aligned}$$


---

#### 2.4.1.2 Measuring with a Wheatstone Bridge

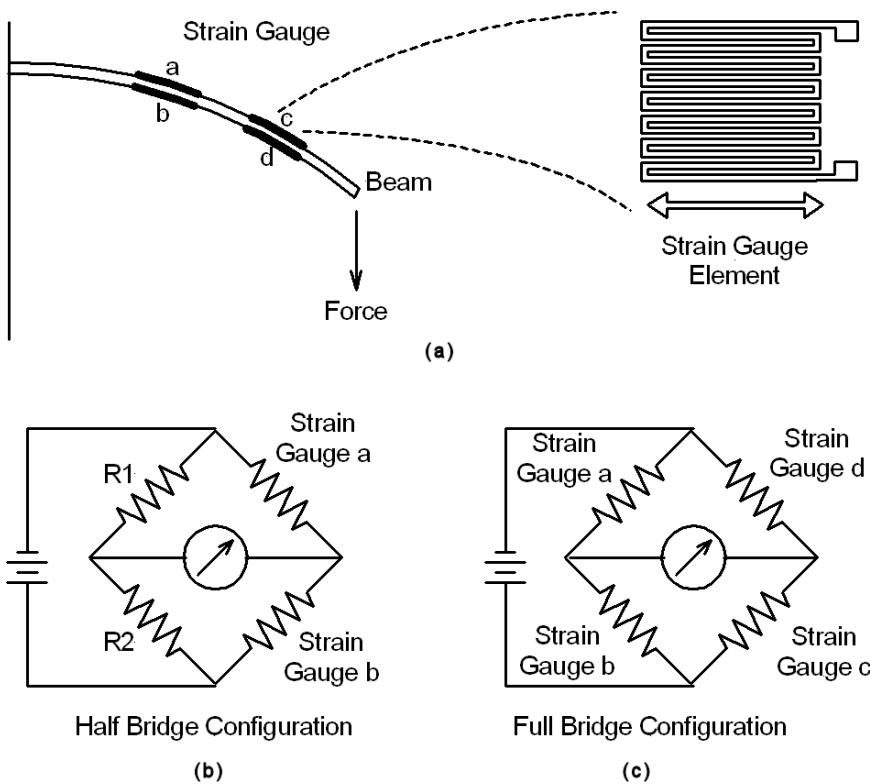
Because the change in resistance is small, measurement can be quite challenging, and a Wheatstone bridge configuration shown in Figure 2-14 is generally used. It should be noted that temperature changes can result in resistance changes of the same order as those caused by strain, so temperature compensation is essential if good accuracy is required.

A typical example of strain gauge based displacement measurement is shown in Figure 2-15. In this application, the strain gauges are attached to the upper and lower faces of a cantilever beam. When the beam deflection is downward, the length of the strain gauges on the upper surface increases, and the length on the lower surface decreases.

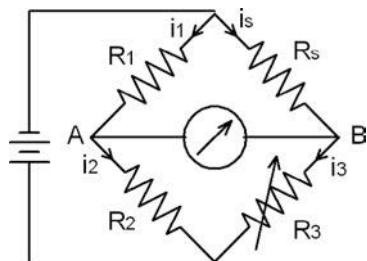
It is possible to use a single strain gauge element and three fixed resistors to make up the bridge, but it is more common to use two or even four elements that provide increased sensitivity and also compensate for temperature variations (if all the elements on the beam remain at the same temperature).

Two different measurement modes are possible; they are the static balanced mode and the dynamic unbalanced mode (Alciatore and Hinstand, 2003).

In the static balanced mode,  $R_1$  and  $R_2$ , are precision resistors, while  $R_3$  is a precision potentiometer with an accurate scale displaying its resistance scale.  $R_s$  is the resistance of the strain gauge in the circuit as seen in Figure 2-16.



**FIGURE 2-15** ■ Strain gauges used to measure the deflection of a cantilever beam.  
 (a) Mounting of strain gauge on cantilever.  
 (b) Circuit for half bridge configuration.  
 (c) Circuit for full bridge configuration.



**FIGURE 2-16** ■ Static balanced bridge circuit.

To balance the bridge, the potentiometer is adjusted until the voltage between nodes A and B is zero. The voltage at A must therefore equal the voltage at B and

$$i_s R_s = i_1 R_1 \quad (2.14)$$

Assuming that a high impedance voltmeter that draws no current is used to measure the voltage, then

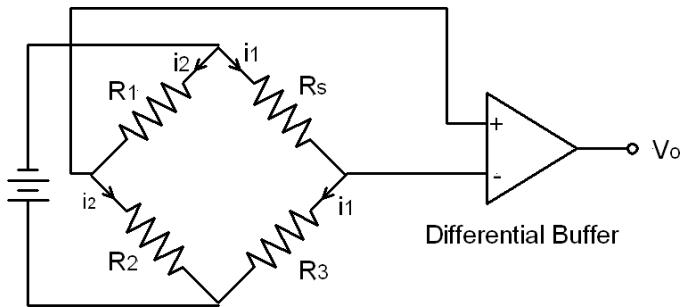
$$i_s = i_3 = \frac{V}{R_s + R_3} \quad (2.15)$$

and

$$i_1 = i_2 = \frac{V}{R_1 + R_2} \quad (2.16)$$

where  $V$  is the DC excitation voltage applied to the bridge.

**FIGURE 2-17** ■ Dynamic unbalanced bridge circuit.



Substituting into equation (2.14) and simplifying gives

$$\frac{R_s}{R_3} = \frac{R_1}{R_2} \quad (2.17)$$

If  $R_1$  and  $R_2$  are known and  $R_3$  can be determined accurately from the angular displacement of the potentiometer, then  $R_s$  can be determined

$$R_s = \frac{R_1 R_3}{R_2} \quad (2.18)$$

Note that this result is independent of the excitation voltage.

In the dynamic deflection mode shown in Figure 2-17, the bridge is first balanced under no-load conditions, and then the measured output voltage can be used to determine the strain gauge resistance as a load is applied.

The output voltage can be written in terms of the volt drop from the +ve terminal and also in terms of the volt drop from the negative terminal

$$V_o = i_1 R_s - i_2 R_1 = i_2 R_2 - i_1 R_3 \quad (2.19)$$

and the applied excitation voltage can also be written in terms of the current and resistance in each arm

$$V = i_1(R_s + R_3) = i_2(R_1 + R_2) \quad (2.20)$$

Solving for  $i_1$  and  $i_2$  in terms of  $V$  and substituting into equation (2.19) gives

$$V_o = V \left[ \frac{R_s}{R_s + R_3} - \frac{R_1}{R_1 + R_2} \right] \quad (2.21)$$

If the bridge is balanced under no-load conditions so that  $V_o = 0$  in this case, then as the strain gauge is loaded the voltage change  $\Delta V_o$  will be related to the new resistance  $R_s + \Delta R_s$  according to

$$\frac{\Delta V_o}{V} = \left[ \frac{R_s + \Delta R_s}{R_s + \Delta R_s + R_3} - \frac{R_1}{R_1 + R_2} \right] \quad (2.22)$$

This equation can be rearranged to give the relationship between the relative change in resistance and the measured output voltage.

Let

$$A = \left[ \frac{\Delta V_o}{V} + \frac{R_1}{R_1 + R_2} \right] \quad (2.23)$$

Then by simplifying equation (2.22) it can easily be shown that

$$\frac{\Delta R_s}{R_s} = \frac{AR_3}{(1-A)R_s} - 1 \quad (2.24)$$

Therefore, by measuring the change in output voltage  $\Delta V_o$  the change in strain gauge resistance  $\Delta R_s$  can be determined using equations (2.23) and (2.24). Finally, the actual strain is determined using equation (2.13).

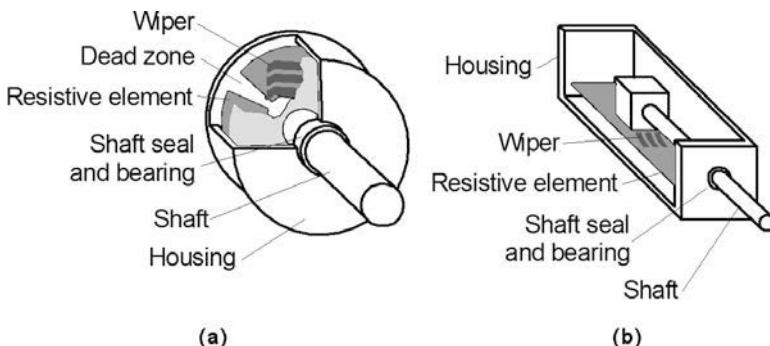
A similar analysis can be performed for the bridge with two or four strain gauge elements either mounted on the flexible beam as shown in Figure 2-14 or as dummy elements to be maintained at the same temperature as the active element.

A more compliant structure that has found applications in biomedical instrumentation is the liquid metal strain gauge. Instead of using a solid electrical conductor, a mercury-filled compliant silicone rubber tube is used. As the tube is stretched, the length increases and the cross sectional area decreases as with a conventional strain gauge (Bronzino, 2006).

#### 2.4.1.3 Potentiometers

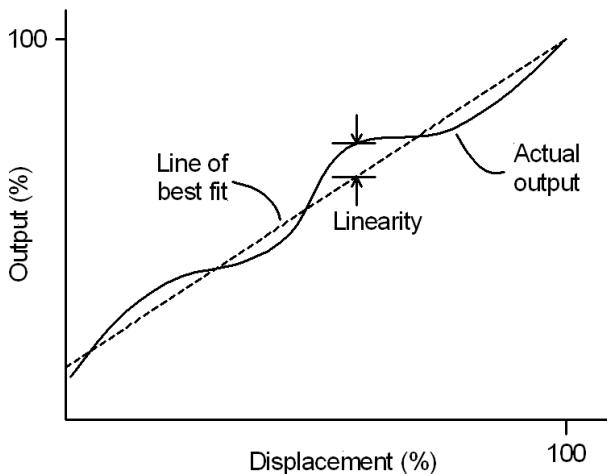
Larger displacements can be measured using precision potentiometers. These are available in rotary, linear-motion, and string pot forms. String pots are also known as cable pots, yo-yo pots, or draw-wire transducers and measure the extended length of a spring-loaded cable. Rotary pots are single or multiturn (commonly 3, 5, or 10 turns), whereas linear-motion pots are available with maximum strokes ranging from 5 mm to over 4 m (Webster, 1999). Schematic cutaways showing how rotary and linear pots operate are shown in Figure 2-18.

Potentiometers may be characterized as wire-wound or non-wire-wound. The former are made from tightly wound resistive wire that quantizes measurements, whereas the latter are made from a sheet of resistive material that is, in theory, capable of unlimited resolution. Wire-wound pots offer good temperature stability and high power handling capability but have a limited life. However, as most measurement applications are low power, potentiometers made from conductive plastic film are the components of choice. They are low friction and low noise and have a long operational life. Unfortunately, they are sensitive to temperature and other environmental factors.



**FIGURE 2-18 ■**  
Schematic cutaways of two common potentiometer types.  
(a) Rotary-motion.  
(b) linear-motion.  
[Adapted from (Webster 1999).]

**FIGURE 2-19** ■ Definition of potentiometer linearity.



For displacement measurements, most potentiometers have a linear taper in that the output varies linearly with wiper motion. The linearity is defined as the maximum deviation of the output function from a straight line, as illustrated in Figure 2-19. With laser trimming, overall linearities of better than 0.1% of full scale output are routinely achieved.

To avoid electrical loading, the pot wiper should drive into a high impedance. This is generally achieved by making the potentiometer two arms of a Wheatstone bridge and measuring the difference voltage using an operational amplifier with a high input impedance, as shown in the circuit in Figure 2-16.

From a mechanical perspective, potentiometers add inertia and friction to the moving parts of the system they are measuring. As a result, they increase the force required to move these parts. Rotary pot manufacturers commonly list the equivalent mass moment of inertia of the rotating part, the dynamic torque required to maintain rotation at a constant angular rate, and the starting torque required to initiate motion.

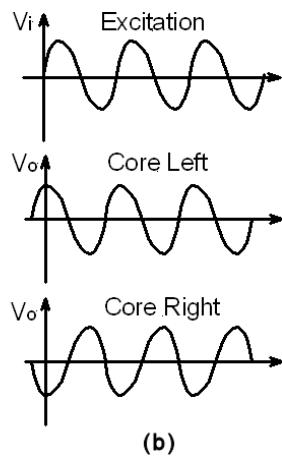
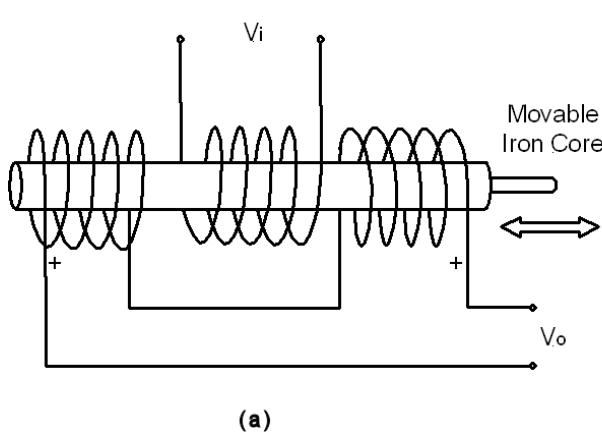
Despite constant mechanical wear, a conductive plastic potentiometer may last hundreds of millions of cycles if the conditions are good and if abrasive dirt is kept off the film. Wire-wound pots do not usually last as long, but they can be expected to survive over 1 million cycles.

## 2.4.2 Inductive Displacement Sensors

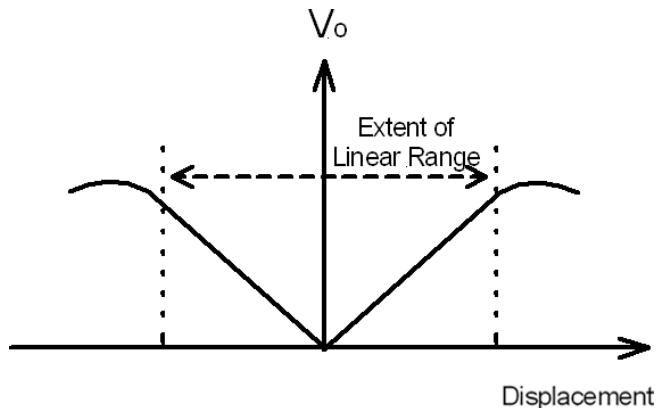
Mutual inductance between two coils can be used to provide a sample displacement sensor. If one coil is driven by an AC signal, then the induced voltage in the other will be proportional to the distance between coils. A variation of this would be to have two fixed coils coupled together by a movable core in which the position of the core controlled the mutual inductance, and hence the induced voltage (Bronzino, 2006).

### 2.4.2.1 Linear Variable Differential Transformer

Probably the most common inductive device for measuring linear displacement is the linear variable differential transformer (LVDT). It consists of a primary and two secondary cores mounted axially over a movable iron core as shown in Figure 2-20. The two secondary coils are generally connected in series-opposing configuration, and their combined output describes the magnitude and direction of the core motion when the primary coil is excited by an AC signal.



**FIGURE 2-20** ■ Linear variable differential transformer. (a) Transformer winding configuration. (b) Excitation and output waveforms.

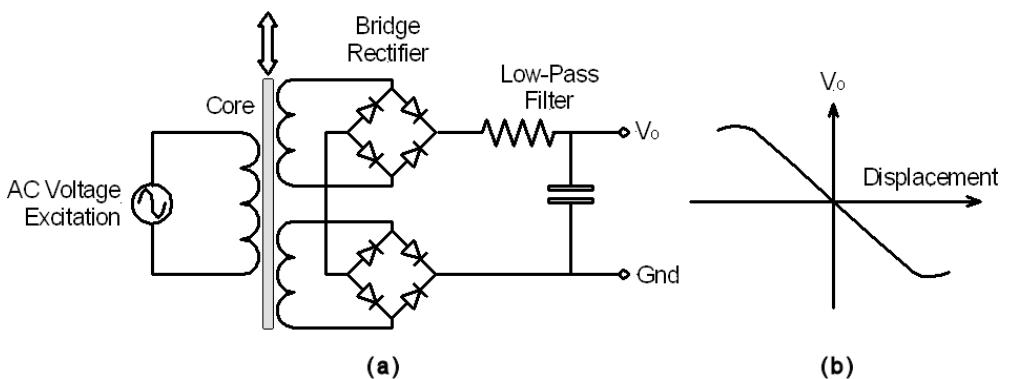


**FIGURE 2-21** ■ RMS voltage output from a LVDT as a function of core displacement.

When the core is positioned midpoint between the symmetrical windings, each of the secondary windings provides signals with the same amplitude but is phase shifted by  $180^\circ$  with the result that the output voltage,  $V_o$ , is zero. As the core is displaced away from the balance point, the induced voltage in one of the secondary windings will increase, and that in the other will decrease, with the result that the magnitude of the output signal will grow in a fairly linear fashion as shown in Figure 2-21.

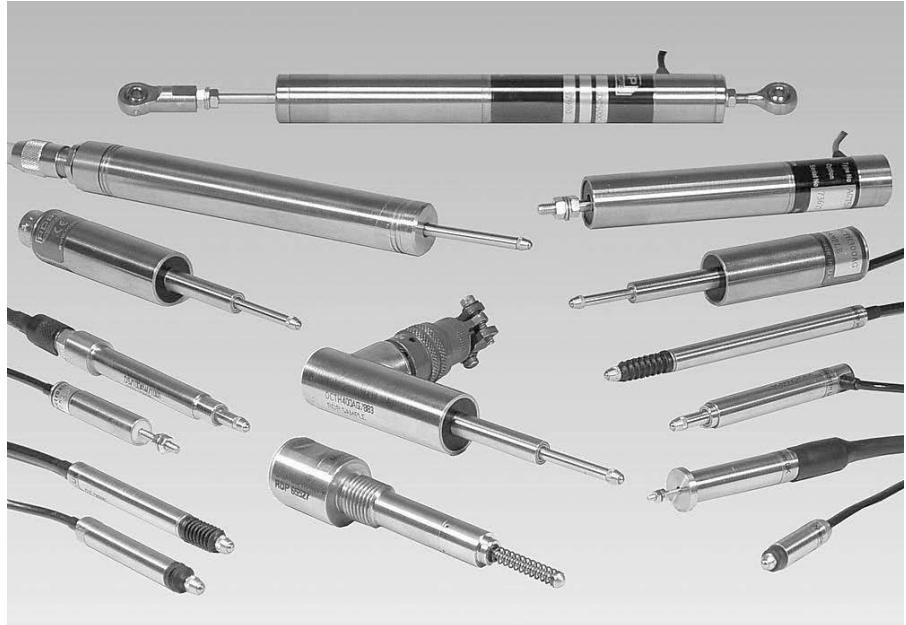
Determining the direction of core displacement requires that the phase of the output signal be taken into account. This can be found using a phase-sensitive detector, a synchronous detector, or the demodulation scheme shown in Figure 2-22. The bridge rectifiers produce a positive or negative rectified sine wave depending on which side of the null the core is located, and a low-pass filter then smooths the output to provide a DC voltage proportional to the displacement.

Commercial LVDTs are available with different diameters, lengths, and strokes. They mostly include internal electronics to provide the AC signal and perform the demodulation so that the output is a DC voltage proportional to the displacement. Good accuracy over the linear range is provided, and the analog output generally does not require amplification. LVDT nonlinearities are typically 0.25% of full scale with precision units being as low as 0.05%. The units shown in Figure 2-23 manufactured by RDP Electronics in the United Kingdom have a specified accuracy of 0.1%, infinite resolution, and strokes from 0.5 up to 470 mm.



**FIGURE 2-22** ■ Demodulation of LVDT output. (a) Circuit diagram. (b) Relationship between displacement and output voltage.

**FIGURE 2-23** ■  
Photograph of a range of LVDTs.  
(Courtesy of RDP Electronics <http://www.rdp.com/>.)



LVDTs are also less sensitive to variations in temperature than the other linear encoders discussed so far, but their main disadvantage is a limited frequency response. This is determined by the inertial effects associated with the core's mass, the choice of excitation frequency, and the bandwidth of the low-pass filter at the output (Alciatore and Histand, 2003).

Measurement of angles using the same principle is by means of a rotary variable differential transformer (RVDT). In this case a rotary ferromagnetic core is used with linear ranges of  $+/-40^\circ$  possible, with a nonlinearity error of about 1% (Fraden, 1996).

#### 2.4.2.2 Inductosyns and Resolvers

Other inductive angle measurement systems include resolvers and inductosyns. Resolvers resemble small motors and are essentially rotary transformers designed so that the

coefficient of coupling between rotor and stator varies with the shaft angle. Fixed windings are placed on a laminated iron stack to form the stator, and movable windings are placed on a similar structure to form the rotor. Usually, resolvers have a pair of windings on a rotor and a second pair on the stator, positioned at right angles to each other. When a rotor winding is excited with an ac reference signal, stator windings produce AC voltage outputs that vary in amplitude according to the sine and cosine of shaft position.

Connection to the rotor is made either by brushes and slip rings or inductive coupling. Resolvers using the inductive method are referred to as brushless types. The inductive (brushless) resolvers offer up to 10 times the life of brush types and are insensitive to vibration and dirt; therefore, they are used in the majority of industrial applications.

The stator signals from a resolver are routed to a specialized type of analog-to-digital converter system known as a resolver-to-digital converter (RDC).

The inductosyn can be a linear or rotary motion measurement device consisting of two parts: a fixed serpentine winding with a small pitch, typically about 2 mm and a movable winding known as a *rotor* or *slider*, depending on whether rotary or linear motion is being measured. The rotor or slider has a pair of windings having the same pitch as the fixed winding that are mutually offset by a quarter pitch so both sine and cosine waves are produced by movement. One slider winding is adequate for counting pulses but provides no direction information. The two-phase windings provide direction information in the phasing of the sine and cosine waves. Movement by one pitch produces a cycle of sine and cosine waves while multiple pitches produce a train of waves.

### 2.4.3 Magnetic Displacement Sensors

Magnetic field strength is most commonly measured using the Hall effect and magnetoresistance (MR).

In the Hall effect, a voltage,  $V_H$ , appears across a conductor when a magnetic field is applied at right angles to the current flow. The direction is perpendicular to both the current and the magnetic field. The magnitude of the Hall voltage,  $V_H$  (V), is proportional to the magnetic flux density,  $\beta$  (Weber/m<sup>2</sup>) and the current,  $I$  (A)

$$V_H = \frac{K_H \beta I}{z} \quad (2.25)$$

where  $K_H$  is the Hall constant, and  $z$  (m) is the thickness of the conductor (Webster, 1999).

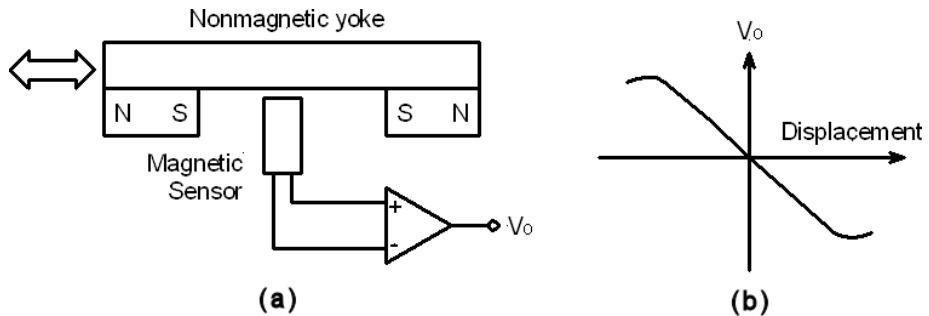
MR is a property of most magnetic materials. A decrease in electrical resistance occurs when a magnetic field is applied perpendicular to the direction of current flow. The resistance decreases as the magnetic flux density increases until the material reaches magnetic saturation. The total change in resistance is about 0.3% for iron and 2% for nickel. Giant magnetoresistance (GMR) was discovered in 1988. It is the phenomenon where the resistance of certain layered materials drops dramatically as a magnetic field is applied. It is described as giant since it is a much larger effect than had ever been previously seen in metals. With GMR, changes in resistance of up to 10% have been measured.

As with other sensors that use changes in resistance, MR and GMR sensors are often incorporated into a Wheatstone bridge configuration to reduce temperature dependence and to increase sensitivity. In these configurations, two of the elements are shielded from the applied magnetic field, and the opposite pair are exposed to it. Commercial sensors are available with outputs that vary by more than 5% of the applied voltage.

**FIGURE 2-24 ■**

Magnetic linear position sensor.

- (a) Schematic diagram of hardware.  
 (b) Relationship between displacement and output voltage.



For short distances, linear displacement measurement can be achieved using a pair of opposing magnets and a single magnetic sensor, as illustrated in Figure 2-24. This configuration partially solves two of the more annoying problems commonly encountered in magnetic distance sensing schemes: (1) lack of linearity; and (2) lack of a well-defined reference point. The pair of opposing magnets mounted on a movable, nonmagnetic yoke creates a magnetic null point exactly halfway between the two magnets. This is convenient as a stable reference point. Another benefit offered by opposed magnets is that the sensed field versus displacement is a nearly linear function over a significant range of travel.

A rotary sensor can be made using the same principles, but with a plastic yoke formed into an annulus, as shown in Figure 2-25. In this case, when the magnets are rotated around the sensor, it sees a sinusoidally varying field with good linearity available over a span of  $\pm 30^\circ$  of rotation. Multiple sensors and linearization processes can be used to extend this technique to operate over the full  $360^\circ$ .

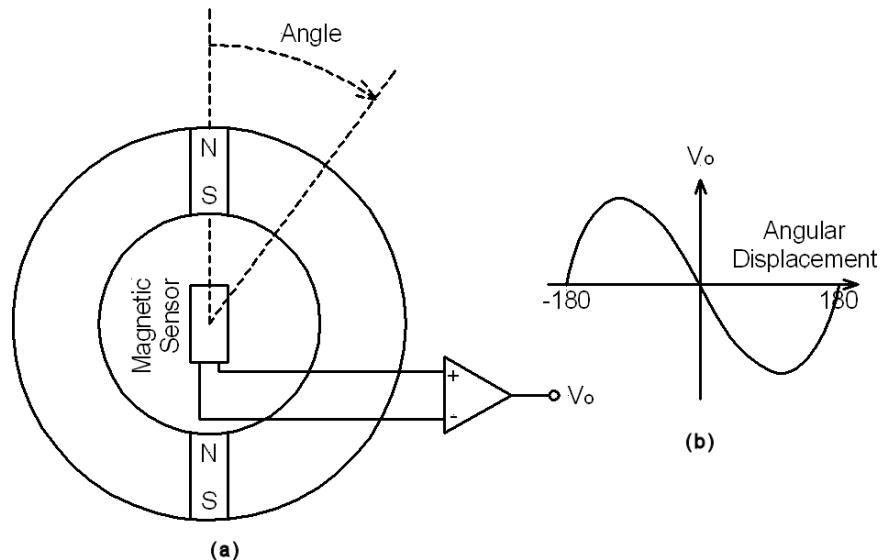
Most magnetic linear or angular displacement encoders use magnets with multiple poles to extend the maximum displacement that can be measured or to improve the resolution of angular encoders. The principles involved are similar to those used by the more common optical encoders described later in this chapter.

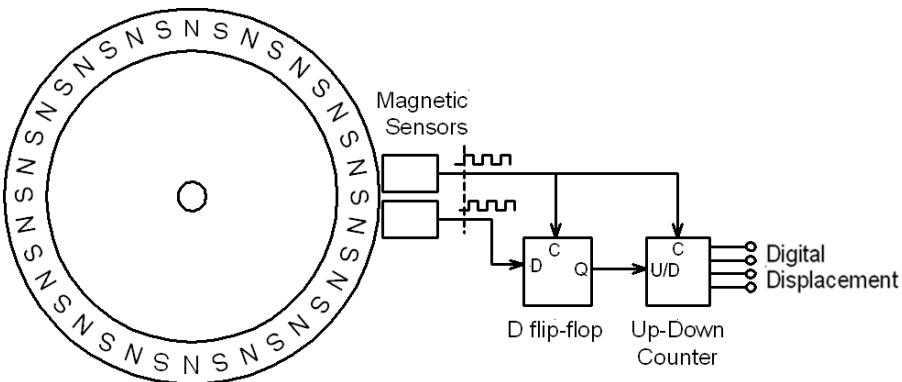
Encoders can be absolute or incremental. At each measurement position, absolute encoders include a binary word made up of a number of magnetic elements that encode

**FIGURE 2-25 ■**

Magnetic rotary position sensor.

- (a) Schematic diagram of hardware.  
 (b) Relationship between angular displacement and output voltage.





**FIGURE 2-26 ■**  
Magnetic  
incremental encoder  
for rotation.

for the complete displacement or angle, whereas incremental encoders include only a single bit and rely on the associated electronics to maintain a count from some datum point. Incremental encoders are simpler and cheaper, but if power is lost then position information is lost and it is necessary to return to the datum point.

To increase accuracy and to enable incremental encoders to determine the direction of travel requires that two magnetic sensors are used. The one is displaced from the other along the axis of travel by  $90^\circ$  so that the outputs are in quadrature. This allows a simple D type flip-flop to determine the direction of travel and to control an up-down counter to produce a quantized measure of the displacement, as shown in Figure 2-26.

By “unwinding” the ring magnet used in the previous application, a device that can measure linear motion is produced. While a rigid rod magnet could be used to provide the required alternating pole pattern, flexible magnetic materials have made this sensor much easier to implement, especially for long linear runs. For low-resolution applications where the sensor can be very close to the magnetic strip, flexible ferrite materials offer good resolution. These materials can be purchased in rolls of varying thickness and width, often with pressure-sensitive adhesive backing for easy installation. For more demanding applications, either in terms of number of poles per mm or working distance from sensor to magnet strip, rare-earth materials (neodymium-iron-boron or samarium-cobalt) mixed with plastic binders can be used.

An example of a linear incremental encoder is the LM10 manufactured by Renishaw, shown in Figure 2-27. This is a solid-state, noncontact encoder featuring a compact IP68 read head that rides between 0.1 and 1.5 mm above a self-adhesive magnetic strip. Devices



**FIGURE 2-27 ■**  
Photograph of the  
LM10 linear  
magnetic encoder.  
(Courtesy of  
Renishaw  
<http://www.renishaw.com/>.)

can operate with a maximum velocity of 25 m/s for sinusoidal outputs, whereas digital models provide resolutions of 100, 50, and 10  $\mu\text{m}$  at 25 m/s, improving to 5  $\mu\text{m}$  at 20 m/s, and 1  $\mu\text{m}$  at 4 m/s. Resistant to shock, vibration, and pressure, they operate over a temperature range of -20 to 85 °C.

#### 2.4.4 Capacitive Displacement Sensors

The capacitance of a parallel plate capacitor is directly proportional to the area of the two plates and inversely proportional to the distance between them. Capacitance is also directly proportional to the dielectric constant of the material between the plates. A simple, albeit very short range sensor could be made by having one fixed plate and one plate that moves. However, more sophisticated capacitance-based sensors use a pair of coaxial capacitors with a movable dielectric sleeve that fits between the outer and inner electrode of the one capacitor. The proportion of the area between the two electrodes that is filled with dielectric determines the capacitance (Fraden, 1996).

These sensors are ratiometric because they measure the ratio of two capacitors maintained under the same conditions of temperature and humidity to minimize drift. This ratio can be determined using an AC equivalent of the Wheatstone bridge described earlier or by incorporating the capacitors into a pair of radio frequency (RF) LC oscillators and measuring the frequency difference between the two.

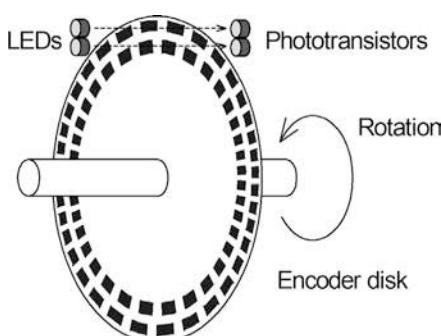
#### 2.4.5 Optical Displacement Sensors

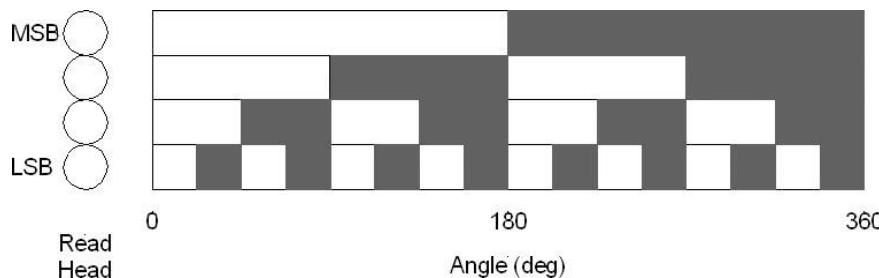
##### 2.4.5.1 Digital Optical Encoders

Digital optical encoders are probably the most common of all the sensors used for angle and angular rate measurement. They consist of a light source and a number of photo transistors separated by a rotating mask made from alternating opaque and transparent radial bands. As with magnetic encoders, rotary optical encoders come in two types: (1) absolute encoders, in which a unique digital word corresponds to every rotational position (within the quantization level); and (2) incremental encoders, which produce a sequence of digital pulses as the shaft rotates, allowing measurement of the relative displacement of the shaft.

Rotary encoder discs can be made from glass or plastic onto which has been deposited a radial pattern organized into tracks or, for more robust applications, a metal disk through which holes have been cut. Astride this disk is placed the light-emitting diode (LED) source on one side and a number carefully aligned photo transistors on the other. As the disk rotates, it will periodically interrupt the beam between the emitter-receiver pair to produce a change in the digital level as shown in Figure 2-28.

**FIGURE 2-28** ■  
Schematic diagram  
of a generic rotary  
optical encoder.





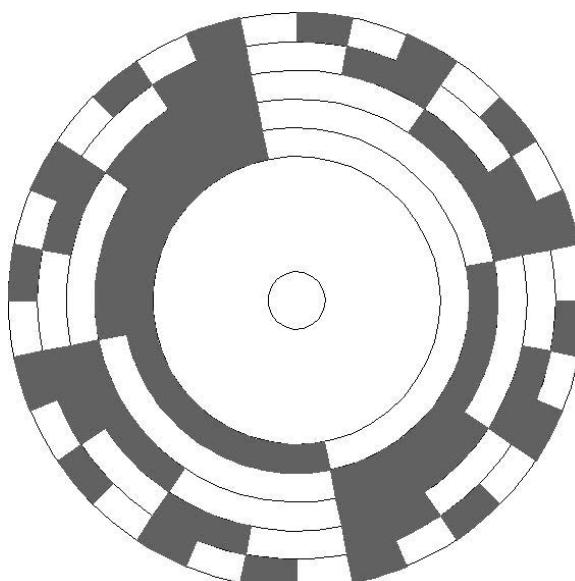
**FIGURE 2-29 ■**  
Encoder pattern for a four-bit binary disc.

The optical disk of an absolute encoder is designed to produce a digital word that distinguishes  $N$  distinct positions on the shaft. For example, if there are eight tracks, then the encoder is capable of measuring  $2^8 = 256$  distinct angles, each corresponding to an angular resolution of  $360^\circ/256 = 1.406^\circ$ .

Most absolute encoder discs are encoded with either binary or Gray codes. Figure 2-29 shows the encoder pattern for a four-bit binary disc, which is the obvious method to encode for angle. Unfortunately, if the level transitions are not simultaneous, large errors will occur in the digital output. For example, in the transition from  $360^\circ$  to  $0^\circ$ , all four bits must change simultaneously; otherwise, any output between 1 and 14 could occur for a brief period during the transition.

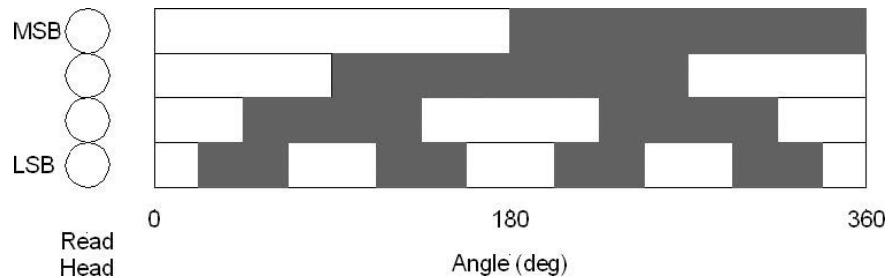
The actual encoder disk will obviously be circular, with the inner rings that encode for the more significant bits being smaller, as can be seen from Figure 2-30. However, because the transitions are aligned along the radial and the individual read heads are smaller than the size of a bit on the innermost ring, the readout should be accurate. However, as the number of bits increases, the size of the outermost track becomes smaller and the devices become more expensive and less robust.

The Gray code is designed so that only one bit changes state for each count transition; therefore, the largest error can be a single count as can be seen if the transitions shown in Figure 2-31 are considered.



**FIGURE 2-30 ■**  
Encoder disk pattern for a five-bit binary encoder.

**FIGURE 2-31 ■**  
Encoder pattern for a four-bit Gray code disc.



Digital electronics uses the binary system; therefore, to be useful the Gray code outputs of the encoder need to be converted to binary. This is easy to achieve using a number of exclusive-OR gates,  $\oplus$ , because the Boolean expression that relates binary bits to Gray code bits are as follows for a four-bit word:

$$\begin{aligned} B_3 &= G_3 \\ B_2 &= B_3 \oplus G_2 \\ B_1 &= B_2 \oplus G_1 \\ B_0 &= B_1 \oplus G_0 \end{aligned}$$

This sequence can be extended to any number of bits starting with the premise that  $B_n = G_n$ .

As the number of bits increases in an absolute encoder, the number of tracks increases. For example, a 12-bit encoder would have 12 individual tracks. Because each position has a unique signal, the absolute encoder will not lose positional information if power is interrupted.

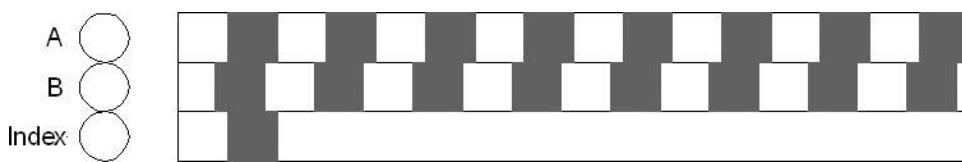
A common application for absolute encoders is on a prosthetic arm, where accurate positional information is paramount to the operation's accuracy and safety. Located at the joints or pivot points of an articulated prosthesis, the absolute encoder can monitor exact angular position, direction, and speed of arm travel. Another critical application for absolute encoders is in aircraft, where they can be mounted to the plane's rudder and aileron drive shafts to provide exact positional and absolute data during flight.

Absolute multturn encoders incorporate standard absolute technology but rely on an additional internal counting process to monitor and track the number of rotations. Some multturn encoders incorporate a gear-driven system that can be quite accurate but that is complex and expensive to manufacture and prone to breakage and wear. Alternative noncontact multturns offer significantly increased encoder life by using a two-pole magnet and an array of reed switches to monitor revolutions and directional information.

Like single-turn absolute encoders, multturns have a unique code for each position within  $360^\circ$  of rotation, dependent on the encoder's resolution but providing unique codes for each revolution. As a result of these unique codes, an absolute multturn encoder will not lose its revolution count or angular position if power fails.

The newest absolute multturn encoders are available with 36-bit resolution—18 bits over  $360^\circ$  and another 18 bits for counting revolutions. With this resolution, multturns can offer 262,144 positions over 262,144 revolutions, allowing the encoder to track 68,719,476,736 unique angular positions.

The simplest encoder technology available, the incremental encoder, is designed to monitor distance, speed, and direction. It features a pulse disk that consists of a double



**FIGURE 2-32** ■  
Encoder pattern for an incremental encoder.

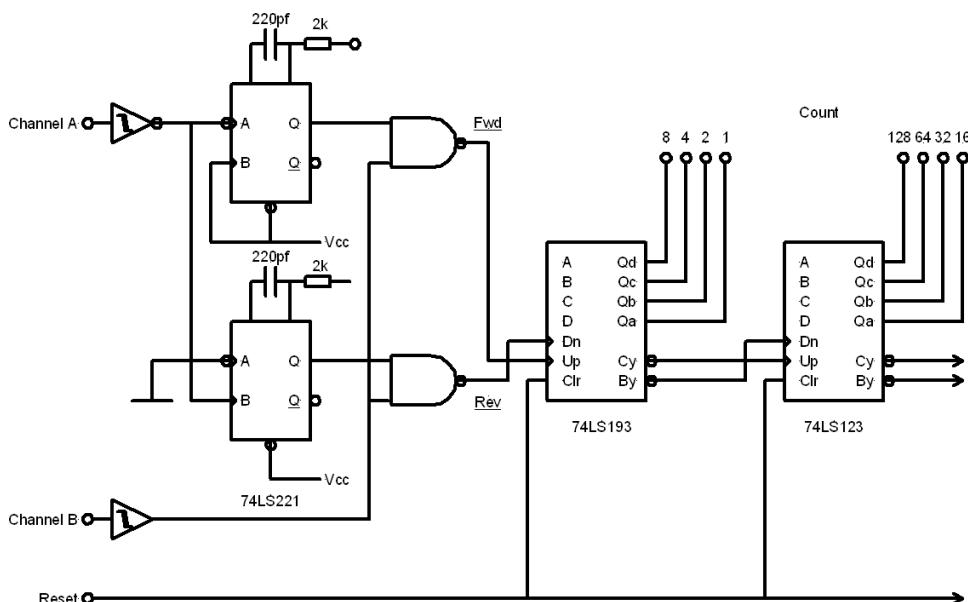
track of evenly spaced clear and opaque segments displaced by  $90^\circ$  electrical, as shown in Figure 2-32. Defined as pulses per revolution (ppr), resolution is dictated by the number of clear and opaque lines on the disk. An incremental encoder with a resolution of 256 ppr would have 256 distinct lines on the disk.

With one pulse track on the encoder disk, incremental encoders deliver only on-off outputs that provide speed, distance, and directional data to a monitoring or control device through the quadrature output. To measure direction, the controller's electronics monitor the A and B channels to determine which channel arrives first. The controller then verifies whether the encoder is turning clockwise or counterclockwise. Distance is measured by assigning each pulse interval a distance value, and counting the number of pulses with an up-down counter, as discussed earlier in this chapter.

In reality, incremental encoders do not have two separate rings for the A and B channels; instead, the two phototransistors are displaced by  $90^\circ$  electrical from each other to give the same effect. If absolute position is required, there will always be a separate index pulse. Figure 2-33 shows a circuit that is used to output the absolute position derived from an incremental encoder. In general similar circuitry is built into the more complex motion-control electronics modules used to drive motors.

In standard mode, an encoder counts the leading edge of one square wave signal. By counting both the leading and trailing edge of the signal, resolution is doubled, a process called 2 times interpolation.

By counting both edges on the A and B channels simultaneously, output increases to four times the ppr, also known as 4 times interpolation. So, by interpolating a quadrature output by a factor of four, a 256 ppr encoder can offer 1024 ppr.



**FIGURE 2-33** ■  
Counter circuitry for incremental encoders.

**FIGURE 2-34 ■**  
Photograph of an incremental encoder with encoder disk exposed. (Courtesy of GPI <http://www.gpi-encoders.com/>.)



With this technology, resolutions of 40,000 ppr are possible.

For even higher resolution, some encoders replace square wave signals with sine wave outputs that allow for interpolation factors as high as 10 times. Through 10 times interpolation of a sine wave, a 256 ppr encoder is capable of 2,560 ppr.

To illustrate the benefits of particular resolutions, one could use an example of an incremental encoder in monitoring conveyor position. An encoder could be mounted to a conveyor drive wheel with a circumference of 150 mm at a 1:1 ratio such that each revolution of the encoder is equivalent to 150 mm of overall conveyor travel. By using an output of 40,000 ppr, the encoder is able to provide positional feedback accurate to 0.012 mm.

Unfortunately, in the event of a power failure incremental encoders will usually lose positional information, and a reset or homing cycle must be performed to synchronize the encoder with the control device.

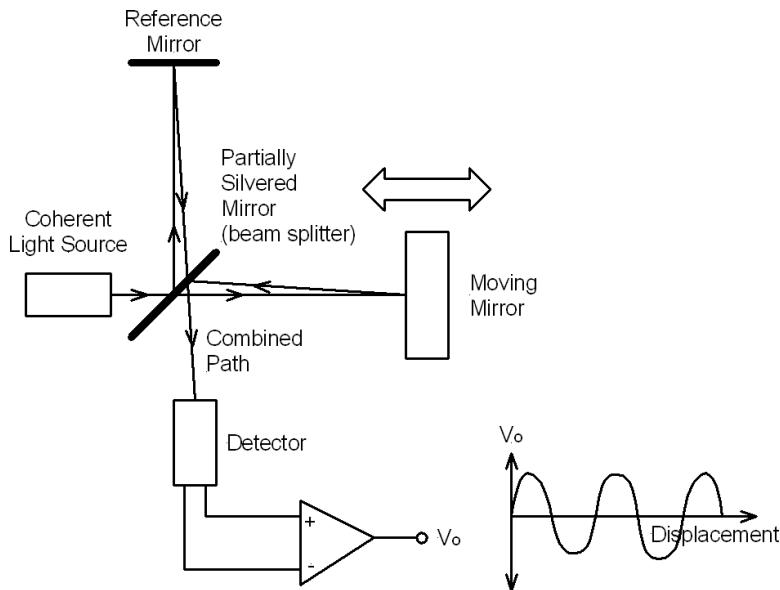
Figure 2-34 shows an example of the 7700 optical encoder and disk manufactured by Gurley Precision Instruments (GPI). It is 13 mm tall by 43 mm diameter with a resolution of up to 5000 ppr.

## 2.4.6 Ranging Sensors

Noncontact distance measurement methods can be classified into three categories: (1) interferometry; (2) triangulation; and (3) time of flight. The method used by a particular sensor usually depends on the maximum range and the measurement accuracy required. For example, interferometric methods can be extremely accurate but are prone to range ambiguity, whereas time-of-flight methods operate at longer ranges with poorer accuracy; triangulation sits somewhere in the middle.

### 2.4.6.1 Interferometry

Optical interferometry operates using the superposition of two monochromatic light beams so that extremely small displacements can be measured. The operational principle is best explained in conjunction with Figure 2-35.



**FIGURE 2-35 ■**  
Using a Michelson interferometer to measure displacement.

Typically, an incoming beam of light will be split into two identical beams by a grating or a partial mirror. Each of these beams will travel a different path before they are recombined at a detector. The difference in the distance traveled by each beam creates a phase difference between them. This introduced phase difference creates the interference pattern between the initially identical waves. If one of the paths is held constant as a reference, then any change in the distance to the mirror in the measurement path will result in a change in the relative phase of the two beams back at the detector.

Interferometers are ambiguous over distances of  $\lambda/2$ , but as with incremental encoders a count can be made of the number of electrical cycles out of the detector to accommodate extremely accurate displacement measurement.

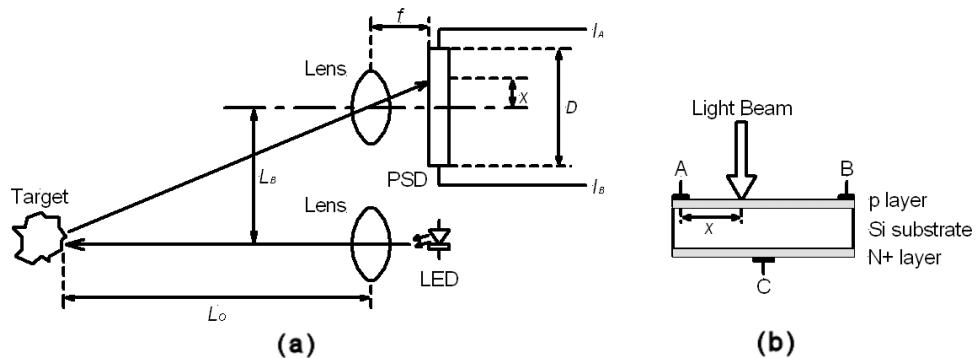
### 2.4.6.2 Triangulation Sensors

As discussed by Brooker (2008), triangulation sensors make good noncontact devices for measuring ranges from a few centimeters up to about a meter with reasonable accuracy. One of the most common uses a collimated LED source and a position-sensitive detector (PSD) to measure the direction of arrival of the reflected light beam. As shown in Figure 2-36, the relationship between the range to the target,  $L_o$ , and the distance between the transmit and receive apertures,  $L_B$ , is a function of the focal length,  $f$ , and the displacement,  $x'$ , from the center of the PSD.

$$L_o = f \frac{L_B}{x'} = f \frac{L_B}{D/2 - x} \quad (2.26)$$

An LED or laser source emits a narrow beam of infrared light in the direction of the target. The small amount of reflected light is focused onto the sensitive surface of the PSD, which generates two output currents,  $I_A$  and  $I_B$ , that are each proportional to the displacement of the light spot from the center of the device.

**FIGURE 2-36** ■ Measuring range using a PSD-based sensor.  
 (a) Operational principles.  
 (b) Cutaway view of PSD. (Brooker 2008.)



Though the current generated by the PSD depends on the intensity of the incident radiation, the ratio of the two currents does not, so the distance to the target can be determined with good accuracy. The PSD operates by exploiting the change in resistance of a doped silicon semiconductor, which is proportional to the intensity of the incident light. The device is fabricated from a thin slab of high-resistance silicon with the top and bottom layers doped p and n+, respectively. Two electrodes, A and B, are placed at opposite ends of the upper layer, and a common electrode, C, covers the bottom. The distance between the upper electrodes is  $D$ , and the corresponding resistance is  $R_D$ .

If the beam strikes the PSD at a distance  $x$  from electrode A, the resistance between that point and the electrode is  $R_x$ , and a photocurrent,  $I_o$ , proportional to the intensity of the light, will flow. The amount that flows to the electrodes A and B will be proportional to the relative distances from the incident beam; therefore,

$$I_A = I_o \frac{R_D - R_x}{R_D} = I_o \frac{D - x}{D} \quad (2.27)$$

and

$$I_B = I_o \frac{R_x}{R_D} = I_o \frac{x}{D} \quad (2.28)$$

Taking the ratio of the two currents  $I_A$  and  $I_B$

$$S = \frac{I_A}{I_B} = \frac{D}{x} - 1 \quad (2.29)$$

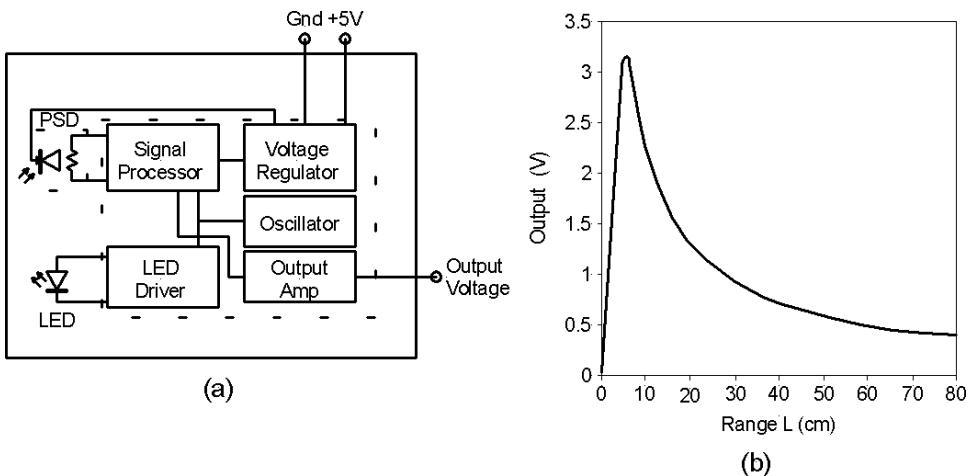
The distance,  $x$ , can then be written as

$$x = \frac{D}{S + 1} \quad (2.30)$$

Substituting into (2.26) gives

$$L_o = f \frac{2L_B}{D} \frac{S + 1}{S - 1} = k \frac{S + 1}{S - 1} \quad (2.31)$$

where  $k$  is the module geometrical constant. Therefore,  $L_o$ , the distance from the sensor to the target, can be determined in terms of the ratio of the two currents out of the PSD (Fraden, 2003).



**FIGURE 2-37 ■**  
**Sharp GP2Y0A21YF**  
**distance**  
**measurement**  
**(a)** Schematic block  
**diagram.** (b) Voltage  
**output as a function**  
**of range.** (Courtesy  
**of Sharp Electronics**  
[http://www.  
sharpusa.com/\).](http://www.sharpusa.com/)

A common IC that operates using this principle is the Sharp Electronics GP2Y0A21YF, which has an operational range between 10 and 80 cm with the characteristics as illustrated in Figure 2-37.

In more critical applications, a laser source is used because of its brightness and good spatial coherence, and a charge-coupled device (CCD) array is often used as a receiver. On transmit, the width of the laser beam is diffraction-limited by the size of the exit aperture. On scattering from the target, the coherence of the laser beam is lost, so to produce the smallest spot on the CCD array it must be placed at the focal plane of a lens (Amman, 2001).

Longer-range triangulation sensors can be used to obtain two-dimensional (2-D) or three-dimensional (3-D) images of the terrain by scanning the beam. Such scanning mechanisms must be constructed to ensure that both the transmitter and the receiver can see each point, so that as the transmitted spot is swept across the target a corresponding image is reflected onto the CCD array (Probert-Smith, 2001).

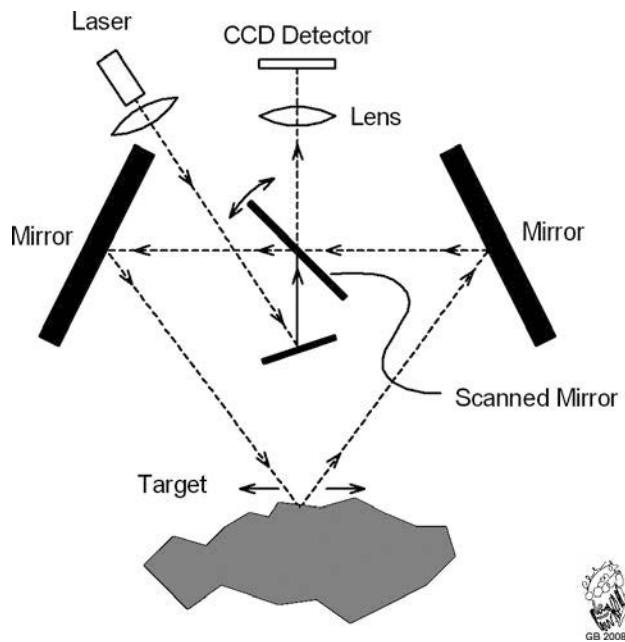
Figure 2-38 shows a diagram of a sensor developed for 2-D scanning (Livingstone and Rioux, 1986).

## 2.4.7 Time-of-Flight Ranging

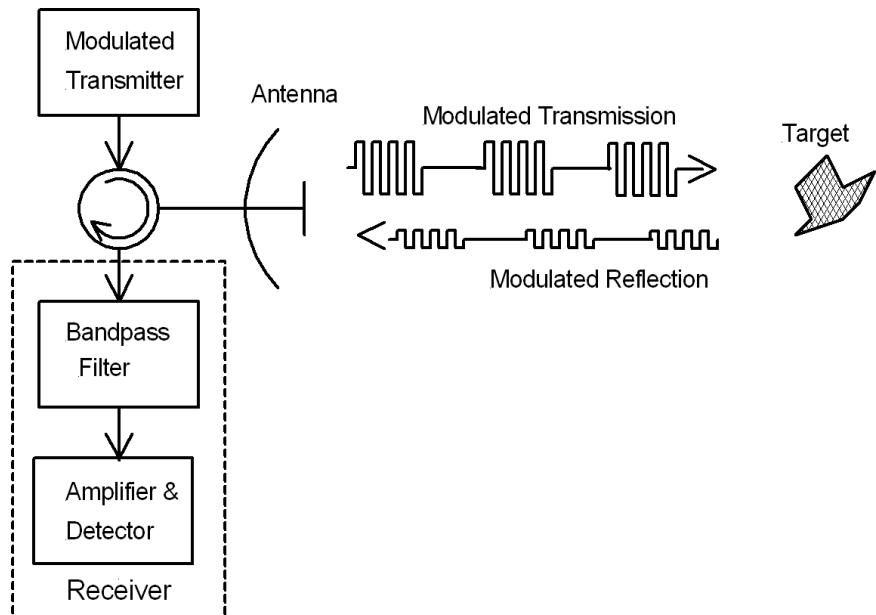
The basic principles of active noncontact range-finding are similar for electromagnetic (e.g., radar, laser) and active acoustic sensing. A signal is radiated toward an object or target of interest, and the reflected or scattered signal is detected by a receiver and used to determine the range.

As shown in Figure 2-39, a source of radiation is modulated and fed to a transmit antenna, or aperture, which is usually matched to the impedance of the transmission medium to maximize power transfer. This can take the form of a horn for acoustic or radar sensors or an appropriately coated lens for a laser. The antenna also operates to concentrate the radiated power into a narrow beam to maximize the operational range and to minimize the angular ambiguity of the measurement. When the transmitted beam strikes the target, a portion of the signal is reflected or scattered because the target has a different impedance, or refractive index, to the medium through which the signal is propagating. A small percentage of the reflected power travels back to the receiver (which is often colocated with the transmitter), where it is captured by the receiver antenna and

**FIGURE 2-38** ■ Line-scanned triangulation-based line scanner.



**FIGURE 2-39** ■ Operational principles of a generic time-of-flight sensor.



converted to an electrical signal that can be filtered to remove extraneous noise before being amplified and detected.

The time between the transmission of a pulse and the reception of an echo is used to provide range. Because the round-trip time is measured, there is a factor of two in the formula

$$R = \frac{v\Delta T}{2} \quad (2.32)$$

where  $R$  is the range (m),  $v$  is the propagation velocity (m/s), and  $\Delta T$  is the round-trip time (s).

Most ranging sensors transmit and receive from the same aperture or from adjacent apertures, requiring special techniques that allow the transmission and receipt of signals simultaneously. In general the transmit pulse is made sufficiently short to have been radiated before the echo is received. This limitation restricts the minimum distance that can be measured and requires that very short pulses are transmitted.

Because the velocity of sound is so much smaller than the speed of light, measuring short ranges with acoustic systems is not difficult, but for optical or microwave time-of-flight sensors the minimum range is usually a couple of meters.

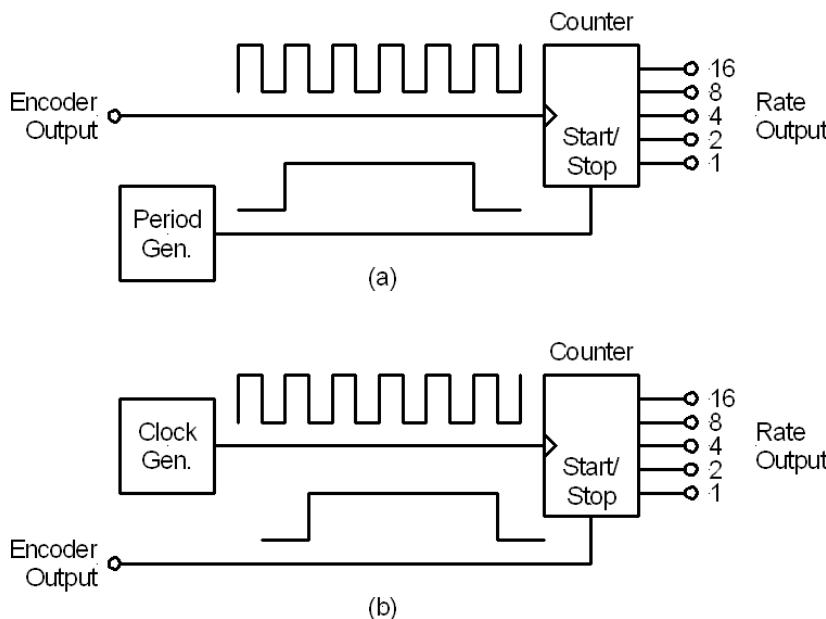
## 2.4.8 Measuring Rate and Angular Rate

### 2.4.8.1 Incremental Encoders

Magnetic and optical incremental encoders and interferometers output a continuous stream of pulses at a frequency that is a function of the rate. Good estimates of the rate can be obtained by measuring the frequency over a reasonable interval. This is achieved by counting cycles. If the speed is low, better estimates are made by running a high-speed clock and measuring the interval between pulses, as shown in Figure 2-40.

### 2.4.8.2 Tachogenerators

Tachogenerators are small AC or DC generators that output a voltage in proportion to the rotational speed of a shaft. They are capable of measuring speed and direction of rotation but not position. They convert a rotational speed into an isolated analog voltage signal that is suitable for remote indication and control applications. These generators are often used



**FIGURE 2-40 ■**  
Estimating rate from the output of an incremental encoder for (a) high speeds and (b) low speeds.

in servo systems to supply velocity or damping signals and may be mounted on or in the same housing as a servo motor.

There are two basic types of tachogenerators—AC and DC—both of which develop an output voltage proportional to rotational speed. In DC generators the polarity depends on the direction of rotation, whereas in AC generators the relationship between phases changes.

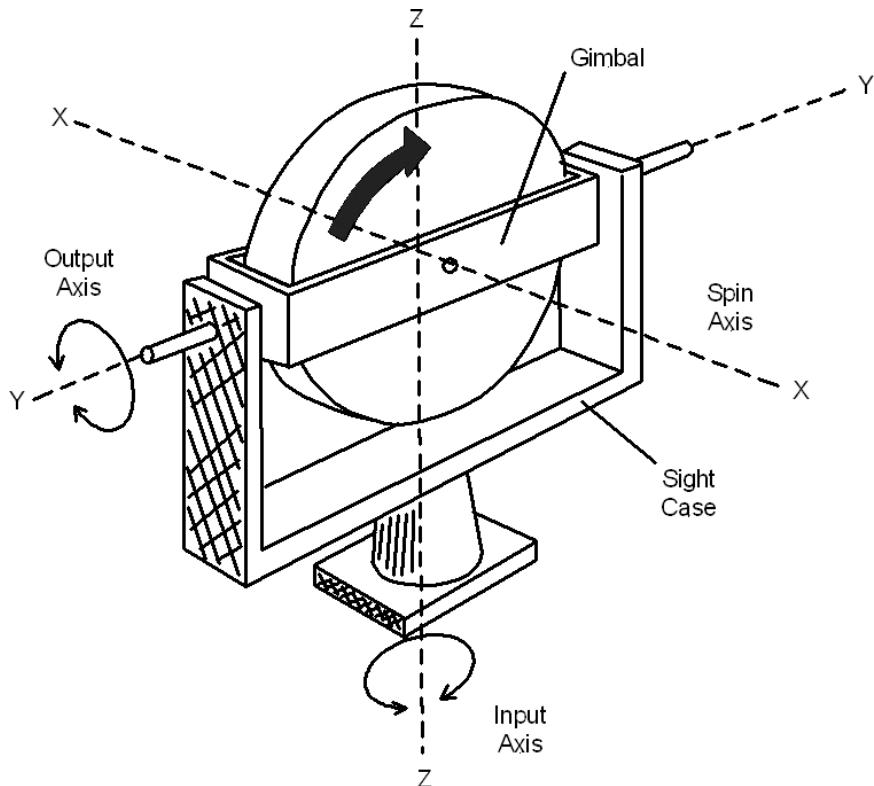
Performance specifications for these devices include generated voltage, accuracy, maximum speed, and ripple. Generated voltage is measured in volts (V) per revolution per minute (rpm) over a range of speeds. Accuracy or linearity is the deviation of the voltage output signal expressed as a percentage, and ripple is determined as a percentage of the output voltage.

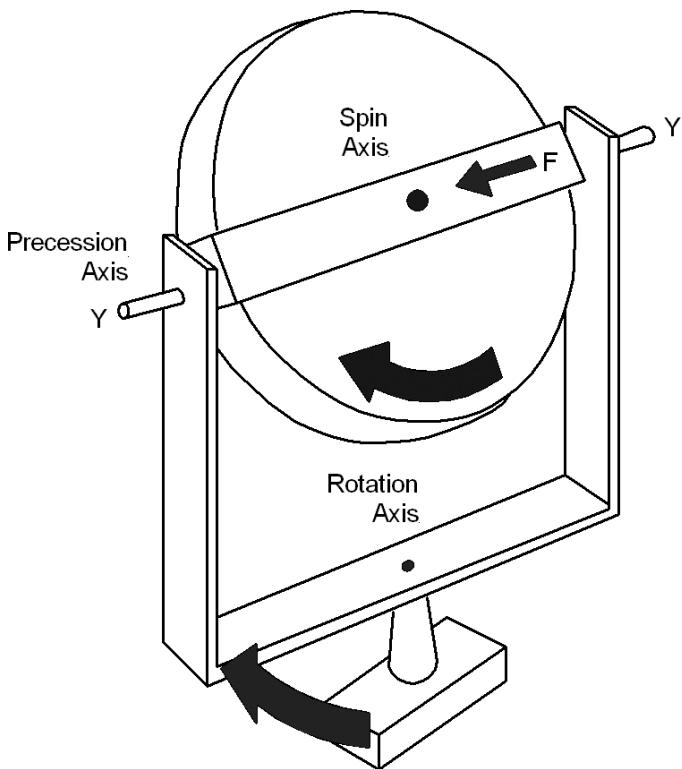
#### 2.4.8.3 Rate Gyros

In many applications, angular rate must be measured with reference to an inertial frame rather than relative to some physical object. Rate gyroscopes, commonly called rate gyros, use the conservation of angular momentum to keep one or more inertial axes pointed in one direction as the external frame translates and rotates. An output voltage is produced that is proportional to the rate of rotation of an axis perpendicular to the axis of the gyro.

Many high-accuracy rate gyros still consist of a spinning rotor mounted in a single gimbal, as shown in Figure 2-41. A gyro mounted in this manner has 1 degree of freedom; that is, it is free to tilt in only one direction. The rotor in a rate gyro is generally restrained

**FIGURE 2-41** ■  
Rate gyro with a single degree of freedom. [Adapted from Neets, Electrical Engineering Training Series <http://www.tpub.com/content/neets/14187/>.]



**FIGURE 2-42 ■**

Rate gyro precession.

[Adapted from Neets, Electrical Engineering Training Series <http://www.tpub.com/content/neets/14187/>.]

from precessing by some means, usually a spring arrangement. This is done to limit precession and to return the rotor to a neutral position when there is no angular change taking place. The amount of precession of a gyro is proportional to the force that causes the precession.

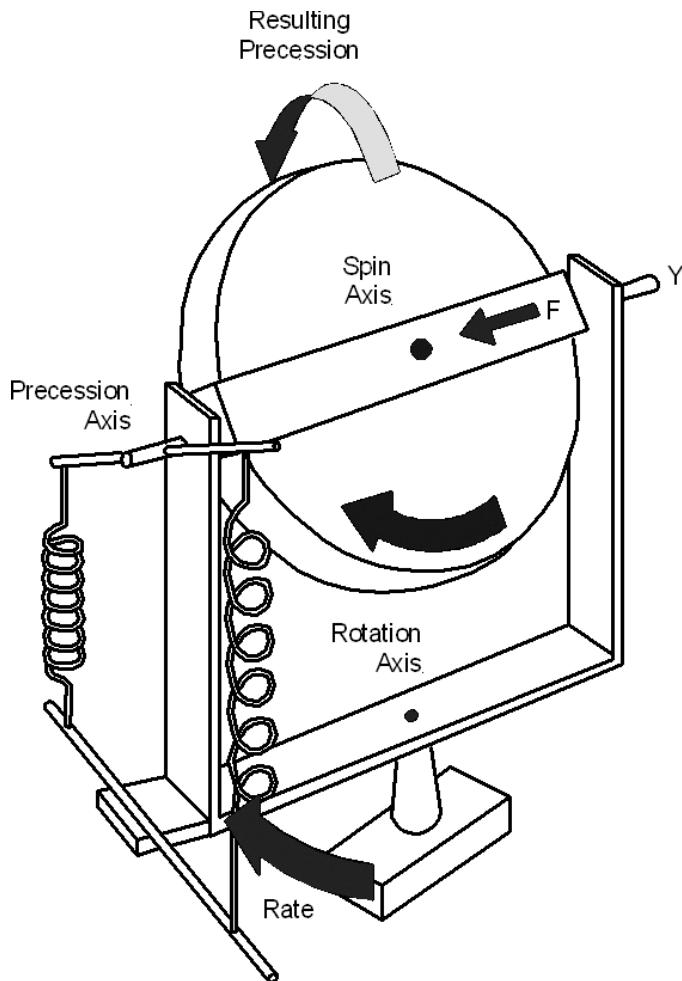
If the gyro's plane of rotor spin is changed by rotating the case about the input axis, the gyro will precess as shown in Figure 2-42 because turning the gyro case has the same effect as applying a torque on the spin axis. This is illustrated by arrow F in Figure 2-42. The direction of precession can be determined by using the right-hand rule.

The force applied at F will cause the gyro to precess at right angles to the force. Likewise, attempting to turn the gyro case will have the same result. The gyro will precess, as shown by the arrows, around the Y-Y axis (output axis). Since the rate of precession is proportional to the applied force, the precession can be increased by increasing the speed with which the gyro case is moved.

The amount of precession is proportional to the rate at which the gyro base is turned. This characteristic of a gyro, when properly used, fits the requirements needed to sense the rate of motion about any axis. Figure 2-43 shows a method of restraining the precession of a gyro to permit the calculation of an angle. Springs have been attached to the cross-arm of the output shaft. These springs restrain the free precession of the gyro.

As the gyro precesses, it exerts a precessional force against the springs that is proportional to the momentum of the spinning wheel and the applied force. This force will result in the precession increasing until it is balanced just by the force on the springs, and it will remain in the precessed position as long as the gyro base is rotated at the same constant speed. When rotation ceases, the force at F is removed and the gyro will stop

**FIGURE 2-43 ■**  
**Precession of a spring retained rate gyro.** [Adapted from Neets, Electrical Engineering Training Series <http://www.tpub.com/content/neets/14187/>.]

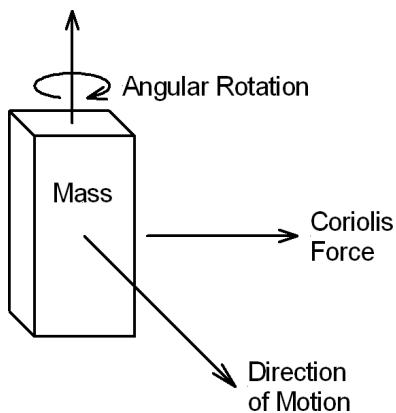


trying to precess, so the spring returns the cross-arm to the neutral point. A low-friction device to measure the precession angle and output an electrical signal completes the rate gyro.

Spinning mass rate gyros are used in several areas, such as the guidance of intercontinental ballistic missiles (ICBMs) and submarine-launched ballistic missiles (SLBMs). They are also the technology of choice in low-jitter pointing for the Hubble telescope and control and stabilization of satellites. In addition, they can be found in legacy systems such as aircraft navigators and tactical missiles that were built before the advent of solid-state equivalents.

Most modern low-cost microelectromechanical systems (MEMSs) rate gyros operate using the Coriolis force. In the 1830s, Gaspard-Gustave de Coriolis discovered that an object moving in a rotating frame would cause the observer on the rotating frame to see an apparent acceleration of the object. In other words, if an object is moving in a straight line and is subject to a rotation, its path will deviate from that straight line.

In this type of rate gyro, a MEMS structure is caused to vibrate in one plane. If the structure is also rotating, then the Coriolis force will be generated at right angles to both the direction of vibration and the axis of rotation as shown in Figure 2-44.



**FIGURE 2-44** ■ The Coriolis effect.

If  $V$  (m/s) is the instantaneous velocity of the vibrating structure of mass  $m$  (kg), which is rotating at an angular rate  $\omega$  (rad/s), then the Coriolis force is given by

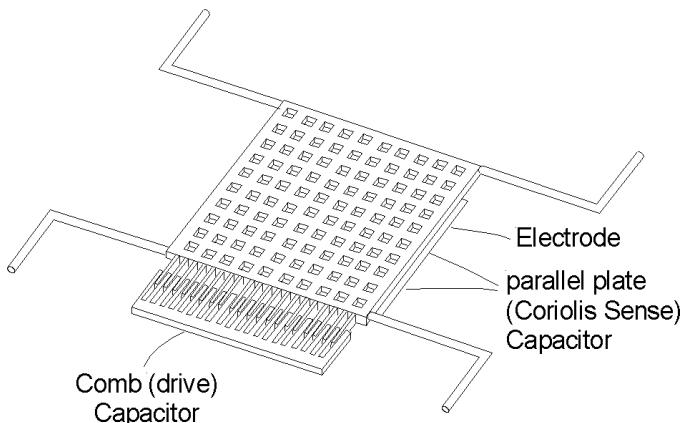
$$F = 2m\omega V \quad (2.33)$$

As the velocity of the vibrating element is sinusoidal, the Coriolis force will introduce a lateral vibration at the same frequency. This can be converted to an electrical signal using a piezoelectric stress/bending mechanism or a variable capacitance.

One implementation of a MEMS tuning fork gyro is shown in Figure 2-45. A proof mass attached to springs is forced to oscillate in the horizontal plane.

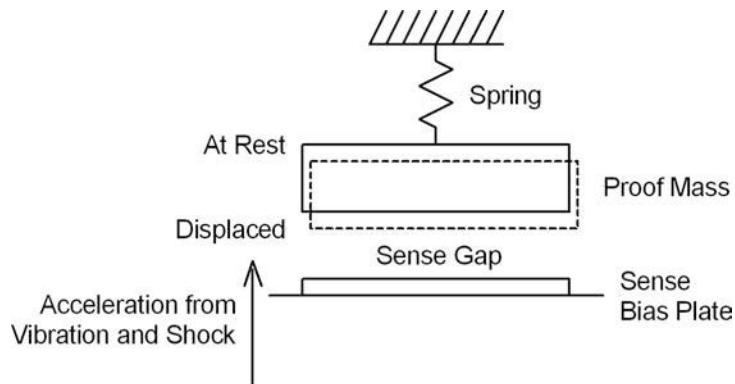
A voltage is applied to a sensing electrode (sense plate) below the proof mass, creating an electrical field. The Coriolis force imparted by angular rotation causes the proof mass to oscillate vertically, which in turn changes the gap between the proof mass and the sense plate, as shown in Figure 2-46. This motion generates an AC current with amplitude proportional to the rotation rate.

Unfortunately, the bias voltage also generates an electrostatic force between the proof mass and the sense plate. This force acts to pull the proof mass toward the sense plate. In addition, since the gyro is a mechanical device, it is sensitive to external stimuli including vibration, acoustic excitation, and acceleration due to mechanical shock. In some applications the acceleration levels at the gyro can be as large as 500 g. These stimuli will



**FIGURE 2-45** ■ MEMS rate gyro with capacitive coupling.

**FIGURE 2-46 ■**  
Capacitive mechanism to measure rate using the Coriolis force.



increase or decrease the sense gap depending on amplitude and frequency and must be considered when the rate gyro is used.

The main advantages of MEMS rate gyros are their small size and the fact that they are low power. In addition, they are extremely reliable and have a long mean time between failures (MTBF), and because the mass of the vibrating element is so small they are reasonably insensitive to external environmental factors like shock, vibration, and acceleration.

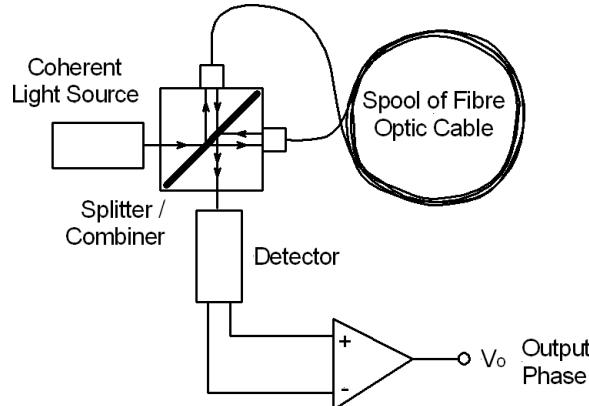
MEMS rate gyros do suffer from drift and offset, and their sensitivity is lower than some of the other types. Because drift is often a function of temperature, many packaged gyros include temperature compensation tables.

Another robust technology is the fiber optic gyro (FOG) rate sensor. These rate sensors use a fiber optic ring and a solid-state laser to measure rotation rates using the Sagnac effect. In 1913, Georges Sagnac showed that light sent around a closed loop in different directions would show a phase difference between the two beams when the loop was rotated (Fraden, 1996).

The detection process is similar to that illustrated for the Michelson interferometer. As shown in Figure 2-47, a laser provides a coherent beam that is split and sent in opposite directions around a fiber optic ring.

If the ring gyro is static, then the path lengths back to the detector are identical and the beams interfere constructively. However, if the ring is rotating, the path traveled in the one direction is slightly shorter than that traveled in the other because the position of the

**FIGURE 2-47 ■**  
FOG rate sensor.



detector has moved, with the result that the signals are no longer in phase. The phase difference is proportional to the difference in the two distances and the wavelength of the laser.

The Sagnac phase shift,  $\Delta S$ , is

$$\Delta S = \frac{8\pi n A \omega}{c \lambda} \quad (2.34)$$

where  $A$  ( $\text{m}^2$ ) is the cross sectional area enclosed by the fiber optic coil,  $n$  is the number of turns of fiber optic cable around the ring,  $\omega$  ( $\text{rad/s}$ ) is the angular rate at which the ring is rotating in the sensitive plane,  $c$  ( $\text{m/s}$ ) is the speed of light, and  $\lambda$  ( $\text{m}$ ) is the wavelength of the light.

Because all of the other terms are known, FOG rate sensors can easily convert a measured phase shift,  $\Delta S$ , into an angular rate. This technique is very stable with temperature, and the drift levels are very low. There are no moving parts, and with the good reliability of modern laser diodes and detectors the MTBF is very high.

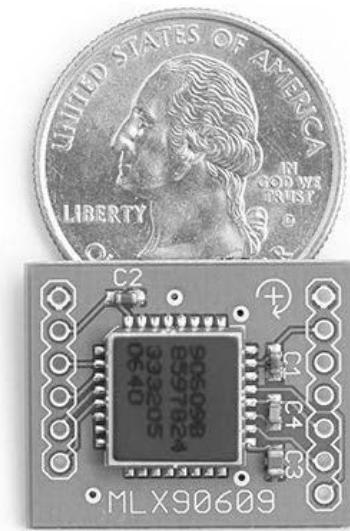
Table 2-5 lists many manufacturers of rate gyros, which range from precision military devices costing thousands of dollars to mass-produced MEMS devices for consumer electronics such as the Melexis MLX90609 single-chip solution with a  $20 \times 20$  mm footprint and a cost of \$50.00 (Figure 2-48). The specifications of this gyro include a  $+/-5\%$  drift in scale for a temperature range of  $-40^\circ\text{C}$  to  $+85^\circ\text{C}$  and a noise power spectral density of  $0.03^\circ/\text{s}/\sqrt{\text{Hz}}$ . The latter equates to an RMS noise of  $0.25^\circ/\text{s}$  for the maximum gyro bandwidth of 75 Hz. Other MEMS rate gyro manufacturers quote biases over temperature of between 0.3 and  $3^\circ/\text{s}$  for their commercial-grade products. In contrast, the long-term drift quoted for good FOGs is quoted to be around  $3^\circ/\text{hr}$ , more than 1000 times lower than that achievable from MEMS gyros.

In most applications, rate gyros are used not in isolation but usually as part of an inertial measurement unit (IMU), which uses sensor modeling and filtering techniques to improve overall sensing accuracy.

**TABLE 2-5 ■** Manufacturers of Commercial Rate Gyros

Manufacturer	Manufacturer
Akustica	JAE
Analog Devices	Kenyon Laboratories
Applanix Corp	KVH Industries
Applied Technology Associates	Litton Industries
Atlantic Inertial Systems	MEMSCAP
Cloud Cap Technology	MEMSense
Colibrys	Micro Aerospace Solutions
Crossbow Technology	Murata Electronics
Flightline Electronics	Northrop Grumman
Flightline Systems	Sagebush Technology
Goodrich Sensor Systems	Silicon Sensing Systems
Honeywell Sensing and Control	Sagem Avionics
iMAR GmbH	Systron Donner
Instrumented Sensor Technology	True North Technologies
IMU Solutions	TSS International
InterSense	Ultra Electronics
InvenSense	Watson Industries

**FIGURE 2-48 ■**  
**MLX90609 rate gyro.**  
(Courtesy of  
Melexis.)



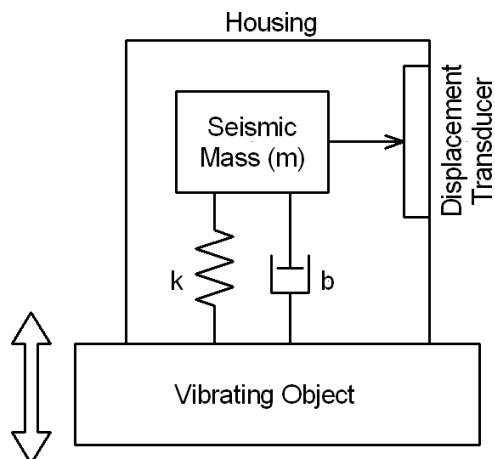
## 2.4.9 Accelerometers

Accelerometers are sensors designed to measure continuous mechanical vibration such as bearing vibration, transitory vibration from blasts or impacts, or the lower-frequency acceleration generated by bodies in motion. These devices are generally bonded firmly to a structure and aligned in the direction of interest as they are sensitive to motion along one axis only.

### 2.4.9.1 Theory

All accelerometers are based on the inertial effects associated with a mass connected to a moving object through a spring, damper, and a displacement sensor as shown in Figure 2-49. When an object accelerates, there is relative motion between the housing and the seismic mass, which is measured by the displacement sensor (Alciatore and Histand, 2003).

**FIGURE 2-49 ■**  
Accelerometer  
construction.



The important characteristics of an accelerometer are its sensitivity and dynamic range as well as its frequency response. The sensitivity determines the smallest acceleration that can be reliably measured. The dynamic range can then be used to find the maximum acceleration if the response must remain linear. The frequency response gives the range of frequencies that an accelerometer can measure.

If the relative displacement,  $x_r$  (m), between the seismic mass and the object measured by a position transducer mounted between the seismic mass and the housing is given by

$$x_r = x_o - x_i \quad (2.35)$$

The spring force,  $F_k$  (N), can then be expressed in terms of this displacement and the spring constant  $k$

$$F_k = k(x_o - x_i) = kx_r \quad (2.36)$$

and the damper force can be expressed in terms of the time derivatives,  $\dot{x}_i$  and  $\dot{x}_o$

$$F_b = b(\dot{x}_o - \dot{x}_i) = b\dot{x}_r \quad (2.37)$$

Applying Newton's second law

$$-F_k - F_b = m\ddot{x}_o \quad (2.38)$$

These forces have negative signs because they are in the opposite direction to the reference direction  $x_o$ .

Substituting equation (2.36) and (2.37) into (2.38)

$$-kx_r - b\dot{x}_r = m\ddot{x}_o \quad (2.39)$$

The output acceleration,  $\ddot{x}_o$ , can be written as

$$\ddot{x}_o = \ddot{x}_r + \ddot{x}_i \quad (2.40)$$

And this can be substituted into equation (2.39) and rearranged to produce a second-order differential equation

$$m\ddot{x}_r + b\dot{x}_r + kx_r = -m\ddot{x}_i \quad (2.41)$$

This can be rewritten in terms of the natural frequency and the damping ratio as

$$\frac{1}{\omega_n^2}\ddot{x}_r + \frac{2\zeta}{\omega_n}\dot{x}_r + x_r = -\frac{1}{\omega_n^2}\ddot{x}_i \quad (2.42)$$

where the natural frequency is

$$\omega_n = \sqrt{\frac{k}{m}} \quad (2.43)$$

and the damping ratio is

$$\zeta = \frac{b}{2\sqrt{km}} \quad (2.44)$$

If the input displacement is limited to a sinusoidal term with amplitude  $X_i$ , then because the system is linear the output displacement will also be sinusoidal with the same frequency and with a phase shift,  $\phi$ .

$$\begin{aligned} x_i(t) &= X_i \sin \omega t \\ x_o(t) &= X_r \sin(\omega t + \phi) \end{aligned} \quad (2.45)$$

From the known characteristics of second-order systems, the amplitude ratio can be written as

$$\frac{X_r}{X_i} = \frac{(\omega/\omega_n)^2}{\left( \left[ 1 - \left( \frac{\omega}{\omega_n} \right)^2 \right]^2 + 4\xi^2 \left( \frac{\omega}{\omega_n} \right)^2 \right)^{1/2}} \quad (2.46)$$

and the phase shift will be

$$\phi = -\tan^{-1} \left( \frac{2\xi(\omega/\omega_n)}{1 - \left( \frac{\omega}{\omega_n} \right)^2} \right) \quad (2.47)$$

To determine the output displacement in terms of the input acceleration, equation (2.45a) is differentiated twice

$$\ddot{x}_i(t) = -X_i \omega^2 \sin \omega t \quad (2.48)$$

Equation (2.46) can be rewritten as

$$H_a(\omega) = \frac{X_r \omega_n^2}{X_i \omega^2} = \frac{1}{\left( \left[ 1 - \left( \frac{\omega}{\omega_n} \right)^2 \right]^2 + 4\xi^2 \left( \frac{\omega}{\omega_n} \right)^2 \right)^{1/2}} \quad (2.49)$$

It is therefore possible to relate the measured output displacement amplitude in terms of the input acceleration amplitude as

$$X_r = \left( \frac{1}{\omega_n^2} \right) H_a(\omega) (X_i \omega^2) \quad (2.50)$$

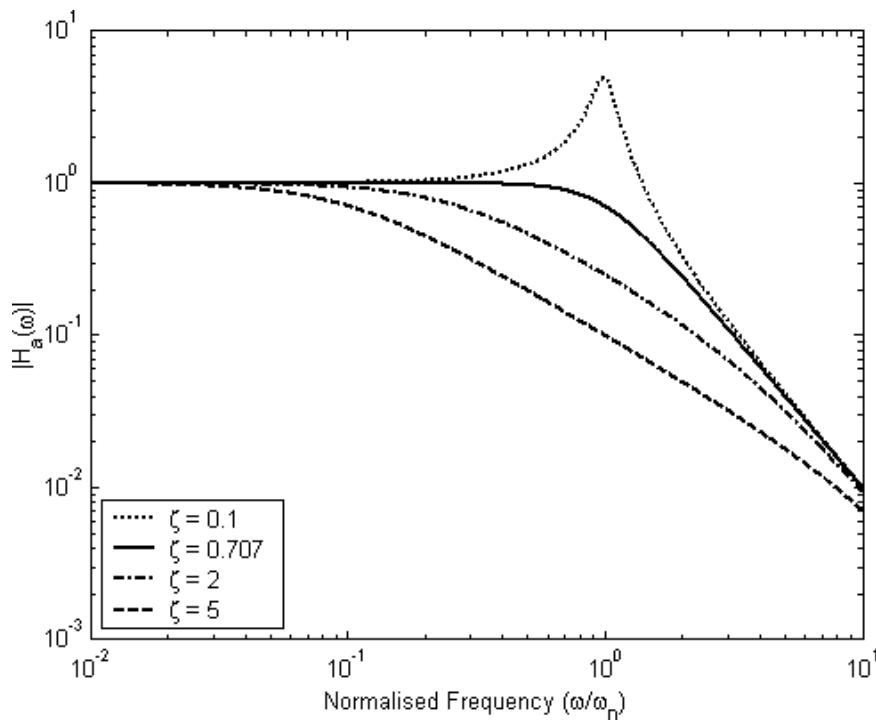
where  $(X_i \omega^2)$  is the input acceleration amplitude. It can be expressed in terms of the transfer function,  $H_a(\omega)$ , and the output amplitude,  $X_r$

$$(X_i \omega^2) = \frac{X_r \omega_n^2}{H_a(\omega)} \quad (2.51)$$

If the accelerometer is designed so that  $H_a(\omega) = 1$  over a large frequency range, then the input acceleration amplitude is given by the relative displacement amplitude scaled by a constant factor  $\omega_n^2$

$$(X_i \omega^2) = \omega_n^2 X_r \quad (2.52)$$

It can be seen from Figure 2-50 that the most linear response occurs with a damping ratio  $\xi = 0.707$  and the natural frequency  $\omega_n$  as large as possible. Looking back at

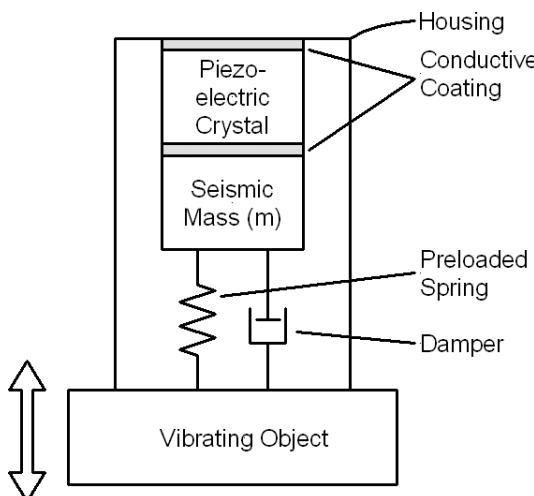


**FIGURE 2-50** ■ Ideal accelerometer amplitude response.

equation (2.42) it can be seen that  $\omega_n$  can be made large by choosing a small seismic mass and a large spring constant.

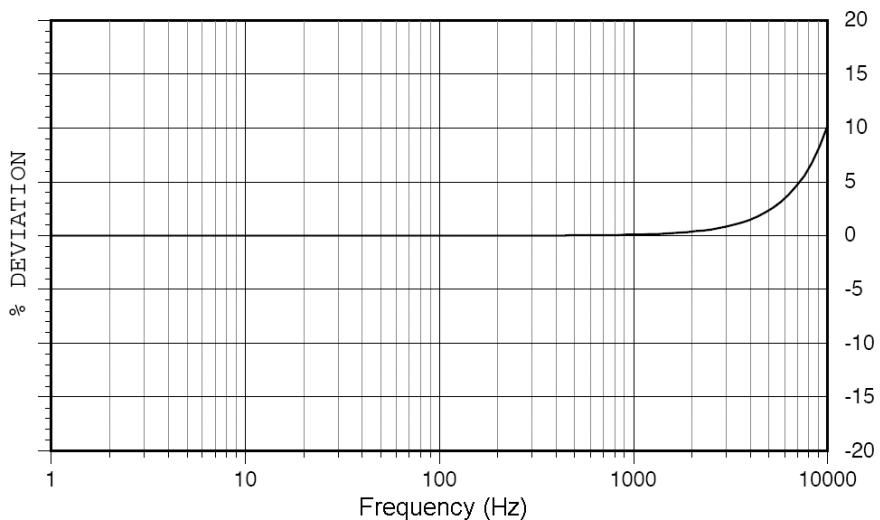
#### 2.4.9.2 Piezoelectric Accelerometers

Piezoelectric accelerometers use piezoelectric crystals to measure the force exerted by the seismic mass. Deformation of these crystals results in the generation of a charge across the opposite faces. As illustrated in Figure 2-51, a piezoelectric accelerometer includes a crystal in contact with the mass. In addition to the natural damping inherent in the crystal, additional damping is sometimes included by filling the housing with oil.



**FIGURE 2-51** ■ Piezoelectric accelerometer.

**FIGURE 2-52 ■**  
Frequency response  
of an Endevco  
2273A piezoelectric  
accelerometer.



When the object is accelerated, relative displacement occurs between the housing and the seismic mass due to the inertia of the mass. The resulting strain within the piezoelectric crystal causes a charge to be developed on the two faces of the crystal that can be tapped by the conductive coating and measured using a charge amplifier (discussed in Chapter 5).

In general, piezoelectric accelerometers cannot measure constant or slowly changing accelerations because the crystal can measure changes only in strain. Low-frequency response is typically a few hertz, while the high-frequency response is determined by the mechanical characteristics of the accelerometer and the mounting stiffness. This is normally close to the natural (resonant) frequency of the system, which is usually in the kHz range. Manufacturers generally specify this as a deviation from the nominal, as shown in Figure 2-52.

Accelerometers come in all sizes and shapes for any conceivable application, as can be seen from Figure 2-53.

In addition to piezoelectric accelerometers, most manufacturers make piezoresistive types that can operate down to DC and are therefore useful for measuring static acceleration (gravity). These devices use a bridge configuration to output a voltage proportional to the acceleration.

Manufacturers of accelerometers are listed in Table 2-6.

**FIGURE 2-53 ■** An assortment of accelerometers.  
(Courtesy of Endevco  
<http://www.endevco.com/>.)



**TABLE 2-6** ■ Manufacturers of a Selection of Accelerometers

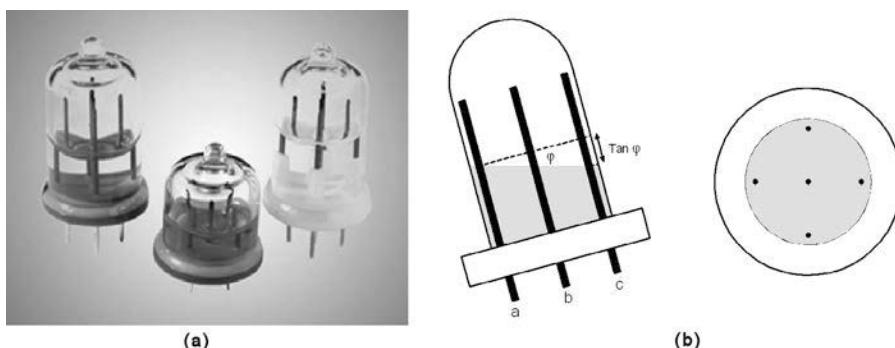
Manufacturer	Product Range
Columbia Research Labs	99
Dataforth Corporation	1
Dytran Instruments	111
Endevco Corporation	104
Extech Instruments	1
Honeywell Sensing and Control	103
Instrumented Sensor Technology	43
Kistler Instrument Corporation	161
Measurement Specialities	187
MEMSTech	3
Ricker	9
Soltec Corporation	49
Wilcoxon Research	45

### 2.4.10 Tilt Sensors

Pinball players all know at least one application of tilt sensors. At their most basic, tilt sensors consist of a pool of mercury and two contacts within a hermetically sealed glass container that complete a circuit if the device is tilted.

More sophisticated tilt sensors can be manufactured from accelerometers mounted on three orthogonal axes and equating the relative gravity vector in each. However, most low-cost tilt sensors exploit the bubble-level principle. Figure 2-54 shows one axis of a fluid-filled sensor tipped at  $\sim 15^\circ$ . As the sensor tilts, the surface of the fluid remains level due to gravity. The fluid is electrically conductive, and the conductivity between the two electrodes is proportional to the length of electrode immersed in the fluid. At the angle shown, for example, the conductivity between pins *a* and *b* would be greater than that between *b* and *c*.

Electrically, the sensor is similar to a potentiometer, with resistance changing in proportion to tilt angle; therefore, the output can be determined using a Wheatstone bridge in its dynamic unbalanced configuration. For small angles, typically less than  $20^\circ$ , the output voltage is proportional to the tangent of the tilt angle. However, as the angle increases, nonlinearities become more pronounced, and more sophisticated microcontrollers are required to apply corrections to the measurements to maintain accuracy.



**FIGURE 2-54** ■ Electrolytic tilt sensor.  
 (a) Photographs.  
 (b) Schematic showing fluid and electrodes.

DC induces electrolysis in the fluid as positive ions in the fluid migrate to the cathode, where they combine with excess electrons and lose some of their charge. In a similar fashion, negative ions in the fluid propagate to the anode and combine with excess protons to lose their charge. If this is allowed to proceed, the reaction will eventually reduce the conductivity of the fluid to the extent that the sensor is no longer operational.

To prevent electrolysis, AC must be used to excite the sensor. The required frequency and symmetry of the AC waveform depend on the chemistry of the fluid and composition of the electrodes. The frequency must be high enough to make the aforementioned process reversible, which for some electrolytes can be as low as 25 Hz and in other solutions can be a minimum of 1000 Hz to 4000 Hz.

Dual-axis sensors exhibit the same fluid characteristics as single-axis devices but have the added complexity of interaction between the axes. Both axes share the center electrode, with the four outer electrodes placed at the four corners of a square. Misalignment among electrodes gives rise to cross-axis coupling, which can result in significant errors.

Two techniques can be used to derive independent measurements for each axis. The first is to excite only one axis at a time, alternating between pitch and roll at an appropriate rate, and the second technique requires two excitation frequencies, one twice the other. Here, all four pins are driven simultaneously, and the phase of the excitation determines which axis is being measured (Powell and Pheifer, 2006).

### 2.4.11 Pressure Measurement

Pressure is defined as the normal force per unit area exerted by a fluid (liquid or gas) on any surface. Only the force normal to the surface needs to be considered (Webster, 1999). Blood pressure is one physiological variable that is easily measured; hence, it has been used as an indication of general health, particularly the health of the cardiovascular system, for many years. The measurement of blood pressure, particularly the timely diagnosis of dangerously high blood pressure or hypertension, has saved many thousands of lives.

Three types of pressure measurements are commonly performed:

- Absolute pressure: Represents the pressure difference between the point of measurement and a perfect vacuum where the pressure is zero.
- Gauge pressure: The pressure difference between the point of measurement and the ambient (atmospheric) pressure.
- Differential pressure: The difference in the pressure between two points, one of which has been chosen to be the reference.

A number of common units are used for pressure measurement: kilopascal (kPa); pound per square inch (psi); inches or cm of water at 4 °C (cm H<sub>2</sub>O); inches or mm of mercury at 0 °C (in Hg); and millibar (mbar). Table 2-7 lists the conversion factors between these units.

A column of fluid is commonly used to measure pressure because there is a simple relationship between the pressure,  $P$ ; the density of the fluid,  $\rho$ ; acceleration due to gravity,  $g$ ; and the height of the column,  $h$ .

$$P = \rho gh \quad (2.53)$$

Biomedical pressure sensors are used mostly to measure blood or air pressure. They can also be used to measure the pressure in pneumatic and hydraulic actuators for control or monitoring purposes.

**TABLE 2-7** ■ Conversion Among Different Pressure Measurement Units

Unit	kPa	psi	in H <sub>2</sub> O	cm H <sub>2</sub> O	in Hg	mm Hg	mbar
kPa	1	0.145	4.015	10.2	0.2593	7.501	10
psi	6.895	1.0	27.68	70.31	2.036	51.72	68.95
in H <sub>2</sub> O	0.2491	0.03613	1.0	2.54	0.07355	1.868	2.491
cm H <sub>2</sub> O	0.09806	0.01422	0.3937	1.0	0.02896	0.7355	0.9806
in Hg	3.386	0.4912	13.6	34.53	1.0	25.4	33.86
mm Hg	0.1333	0.01934	0.5353	1.36	0.03937	1.0	1.333
mbar	0.10	0.01450	0.04015	1.020	0.02953	0.7501	1.0

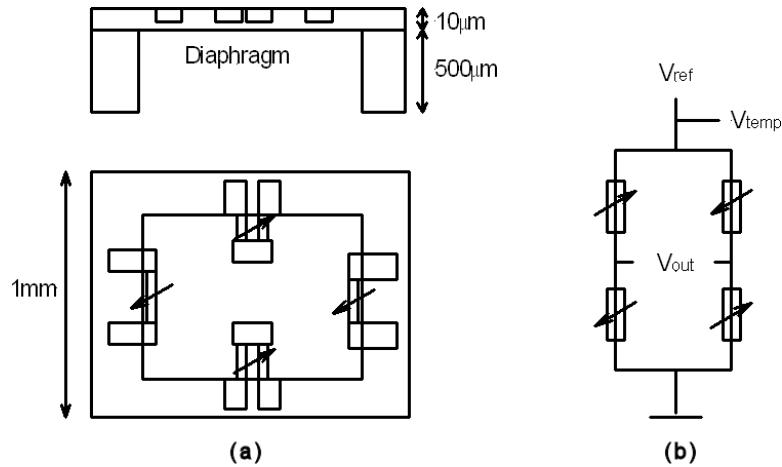
The conventional sphygmomanometer is an indirect method of measuring blood pressure. It consists of an inflatable cuff and a mercury manometer and is considered to be the gold standard for blood pressure measurement because it cannot be in error if operated correctly. However, its main disadvantage is that it is unable to provide a continuous reading of pressure variations, and its update rate is limited. In addition, only the systolic and diastolic arterial pressures can be measured. The operational principles of the sphygmomanometer are quite simple. A cuff is placed around the upper arm and inflated, and arterial blood can flow past it only when the arterial pressure exceeds that of the cuff. At this stage, blood squeezing through the reduced aperture of the brachial artery generates turbulence that can be heard by a practitioner with a stethoscope. As the pressure in the cuff is reduced and monitored using a manometer, the first indication of this sound (called Korotkoff sounds) gives an indication of the systolic pressure. This reduction in cuff pressure continues until Korotkoff sounds disappear, and that is an indication of the diastolic pressure.

Direct pressure measurement devices consist of a chamber with a flexible diaphragm making up a portion of one wall with the other side of the diaphragm at atmospheric pressure. A pressure differential across the diaphragm will cause it to deflect, and this deflection can be measured using a displacement sensor. Capacitance was the method of choice for measuring small displacements. If a second plate was placed within a few hundred microns of the diaphragm, changes in the relative separation resulted in changes in capacitance, as discussed earlier in this chapter.

Today, such small displacements are generally measured using strain gauges in a bridge configuration. In the past, a mechanical assembly comprising the diaphragm, a force rod, and four unbonded strain gauges was used. With increasing deflection, two of the strain gauges become progressively more relaxed and two are stretched (Gregory, 1975). Modern pressure transducers are generally manufactured onto a single MEMS silicon chip where a portion of the chip is formed into a diaphragm and semiconductor strain gauges (piezoresistive bridge) are incorporated directly onto the diaphragm to form small inexpensive pressure sensors, as shown in Figure 2-55. Such sensors are sufficiently low cost, so they can be used as disposable, single-use devices for measuring blood pressure without need for additional sterilization (Bronzino, 2006).

Wire strain gauges usually have gauge factors of between two and four, while the semiconductor based units have gauge factors of between 50 and 200. For silicon, the gauge factor is typically 120. However, in measuring pressure sensitivities are generally quoted in microvolts per volt (applied to the strain gauge) per millimeter of mercury (Cromwell, Weibell et al., 1973). Signal conditioning and display instruments for these sensors consist of a method of exciting the strain gauge bridge, a method of zeroing or balancing it, followed by an amplifier and a display device or data logger.

**FIGURE 2-55 ■**  
MEMS strain gauge pressure transducer.  
(a) mechanical layout. (b) Circuit diagram.



In the pressure transducer shown in Figure 2-55, the total force,  $F$  (N), applied to the membrane is equal to the product of differential pressure,  $\Delta P$  (Pa), and the area of the membrane,  $A$  ( $\text{m}^2$ )

$$F = \Delta P A \quad (2.54)$$

The stress introduced in the bridge elements is directly proportional to the applied force, and its effect on the semiconductor resistors is to alter their resistance by  $\Delta R$

$$\frac{\Delta R}{R} = (\alpha_l \sigma_l + \alpha_t \sigma_t) \quad (2.55)$$

where  $R(\Omega)$  is the initial resistance,  $\alpha_l$  and  $\alpha_t$  are the piezoresistive coefficients in the longitudinal and transverse directions, respectively, and  $\sigma_l$  and  $\sigma_t$  are the stresses in these directions, respectively (Fraden, 1996).

The magnitudes of these coefficients depends on the orientation of the silicon crystal, and the resistors are positioned on the diaphragm in such a manner (as shown in Figure 2-55) as to have the longitudinal and transverse coefficients of opposite polarities. This results in the resistances changing in opposite directions as the pressure alters.

$$\alpha_l = -\alpha_t = \frac{1}{2}\alpha \quad (2.56)$$

Therefore

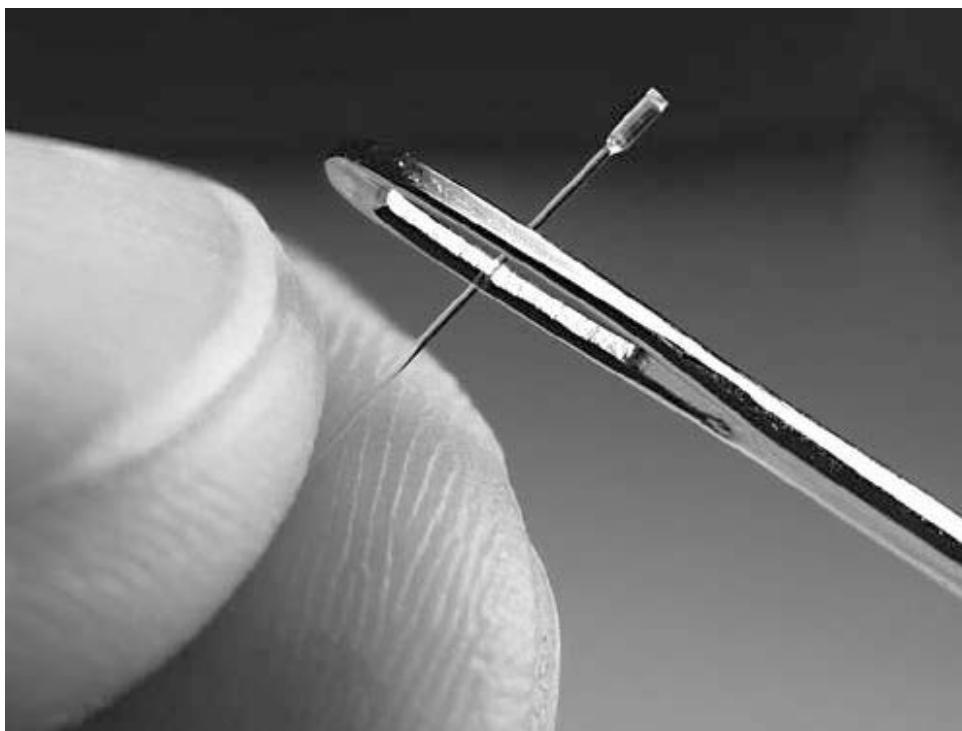
$$\frac{\Delta R_1}{R_1} = -\frac{\Delta R_2}{R_2} = \frac{1}{2}\alpha(\sigma_{1y} - \sigma_{1x}) \quad (2.57)$$

where  $\sigma_{1x}$  and  $\sigma_{1y}$  are the longitudinal stresses in the  $x$  and  $y$  directions, respectively.

For the full-bridge configuration shown in Figure 2-55 excited by a voltage  $V_{ref}$ , the output voltage will be

$$V_{out} = \frac{V_{ref}}{2}\alpha(\sigma_{1y} - \sigma_{1x}) \quad (2.58)$$

Because  $\alpha$  is a function of temperature, the output voltage will change with temperature. The substrate temperature must therefore be measured and the appropriate corrections applied.



**FIGURE 2-56 ■**  
Fiber optic pressure sensor from FISO Technologies employs a MEMS-based sensing catheter for in vivo physiological measurements.  
(Courtesy <http://www.fiso.com/>.)

For measuring blood pressure, the chamber containing the diaphragm is coupled via a thin plastic catheter to an artery. The catheter is filled with saline solution so that the arterial blood pressure is coupled directly to the diaphragm outside the patient. However, in this configuration it is important to realize that both ends of the catheter must be at the same height to avoid hydrostatic effects and that the tube must be sufficiently stiff to minimize compliance effects on the frequency response of the sensor. Air bubbles in the catheter and obstructions due to clotted blood or other materials can introduce distortions into the measured waveform due to resonance and damping effects.

It is now possible to obtain miniature pressure sensors that are located at the tip of the catheter and to measure pressure within the blood vessel (Allen, 2002; Cromwell, Weibell et al., 1973). As illustrated in Figure 2-56, these devices are sufficiently small to be inserted into an artery or a vein through the needle of a hypodermic syringe, and because of their small size they have an insignificant effect on the pressure or flow at the measurement site.

Long-term stability of pressure monitors is a problem, and drift, particularly when measuring low pressures such as venous blood or cerebrospinal fluid, can result in significant errors. Pressure transducers need regular recalibration, which can pose significant problems, particularly if the sensor is implanted.

Table 2-8 shows a comparison of the properties of some pressure sensing technologies.

Measurement of respiratory processes is generally achieved by measuring expired volume using a spirometer or flow rate using a pneumotachograph, as discussed in Chapter 10. However, air pressure is measured as part of the operation of negative and positive pressure respiratory devices. Sensors for measuring air pressure in these devices need not be miniaturized but must be capable of measuring pressures from less than 2 cm of H<sub>2</sub>O up to 40 cm H<sub>2</sub>O to assist or provide full ventilation.

**TABLE 2-8** ■ Comparison of Properties of Some Pressure Sensing Technologies

Property	Capacitive	Resistive	Piezoelectric
Maximum range	Good	Excellent	Good
Minimum size	Good	Excellent	Poor
Sensitivity	Excellent	Poor	Good
Repeatability	Excellent	Good	Good
Temperature stability	Excellent	Excellent	Poor
Hysteresis	Low	High	Medium
Installation	Easy	Hard	Easy

Source: Pons, J. (Ed.), *Wearable Robots—Biomechatronic Exoskeletons*, Chichester, UK: John Wiley & Sons, 2008.

Differential pressure sensors are often used to measure air flow using the obstructive-flow technique employed by some pneumotachographs, as discussed in the following section. A good example of such a sensor is the SM5822 manufactured by Silicon Microstructures, which is designed to operate at pressure ranges from 0 to 0.6 PSI (0 to 42 cm H<sub>2</sub>O) full scale.

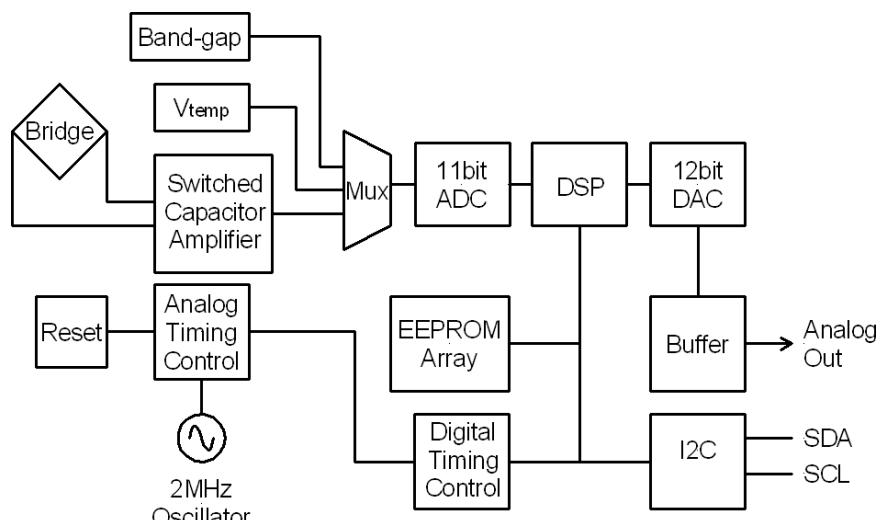
As shown in Figure 2-57, the piezoresistive bridge is only the beginning of the pressure measurement process. The bridge output is filtered and then digitized along with the temperature and a band-gap reference voltage. These parameters are then processed by the onboard digital signal processing (DSP) to correct for pressure and temperature nonlinearities before conversion back to an analog voltage and output. An I<sup>2</sup>C digital interface is also provided.

Finally, as with all sensors in contact with body fluids or inserted within the body, pressure sensors must remain inert and mechanically stable within corrosive environments.

### 2.4.12 Sound Pressure

Microphones play an essential role in making hearing aids and cochlear implants possible by converting acoustic signals propagating in the air as variations in sound pressure

**FIGURE 2-57** ■ Block diagram of the SM5822 pressure sensor processing structure (Courtesy Silicon Microstructures <http://www.si-micro.com/>.)



into electrical signals. At higher frequencies, *microphones*, commonly known as ultrasound transducers, are used in sonar devices for sensory substitution systems discussed in Chapter 7 and in medical Doppler and imaging systems.

A large variety of physical mechanisms can be harnessed to perform the microphone function. These include electromagnetic induction discussed in Chapter 3 as well as optical interferometry and the piezoelectric effect discussed earlier in this chapter. However, since the advent of MEMSs, millions of low-cost microphones have been produced that use changes in the capacitance between a fixed plate and a vibrating diaphragm to produce a varying electrical output.

### 2.4.13 Flow

The most common medical applications of flow measurement include blood flow through arteries and veins and airflow into and out of the lungs. It also includes the measurement of drug dispensing through catheters and gas flow during anesthetics, among others. A great number of different sensors can be used to measure flow by determining the rate of displacement of either mass or volume. However, whichever sensor is used, inherent difficulties in the measurement make the process complicated, and it is necessary to consider the natural characteristics of both the environment and the medium (e.g., pipe shapes and materials; temperature, pressure, and viscosity of the fluid; Fraden, 1996).

#### 2.4.13.1 Differential Pressure Flowmeter

Flow is commonly measured using a differential pressure flowmeter across an orifice plate, a Venturi tube, or a nozzle. The operation of these flowmeters is based on an observation by Bernoulli that if an annular restriction was placed in a pipe, then the velocity of the fluid through the restriction was increased. The increase in velocity at the restriction then causes the static pressure to decrease, and a pressure difference is built up. The difference between the pressure upstream and the pressure downstream of the restriction is related to the rate of flow through it and therefore through the pipe. A differential pressure-based flow transducer consists of an obstruction to cause a pressure drop and a method of measuring the differential pressure across the obstruction (Webster, 1999).

Bernoulli's equation, which defines the relationship between fluid velocity,  $v$ , fluid pressure,  $p$ , and height,  $h$ , above some fixed point for fluid flowing through a pipe of varying cross section is the fundamental to quantifying the differential pressure measurement technique.

For the inclined, tapered pipe shown in Figure 2-58, Bernoulli's equation states that

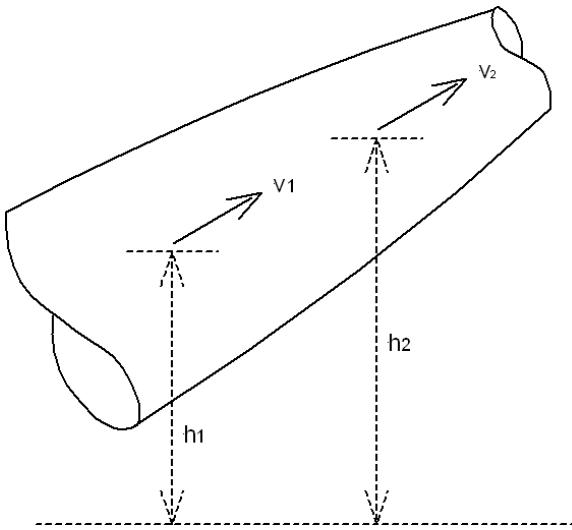
$$\frac{p_1}{\rho g} + \frac{v_1^2}{2g} + h_1 = \frac{p_2}{\rho g} + \frac{v_2^2}{2g} + h_2 \quad (2.59)$$

The sum of the pressure head, the velocity head, and the potential head is constant along a flow streamline. It is assumed that the flow is frictionless (no viscosity) and of constant density (incompressible).

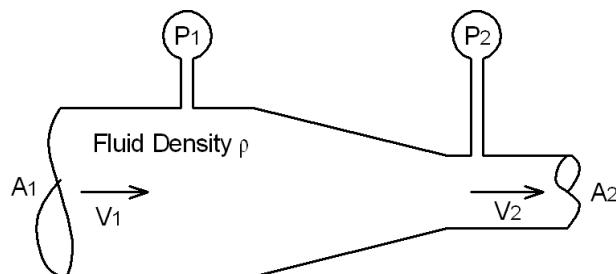
If there is a restriction in a piece of pipe as shown in Figure 2-59, then since  $h_1 = h_2$  Bernoulli's equation can be rewritten as

$$\frac{p_1 - p_2}{\rho} = \frac{v_1^2 - v_2^2}{2} \quad (2.60)$$

**FIGURE 2-58 ■**  
Flow through an inclined tapered pipe.



**FIGURE 2-59 ■**  
Using a restriction in a piece of pipe to measure flow.



and the conservation of mass requires that

$$v_1 A_1 \rho = v_2 A_2 \rho \quad (2.61)$$

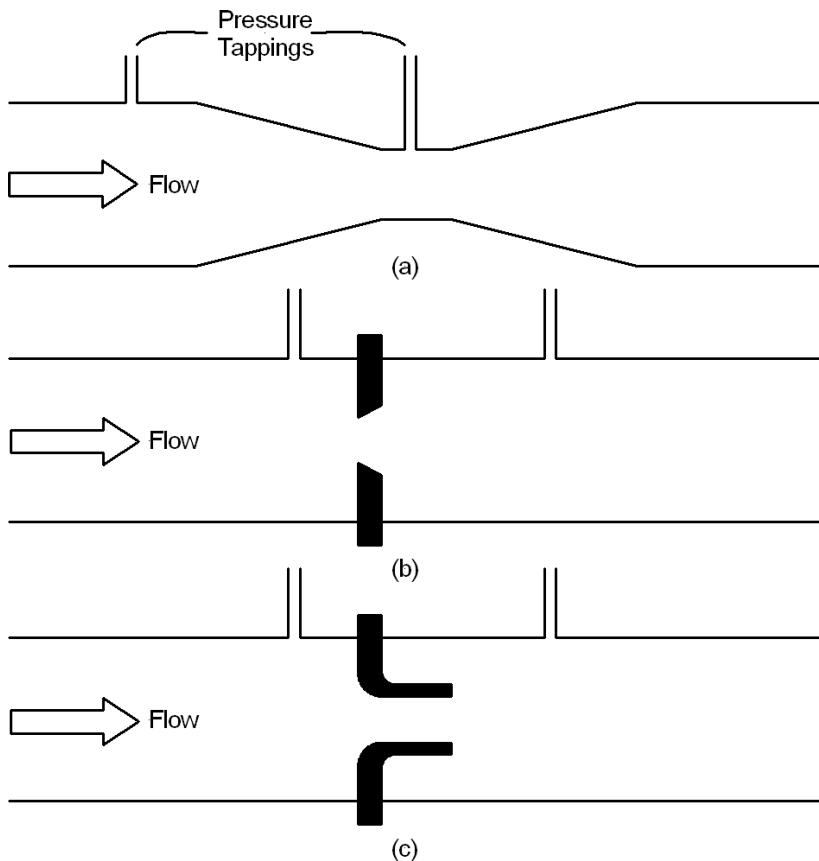
Rearranging and substituting for  $v_2$  in equation (2.61) gives the volumetric flow rate,  $Q$ , in terms of the drop in pressure across the restriction in the pipeline

$$Q = v_1 A_1 = \frac{A_2}{\sqrt{1 - \left(\frac{A_2}{A_1}\right)^2}} \sqrt{\frac{2(p_1 - p_2)}{\rho}} \quad (2.62)$$

The assumption of incompressibility is reasonable for a liquid but not a gas, but the assumption that the fluid has no viscosity, resulting in a flat velocity profile, is generally not justified. This requires the application of correction factors to compensate.

The orifice plate, shown in Figure 2-60b, is the simplest and cheapest type of differential pressure flowmeter. It is simply a plate with a hole of the specified diameter that is clamped into the pipe. Unfortunately, after passing through the plate the jet continues to contract until it reaches a minimum diameter before expanding again. Because  $A_2$  should be this diameter, which is unknown, correction factors must be applied to equation (2.62). This modified equation is

$$Q = \frac{C}{\sqrt{1 - \beta^4}} \epsilon \frac{\pi}{4} d^2 \sqrt{\frac{2(p_1 - p_2)}{\rho}} \quad (2.63)$$



**FIGURE 2-60 ■**  
Common differential pressure flowmeters.  
(a) Venturi. (b) Orifice plate. (c) Nozzle.

where  $\rho$  is the density of the fluid upstream from the plate,  $d$  is the diameter of the hole in the orifice plate, and  $\beta = d/D$ , where  $D$  is the upstream diameter of the interior of the pipe. The empirically determined correction factors are  $C$ , which is affected by, for example, diameter ratio, Reynolds number, and pipe roughness, and  $\varepsilon$ , which is the expansibility factor.

The Venturi tube, shown in Figure 2-60a, is the oldest type of differential pressure flowmeter. Because the change in diameter is more gradual, this better approximates the theoretical result,  $C$ , which in this case is close to 0.95. Its major disadvantages are the lower differential pressure for a given diameter ratio and the expense of manufacture.

The nozzle method shown in Figure 2-60c combines the best features of the other two techniques, making it both reasonably cheap to manufacture and accurate. Pneumotachographs, discussed in Chapter 10, use a variation of the nozzle method in which a restrictive element made from a perforated plate, or a large number of narrow-bore pipes, is used. These offer very little resistance to the airflow, and hence the pressure drop is small.

It follows from equation (2.63) that the pressure gradient technique requires the use of either a single differential pressure sensor or a pair of absolute sensors. Additionally, because the flow rate is proportional to the square root of the pressure difference, if a linear representation of the flow is required then a square root extraction circuit or algorithm must be applied.

The primary advantage of this technique is the absence of any moving components and the ready availability of standard pressure sensors. The main disadvantage is the requirement that the flow may be affected by the use of a restrictive device.

### 2.4.13.2 Temperature Flowmeters

An improved method of measuring flow would be to mark the fluid in some way and then to detect the movement of the mark. Marking techniques could be as simple as mechanical floats or as complicated as radioactive elements or a dye that changes the optical properties of the medium. In medicine, the dye dilution method is used for studies in hemodynamics. Under most circumstances, however, placing any foreign material into the fluid is impractical or forbidden.

A noninvasive marker is temperature, and sensors called hot-wire or thermoanemometers use this principle. One sensor type often used to measure blood flow consists of a small isolated heating element immersed in the fluid between two similarly immersed temperature probes. If the medium is not flowing, then diffusion will result in the two probes reading the same temperature, but if there is flow, the downstream probe will be warmer than the upstream one. In general, the heating element is placed closer to the downstream element to improve sensitivity, and it must always be heated to above the highest temperature reached by the fluid under normal conditions. In another method, a thermistor placed in the blood stream is kept at a constant temperature using current feedback. The amount of energy required to maintain the temperature is proportional to the flow rate because the temperature of the blood and its thermal conductivity remain constant (Cromwell, Weibell et al., 1973).

### 2.4.13.3 Ultrasound Flowmeters

Another nonintrusive method to measure flow is ultrasound. Sensors can measure changes in the effective velocity in the medium, or they can use the Doppler shift. In Figure 2-61, two transducers are placed on either side of a pipe directed through the fluid toward each other. If a pulse is transmitted from one, the time,  $T$  (s), taken to reach the other will be

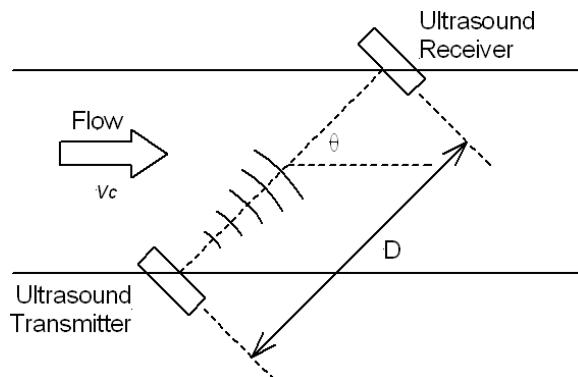
$$T = D/c \quad (2.64)$$

where  $D$  (m) is the distance between the transducers, and  $c$  (m/s) is the speed of sound in the medium if the fluid is still.

If the fluid is flowing, then the propagation speed is altered by the flow, and equation (2.64) becomes

$$T = \frac{D}{c \pm v_c \cos \theta} \quad (2.65)$$

**FIGURE 2-61** ■  
Ultrasonic flowmeter.



where  $v_c$  (m/s) is the average fluid velocity, and the  $\pm$  refers to the direction of flow relative to the direction of the ultrasound signal.

It has been shown that for laminar flow  $v_c \approx 4v_a/3$  and for turbulent flow  $v_c \approx 1.07v_a$ , where  $v_a$  (m/s) is the flow averaged over the cross sectional area (Fraden, 1996).

By taking the time difference,  $\Delta T$ , between the downstream and upstream measurements

$$\Delta T = \frac{2Dv_c \cos \theta}{c^2 + v_c \cos^2 \theta} \approx \frac{2Dv_c \cos \theta}{c^2} \quad (2.66)$$

which is true for most situations because  $c \gg v_c \cos \theta$ .

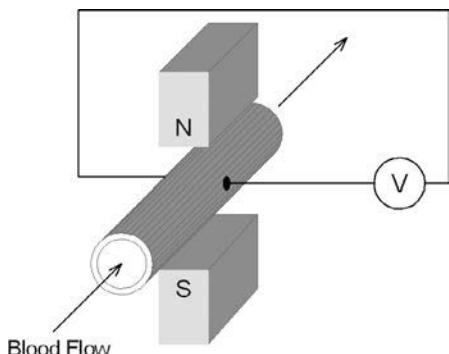
Typically, sensors of this type transmit short pulses at high frequency ( $f = 3$  MHz) and alternate between the transmitter being the upstream and the downstream transducer. An alternative is to measure the phase difference between the transmitted and the received signals.

#### 2.4.13.4 Magnetic Flowmeters

Magnetic flowmeters are based on the principle of magnetic induction. When an electrical conductor cuts through a magnetic field, a voltage in the conductor is induced proportional to its velocity. This principle can be applied when the conductor is a column of conductive fluid flowing through a tube in a magnetic field. For measuring blood flow, a magnet that generates a magnetic field orthogonal to the direction of flow is placed around the blood vessel. Electrodes measure the voltage on either side of the blood vessel orthogonal to both the magnetic field and the direction of flow. The measured voltage is directly proportional to the flow rate so long as the conductivity of the blood remains constant. Figure 2-62 shows an example of such a flow measurement sensor.

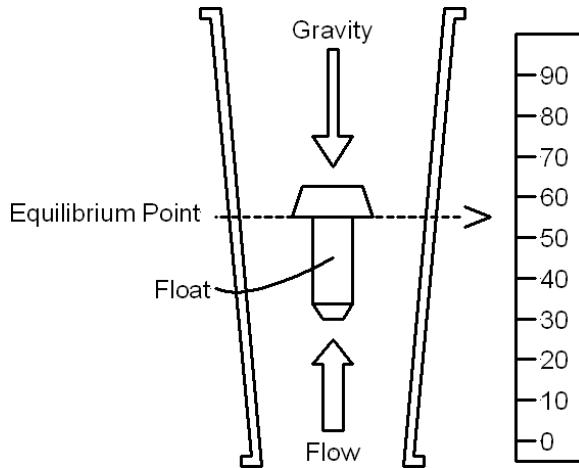
#### 2.4.13.5 Target Flowmeters

Target meters measure the force caused by liquid impacting on a target or drag disk suspended in the liquid stream. A direct indication of the liquid flow rate is obtained by measuring the force exerted on the target. In its simplest form, the meter consists only of a hinged, swinging plate that moves outward, along with the liquid stream. A more sophisticated version uses a precision force sensing element. For example, the small displacement of the target caused by the liquid flow can be sensed by a strain gauge. The output signal from the gauge is indicative of the flow rate. Target meters are useful for measuring flows of dirty or corrosive liquids.



**FIGURE 2-62** ■  
Schematic diagram  
of a magnetic blood  
flow sensor.

**FIGURE 2-63 ■**  
Flow measurement  
using a variable-area  
flowmeter  
(rotameter).



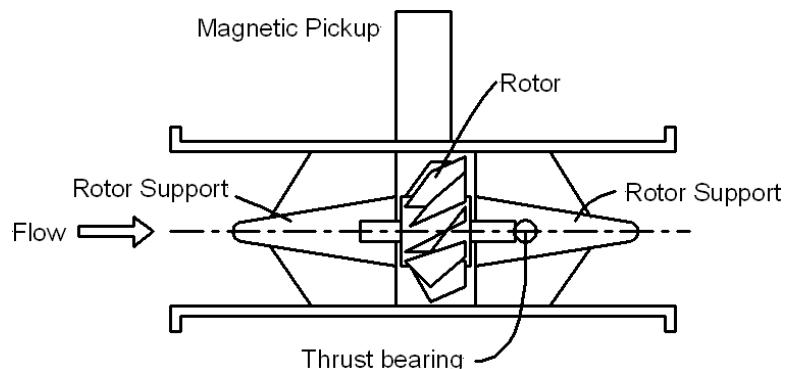
Variable-area meters, often called rotameters, consist of a tapered tube and a float, as shown in Figure 2-63. Although classified as differential pressure units, they are, in reality, constant differential pressure devices. Flanged-end fittings provide an easy means for installing them in pipes. When there is no liquid flow, the float rests freely at the bottom of the tube. As liquid enters the bottom of the tube, the float begins to rise. The position of the float varies directly with the flow rate. Its exact position is at the point where the differential pressure between the upper and lower surfaces balances the weight of the float.

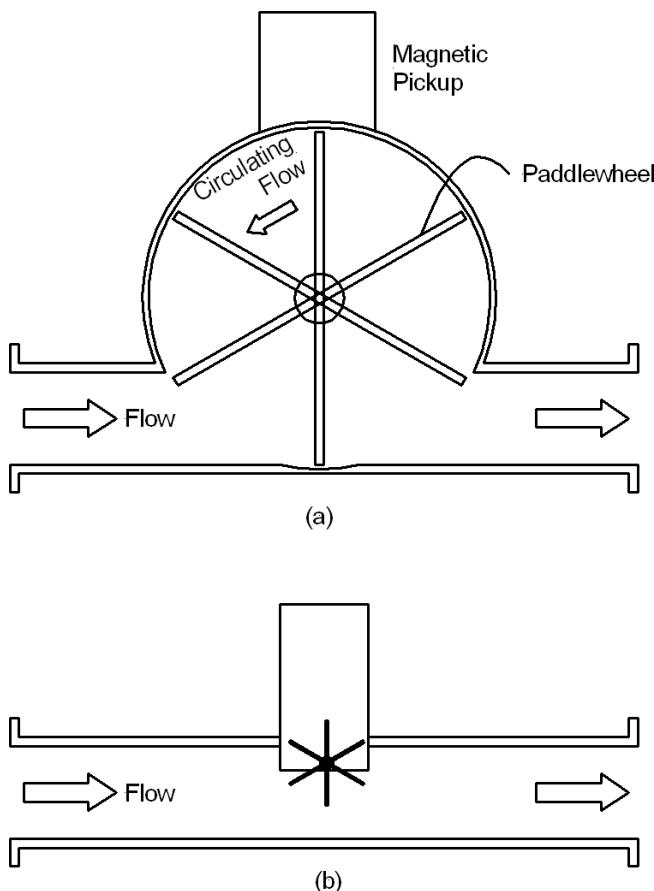
Because the flow rate can be read directly on a scale mounted next to the tube, no secondary flow-reading devices are necessary. However, if desired, a sensor can be used to measure the float's level and transmit a flow signal.

#### 2.4.13.6 Turbine Flowmeters

Many medical appliances still use the basic turbine or vane to measure flow because if properly installed and calibrated they provide the highest accuracies attainable for any currently available flowmeter for both liquids and gases. The unit consists of a multiple-bladed rotor mounted with a pipe perpendicular to the liquid flow. The rotor spins as the liquid passes through the blades, as shown in Figure 2-64. The rotational speed is a direct function of flow rate and can be sensed by magnetic pickup, photoelectric cell, or gears. Alternatively, a tachometer can be used. The number of revolutions or electrical pulses counted for a given period of time is directly proportional to flow volume. Turbine meters,

**FIGURE 2-64 ■**  
Turbine flowmeter.





**FIGURE 2-65 ■**  
Diagram of two impeller flowmeters.  
(a) In-line.  
(b) Insertion.

when properly specified and installed, have good accuracy, particularly with low-viscosity liquids.

A major concern with turbine meters is bearing wear. Bearingless options have been developed using the liquid itself for lubrication. These are particularly important in medical applications, particularly implanted ventricular-assist pumps, discussed in Chapter 9.

#### 2.4.13.7 Impeller Flowmeters

Impeller flowmeters, also called paddle-wheel meters, are one of the more commonly used types. They are a direct offshoot of the old undershot water wheels. In-line meters are constructed as a unit that includes inlet and outlet orifices, whereas the insertion type can be installed into existing pipes through a round hole, as shown in Figure 2-65.

In-line meters are more sensitive than axial turbine flowmeters at low flow rates because the blade incidence angle is much larger. They are also relatively insensitive to the flow regime (laminar or turbulent).

As with the turbine meters, paddle rotation is usually sensed using a magnetic pickup, and the flow rate is determined by measuring the pulsation rate.

#### 2.4.14 Temperature Sensors

The measurement of temperature is probably one of the most common sensing requirements in the medical field. The human organism can function effectively over only a small

range of temperatures, so almost all processes associated with the organism are temperature controlled. These extend from incubators for premature babies through heart–lung machines and even respirators.

Mechanical measurement of temperature relies on the expansion of the working material. This can be a liquid in the case of a normal thermometer or a solid in the case of a bimetallic strip.

#### 2.4.14.1 Bimetallic Thermostats

It can be shown (Webster, 1999) that the radius of curvature of a uniformly heated bimetallic strip will change from a radius  $R_o$  to  $R$  (m) with a temperature change from  $T_o$  to  $T$  ( $^{\circ}$ C) is

$$\frac{1}{R} - \frac{1}{R_o} = \frac{6(1+m)^2(\alpha_2 - \alpha_1)(T - T_o)}{t [3(1+m)^2 + (1+mn)(m^2 + 1/mn)]} \quad (2.67)$$

where

$1/R_o$  initial curvature of the strip at temperature  $T_o$

$\alpha_1$  and  $\alpha_2$  coefficients of expansion of the two materials  $\alpha_1 > \alpha_2$

$n$   $E_1/E_2$  respective Young's moduli of the two materials

$m$   $t_1/t_2$  respective thicknesses of the two materials

$t$   $t_1 + t_2$  total thickness of the strip

It can be seen that the ratio of the Young's moduli hardly makes any difference; therefore,  $n = 1$ . In addition, most industrial bimetallic strips are made from strips in which the materials are the same thickness; therefore,  $m = 1$ . Finally, if the strip starts out flat, then

$$\frac{1}{R} = \frac{3(\alpha_2 - \alpha_1)(T - T_o)}{2t} \quad (2.68)$$

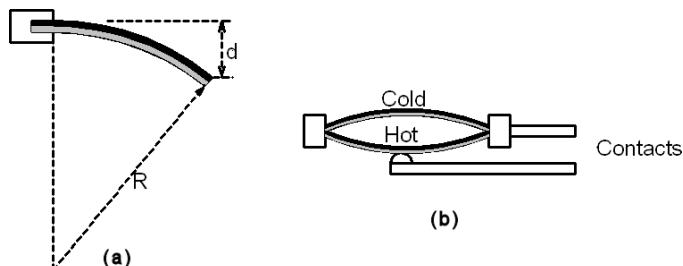
Bimetallic strips are mostly used in thermostats, which are in turn used for temperature control by switching heaters or refrigeration plants on or off as illustrated in Figure 2-66.

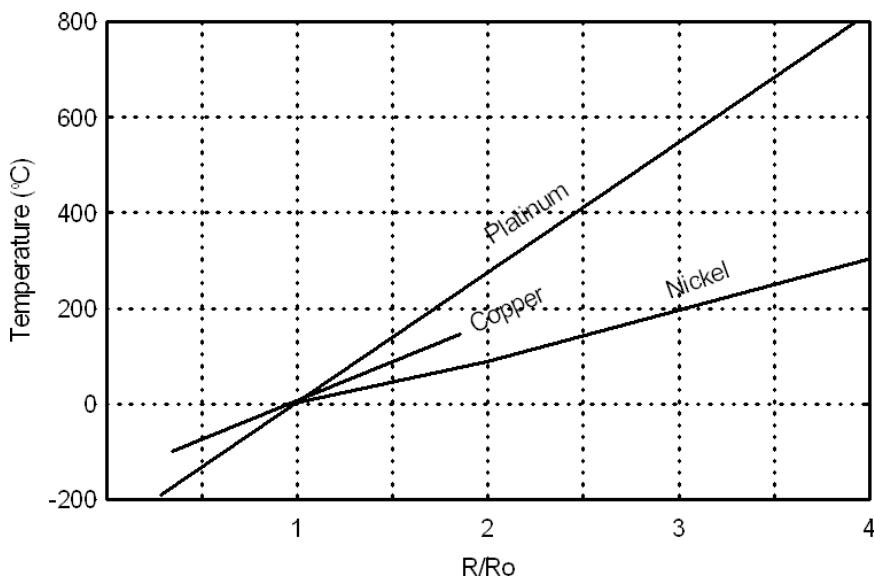
#### 2.4.14.2 Resistance Temperature Detector

Most other thermometer types are based on materials whose electrical properties change with temperature as shown in Figure 2-67. Possibly the most accurate of these are resistance temperature detectors (RTD). Standard platinum RTDs can be manufactured with accuracies of  $+/- 0.0001$   $^{\circ}$ C, whereas simple industrial RTDs are generally accurate to about  $+/- 0.1$   $^{\circ}$ C.

The sensitive portion of an RTD is a coil of high-purity wire or a thin film deposited onto a ceramic substrate. The metal used is usually platinum, copper, or nickel depending

**FIGURE 2-66** ■  
Bimetallic  
thermostat.  
(a) Cantilever  
operation. (b) Snap  
action switch.





**FIGURE 2-67** ■ Resistance characteristics of a number of different RTDs.

on the temperature range and accuracy requirements. Platinum is the material of choice because it doesn't oxidize even when subjected to very high temperatures.

In operation, a constant current (0.8 mA to 1 mA typically) is passed through the coil. As the temperature increases, the resistance will increase (the temperature coefficient is positive), and the voltage that is developed across the coil will reflect this. An accurate voltmeter calibrated to measure temperature can be used to display this value. Most modern temperature sensors use an accurate analog-to-digital converter to digitize the voltage so that it can be further processed (curve fitting) and then incorporated into some control or display application.

Platinum RTDs can be operated from about  $-184^{\circ}\text{C}$  to  $+649^{\circ}\text{C}$ . They have good temperature to resistance linearity, good repeatability, and long life. Nickel can be used over a smaller range because the relationship between temperature and resistance becomes very nonlinear at temperatures above  $300^{\circ}\text{C}$ . It is, however, the most sensitive of the commonly used materials. Copper has a very linear resistance to temperature relationship, but it oxidizes at moderate temperatures and thus cannot be used above  $150^{\circ}\text{C}$ .

The temperature coefficient of resistance,  $\alpha$ , of copper is 0.00427, that of nickel is 0.00672, and that of platinum is 0.003902, where  $\alpha$  is defined as the change in resistance per degree C per ohm over the range from  $0^{\circ}\text{C}$  to  $100^{\circ}\text{C}$

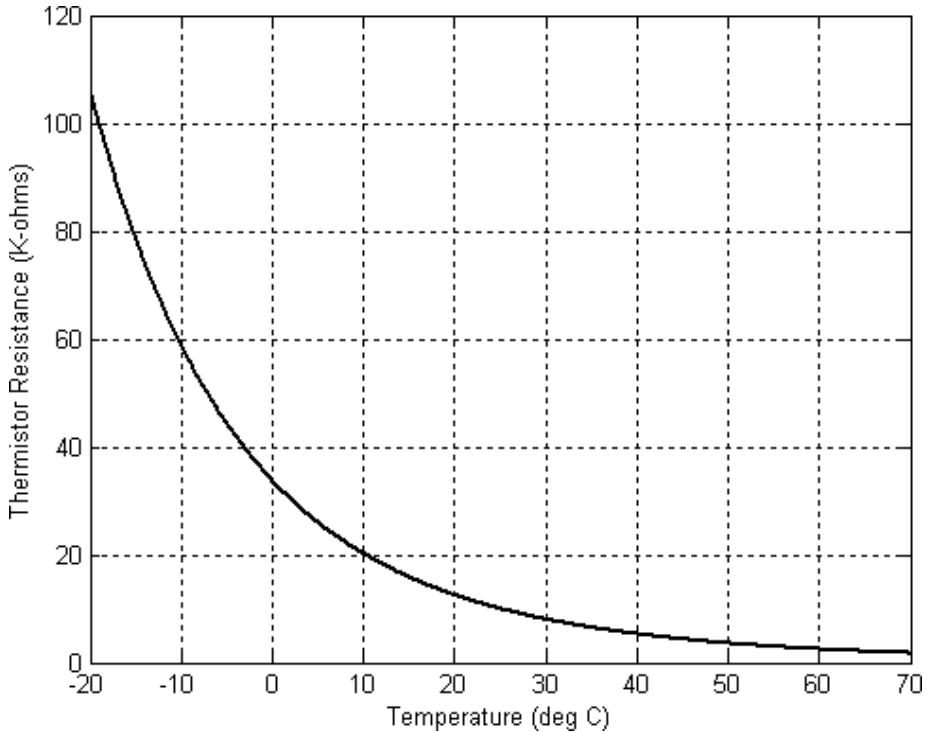
$$\alpha = \frac{R_{100} - R_0}{100R_0} \quad (2.69)$$

#### 2.4.14.3 Thermistors

Thermistors are generally semiconductor materials that exhibit a higher temperature coefficient of resistance than pure metals, but their linearity is generally worse. Silicon positive temperature coefficient (PTC) thermistors rely on the bulk properties of doped silicon and have temperature coefficients of between 0.07 and 0.08. These are often built into electronics for temperature compensation.

Negative temperature coefficient (NTC) thermistors are made from metal oxides, and they exhibit a monotonic decrease in resistance with increasing temperature. They are very nonlinear, as can be seen from Figure 2-68 (Maxim, 2001).

**FIGURE 2-68** ■ Resistance characteristics of a typical NTC thermistor ( $R_{25} = 10 \text{ k}\Omega, \beta = 3965 \text{ K}$ )



The standard formula that describes the resistance of a NTC thermistor as a function of temperature is

$$R_T = R_{25} e^{\left\{ \beta \left[ \frac{1}{T+273} - \frac{1}{298} \right] \right\}} \quad (2.70)$$

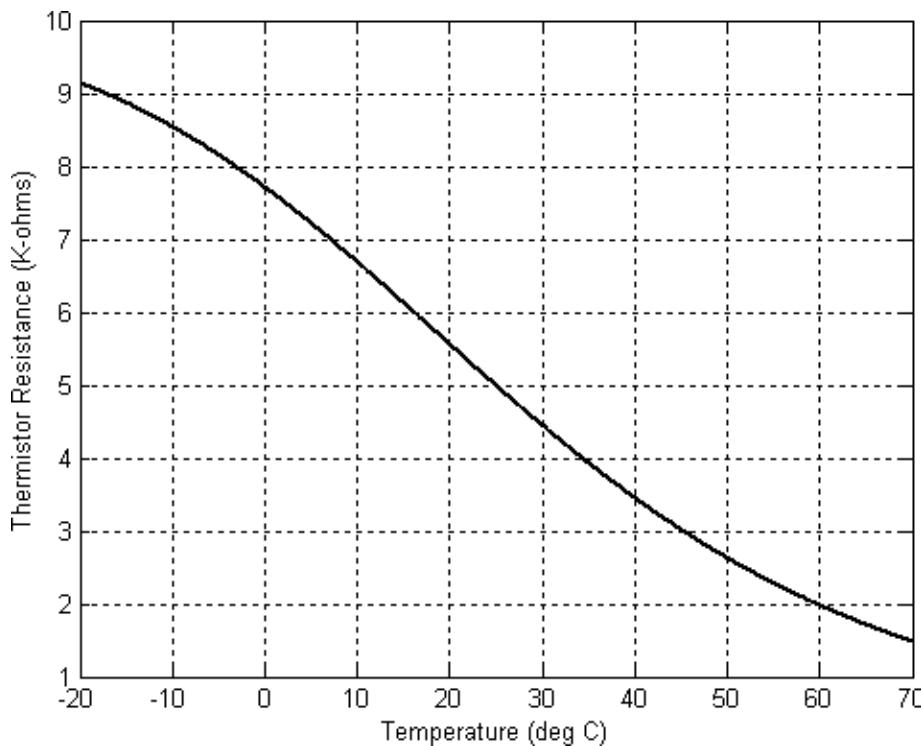
where  $R_{25}$  is the resistance at  $25^\circ\text{C}$ ,  $\beta$  is the thermistor's material constant (K), and  $T$  is the actual temperature of the thermistor ( $^\circ\text{C}$ ).  $R_{25}$  and  $\beta$  are generally published in the manufacturer's data sheet;  $R_{25}$  can range from  $22 \Omega$  to  $500 \text{ k}\Omega$  and  $\beta$  typically from 2500 to 5000 K.

Linearization of this function can be achieved using resistance- or voltage-mode techniques. Resistance-mode linearization involves placing a normal resistor parallel to the thermistor. If the resistance is chosen to equal  $R_{25}$ , then the region of relatively linear resistance operation will be symmetrical around room temperature, as shown in Figure 2-69.

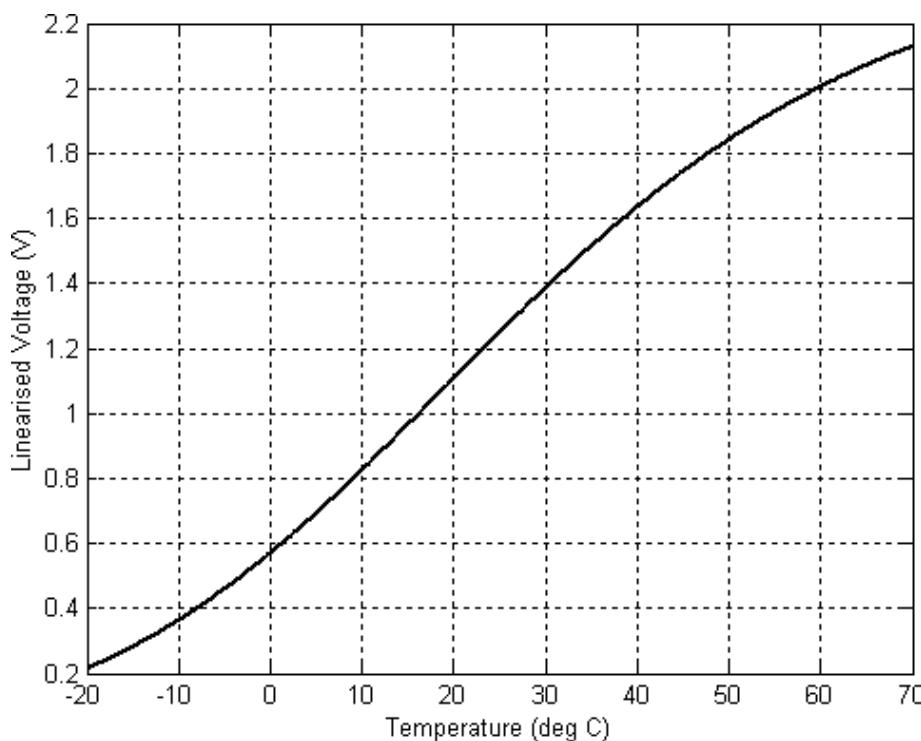
In voltage-mode linearization, the thermistor is connected in series with a normal resistor to form a voltage divider. For a fixed bias voltage and a resistor equal to  $R_{25}$ , the region of linear voltage will be symmetrical around room temperature. The results of this linearization process are shown in Figure 2-70.

#### 2.4.14.4 Semiconductor Sensors

A semiconductor PN junction of a diode or transistor exhibits quite a strong temperature dependence. If the junction is connected to a constant current source, the voltage becomes a measure of the junction temperature. This relationship is very linear and is therefore used for accurate three-terminal temperature-sensing integrated circuits.



**FIGURE 2-69** ■  
Resistance-mode  
linearization of a  
thermistor.



**FIGURE 2-70** ■  
Voltage-mode  
linearization of a  
thermistor.

The LM35Z temperature sensor has a linear output internally trimmed for the Celsius scale with a sensitivity of 10 mV per  $^{\circ}\text{C}$  and a nonlinearity error confined within  $+/- 0.1^{\circ}\text{C}$ . The output can therefore be modeled as

$$V_{out} = V_0 + \alpha T \quad (2.71)$$

where  $T$  ( $^{\circ}\text{C}$ ) is the temperature, and  $V_0$  (volts) should equal zero but can be as large a 10 mV resulting in an offset error of up to  $1^{\circ}\text{C}$ . The slope  $\alpha$  (mV per  $^{\circ}\text{C}$ ) can vary between 9.9 and 10.1.

#### 2.4.14.5 Thermocouples

A thermocouple consists of a combination of two different materials bonded together which will generate a potential difference,  $\Delta V$ , proportional to the temperature difference,  $\Delta T$ , between the hot and cold terminals. This characteristic is known as the Seebeck effect (after Thomas Seebeck) and can be mathematically expressed as

$$\Delta V = \alpha_s \Delta T, \quad (2.72)$$

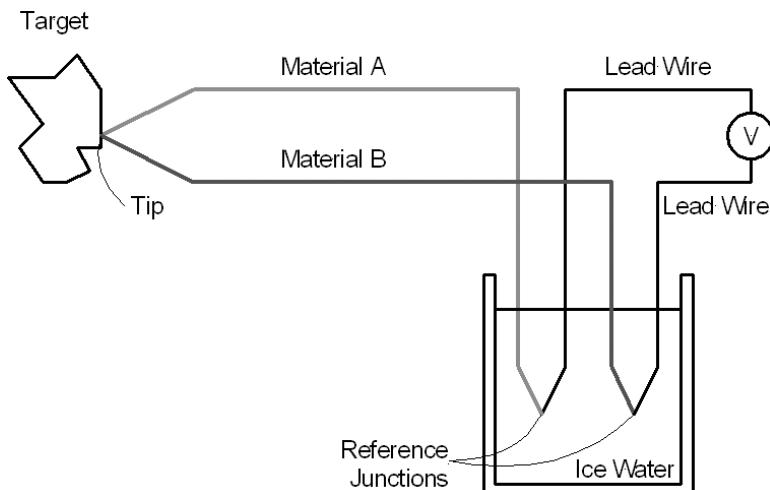
where  $\alpha_s$  is the Seebeck or thermoelectric coefficient, and its value depends on the difference between the thermoelectric coefficients of the two materials used.

A typical thermocouple consists of wires made from the two materials and a method of measuring the potential difference as shown in Figure 2-71. The designated reference temperature  $T_{ref}$  is usually supplied by an ice-water bath in a laboratory, but in the field that is not practical so the ambient temperature is used as the reference. Where the reference junction cannot be held at  $0^{\circ}\text{C}$ , the observed value must be adjusted by the temperature difference.

If the thermoelectric coefficients,  $\alpha_{SA}$  and  $\alpha_{SB}$ , are nearly constant across the targeted temperature range, then the temperature at the probe tip  $T_{tip}$  can be determined from the reference temperature  $T_{ref}$  and the difference between the coefficients of the two materials.

$$V_{out} = (\alpha_{SA} - \alpha_{SB})(T_{tip} - T_{ref}) \quad (2.73)$$

**FIGURE 2-71** ■ Single temperature reference junction thermocouple.



**TABLE 2-9** ■ Thermoelectric Coefficients of Some Common Materials at 0 °C

Material	Thermoelectric Coefficient, $\alpha_S(\mu\text{V K}^{-1})$	Material	Thermoelectric Coefficient, $\alpha_S(\mu\text{V K}^{-1})$
Aluminium	3.5	Nichrome	25
Antimony	47	Nickel	-15
Bismuth	-72	Platinum	0
Cadmium	7.5	Potassium	-9
Carbon	3.0	Rhodium	6
Constantan	-35	Selenium	900
Copper	6.5	Silicon	440
Germanium	300	Silver	6.5
Gold	6.5	Sodium	-2
Iron	19	Tantalum	4.5
Lead	4	Tellurium	500
Mercury	0.6	Tungsten	7.5

Therefore, the temperature of the probe tip can be determined as

$$T_{tip} = T_{ref} + \frac{V_{out}}{\alpha_{SA} - \alpha_{SB}} \quad (2.74)$$

In practice, manufacturers generally provide calibration functions for their products. These functions are usually high-order polynomials and are calibrated with respect to a certain reference temperature. Suppose that the coefficients of the calibration polynomials are  $a_0, a_1, a_2, \dots, a_n$ . The temperature at the probe tip can then be related to the voltage output as

$$T_{tip} = a_0 + a_1 V_{out} + a_2 V_{out}^2 + \dots + a_n V_{out}^n \quad (2.75)$$

Note that equation (2.75) is effective only if the reference temperature,  $T_{Ref}$ , in the measurement remains equal to the reference temperature specified on the data sheet.

Materials should have a high thermoelectric coefficient, low thermal conductivity, and low resistivity. Unfortunately, as shown in Table 2-9, materials like gold, silver, and copper that have low resistivity also have poor thermoelectric coefficients, whereas those with high thermoelectric coefficients, like bismuth (Bi) and antimony (Sb), have high resistivities (Fraden, 2003).

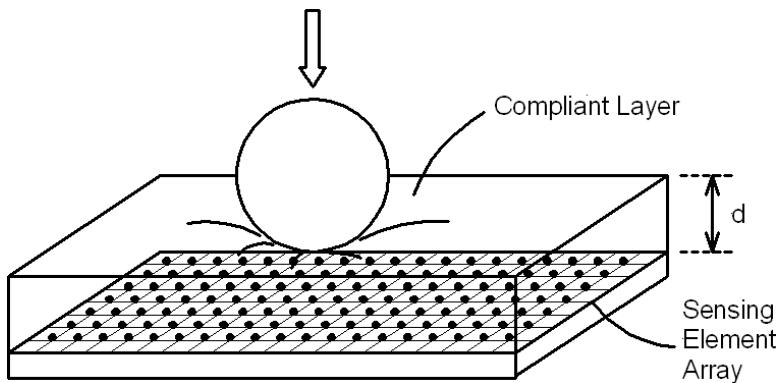
#### 2.4.15 Tactile Sensing

Tactile feedback is one of the critically important aspects of any successful hand prosthesis. In addition, it also plays a role in improving haptic feedback in remotely monitored medical examinations and surgery (Pawluk, Son et al., 1998).

Tactile feedback relies on contact-based effects including contact stresses, slippage, heat transfer, and hardness. These properties, in a grasped object, can be classified into geometric and dynamometric types (Webster, 1999). Among the geometric properties are presence, location in relation to the end-effector, shape, dimensions, and surface conditions. Among the dynamometric parameters associated with grasping are force distribution, slippage, elasticity, and hardness as well as friction.

Though tactile sensing requires sophisticated sensors, it is also reliant on the processes through which the device interacts with the explored object. These include controlling

**FIGURE 2-72** ■  
Interaction between a rigid object and the compliant covering of a tactile sensor.



contact force and end-effector position and orientation. This leads to active tactile sensing, which requires a high degree of complexity in the acquisition and processing of tactile information.

Tactile sensing generally involves the interaction of a rigid object with the compliant cover layer of the tactile sensor, as shown in Figure 2-72. The indentation of the tactile layer can be analyzed from two distinct perspectives. The first is the measurement of the contact stresses (force distribution) in the layer, which is relevant to controlling manipulation tasks. The second is the deflection profile of the layer, which is important in recognizing geometrical features of the object.

In 1980, a survey was conducted to determine the general specifications for tactile sensors (Harmon, 1982), which are used by many tactile sensor designers:

- Spatial resolution 1 to 2 mm
- Array sizes of  $5 \times 10$  to  $10 \times 20$  elements
- Sensitivity between  $0.5 \times 10^{-2}$  N and  $1 \times 10^{-2}$  N for each sensing element
- Dynamic range 1000:1
- Stable behavior with no hysteresis
- Monotonic response but not necessarily linear
- Compliant interface, rugged, and inexpensive

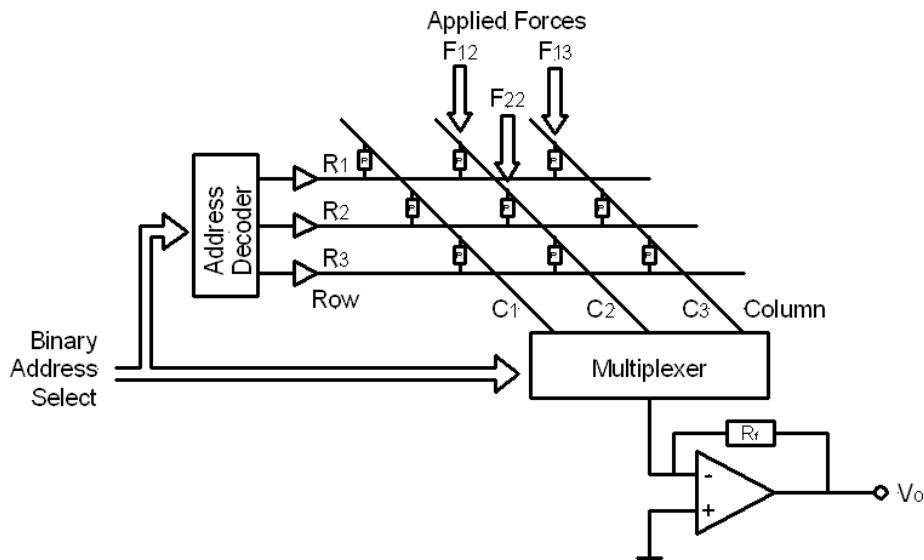
#### 2.4.15.1 Resistive Tactile Sensors

These sensors rely on materials whose resistance changes with increases in applied force. Conductive elastomers manufactured by embedding conductive particles in natural or silicone rubber were among the first. These can operate using changes in the bulk resistance of the material or changes in the contact resistance. Piezoresistive materials have also been used, as have embedded pressure sensors.

A typical array will include conductive strips connecting the rows and columns of the grid so that the individual element resistances may be sensed, as shown in Figure 2-73.

A binary address selection connects a voltage,  $V_i$ , to a single row of resistive elements, while simultaneously the address controls a multiplexer that connects a single column to an op amp inverter. The voltage output,  $V_o$ , is proportional to the resistance of a single element

$$V_o = -\frac{R_f}{R_{RC}} V_i \quad (2.76)$$



**FIGURE 2-73** ■ Configuration of a resistive tactile sensor.

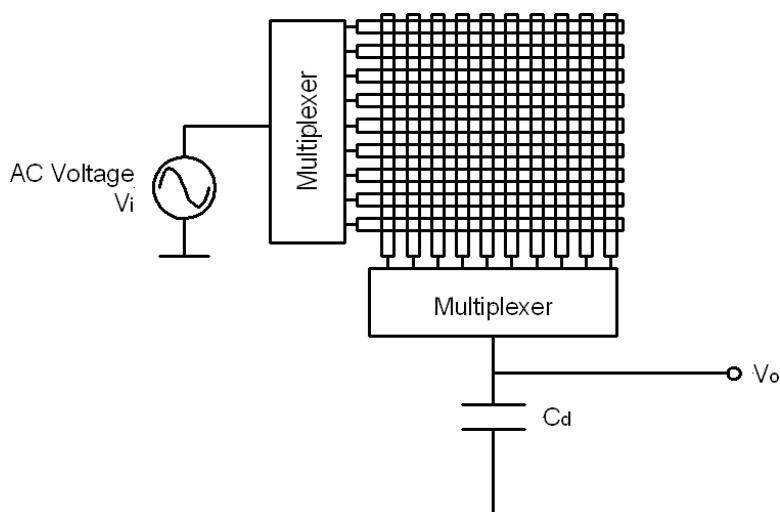
where  $R_{RC}$  is the resistance of the element at row  $R$  and column  $C$ , and  $R_f$  is the resistance of the feedback resistor.

#### 2.4.15.2 Capacitive Tactile Sensors

If the row and column conductive strips are placed on either side of a compliant dielectric material, then the intersection of each forms a tiny capacitor. The capacitance of a parallel plate capacitor is directly proportional to the product of the plate area,  $A$ , and the dielectric constant,  $\epsilon$ , and inversely proportional to the distance between them,  $h$ .

$$C = \frac{\epsilon A}{h} \quad (2.77)$$

An applied force that reduces the distance between the plates will result in an increase in capacitance that can be measured by an external circuit as shown in Figure 2-74.



**FIGURE 2-74** ■ Configuration of a capacitive tactile sensor.

If stray capacitances between elements are ignored and the multiplexers are ideal, then the capacitors function as an AC voltage divider and  $V_o$  will be a function of the element capacitance

$$V_o = \frac{C_{RC}}{C_d + C_{RC}} V_i \quad (2.78)$$

where  $C_d$  is the detector capacitance, and  $C_{RC}$  is the capacitance of the element addressed by row  $R$  and column  $C$ .

#### 2.4.15.3 Piezoelectric Tactile Sensors

As discussed earlier in this chapter, a piezoelectric material is one that will develop a charge,  $Q$ , across opposite faces when subjected to a force or deformation. Because the material is an insulator and a dielectric, it also forms a capacitor so a voltage can be measured across each element.

$$V = \frac{Q}{C} = \frac{Qh}{\epsilon A} \quad (2.79)$$

where  $h$  is the thickness of the piezoelectric material,  $A$  is the area of each element, and  $\epsilon$  is the dielectric constant.

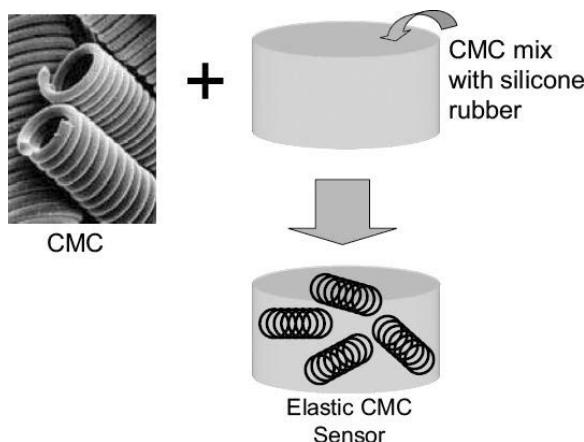
The piezoelectric material most commonly used is PVF2 because it has a strong piezoelectric effect; therefore, it can be made very sensitive. It is also flexible and can be made into small sensor elements. Unfortunately, it is sensitive to temperature, and because the output occurs only when the elements are flexed the response does not extend right down to DC.

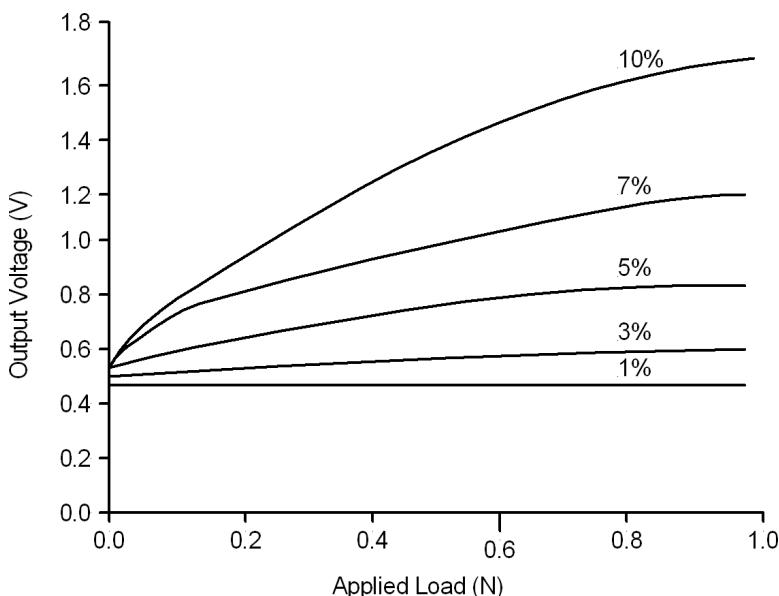
Sliding force sensors can be constructed using thin slivers of piezoelectric materials orientated in different directions. These are situated within a silicone rubber skin and will respond differently as the sensor is moved in different directions across a rough surface. According to Fraden (1996), they are capable of detecting surface discontinuities or bumps as small as  $50 \mu\text{m}$  high.

#### 2.4.15.4 Other Tactile Sensors

Exotic materials such as carbon nanotubes and carbon microcoils (CMCs) can also be embedded into a rubber matrix to make tactile sensors, as shown in Figure 2-75.

**FIGURE 2-75 ■**  
Tactile sensor made from carbon microcoils (Nishikawa, Young-Kwang et al., 2005), with permission.





**FIGURE 2-76 ■**  
Measured sensitivity of a CMC based tactile sensor with different concentrations of CMC [adapted from (Nishikawa, Young-Kwang et al., 2005).]

Unlike normal resistive sensors, because of the coiled nature of the conductive material a significant inductive component also exists. These sensors are therefore excited by an AC signal at a frequency of between 100 kHz and 400 kHz to give the best response. In addition, the sensitivity can be controlled by adjusting the concentration of the CMC material compared with the silicone rubber, as shown in Figure 2-76 (Nishikawa, Young-Kwang et al., 2005).

In addition to these resistive methods, various other mechanisms are used including optical and pneumatic methods.

#### 2.4.16 Chemical Sensors

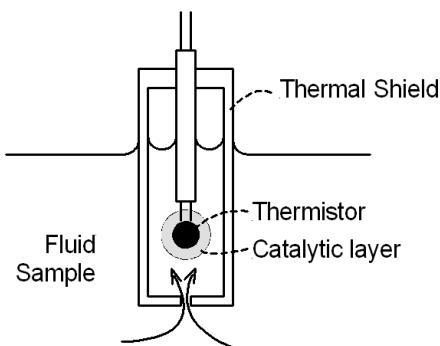
Chemical sensors are sensitive to stimuli produced by various chemical compounds or elements. Their most important property is selectivity (the ability to discriminate between types), while a secondary one is their relationship to concentration. Chemical sensors are very important in medical applications; for instance, the ability to monitor oxygen concentration in the air or in solution is crucial during surgery.

Biosensors are considered a special class of chemical sensors because they have a much greater selectivity and sensitivity than other chemical sensors. Man-made biosensors often use enzymes that have evolved over millions of years to be extremely sensitive to specific molecules. For these sensors, it is critically important that the active material on the sensing element be immobilized on the physical transducer and that it remains active for the lifetime of the sensor.

##### 2.4.16.1 Enzyme and Catalytic Sensors

Enzymes have two remarkable properties: They are extremely selective to a given substance and are very effective in increasing the rate of chemical reactions. The sensing element can be a heated probe, an electrochemical sensor, or an optical one. The enzyme is immobilized into a hydrogel into which the chemical diffuses. The enzyme-assisted reaction alters the characteristics of the hydrogel in a manner that can be detected by the associated sensor.

**FIGURE 2-77 ■**  
Catalytic thermal sensor [Adapted from (Fraden 1996).]



Catalytic sensors are a subset of enzyme sensors in which the catalytic reaction releases heat, and the temperature change can be measured by the sensor. These sensors have been developed specifically to measure low concentrations of inflammable gases (particularly in mines). As shown in Figure 2-77, a typical catalytic sensor comprises a temperature sensor such as an RTD or a thermistor surrounded by the appropriate catalyst. As the concentration of the target molecule increases, the rate of conversion increases, as does the temperature.

In general, a pair of sensors is used in a Wheatstone bridge configuration, as discussed earlier in this chapter. One sensor is immersed in the fluid containing traces of the molecule to be detected, and the other is placed in an inert solution. The two solutions are in physical contact so the initial temperatures of the two thermistors start out equal.

#### 2.4.16.2 Electrochemical Sensors

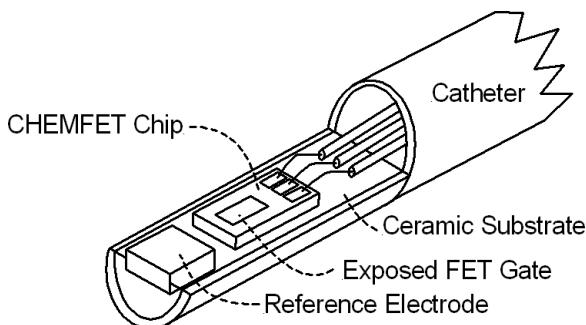
These are the most versatile and are the best developed of all of the chemical sensor types. They measure voltage, current, or resistivity and generally consist of a pair of electrodes as part of a closed circuit.

In voltage-based sensors, a redox reaction at the electrode–electrolyte interface of the electrodes results in a half-cell potential developing at each. One of the reactions involves the molecule of interest, whereas at the other a known reaction occurs. To maintain equilibrium, the current flow should be minimized, so a very high-impedance amplifier is used to measure the potential difference across the two electrodes. This potential difference is a function of the concentration of the molecule of interest. These sensors can be made using field-effect transistors (FETs) and are known as CHEMFETs. They are sensors in which the gate of the transistor is coated with an appropriately sensitive gate insulator material that is, in turn, exposed to the electrolyte as illustrated in Figure 2-78. As the charge around the sensing area changes, the conductivity of the FET is altered, and this can be measured. Ion selective membranes can be deposited on top of the gate insulator to provide a large selection of different chemical sensors.

In resistive sensors (also known as conductive sensors), the conductivity of the electrolyte is dependent on the chemical concentration of the molecule of interest. As with the voltage types, these sensors are incorporated into one arm of a Wheatstone bridge.

#### 2.4.16.3 Resistive Chemical Sensors

To detect the presence of a liquid, the sensor must be specific to a particular molecule at a specific range of concentrations. A typical resistive sensor discussed by Fraden (1996)



**FIGURE 2-78 ■**  
Catheter tip with a CHEMFET pH sensor for medical applications.  
[Adapted from (Fraden 1996).]

consists of a silicone rubber–carbon mixture in a thin film to maximize its surface-to-volume ratio. In the presence of a polar solvent (hydrocarbon fuel), the polymer matrix swells, pushing the carbon particles further apart and increasing the resistance significantly. This change can be measured in the normal way.

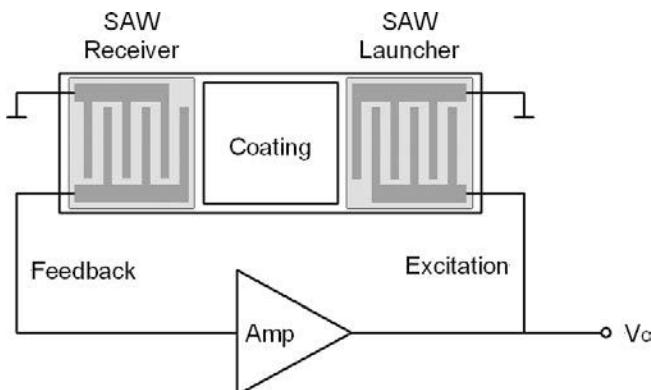
#### 2.4.16.4 Oscillating Chemical Sensors

Measurements of minute amounts of mass cannot use conventional scales because they are not sufficiently sensitive. Oscillating sensors have the required sensitivity because they measure the shift in resonant frequency of a small piezoelectric crystal. A piezoelectric crystal cut to oscillate at ultrasonic frequencies will have a resonant mode that is a function of its mass. The surface of the crystal is covered by a thin layer of material that has an affinity for the molecule of interest, so if any is present it will attach to the surface and alter the mass—and consequently the resonant frequency.

Oscillating sensors can be extremely sensitive, typically  $5 \text{ MHz cm}^2/\text{kg}$ . This means that a 1 Hz frequency shift corresponds to about  $17 \text{ ng/cm}^2$  added weight. The response is very linear with mass, and the dynamic range is quite broad—up to  $20 \mu\text{g/cm}^2$ .

A second type of oscillating sensor uses surface acoustic waves (SAWs). These sensors comprise a SAW transmission line (membrane) covered by a chemically sensitive coating situated between an acoustic launcher and a receiver, as shown in Figure 2-79.

An oscillator drives the launcher, which stimulates the production of a mechanical wave that travels along the transmission line to the receiver where it is converted back to an electrical signal for analysis. In the gas sensor shown in Figure 2-79, oscillation is maintained by feedback from the receiver, which will be a function of the transmission time.



**FIGURE 2-79 ■**  
Flexing plate gas sensor. [Adapted from (Fraden 1996).]

**TABLE 2-10 ■ SAW Chemical Sensors**

Compound	Chemical Coating	SAW Substrate
Organic vapor	Polymer film	Quartz
SO <sub>2</sub>	Triethanolamine (TEA)	Lithium niobate, silicon
H <sub>2</sub>	Pd	Quartz
NH <sub>3</sub>	Pt	Lithium niobate
H <sub>2</sub> S	WO <sub>3</sub>	Lithium niobate
Water vapour	Hygroscopic material	Lithium niobate, quartz
NO <sub>2</sub>	Phthalocyanine	Lithium niobate
NO <sub>2</sub> , NH <sub>3</sub> , SO <sub>2</sub> , CH <sub>4</sub>	Phthalocyanine	Lithium niobate
Vapor explosives, drugs	Polymer	Quartz
SO <sub>2</sub> , CH <sub>4</sub>	None—thermal conductivity	Lithium niobate

*Source:* Fraden, J., *Handbook of Modern Sensors*, New York: AIP Press, Springer-Verlag, 1996, with permission.

This is in turn sensitive to the mass of the membrane, which depends on the concentration of the molecule of interest.

The theoretical sensitivity of such sensors operating at 2.6 MHz is of the order of 900 cm<sup>2</sup>/g, so if a sensor with a sensitive area of 0.2 cm<sup>2</sup> captures 10 ng the oscillator frequency is shifted by

$$\begin{aligned}\Delta f &= -900 \times 2.6 \times 10^6 \times 10^{-8} / 0.2 \\ &= -117 \text{ Hz}\end{aligned}$$

SAW sensors are very versatile and can be adapted to measuring a wide variety of chemical compounds just by changing the coating on the membrane. Some of these are listed in Table 2-10.

#### 2.4.16.5 Microbalance Odor Sensors

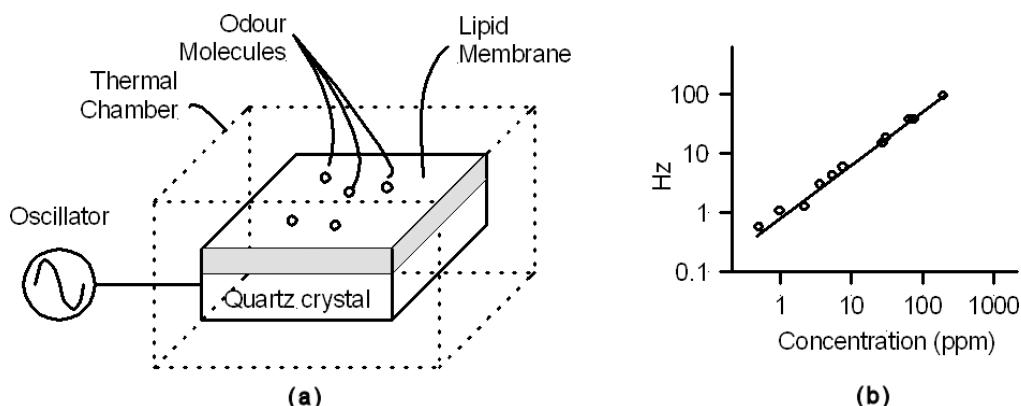
Odor sensors are gas sensors that have sensitivities approaching that of the human nose. Human odor sensors are covered with a phospholipid bilayer membrane, and it is believed that odorant adsorption into the membrane induces nerve impulses that travel to the brain for interpretation. Man-made odor sensors use a composite sensor made from polyvinyl chloride (PVC), a plasticizer, and a synthetic lipid. The lipid molecules are randomly oriented in the polymer matrix.

A 200  $\mu\text{m}$  thick membrane of this polymer composite is spread on one side of the quartz crystal oscillator as shown in Figure 2-80, which is designed, along with the membrane blend, to have a resonant quality factor,  $Q > 5 \times 10^4$ .

As with the other oscillating sensors, odor molecules embed into the membrane and alter the resonant frequency because of the increase in mass. Experimental results indicate a sensitivity starting at 1 ppm (approximately the human threshold) and extending in a linear fashion up to 3000 ppm.

#### 2.4.17 Optical Chemical Sensors

Optical methods are among the oldest techniques for sensing biochemical concentration. In principle, these sensors consist of a light source tuned to a frequency that will interact with the molecule of choice, a method of directing the light into the medium, and a photodetector for processing the optical signal.



**FIGURE 2-80** ■ Microbalance odor sensor and the transfer function for amyacetate gas. (a) Construction schematic. (b) Relationship between gas concentration and frequency change. [Adapted from (Fraden 1996).]

Optical sensors are usually based on optical fibers or planar waveguides, and there are three main methods of sensing at the surfaces of these devices:

- The molecule directly affects the optical properties of the waveguide such as evanescent waves<sup>1</sup> or surface plasmons.<sup>2</sup>
- An optical fiber is used to convey light to and from the sample. Changes in the optical properties of the medium containing the molecule of interest are sensed by an external spectrophotometer.
- An indicator or chemical reagent placed near the tip of the optical fiber reacts with the molecule of interest to produce an optical signature that can be detected. These include absorption spectroscopy and fluorimetry.

Light sources range from a wide variety of coherent laser sources from LEDs extending from the infrared (IR) to the visible as well as broad spectral incandescent lamps.

Optical elements include optical fiber, lenses, prisms, beam splitters, polarizers, and diffraction gratings. Optical fibers and lenses convey light to and from the test area, whereas prisms and diffraction gratings provide wavelength selection for frequency selective processing. Polarizers are useful in identifying or selecting the light polarization for transmission and reception.

Photodetection is most often performed using photodiodes that are sensitive to the frequency band of interest. Most systems use a pair of identical diodes: one in the active channel; and the other to function as a reference.

#### 2.4.17.1 Evanescent Wave Spectroscopy

When light propagates along an optical fiber, it is not completely confined to the core region but penetrates a short distance (typically one wavelength) into the cladding. Therefore,

<sup>1</sup>Electromagnetic waves generated in the medium outside the optical waveguide when light is reflected from within.

<sup>2</sup>Resonances induced by an evanescent wave in a thin film deposited on a waveguide surface.

any compounds close to the surface will have an effect on this wave. Typically, a section of fiber is left unclad and in contact with a solution of the molecule of interest or with a reagent that reacts with the molecule. Multiple internal reflections along this section of the fiber contribute to the overall absorption or fluorescence, which can be measured at the end of the fiber.

#### 2.4.17.2 Surface Plasmon Resonance

In this technique, the fiber is clad with a metallized layer, typically gold or silver. When monochromatic light from the laser reaches this layer, it is absorbed by the plasma generated by the conduction electrons of the metal. This results in a phenomenon called surface plasmon resonance (SPR), which leads to absorption of the light. The resonance depends on the angle, wavelength, and polarization state of the incident light and the refractive index of the metal film and material adjacent to it. Therefore, SPR is an extremely sensitive method of measuring changes in the refractive index of the medium.

#### 2.4.17.3 Optical Fiber Sensors

Optical fibers are small and low cost. In addition, because electrical potentials are not involved, there is no electrical risk to the patient, and interference from electric and magnetic fields does not occur.

Most measurements are based on spectral or absorption changes in the medium determined by the active molecule, either directly or through an indicator mediated reaction. Most chemicals of medical interest such as hydrogen, oxygen, carbon dioxide, and glucose require the use of reagents.

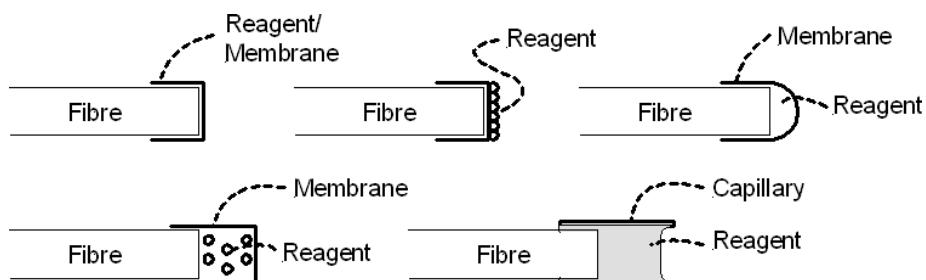
In optical fiber-based sensors, light travels efficiently to the end of the fiber where it exits into the medium and interacts with the molecule or reagent before returning via the same or a different optical fiber to a detector for analysis and interpretation. Typical reagent mediated optical fiber-based sensors are shown in Figure 2-81.

#### 2.4.17.4 Measurement of Blood Oxygen Concentration

This is commonly known as oximetry and generally relies on a color change in the blood as the relative amounts of deoxyhemoglobin (Hb) and oxyhemoglobin ( $\text{HbO}_2$ ) vary.

Measurements are performed at two specific wavelengths, the first at 660 nm, where there is a large difference between the relative absorption of the two molecules, and the second at  $\approx 805$  nm, where the absorption is independent of blood oxygenation. The oxygen saturation,  $O_{2\text{sat}}$ , can then be determined from the ratio of the two absorption

**FIGURE 2-81** ■ Configurations of different reagent mediated fiber optic sensors. [Adapted from (Fraden 1996).]



levels,  $\gamma_{\lambda 1}$ , and  $\gamma_{\lambda 2}$

$$O_{2sat} = a - b \frac{\gamma_{\lambda 1}}{\gamma_{\lambda 2}} \quad (2.80)$$

where  $a$  and  $b$  are sensor-dependent constants.

In vivo fiber optic oximeters can be manufactured within catheters that are inserted into a vein or artery or as noninvasive sensors clamped to the finger or ear where blood flow is close to the surface.

The latter are referred to as pulse oximeters and monitor differences in the absorption only as a function of time. This relationship is dependent on the blood volume within the sensitive area and will therefore be a function of arterial pressure and heartbeat. A normalization process determines the ratio of the changing component, the AC, and the fixed current (DC) component of the measured signal to produce a stable and reliable signal.

Pulse oximeters consist of a pair of small and inexpensive LEDs, one operating in the red and the other in the infrared band. A single, highly sensitive phototransistor monitors the signal level transmitted through a finger or ear lobe or the backscatter from any other well-vascularized portion of the body.

#### 2.4.17.5 Measuring Other Molecules

Fiber optic sensors have been developed to measure other blood constituents as well as oxygen partial pressure ( $PO_2$ ). These include  $CO_2$  partial pressure ( $PCO_2$ ) and blood pH, both of which are essential to clinical diagnosis and the management of respiratory and metabolic problems. In addition, glucose sensors have been developed, and these are playing an increasingly important role as the incidence of diabetes increases.

## 2.5 | ELECTRODES

---

Biological functions often have some electrical activity associated with them. The activity can be a constant DC potential or a time-varying one. Even though almost all the organs in the body generate some voltage, most of the signals are small and hard to measure. However, a number have proved to be useful for biological assessment and are listed in Table 2-11.

**TABLE 2-11** ■ Useful Bioelectric Signals

Bioelectric Signal	Abbreviation	Source
Electrocardiogram	ECG	Heart—seen from body surface
Cardiac electrogram		Heart—seen from within
Electromyogram	EMG	Muscle
Electroencephalogram	EEG	Brain
Electroocptogram	EOG	Eye dipole field
Electroretinogram	ERG	Eye retina
Action potential		Nerve or muscle
Electrogastrogram	EGG	Stomach
Galvanic skin reflex	GSR	Skin

*Source:* Bronzino, J. (Ed.), *Medical Devices and Systems*, Boca Raton, FL: CRC Press, 2006, with permission.

The mechanism of electrical conductivity in the body involves ions as charge carriers. Therefore, picking up bioelectric signals involves converting the ionic currents into electric currents that will flow through wires. This conversion process is carried out by electrodes that consist of electrical conductors in contact with the aqueous ionic (electrolyte) solutions in the body. At the interface between the electrode and the electrolyte solution, an electrochemical reaction needs to take place for a charge to be transferred.

When no current is flowing between the electrode and the organism, a potential, known as the half-cell potential, exists across the boundary. If current flows, then the potential may drop, an effect known as polarization, and the sensitivity of the electrode may be reduced.

Polarizable electrodes pass a current between the electrode and the electrolytic solution by changing the charge distribution in the solution near the electrode. No actual current crosses the electrode–electrolyte interface. Nonpolarized electrodes allow current to pass freely across the interface without changing the charge distribution in the electrolytic solution.

Electrodes made from the noble metals such as platinum are highly polarizable, and the charge distribution in the electrolyte adjacent the electrode is different from that of the bulk. This limits the ability of such electrodes to measure DC or low-frequency signals. In addition, if the electrode moves with respect to the electrolyte solution, the charge distribution in the solution adjacent to the electrode surface will change, which will generate a voltage change that will appear as a motion artifact in the measurement. For these reasons, nonpolarizable electrodes are preferred for most biomedical measurements.

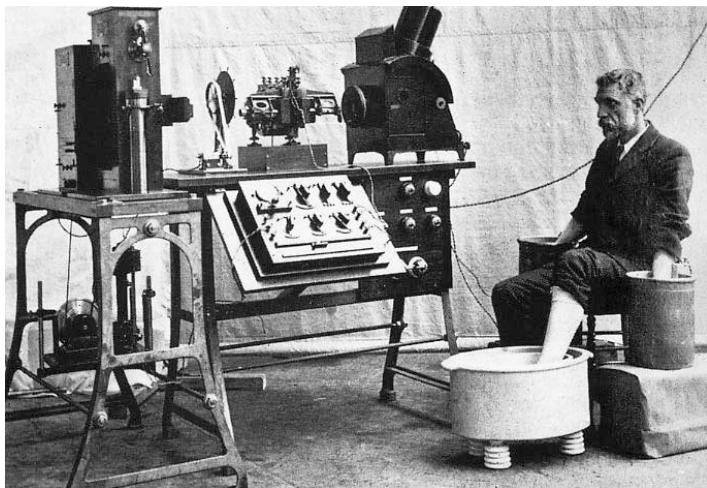
The silver–silver chloride electrode is a good example of a nonpolarizable electrode that is suitable for many biomechanical applications. Such electrodes have a silver base with a silver–silver chloride matrix on the surface. They do not polarize and thus are more capable of conducting low-frequency and DC signals and are also less prone to generating motion artifacts.

A simplified equivalent circuit of an electrode is a parallel resistor and capacitor with the resistor describing the DC and low-frequency impedance of the electrode and the capacitor describing the higher-frequency AC component. Typical surface electrodes have impedances of between 2 and 10 k $\Omega$ , with larger electrodes having lower impedances and needle or microelectrodes having much higher values (Cromwell, Weibell et al., 1973).

### 2.5.1 Body–Surface Biopotential Electrodes

This category of electrodes includes those that can be placed on the body surface for recording bioelectric signals. The integrity of the skin is not compromised, and they can be used for short- or long-duration applications. The earliest bioelectric potential measurements relied on immersion electrodes that were simply buckets of saline solution into which the patient placed a hand and a foot, as shown in Figure 2-82. As expected, this type of electrode presented many difficulties in regard to restrictions on the patient position and movement as well as the possibility of spillage.

Plate electrodes, first introduced in 1917, were a great improvement on immersion electrodes. They were originally separated from the skin by cotton pads soaked in saline to emulate the immersion electrode mechanism. Later, an electrolytic paste was used in place of the pad with the metal in contact with the skin.



**FIGURE 2-82 ■**  
Early ECG system using immersion electrodes. (Aquilina 2006).

### 2.5.1.1 Metal Plate Electrodes

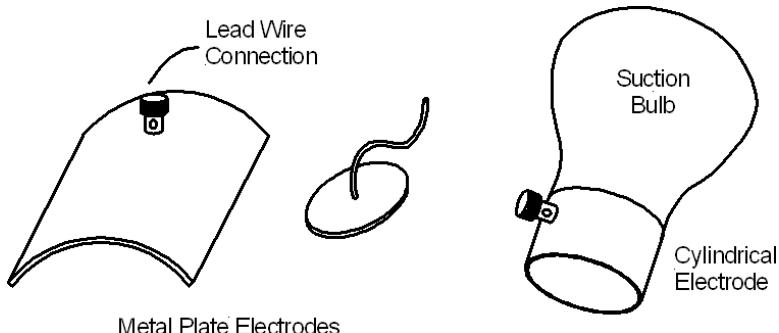
The basic metal plate electrode consists of a flat metal plate with a thin film of conductive electrolyte gel between the plate and the skin to establish contact. These are usually made from German silver (a nickel–silver alloy), platinum, gold, or silver. They can be made from foil so that they are flexible, or they can even be formed with a suction cup to facilitate attachment, as shown in Figure 2-83. These electrode types are primarily used for diagnostic recordings of ECG and Electroencephalogram (EEG) signals.

Conical gold-plated disk electrodes that include a hole at the apex of the cone for the insertion of electrolyte gel are frequently used for EEG monitoring, as the electrolyte can be replaced while the electrode is in situ.

These days, most electrodes are made from silver–silver chloride. They are easy to attach and because they are disposable do not need to be cleaned after use.

### 2.5.1.2 Electrodes for Long Term Use

Long-term monitoring of biopotentials such as the ECG performed by cardiac monitors places a number of constraints on the electrodes. These must provide a stable interface, making the nonpolarizable variety preferable. In addition, mechanical stability, with the resultant reduction of motion artifacts, is important. One method of achieving this is to reduce motion between the electrode and the coupling electrolyte, which can be achieved by recessing the electrode into the bottom of a cupful of electrolyte gel. Movement of the



**FIGURE 2-83 ■**  
Examples of different skin electrodes. [Adapted from (Fraden 1996).]

skin with respect to the electrode is then buffered by several millimeters of electrolyte, with the result that motion artifacts are reduced.

The advantages of this recessed electrolyte design can be realized in a much simpler design that lends itself to mass production. The electrodes consist of a conductive metal electrode bonded to a layer of open-cell foam that is impregnated with a high-viscosity electrolyte gel. Frequently these electrodes are attached to a press stud through an insulated adhesive disk that holds the electrode in place against the skin.

A modification of this design replaces the electrode with a thin film of foil that can deform for a better fit on a curved body. The sponge can then be replaced with a thinner hydrogel film saturated with an electrolyte gel. This electrolyte–hydrogel mix is very sticky, so no further adhesive is required.

Very thin electrode films can be used in place of the foil electrodes if they are backed by a strong polymer film. The advantage of these is that if the metal film is thin enough (typically  $1\ \mu\text{m}$  or less) they are x-ray transparent. This is particularly useful for monitoring premature babies with sensitive skins, as the repeated application and removal of electrodes can be a severe skin irritant (Bronzino, 2006). Some of these electrode types are shown in Figure 2-84.

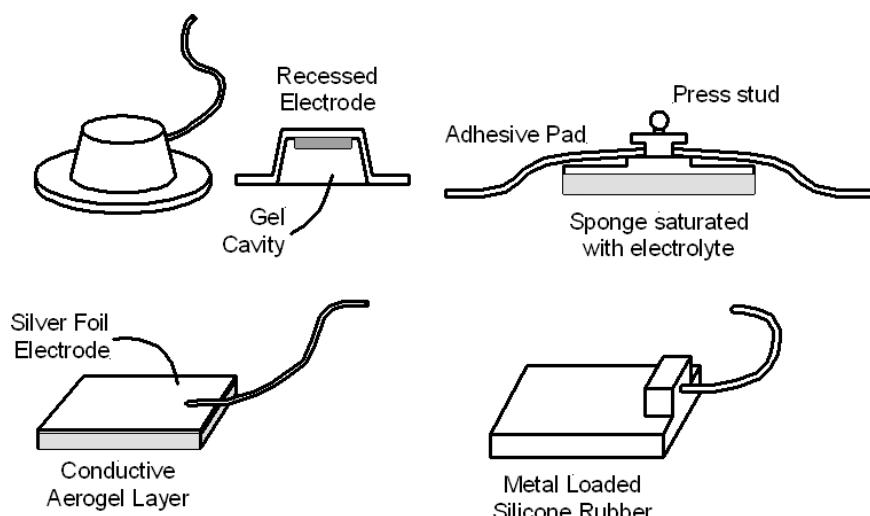
Special problems encountered in monitoring the ECGs of astronauts during long durations in space and under conditions of perspiration and considerable movement led to the development of spray-on electrodes. In these, a small spot of conductive material is sprayed or painted directly on the skin, which has previously been treated with an electrolyte coating (Cromwell, Weibell et al., 1973).

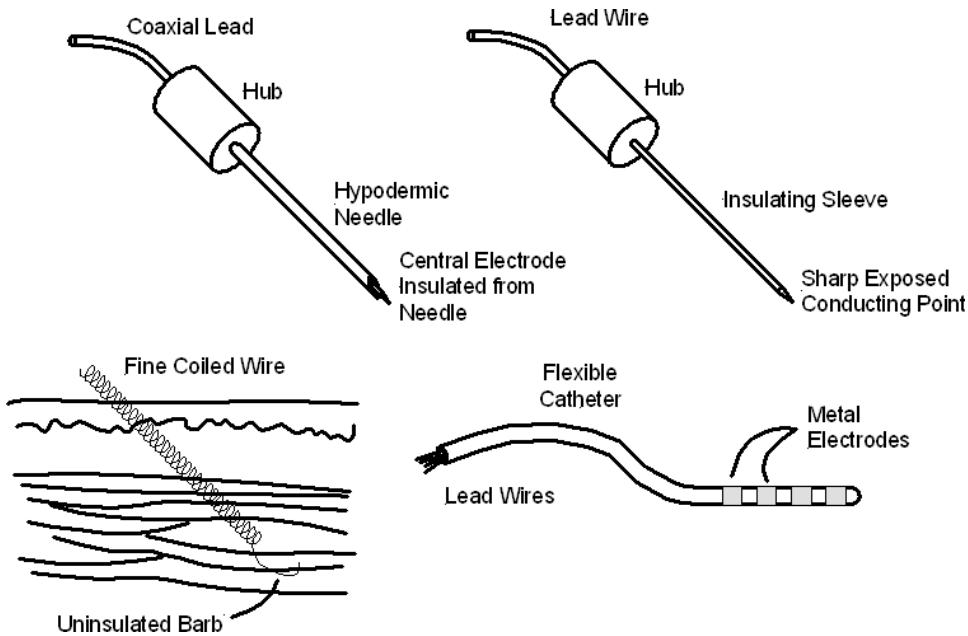
Electrodes that do not use externally applied electrolyte gels are known as *dry*. They can easily be applied and held in place using a rubber band or tape. They are made from graphite or metal-powder-impregnated silicone rubber to produce a flexible conductive material similar to those discussed in the section on tactile sensors. These are applied directly to the skin, and the electrolyte layer is formed by a film of sweat that collects under the contact area.

Dry electrodes are used in home medical monitoring devices and on consumer goods like exercise bikes and treadmills to pick up an ECG signal. The signals picked up by these electrodes are generally much noisier than those obtained from the wet types. This

**FIGURE 2-84 ■**

Electrodes for chronic patient monitoring.  
[Adapted from (Bronzino 2006).]





**FIGURE 2-85 ■**  
Internal electrodes.  
[Adapted from  
(Bronzino 2006).]

requires that more consideration is taken with the signal processing to compensate for these shortcomings.

### 2.5.1.3 Electrodes for Internal Use

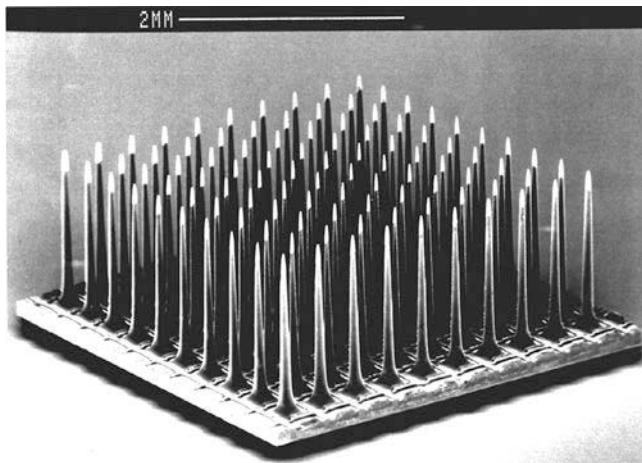
Internal electrodes are generally smaller than their surface counterparts and don't require a gel layer because they are in intimate contact with the organism's own electrolyte solution. These are either needle electrodes introduced through the skin or cavity electrodes that can be implanted surgically or through a catheter as illustrated in Figure 2-85. Good examples of the latter are the electrodes for a cardiac pacemaker, which are introduced into the heart through a vein and are barbed to ensure that good contact is maintained even after millions of contractions.

A needle electrode consists of a solid stainless steel needle with a sharp point. An insulating layer covers most of the needle, exposing only a millimeter or so of the point. When this device is inserted into skeletal muscle, electrical signals are picked up by the exposed tip. Needle electrodes can also be made that are inserted through the hollow shaft of a standard hypodermic needle. These fine electrodes can then remain in place for long periods of time. Fine coiled-wire electrodes have also remained in skeletal muscles for a number of years without adverse effects.

Arrays of surface or needle electrodes are used in a number of biomedical prostheses, including those that are embedded in the retina or directly in the visual cortex as the final output of visual prostheses or within the cochlea to bypass damaged sections of the ear. Some of the most impressive of these arrays are those developed at the University of Utah and shown in Figure 2-86.

Recent advances in electrode design allow complete rice-sized sensors to be implanted into muscles where they communicate wirelessly with an external receiver. A magnetic coil surrounding the implant region (usually the stump of an arm) powers the devices (Weir, Troyk et al., 2009).

**FIGURE 2-86 ■**  
 Photograph showing  
 one of the Utah  
 micro arrays.  
 (Medscape 2008),  
 with permission.



## 2.6 | REFERENCES

- Alciatore, D. and M. Histan. (2003). *Introduction to Mechatronics and Measurement Systems*, 2nd ed. Boston: McGraw Hill.
- Allen, R. (2002). “MEMS: Laying the Foundation for Exciting Applications.” *Electronic Design*. Retrieved June 2008 from <http://electronicdesign.com/Articles/Index.cfm?AD=1&ArticleID=2138>
- Amman, M. (2001). “Laser Ranging: A Critical Review of Usual Techniques for Distance Measurement.” *Optical Engineering* 40: 10–19.
- Aquilina, O. (2006). “A Brief History of Cardiac Pacing.” *Images in Paediatric Cardiology* 27: 17–81.
- Bronzino, J. (Ed.). (2006). *Medical Devices and Systems*. Boca Raton, FL: CRC Press.
- Brooker, G. (2008). *Introduction to Sensors for Ranging and Imaging*. Raleigh, NC: SciTech.
- Buchmann, I. (2005). “Battery University.” Retrieved July 2008 from <http://www.batteryuniversity.com/index.htm>
- Campbell, M. (2010). “Heartbeat Generator Could Power Implanted Sensors.” *New Scientist*. London, Sunita Harrington, May 29.
- Chen, X., X. Shiuou, et al. (2010). “Nanogenerator for Mechanical Energy Harvesting Using PZT.” *NanoLetters* 10(6): 2133–2137.
- Cromwell, L., F. Weibell, et al. (1973). *Biomedical Instrumentation and Measurements*. Englewood Cliffs, NJ: Prentice-Hall, Inc.
- Fett, T., D. Munz, et al. (1999). “Tensile and Bending Strength of Piezoelectric Ceramics.” *Journal of Material Science Letters* 18: 1899–1902.
- Fraden, J. (1996). *Handbook of Modern Sensors*. New York: AIP Press, Springer-Verlag.
- Fraden, J. (2003). *Handbook of Modern Sensors: Physics, Designs and Applications*. New York: Springer Verlag.
- Gregory, B. (1975). *An Introduction to Electrical Instrumentation*. London: Macmillan Press.
- Harmon, L. (1982). “Automated Tactile Sensing.” *International Journal on Robotics Research* 1(2): 196–212.
- Horowitz, P. and W. Hill. (1989). *The Art of Electronics*, 2d ed. Cambridge, UK: Cambridge University Press.
- Johnston, H. (2008). “Knee Brace Harvests ‘Negative Work.’” *physicsworld.com* Retrieved June 2010 from <http://physicsworld.com/cws/article/news/32812>

- Livingstone, R. and M. Rioux. (1986). *Development of a Large Field of View 3-D Vision System. Optical Techniques for Industrial Inspection.*
- Maxim. (2001). "App Note 817: Using Thermistors in Temperature Tracking Power Supplies." Retrieved June 2008 from [http://www.maxim-ic.com/appnotes.cfm/appnote\\_number/817/](http://www.maxim-ic.com/appnotes.cfm/appnote_number/817/)
- Medscape. (2008). "The Neural Interface: The Utah Electrode Arrays." Retrieved September 2008 from [http://www.medscape.com/viewarticle/560817\\_2](http://www.medscape.com/viewarticle/560817_2)
- Neets. (2000). "Module 15-Principles of Synchros, Servos, and Gyros." *Electrical Engineering Training Series*. Retrieved June 2009 from <http://www.tpub.com/content/neets/14187/>
- Nishikawa, K., P. Young-Kwang, M. Aizuddin, K. Yoshinaga, K. Ogura, M. Umezu and A. Takamshi. (2005). "Development of Carbon Microcoils (CMC) Sensor System with High Sensitivity for Effective Acquisition of Tactile Information." In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2005*, August 2–6, pp. 2061–2066.
- Pawluk, D., J. Son, P. Wellman, W. Peine and R. Howe. (1998). "A Distributed Pressure Sensor for Biomechanical Measurements." *Journal of Biomechanical Engineering* 120(2): 302–305.
- Pons, J. (Ed.). (2008). *Wearable Robots—Biomechatronic Exoskeletons*. Chichester, UK: John Wiley & Sons.
- Powell, W. and D. Pheifer. (2006). "The Electronic Tilt Sensor." Retrieved September 2008 from <http://archives.sensorsmag.com/articles/0500/120/index.htm>
- Probert-Smith, P. (Ed.). (2001). *Active Sensors for Local Planning in Mobile Robotics*. Hackensack, NJ: World Scientific.
- Starner, T. and J. Paradiso. (2004). "Human Generated Power for Mobile Electronics." In *Low-Power Electronics*, C. Piguet (Ed.). Boca Raton, FL: CRC Press, pp. 45-1–45-35.
- Webster, J. (Ed.). (1999). *The Measurement, Instrumentation and Sensors Handbook*. Boca Raton, FL: CRC Press.
- Weir, R. F., P. R. Troyk, G. A. DeMichele, D. A. Kerns, J. F. Schorsch and H. Maas. (2009). "Implantable Myoelectric Sensors (IMESs) for Intramuscular Electromyogram Recording." *IEEE Transactions on Biomedical Engineering*, 56(1): 159.



# Actuators

## Chapter Outline

3.1	Introduction .....	91
3.2	Electromechanical Actuators .....	91
3.3	Hydraulic Actuators .....	137
3.4	Pneumatic Actuators .....	139
3.5	Shape Memory Alloy .....	142
3.6	Mechanical Amplification .....	145
3.7	Prosthetic Hand Actuation .....	154
3.8	References .....	157

## 3.1 | INTRODUCTION

Most biomechanical systems involve some sort of motion or an action, which can range from the articulation of a large exoskeleton to the mechanical stimulation of the tiny bones in the middle ear. These actions are created by a force or torque that leads to acceleration and displacement. In most cases these actuators operate by the conversion of electrical power; however, in biomechatronics, pneumatic and hydraulic devices offer some advantages, discussed in this chapter.

Electrical devices that produce or trigger a physical or physiological response by stimulating the body's musculature or nerves are also considered to be actuators. These include pacemakers, defibrillators, transcutaneous electrical nerve stimulation (TENS) devices, and other electrode arrays such as retinal, neural, or cochlear implants. They are discussed in other chapters in this book.

## 3.2 | ELECTROMECHANICAL ACTUATORS

When a current carrying conductor is placed in a magnetic field, a force is produced in a direction perpendicular to both the direction of the current and the magnetic field. This is the Lorentz force law and can be stated in vector form as (Alciatore and Histon, 2003)

$$\vec{F} = \vec{I} \times \vec{B} \quad (3.1)$$

where  $\vec{F}$  (N/m) is the force vector per unit length of conductor,  $\vec{I}$  (A) is the current vector, and  $\vec{B}$  (weber/m<sup>2</sup>) is the magnetic vector or flux density. The relationship among these vectors can be determined using the right-hand rule, which states that if the index finger is pointing in the direction of the current flow and the middle finger is aligned with the magnetic field, then the thumb, extended perpendicular to the two fingers, points in the direction of the force.

The magnitude of the force,  $F$  (N), depends on the total length,  $l$  (m), of the conductor in the field, the magnetic flux,  $B$  (weber/m<sup>2</sup>), and the current,  $I$  (A), if all the components are orthogonal

$$F = BIl \quad (3.2)$$

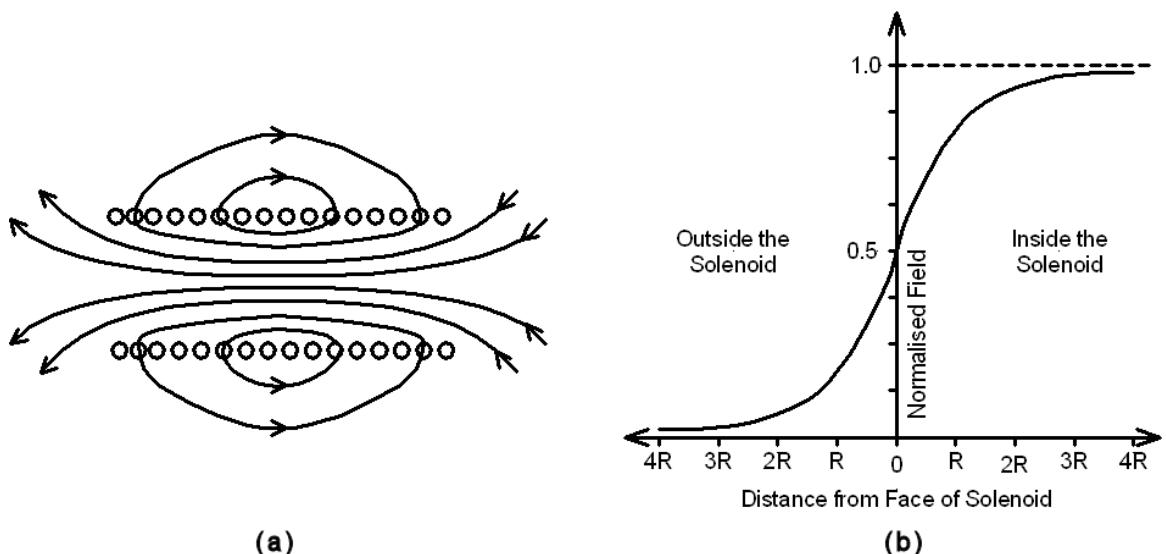
This equation is extended later in this chapter to determine the force generated by a solenoid and the torque generated by an electric motor.

A second consideration of importance is the relationship among the magnetic flux,  $B$ , produced by a conductor wound into a coil (known as a solenoid), the current,  $I$ , flowing through the solenoid, and the number of turns,  $N$ . If the solenoid is long compared with its diameter, then the field in its center can be approximated by

$$B = \mu_o \frac{NI}{L} \quad (3.3)$$

where  $\mu_o = 4\pi \times 10^{-7}$  (weber/amp-m) is the permeability of free space, and  $L$  (m) is the length of the solenoid.

The flux density drops off toward the ends of the coil, and within one radius it reduces to 15% as the flux leaks out. This effect is illustrated in Figure 3-1.



**FIGURE 3-1** ■ Flux density along the length of a coil. (a) Cross-section through coil showing magnetic flux lines. (b) Relationship between the magnetic flux density and the distance from the centre of the solenoid.

**WORKED EXAMPLE**

Consider a 0.5 m long solenoid with a diameter of 30 mm consisting of five layers of windings of 600 turns each and carrying a current of 3 A. The magnetic flux at the centre of the solenoid is

$$\begin{aligned} B &= \mu_o \frac{NI}{L} \\ &= 4\pi \times 10^{-7} \frac{600 \times 5 \times 3}{0.5} \\ &= 0.0226 \text{ weber/m}^2 \end{aligned}$$


---

Note that the solenoid diameter and the number of layers of turns do not enter the equation so long as the length to diameter ratio remains large.

If the air core within the solenoid is replaced with a metal (usually iron), then the magnetic field is increased in proportion to the relative permeability,  $\mu_r$ , of the material

$$\begin{aligned} B &= \mu_r \mu_o \frac{NI}{L} \\ &= \mu \frac{NI}{L} \end{aligned} \tag{3.4}$$

where  $\mu$  is the permeability of the material.

The relative permeability is dependent on the magnetic field and is also a function of frequency and saturation. Typical values of  $\mu_r$  are 200 for magnetic iron and 100 for nickel at a magnetic flux density of 0.002 weber/m<sup>2</sup>. Special alloys can have much higher permeability with permalloy (78.5% Ni and 21.5% Fe) reaching 8000 and mumetal (75% Ni, 2% Cr, 5% Cu, and 18% Fe) reaching 20,000.

If the air core of the solenoid in the previous example is replaced with an iron core with  $\mu_r = 200$ , the magnetic flux at its center will increase to  $0.0226 \times 200 = 4.52$  weber/m<sup>2</sup>.

When a conductor of length  $l$  (m) cuts through a magnetic field with strength  $B$  (weber/m<sup>2</sup>) at a velocity  $v$  (m/s), a voltage  $\varepsilon$  (V) is developed across the conductor if the respective directions of the various components are orthogonal.

$$\varepsilon = Blv \tag{3.5}$$

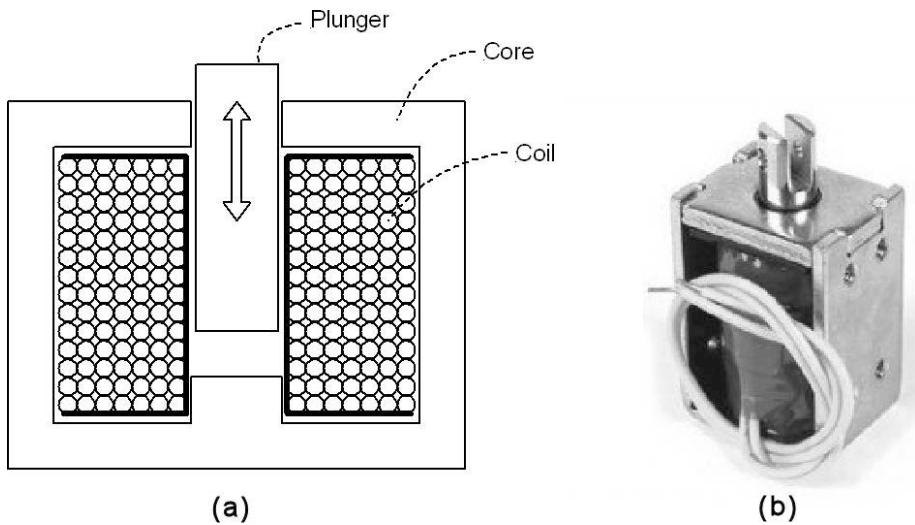
Once again, the direction of the induced current is determined by the right-hand rule, where the direction of the magnetic field is defined by the thumb, the direction of motion by the middle finger, and the direction of current flow (if the circuit is closed) by the index finger.

One final important relationship is Faraday's law, which states that the voltage  $\varepsilon$  (V) induced in a coil is equal to product of the number of turns,  $N$ , and the rate of change of the magnetic flux,  $d\varphi/dt$ , where  $\varphi$  is measured in webers.

$$\varepsilon = N \frac{d\varphi}{dt} \tag{3.6}$$

It is not important whether the change in flux is induced by moving a permanent magnet within the coil or by changing the current through the primary winding of a transformer to induce a changing flux in the secondary.

**FIGURE 3-2 ■**  
**Solenoid actuator.**  
 (a) Schematic cross section.  
 (b) Photograph.



### 3.2.1 Solenoids and Voice Coils

#### 3.2.1.1 Solenoids

Solenoids consist of a fixed core and a movable plunger that complete a magnetic circuit broken by a small gap, as shown in Figure 3-2. When the coil is energized by passing a current through it, the core–plunger combination forms an electromagnet, and the plunger is drawn downward to close the gap.

As a first-order approximation of how this works, a number of assumptions must be made. It is assumed that all of the magnetic flux is constrained within the core and within the gap and that all of the resistance to magnetic flux flow (magnetic reluctance) occurs within the gap. The field is assumed to be uniform in the gap and zero everywhere else and also that the core material does not saturate.

The magnetomotive force (*mmf*) set up by the coil is defined by the number of turns on the coil,  $N$ , and the current  $I$  (A), flowing through it.

$$mmf = NI = \int H \cdot dl \quad (3.7)$$

where  $H$  is the coercive force on the flux path, which is nonzero only in the gap.

Because the field is considered to be uniform and the gap length  $l_g$  (m) is constant, then equation (3.7) can be simplified to

$$mmf = NI = Hl_g \quad (3.8)$$

Therefore the coercive force will be

$$H = \frac{NI}{l_g} \quad (3.9)$$

The energy  $W_{mag}$  (J) stored in the volume of space occupied by a magnetic field is

$$W_{mag} = \frac{1}{2} \int BH dV \quad (3.10)$$

where  $B$  (weber/m<sup>2</sup>) is the flux density and  $V$  (m<sup>3</sup>) is the volume.

Substituting for equation (3.9) and solving for a specific cross sectional area  $A_g$  ( $\text{m}^2$ )

$$W_{mag} = \frac{\mu_o H^2 A_g l_g}{2} = \frac{\mu_o N^2 I^2 A_g}{l_g} \quad (3.11)$$

The mechanical energy is determined by the product of the force and the distance traveled

$$W_{mech} = \int F dl \quad (3.12)$$

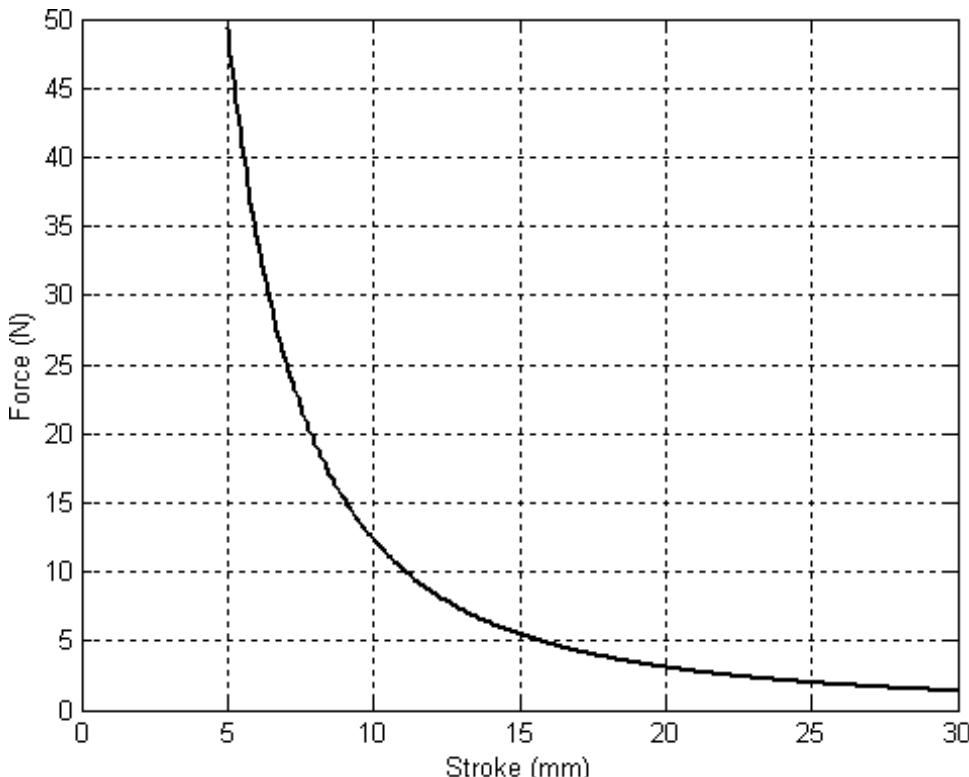
Therefore, the force can be described by the change in energy per unit change in the length of the gap

$$F = \frac{dW_{mech}}{dl_g} \quad (3.13)$$

and this must be equal to the magnetic energy given up. Hence

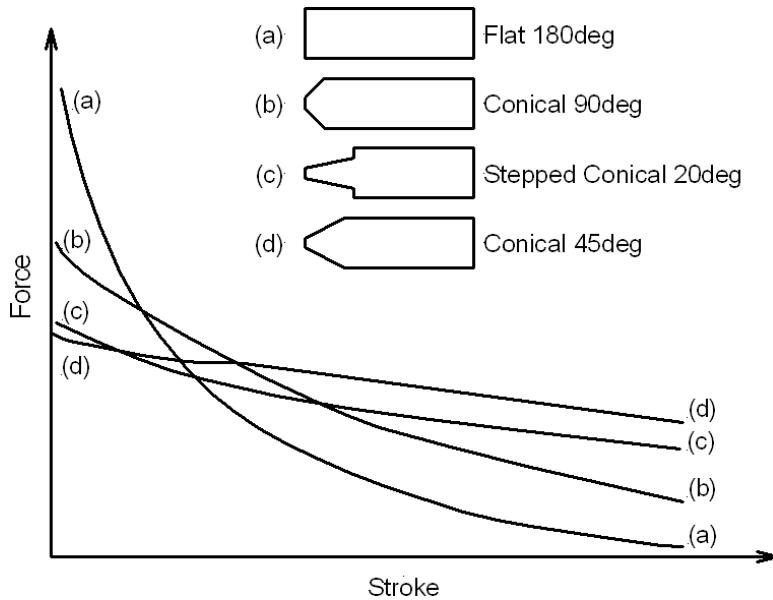
$$F = \frac{dW_{mag}}{dl_g} = \frac{\mu_o N^2 I^2 A_g}{2l_g^2} \quad (3.14)$$

From this result it can be seen that as the gap is reduced, the force increases proportionally with the result that the relationship between the force and the solenoid displacement is very nonlinear. The first-order approximation is shown in Figure 3-3 for a solenoid with  $N = 1000$ ,  $I = 5\text{A}$ , and  $A_g = 7.85 \times 10^{-5} \text{ m}^2$  (plunger diameter of 10 mm).



**FIGURE 3-3** ■ Relationship between the displacement and force of a solenoid.

**FIGURE 3-4 ■**  
Effects on the force stroke curve of a number of differently shaped plungers.



This inverse squared relationship of the solenoid force with distance is not the ideal relationship for an actuator. At the beginning of the stroke, where the load is to be accelerated, there is very little force, and at the end of the stroke, where the plunger should be decelerating, the force reaches a maximum.

The force–distance characteristics of the solenoid can be altered slightly by shaping the gap. For example, if the plunger tip is made in the form of a truncated cone and the pin is made in the form of a cup, as shown in Figure 3-4, when the gap is large, there is very little difference between the flux paths of this design and the square ended design, so the attractive force will be similar. However, as the gap gets smaller and the cone is inserted within the cup, the cross sectional area increases, but because the total flux remains unchanged the flux density is reduced and the force reduces proportionately.

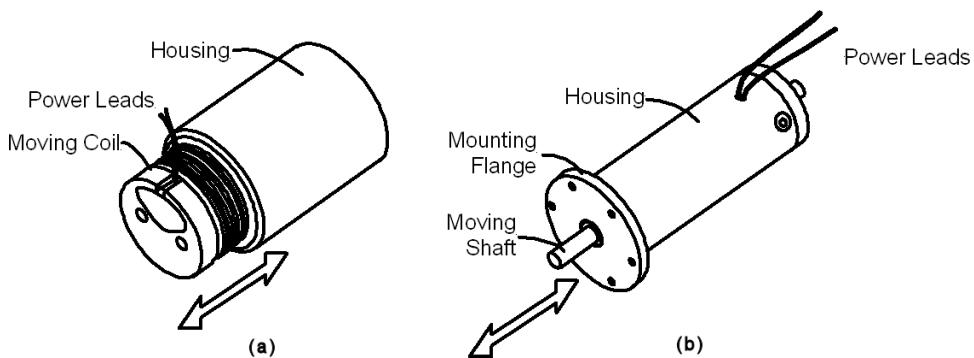
Most solenoids operate against return springs to ensure that once the current has been disconnected the plunger returns to the extended position. To guarantee that this occurs, a residual gap must remain when the solenoid is energized; otherwise, residual magnetism within the core can hold the plunger against the spring.

A frequently attempted experiment is to adjust the current supply to a solenoid to balance the load at some adjustable position. Unfortunately, this does not work—either the solenoid does not move, or it moves all the way to the inside stop. It is sometimes possible to achieve an unstable equilibrium, but there is no correcting force so any motion results in positive feedback and the plunger moves quickly to one of the extremes.

### 3.2.1.2 Voice Coil Actuators

Voice coil actuators are capable of moving an inertial load at extremely high accelerations ( $>20\text{g}$ ) and relocating it to an accuracy of better than  $10^{-5}\text{ mm}$  over a limited travel. Motion can be in a straight line (linear actuators) or in an arc (swing-arm actuators). Most voice coil actuators are used to position the heads of disk drives, but they are also found in shaker tables, medical equipment, lens-focusing applications, and servo systems.

Linear voice coil actuators come in two forms—moving coil and moving magnet types—as illustrated in Figure 3-5. Unlike solenoids, in voice coils a reasonably uniform



**FIGURE 3-5 ■**  
Voice coil actuators.  
(a) Moving coil.  
(b) Moving magnet.

magnetic field is generated across a fixed gap using a powerful permanent magnet. The coil lies within this space and is held in neutral position by a spring mechanism.

When current is applied to the coil, a force is generated that is proportional to the product of the flux density, the length of the conductor in the magnetic field, and the current, as described in equation (3.2). It can be seen that, unlike solenoids, there is a linear relationship between the current and the force and hence between the current and the displacement against a spring. An additional characteristic of voice coil actuators that differs from solenoids is that the direction of motion is determined by the direction of the current flow.

Linear voice coil actuators range from the minute, with peak forces from 0.7 N and strokes of 1 mm, to huge with forces of 2000 N and strokes of up to 50 mm. A typical voice coil actuator in the middle of the range is the LA10-12-027A manufactured by BEI Kimco. Its specifications are shown in Table 3-1.

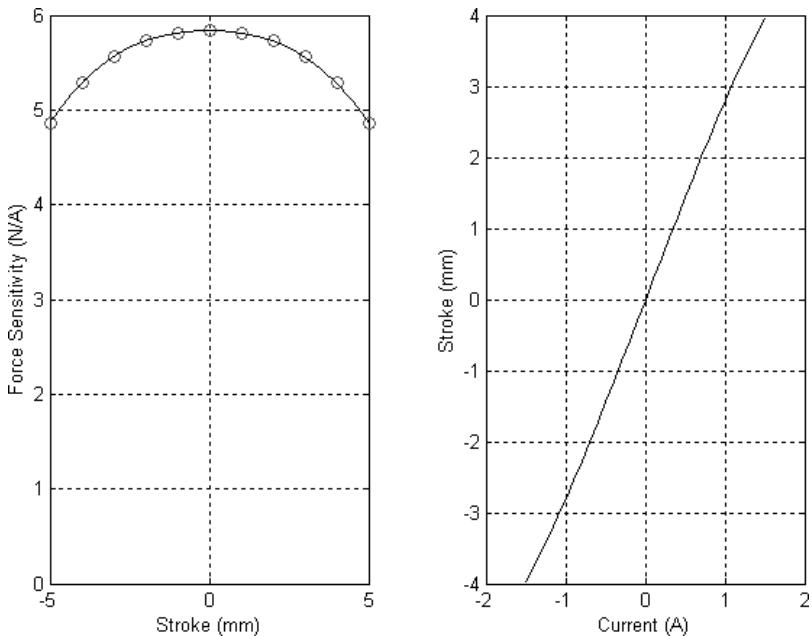
Because of end effects, the force sensitivity,  $K_f$  (N/A), is not completely constant over the stroke and will therefore result in some nonlinearity in the displacement when restrained by a spring, as shown in Figure 3-6.

**TABLE 3-1 ■** Specifications of the BEI Kimco LA10-12-027A Linear Voice Coil Actuator.

Winding Constants	Units	Tolerance	Symbol	Value
DC resistance	Ohms	+/-12.5%	R	11
Voltage	Volts	Nominal	$V_p$	25.1
Current	Amperes	Nominal	$I_p$	2.28
Force sensitivity	Newton/amp	+/-10%	$K_f$	5.87
Back EMF constant	V per m/s	+/-10%	$K_b$	5.87
Inductance	Milli Henry	+/-15@	L	3.05
Actuator parameters	Units		Symbol	Value
Peak force	N		$F_p$	13.3
Actuator constant	$N W^{-0.5}$		$K_a$	1.77
Electrical time constant	Microsecond		$\tau_e$	277
Mechanical time constant	Millisecond		$\tau_m$	4.43
Power from $I^2 R$	Watt		$P_p$	57.2
Stroke	mm		-	4.57

Notes: DC, direct current. EMF, electromotive force.

**FIGURE 3-6 ■ BEI Kimco LA10-12-27A linear voice coil actuator sensitivity for a spring constant  $k_s = 2 \text{ N/mm}$ .**



Because the relationship between the current is slightly nonlinear, really accurate positioning requires the addition of a linear displacement measurement sensor such as an linear variable differential transformer (LVDT) or a linear potentiometer as discussed in Chapter 2. However, for most applications the open-loop linearity is sufficiently good.

This description explains how the actuator can be positioned with good accuracy, but it does not describe the dynamics of reaching that equilibrium point.

When a conductor moves through a magnetic field, a voltage is induced across the conductor in a direction to oppose the motion (if current can flow); this is the back EMF as defined by equation (3.5). A voice coil will continue to accelerate, if the stroke is sufficiently long, until the back EMF is equal to the applied voltage, at which time the voice coil will have reached terminal velocity. In general, however, the velocity is limited to a value far below this.

In the case of the voice coil, the length of conductor is determined by the radius of the coil,  $a$  (m), and the number of turns,  $N$ , making the back EMF,  $\varepsilon_b$  (V),

$$\varepsilon_b = 2\pi NaBv \quad (3.15)$$

When voltage is suddenly applied to a voice coil, the inductance of the windings limits the rate of increase in the current, and the response of the device can be sluggish. To overcome this, a copper sleeve fixed to the magnet structure surrounds the coil. Eddy currents are induced in it by changes in the current through the coil. These currents cause a magnetic field that opposes the original field and cancels much of the apparent coil inductance, with the result that the response is much faster and more linear.

Should the high-frequency response of the voice coil be required when designing a control system, the electrical and mechanical time constants as well as the inductance must be considered.

Figure 3-7 shows a range of voice coil actuators including both linear and rotary types, manufactured by USAS Motion.



**FIGURE 3-7** ■  
Range of voice coil  
actuators  
manufactured.  
(Courtesy of USAS  
Motion.)

### 3.2.1.3 Biomechatronic Applications

As shown earlier in this chapter, the relationship between force and stroke for solenoids is very nonlinear. This limits their use in typical biomechatronic applications to operating latches engaging gears or driving pump diaphragms—applications that require only two states (in or out). In contrast, voice coil actuators provide a reasonably efficient method of generating linear motion, with open-loop positioning accuracies good to fractions of a millimeter. They can therefore be used to actuate prosthetic limbs, drive middle ear implantable hearing devices (MEIHDs), power ventricular assist devices (VADs), and even control the focus in miniature optical prostheses.

### 3.2.2 Direct Current Motors

For a current-carrying conductor in a magnetic field, the magnitude of the force,  $F$  (N), depends on the total length,  $l$  (m), of the conductor in the field, the magnetic flux,  $B$  (weber/m<sup>2</sup>), and the current,  $I$  (A)

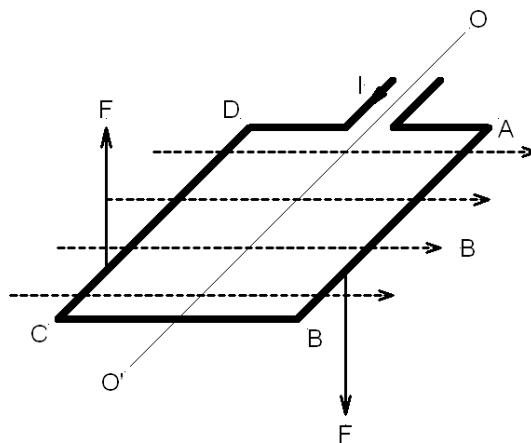
$$F = BIl \sin \alpha \quad (3.16)$$

where  $\alpha$  (rad), is the included angle between  $B$  and  $I$ .

If this single conductor is replaced by a single turn coil mounted on a rotational axis O – O', the interaction of the current in coil segment AB with the magnetic field  $B$  will create a force  $F$ , shown downward in Figure 3-8. In the same way, the coil segment CD will result in a similar force but in the opposite direction because the direction of the current flow is reversed in this case.

Note that the current in the coil segment BC is parallel to the magnetic field making the net force on this section zero. It is also obvious that there is no net force on the coil because the two forces are the same size, but there is a resultant moment (torque) around the rotational axis.

**FIGURE 3-8 ■**  
Application of  
Faraday's law for a  
single-loop coil.



If current continues to flow into the coil, it will rotate about the axis  $O - O'$  by  $90^\circ$  until the coil is orientated vertically and then will stop because the forces in the conductors will pass through the rotation axis and the turning moment will have reduced to zero.

If the current was reversed and the coil was just past the equilibrium point, then it would rotate by a further  $180^\circ$  and then stop again. To allow for continuous motion, some form of switching or commutation is required to automatically reverse the direction of the current at the correct angle. In brushed motors, these are performed using a mechanical method, while in brushless motors the position of the coil is sensed electronically using a Hall switch (see Chapter 2), and commutation also occurs electronically.

### 3.2.2.1 Single-Coil DC Motor

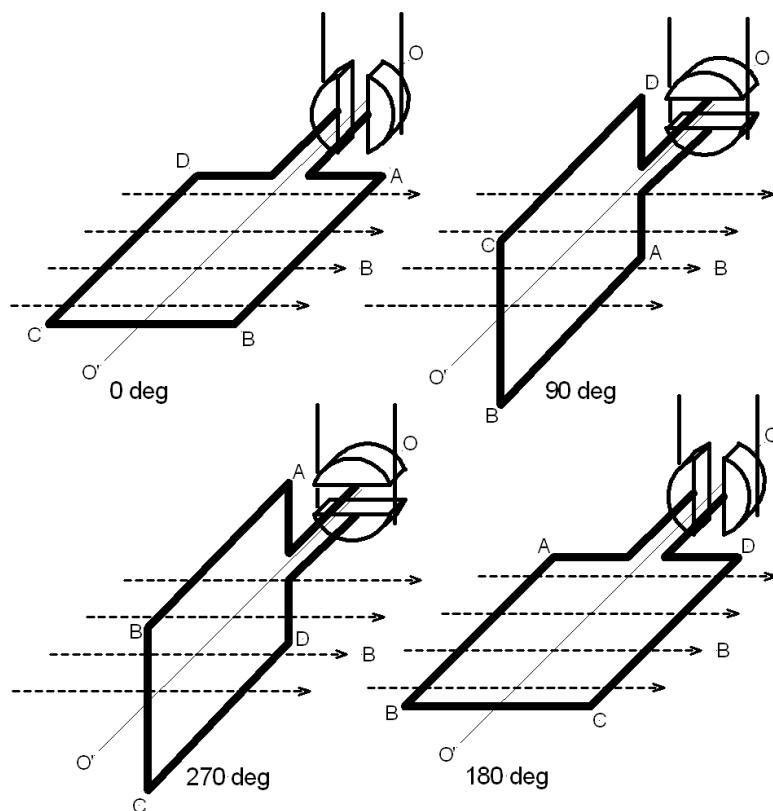
Mechanical commutation using brushes is shown in Figure 3-9. It can be seen that a pair of brushes remain in contact with the coil over angles from  $0^\circ$  to  $90^\circ$ . As with the previous figure, the force on the segment AB is downward, and the force on segment CD is upward. This produces a torque in the clockwise direction. There is a short period in the region around  $90^\circ$  where the brushes are disconnected and the motor must rely on its momentum to carry it past this region. The sense of the current into the coil is now reversed with the force on CD being downward and that on AB being upward, which continues to produce a small torque in the clockwise direction. As the angle increases, the torque reaches a maximum at  $180^\circ$  before decreasing to zero at  $270^\circ$ , where the commutation switches again.

This crude motor is impractical for a number of reasons, the most important of which is the large variation in the torque experienced, as shown graphically in Figure 3-10.

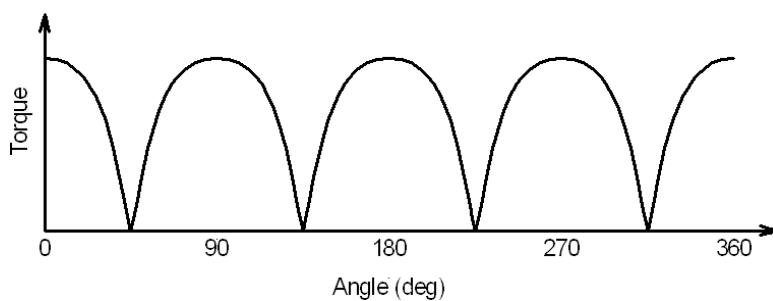
### 3.2.2.2 Multiple-Coil Direct Current Motor

If a second coil with its commutation mechanics is added at  $90^\circ$  to the single coil, the amount of ripple will be reduced significantly, as can be seen in Figure 3-11. In real motors, the number of coils is generally increased well beyond two coil segments with the result that there is very little torque ripple.

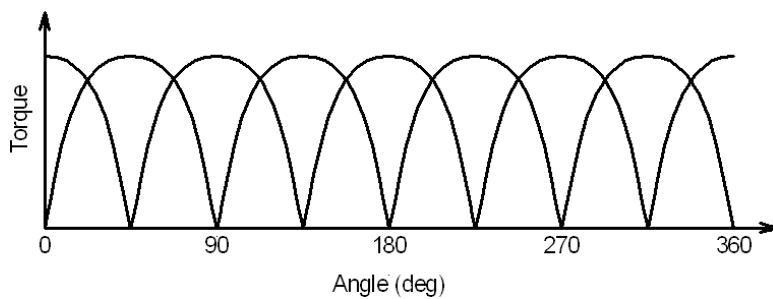
A second problem with these motors is the fact that they rely on a single turn, with the result that the amount of torque generated is limited. Real motors have multiturn coils to maximize the torque output.



**FIGURE 3-9** ■ Commutation of a single-coil DC motor.



**FIGURE 3-10** ■ Torque variation in a single-coil DC motor.



**FIGURE 3-11** ■ Torque variation in a two-coil DC motor.

For a coil with  $N$  turns, the maximum force generated by  $N$  parallel but insulated conductors running along the segment AB is

$$F = NBIl \quad (3.17)$$

The peak torque generated in the coil as a result of this force,  $F$  (N), is the product of the force and the distance from AB to the rotational axis O – O'. If this distance is  $R$  (m), then the peak torque,  $\tau_p$  (Nm), is

$$\tau_p = RNBl \quad (3.18)$$

Because the same amount of torque is developed by conductors running along segment CD, the peak torque will be doubled

$$\tau_p = 2RNBl \quad (3.19)$$

For any manufactured motor,  $R$ ,  $N$ ,  $B$ , and  $l$  are fixed, leaving the user control of the current  $I$  (A) only. Therefore, it is convenient to specify a motor in terms of the relationship between the current and the torque. This is known as the torque constant,  $K_m$  (Nm/A),

$$K_m = \frac{\tau}{I} = 2RNBl \quad (3.20)$$

One further characteristic of DC motors is important. It was shown earlier that when a conductor of length  $l$  (m) cuts through a magnetic field with strength  $B$  (weber/m<sup>2</sup>) at a velocity  $v$  (m/s), an EMF,  $\varepsilon$  (V), is developed across the conductor.

$$\varepsilon = Blv \quad (3.21)$$

If the same assumptions are adopted as were for the development of the motor, it can be shown that an EMF will develop across the brushes if an external torque is applied to rotate the coils physically. The magnitude of this voltage is dependent on the motor configuration and the angular velocity. This allows a constant of proportionality that describes the relationship between the angular velocity and the voltage to be determined. It is generally known as the back EMF constant,  $K_e$  (V per rad/s).

When a direct current (DC) motor that is not driving a load is supplied with a voltage  $V_i$  (V), its speed will increase until the back EMF just equals the applied voltage. This is known as the no-load speed,  $\omega_n$  (rad/s). As the load increases and the motor is required to provide an increasing amount of torque, the motor speed will decrease in a reasonably linear manner until the motor stalls and the torque is at a maximum. This is known as the stall torque,  $\tau_s$  (Nm). In other words, there is a trade-off between how much torque a DC motor can generate and how fast it spins, as shown in Figure 3-12.

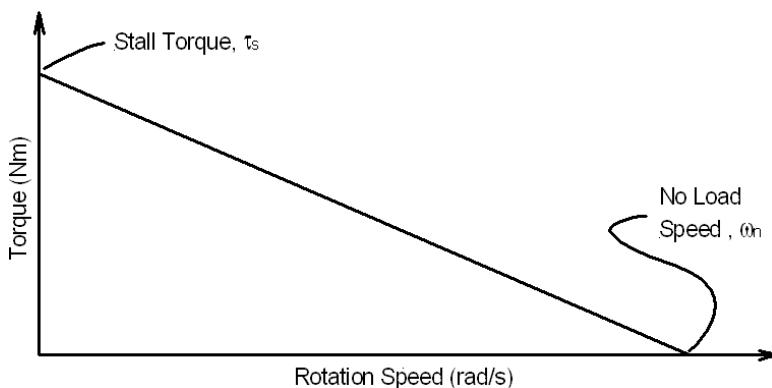
It is important to realize that the graphs shown in Figure 3-12 are for a specific applied voltage. If the applied voltage is altered, then the slope of the line remains the same but is just displaced vertically, as shown in Figure 3-13.

Given the two points at the extremes of the graph, an equation can be written describing the torque in terms of the rotation speed, or the motor speed in terms of the torque

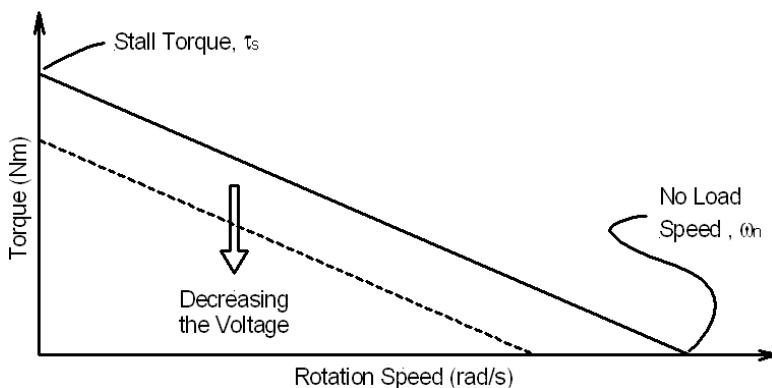
$$\tau_{mot} = \tau_s - \frac{\omega \tau_s}{\omega_n} \quad (3.22)$$

and

$$\omega_{mot} = \frac{(\tau_s - \tau) \omega_n}{\tau_s} \quad (3.23)$$



**FIGURE 3-12** ■ Relationship between motor speed and torque for a DC motor.



**FIGURE 3-13** ■ Effect of changing the applied voltage.

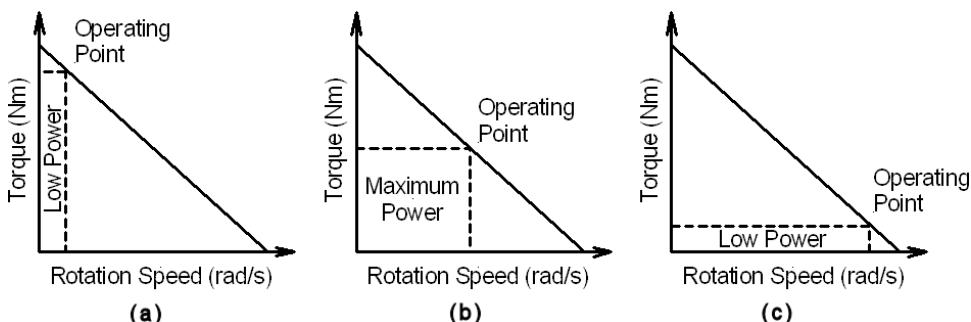
The output power,  $P_{mot}$  (W), supplied by the motor is equal to the product of the torque and the speed.

$$P_{mot} = \omega\tau \quad (3.24)$$

Therefore, at the two extremes no power is output, and a maximum occurs at the center of the graph where  $\omega = \omega_n/2$  and  $\tau = \tau_s/2$ . This relationship, which equates to the area of a rectangle below the curve, is shown graphically in Figure 3-14.

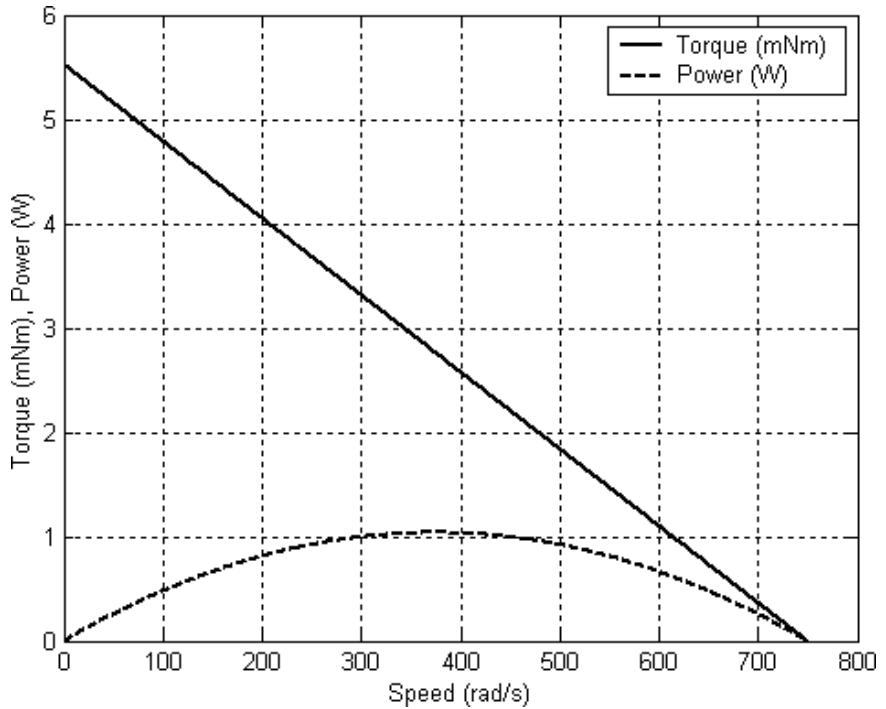
### 3.2.2.3 Real Motor Characteristics

A 2W Maxon motor designed to operate from 15 volts has a no-load speed of 7160 rpm (750 rad/s) and a stall torque of 5.53 mNm. The relationship among the torque, output



**FIGURE 3-14** ■ Relationship among speed, torque, and output power.  
 (a) Low speed.  
 (b) Moderate speed.  
 (c) High speed.

**FIGURE 3-15** ■ Maxon 2 W motor characteristics.



power, and rotation speed is shown in Figure 3-15. Note that the actual output power for this motor reaches a maximum of just over 1 watt when  $\omega = \omega_n/2$  and  $\tau = \tau_s/2$  as predicted.

The governing equation for the power curve can be obtained by substituting equation (3.22) or (3.23) into equation (3.24).

$$P_{mot}(\omega) = -\frac{\tau_s}{\omega_n}\omega^2 + \tau_s\omega \quad (3.25)$$

$$P_{mot}(\tau) = -\frac{\omega_n}{\tau_s}\tau^2 + \omega_n\tau \quad (3.26)$$

The current drawn by the motor is proportional to the output torque. The motor specifications include the no-load current,  $I_o = 4.91$  mA, and the starting current,  $I_a = 281$  mA.

The no-load current is equivalent to the friction torque  $\tau_f$

$$\tau_f = K_t I_o \quad (3.27)$$

where  $K_m$  (Nm/A) is the torque constant described earlier.

Motors develop the highest torque when starting, and this is much larger than the torque available when running. Therefore, the current required on startup is also a maximum

$$\tau_s = K_m I_a \quad (3.28)$$

From the motor specifications the torque constant  $K_m = 19.6$  mNm/A. This is in fact just the ratio of the stall torque and the starting current,  $K_m = \tau_s/I_a$ .

The friction torque can be calculated from the no-load current and the torque constant.

$$\begin{aligned}\tau_f &= K_m I_o \\ &= 19.6 \times 4.91 \times 10^{-3} \\ &= 0.096 \text{ mNm}\end{aligned}$$

The motor efficiency,  $\eta$ , describes the relationship between the mechanical power output by the motor and the electrical power supplied.

$$\eta = \frac{\omega\tau}{V_i I} = \frac{\omega K_m}{V_i} \quad (3.29)$$

It can be seen that the efficiency should increase with increasing speed (decreasing torque) for a fixed supply voltage. However, at low torque (high speed), friction losses become increasingly significant, and efficiency decreases rapidly. Maximum efficiency,  $\eta_{max}$ , is calculated using the starting current and the no-load current and is dependent on the supplied voltage

$$\eta_{max} = \left(1 - \sqrt{\frac{I_o}{I_a}}\right)^2 \quad (3.30)$$

For the specified motor

$$\begin{aligned}\eta_{max} &= \left(1 - \sqrt{\frac{I_o}{I_a}}\right)^2 \\ &= \left(1 - \sqrt{\frac{4.91}{281}}\right)^2 \\ &= 0.75\end{aligned}$$

which is close to the specified efficiency of 76%. As a rule of thumb, the maximum efficiency occurs at roughly one-seventh of the stall torque. This means that maximum efficiency and maximum output power do not occur at the same torque.

In addition to frictional losses (mechanical losses) there are copper losses due to the resistance of the windings and in iron-core motors, there are losses due to magnetization effects (electrical losses). Iron losses do not occur in modern coreless DC motors; therefore, the power balance can be written as

$$P_{el} = P_{mot} + P_{dis} \quad (3.31)$$

where  $P_{dis}$  (W) is all of the electrical losses that do not contribute to the generation of a torque.

This power balance can be expanded

$$V_i I = \omega\tau + I^2 R \quad (3.32)$$

### WORKED EXAMPLE

---

For the 2 W Maxon motor, the resistance of the coil windings is  $R = 53.3 \Omega$ , so the power dissipated can be determined. This is a maximum at startup when the motor is still stalled

$$\begin{aligned}P_{dis} &= I_a^2 R \\ &= 281 \times 10^{-3} \times 53.3 \\ &= 4.2 \text{ W}\end{aligned}$$

It exceeds the power rating of the motor by a factor of two; therefore, the motor cannot be operated in this regime for long periods without overheating.

If the motor operates at the peak power point where  $\tau = \tau_s/2 = 5.53/2 = 2.765$  mNm, the current will be the sum of the load current and the no-load current. The load current is determined from the output torque and the torque constant

$$I = \frac{\tau}{K_m} + I_o \quad (3.33)$$

$$\begin{aligned} I &= \frac{2.765}{19.6} + 4.91 \times 10^{-3} \\ &= 0.1411 + 4.91 \times 10^{-3} \\ &= 0.1460 \text{ A} \end{aligned}$$

This makes the electrical power dissipated in the coil windings

$$\begin{aligned} P_{dis} &= I^2 R \\ &= 0.146^2 \times 53.3 \\ &= 1.13 \text{ W} \end{aligned}$$

The power dissipated in the windings should, in theory, be equal to the power output as mechanical energy at this point. The efficiency can be determined as the ratio of the output power to the total power

$$\begin{aligned} \eta &= \frac{P_{out}}{P_{out} + P_{dis}} \\ &= \frac{1.04}{1.04 + 1.13} \\ &= 0.48 \end{aligned}$$

The temperature rise is determined from the thermal resistance between the windings and the air. For the 2 W motor, this comprises two components: the thermal resistance between (1) the housing and the air,  $R_{th1} = 33^\circ\text{C}/\text{W}$ ; and (2) the windings and the housing,  $R_{th2} = 7^\circ\text{C}/\text{W}$ . Under steady-state conditions the temperature difference between the motor windings and the ambient is

$$\Delta T = P_{dis}(R_{th1} + R_{th2}) \quad (3.34)$$

The temperature of the windings will then be the sum of the ambient air temperature and the temperature rise

$$T_{mot} = T_{amb} + \Delta T \quad (3.35)$$

In the case of the stalled motor

$$\begin{aligned} \Delta T &= P_{dis}(R_{th1} + R_{th2}) \\ &= 4.2 \times (33 + 7) \\ &= 168^\circ\text{C} \end{aligned}$$

The winding temperature will be

$$\begin{aligned} T_{mot} &= T_{amb} + \Delta T \\ &= 20 + 168 \\ &= 188^\circ\text{C} \end{aligned}$$

which is more than 100 °C hotter than the allowed 85 °C. The motor will overheat in this stalled condition, and the windings will burn.

For the motor operating at the peak power point

$$\begin{aligned}\Delta T &= P_{dis}(R_{th1} + R_{th2}) \\ &= 1.13 \times (33 + 7) \\ &= 45^{\circ}\text{C}\end{aligned}$$

the winding temperature will be

$$\begin{aligned}T_{mot} &= T_{amb} + \Delta T \\ &= 20 + 45 \\ &= 65^{\circ}\text{C}\end{aligned}$$

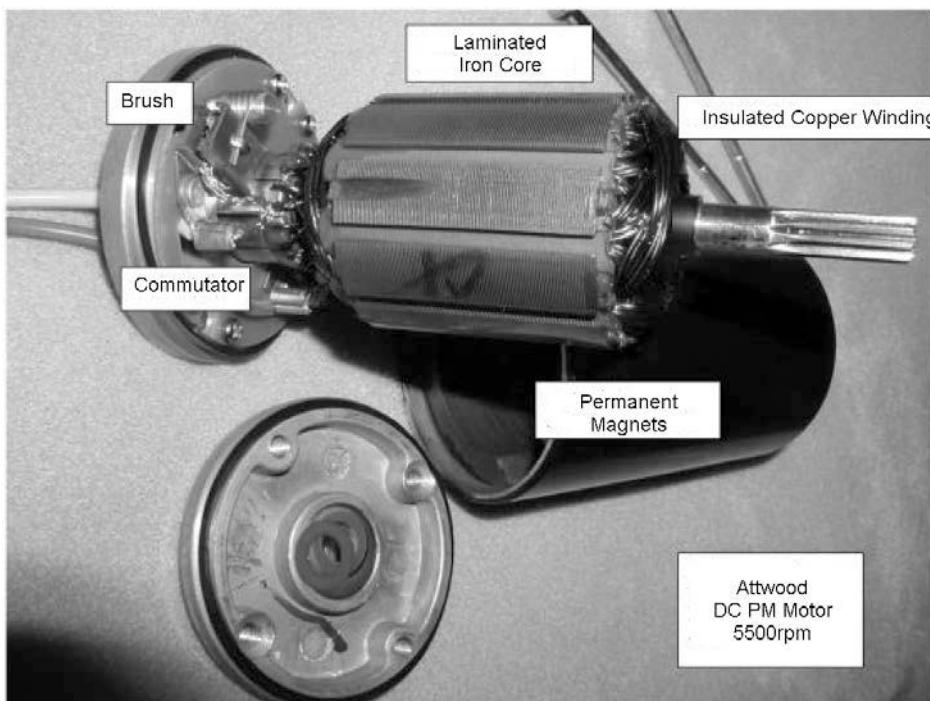
which is more than 20 °C cooler than the allowed 85 °C, so the motor will be fine.

---

### 3.2.2.4 Examples of DC Motor Types

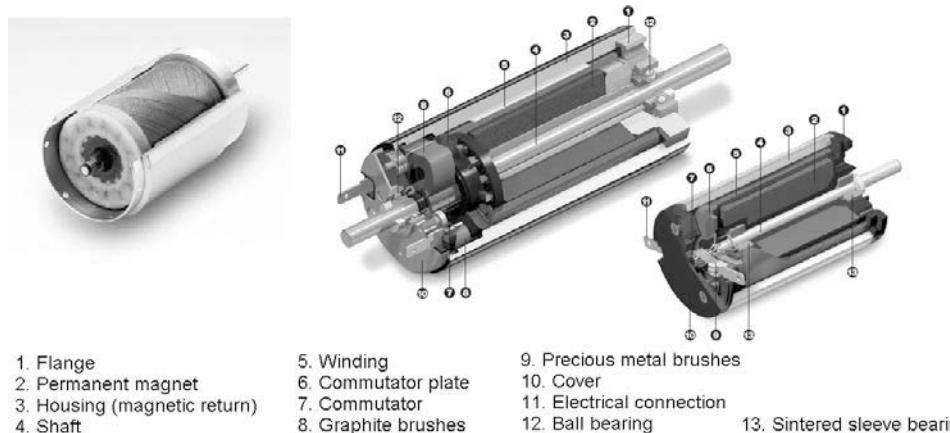
DC motors are probably the most common of all of the actuators used by biomechatronic engineers. This is because they are extremely efficient in converting electric power to mechanical power and there is a wide variety of types from which to choose.

A conventional low-cost iron core permanent magnet (PM) DC motor is shown in Figure 3-16. It can be seen that the armature (rotor) consists of a laminated iron core wound with insulated copper coils that are connected to a multicontact commutator. The stator consists of two ferrite magnets within a cylindrical housing. End plates containing journal bearings hold the rotor between the magnets while allowing free rotation. Attached to one of the end plates is a brush assembly with carbon brushes in contact with the commutator.



**FIGURE 3-16 ■**  
Typical permanent magnet DC motor.  
(Courtesy of Attwood.)

**FIGURE 3-17 ■**  
**Construction of an ironless DC motor.**  
(Courtesy of  
Portescap,  
[http://www.  
portescap.com/](http://www.portescap.com/).)



Modern high-performance motors made by companies like Maxon, Portescap, or Faulhaber are mostly ironless as this reduces the inertia and increases the space available for coil windings. As can be seen from Figure 3-17, the rotor windings are in the form of a cup with the magnetic stator in the center. This construction places the windings close to the outside of the motor for improved heat dissipation.

Mechanical commutation has always been one of the major disadvantages of DC motors. This results in wear and hence limits the operational life of the motor; it generates metal or graphite particles and electromagnetic interference. Graphite brushes are generally used in larger motors for stop–start operation and if the motor controller uses pulse-width modulation (PWM). Precious metal brushes ensure a highly constant and low-contact resistance and so are good for low-power continuous operation or for use in DC tachometers.

### 3.2.2.5 Selecting a DC Motor

Selecting a DC motor for a particular application can be a rather involved process if undertaken comprehensively. However, it is not difficult to obtain a reasonable idea of the performance requirements from which to make a reasonable choice.

The following considerations can be important:

- Constant current operation of a DC motor produces constant output torque regardless of speed.
- Given a constant load (torque requirement), the motor speed is determined solely by the applied voltage.
- Power is the product of speed and torque. Maximum power is produced at an operating point that is defined by half the no-load speed and half the stall torque. A motor will seldom be operated at maximum output due to thermal considerations.
- As a rule of thumb, DC motors are typically operated at 70% to 90% of the no-load speed and from 10% to 30% of the stall torque. This is the region of maximum efficiency.
- For DC motors operated at a constant voltage, the higher the torque output, the lower the speed will be.
- Other factors to consider are size, environmental factors, weight, and required life.

### WORKED EXAMPLE

---

#### Motor for CPAP device

As an example, consider the motor required to drive the turbine for a continuous positive airway pressure (CPAP) air pump. The nominal torque required is 3.5 mNm at a speed of 5000 rpm. A DC supply voltage of 20 V is available.

As a first step, determine the output power. Determine the conversion factor for rotational speed from rpm (n) to rad/s ( $\omega$ ),  $2\pi/60 = 0.1047$ .

$$\begin{aligned} P_o &= 0.1047n\tau \\ &= 0.1047 \times 5000 \times 3.5 \times 10^{-3} \\ &= 1.8 \text{ W} \end{aligned}$$

The motor should be rated at least 1.5 to 2 times the desired output power in relation to the maximum available at the nominal voltage. A motor with a maximum output power of between 2.7 and 3.6 W should be sufficient.

Referring to the Maxon catalog, the smallest motor to achieve this power rating is the RE 16 series (16 mm diameter  $\times$  40 mm long). The 24 V version has the closest operating voltage to the 20 V available. Some of the motor specifications are reproduced in Table 3-2.

**TABLE 3-2 ■** Maxon RE 16 Motor Specifications

No-load speed @ 24V $n_o$ (rpm)	7250
No-load current $I_o$ (mA)	3.11
Terminal resistance $R$ ( $\Omega$ )	42.8
Torque constant $K_m$ (mNm/A)	31.4
Speed constant $K_e$ (rpm/V)	304
Output power $P_o$ (W)	3.2
Maximum winding temperature, $T_{max}$ ( $^{\circ}\text{C}$ )	85
Thermal resistance: housing ambient, $R_{th1}$ ( $^{\circ}\text{C}/\text{W}$ )	30
Thermal resistance: winding housing, $R_{th2}$ ( $^{\circ}\text{C}/\text{W}$ )	8.5

As a first approximation, the no-load speed can be found by taking the specified value and scaling it by the ratio of the voltages

$$\begin{aligned} n_{20} &= \frac{20}{24} n_o \\ &= \frac{20}{24} \times 7250 \\ &= 6041 \text{ rpm} \end{aligned}$$

Since the desired speed is 5000 rpm, this represents 82% of the no-load speed, which falls well within the 70% to 90% range, so the motor should operate at close to optimum efficiency.

The current through the motor is the sum of the no-load current and the load current

$$\begin{aligned} I &= \frac{\tau}{K_m} + I_o \\ &= \frac{3.5}{31.4} + 3.11 \times 10^{-3} \\ &= 114.6 \times 10^{-3} \text{ A} \end{aligned}$$

The volt drop across the motor windings is the product of the motor current and the terminal resistance. The reduction in motor speed is equal to the product of this volt drop and the speed

constant; therefore, the motor speed is

$$\begin{aligned} n &= n_{20} - IRK_e \\ &= 6041 - 114.6 \times 10^{-3} \times 42.8 \times 304 \\ &= 4550 \text{ rpm} \end{aligned}$$

This is reasonably close to the desired value; therefore, the motor will probably be suitable.

One final check is to determine whether the motor will be able to accommodate the heat rise in the coils. The power dissipated in the motor is

$$\begin{aligned} P_{dis} &= I^2 R \\ &= (114.6 \times 10^{-3})^2 \times 42.8 \\ &= 0.6 \text{ W} \end{aligned}$$

The heat rise under steady-state conditions is determined by the product of the dissipated power and the thermal resistance from the rotor coils to the air.

$$\begin{aligned} \Delta T &= P_{dis} (R_{th1} + R_{th2}) \\ &= 0.6 \times (30 + 8.5) \\ &= 23.1 \text{ }^\circ\text{C} \end{aligned}$$

Hence, the motor winding temperature will be the sum of the ambient plus the temperature rise.

$$\begin{aligned} T_{mot} &= T_{amb} + \Delta T \\ &= 20 + 23.1 \\ &= 43.1 \text{ }^\circ\text{C} \end{aligned}$$

This is well below the 85 °C maximum operating temperature for the motor, so the motor will be capable of running indefinitely under these conditions (MicroMo, 2008a).

---

### 3.2.2.6 Powering DC Motors

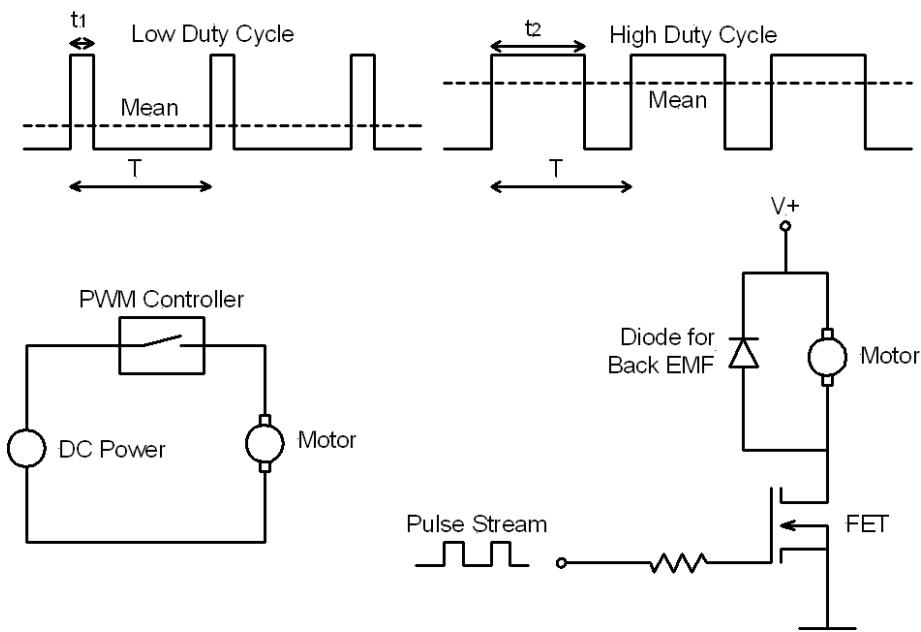
The simplest form of motor control is openloop. The voltage, or sometimes the current, is set to some preset value determined by the required output speed given a specific load torque. The voltage can be provided by a general purpose linear or switched mode regulator as discussed in Chapter 2, but it is more usual to develop specific electronics for motor control. Linear regulators are very inefficient, particularly if the motor speed needs to be controlled over a wide range. Thus, most motor controllers use pulse-width modulation to provide the required average DC voltage, as shown in Figure 3-18.

PWM is achieved by connecting the DC supply to the motor at regular intervals to provide the required mean voltage. Because the switching rate is very high, motor inductance and its rotational inertia ensure that it runs at a constant speed. The relationship among the switching frequency,  $f$  (Hz), the period,  $T$  (s), and the duty cycle,  $\eta\%$ , is as follows:

$$T = \frac{1}{f} \quad (3.36)$$

$$\eta\% = 100 \frac{t}{T} \quad (3.37)$$

where  $t$  (s) is the on period as shown in Figure 3-18.



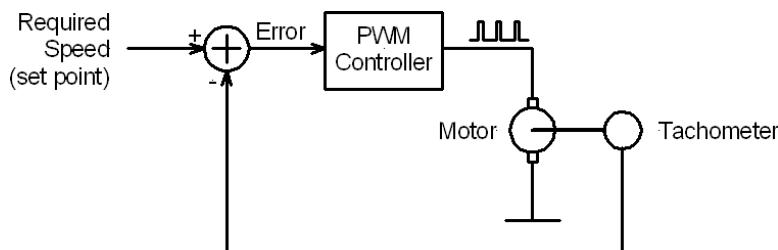
**FIGURE 3-18 ■**  
PWM of a DC motor.

Because of the high switching rate required, typically in excess of 1 kHz PWM, controllers use not physical switches or relays but power transistors, as shown in the partial schematic.

To maintain a constant speed or a specific motion profile that is torque dependent, closed-loop or feedback control is generally used. This involves sensing the motor speed using one of the sensors discussed in Chapter 2 and using that to control the duty cycle of the PWM signal to maintain the required speed. This principle is illustrated in Figure 3-19.

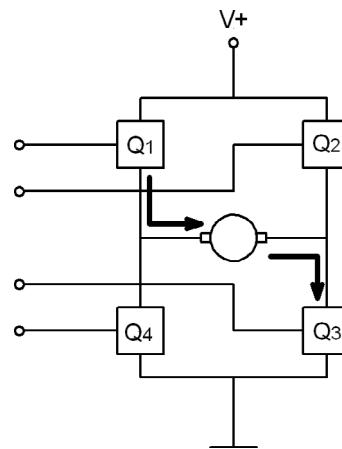
Simple PWM controllers cannot reverse the direction of rotation of a motor unless a dual rail supply is used. This is often inconvenient, particularly if the motor is large and requires a high current, so a more common alternative is to use a switching configuration that can reverse the polarity of the motor supply. This configuration, known as an H-bridge because of its shape, is shown in Figure 3-20.

The H-bridge uses four power transistors that are independently controlled to ensure that the appropriate pulse width is applied to the motor with the correct polarity. If transistors  $Q_1$  and  $Q_3$  are on simultaneously, then current will flow through the motor in the direction shown in Figure 3-20. However, if  $Q_2$  and  $Q_4$  are on, then the current flow is reversed and the motor will turn in the opposite direction. Care needs to be taken to ensure that  $Q_1$  and  $Q_2$  or  $Q_3$  and  $Q_4$  are never conducting simultaneously, or they will provide a direct path to Earth with catastrophic consequences.



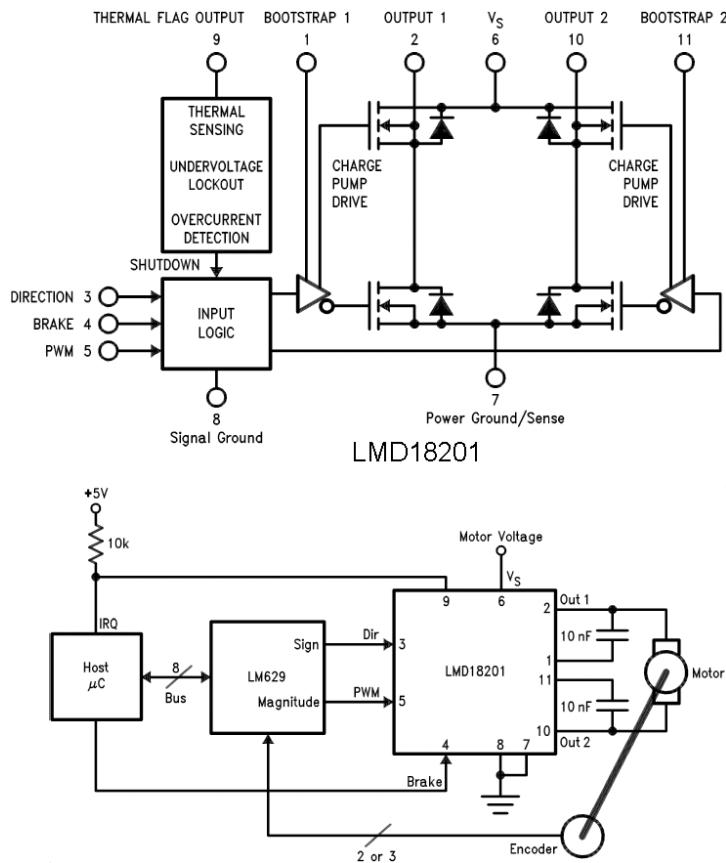
**FIGURE 3-19 ■**  
PWM motor speed control.

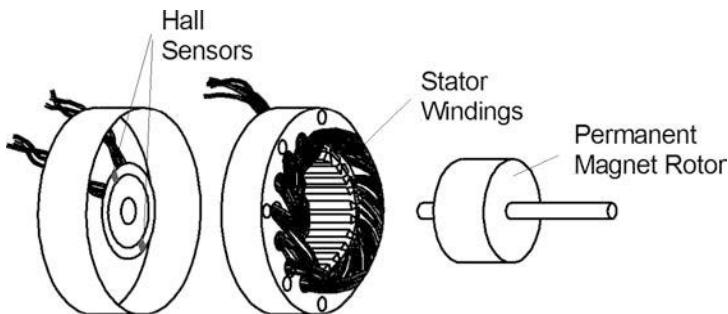
**FIGURE 3-20** ■  
H-bridge control of a motor.



Motor controllers can be built up using discrete components, but a number of available sophisticated integrated circuits (ICs) would probably perform the function much better. A good example is the National Semiconductor LMD18201, shown in Figure 3-21. This H-bridge can handle 3 A continuous and 6 A peak at a motor voltage of up to 55 V. In addition, it has a number of protection circuits including thermal and overcurrent.

**FIGURE 3-21** ■  
National LMD18201  
H-bridge and  
application.  
(Courtesy of National  
Semiconductor.)





**FIGURE 3-22** ■  
Exploded view of a brushless DC motor.  
[Adapted from (Kuphaldt 2008)].

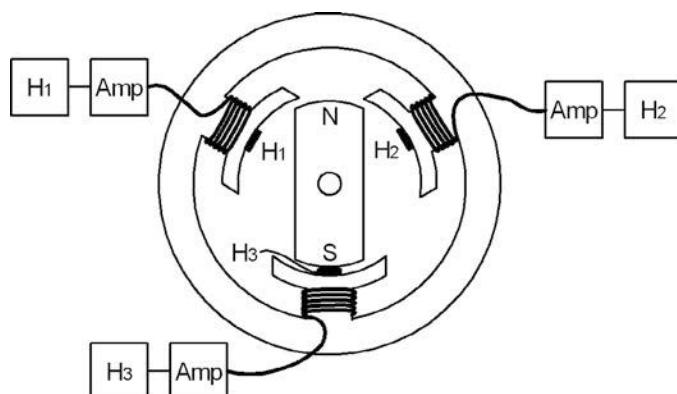
As shown in Figure 3-21, this particular H-bridge requires few external parts and is easily interfaced to an external microcontroller.

### 3.2.3 Brushless DC Motors

Modern brushless DC (BLDC) motors are very similar in construction to permanent magnet alternate current (AC) synchronous motors. As shown in Figure 3-22, the brushless motor comprises a permanent magnet rotor, a stator winding, and a number of Hall or optical sensors to detect the orientation of the rotor. These control current flow into the stator windings to turn the rotor.

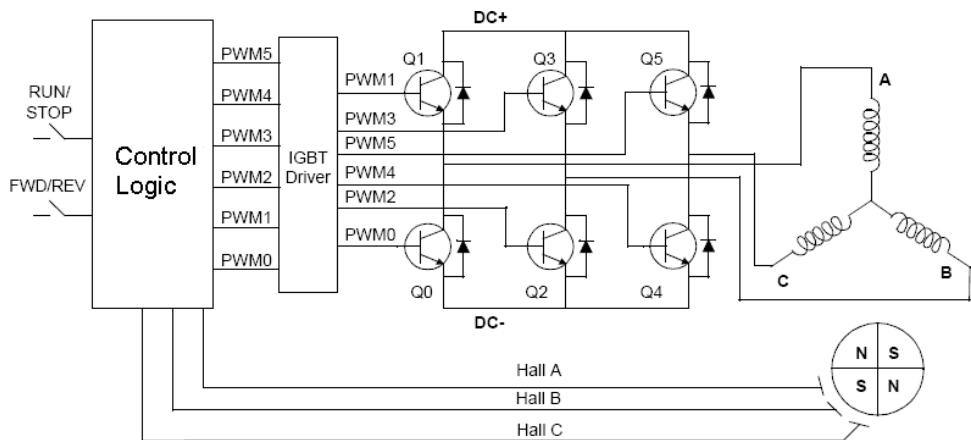
The simple cylindrical three-phase motor shown in Figure 3-23 is commutated by a separate Hall effect device for each of the three stator phases. The changing position of the permanent magnet rotor is sensed by the Hall device as the polarity of the passing rotor pole changes. This Hall signal is amplified so that the stator coils are driven with the proper current.

Most commercial BLDC motors have three-phase windings connected in a star topology, as illustrated in Figure 3-24. A motor with this topology is driven by energizing two phases at a time. The key to efficient operation is to sense the rotor position and then to energize the phases to produce the most torque in the desired direction. The appropriate stator is energized when the rotor is  $120^\circ$  from alignment with the corresponding stator's magnetic field and is deactivated when the rotor reaches  $60^\circ$  from alignment. This process is repeated for the next circuit, and so on, until a full cycle is complete.

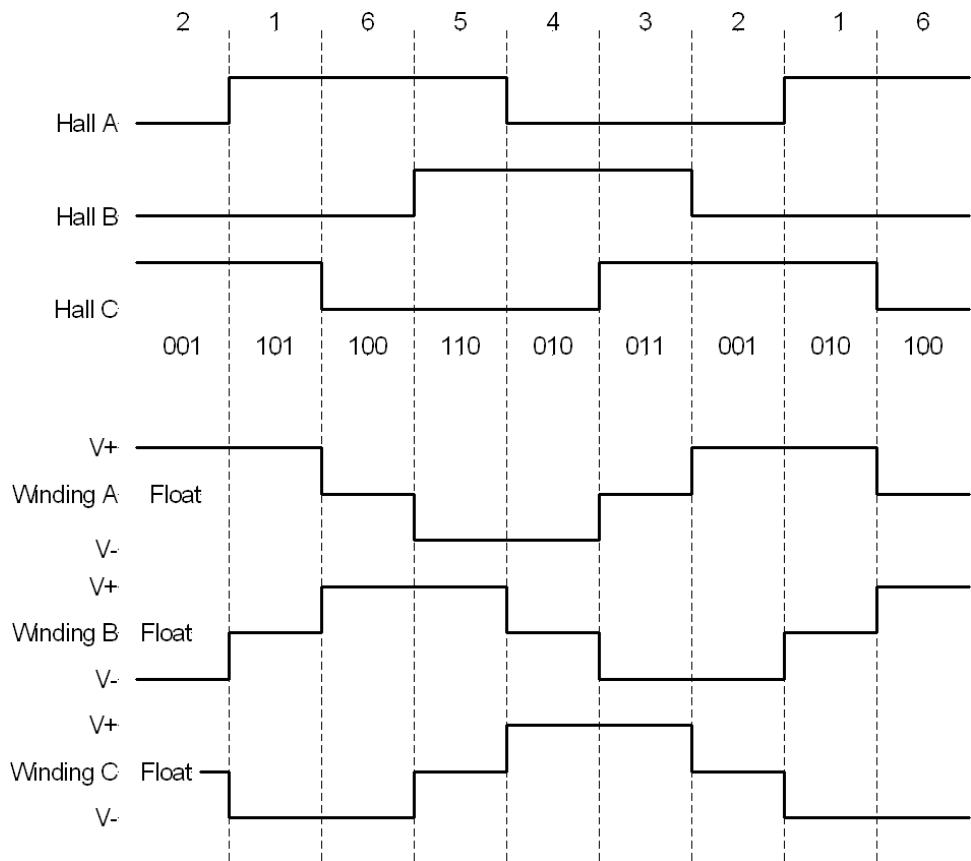


**FIGURE 3-23** ■  
Simple BLDC motor.  
[Adapted from (Kuphaldt 2008)].

**FIGURE 3-24** ■ Controller for a three-phase brushless DC motor. [Adapted from (Yedamale 2003)].

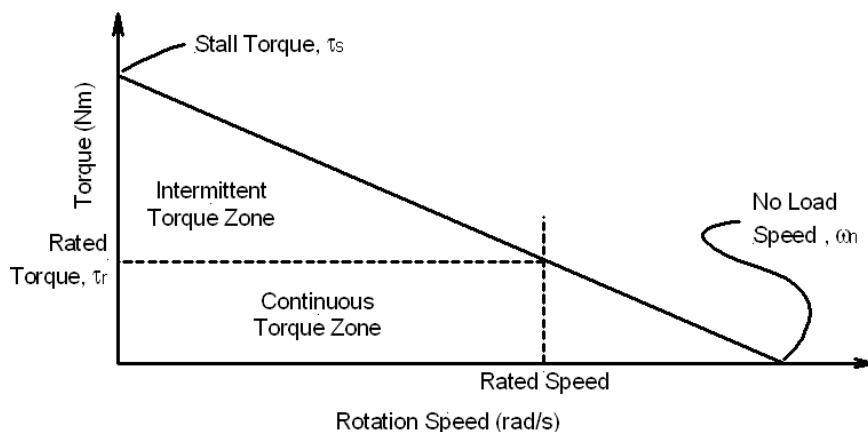


**FIGURE 3-25** ■ Hall sensor and drive timing. (Brown 2002).



The Hall sensors have digital outputs that are high for  $180^\circ$  of rotor rotation and then low for the remainder. They are physically separated by  $60^\circ$  so that an unambiguous digital code is produced for every  $60^\circ$  of rotation, as shown in Figure 3-25.

Note that each drive phase consists of one motor winding driven high, one driven low, and the last one left floating. A necessary precaution with this kind of driver is that both the high- and low-side drivers must never be actuated simultaneously. This is often



**FIGURE 3-26** ■  
Torque–speed  
characteristics for a  
brushless DC motor.

achieved using *dead time* control, in which extra time is allowed for one driver to turn off before the next one is activated. In this configuration there is a floating phase period between high and low drive periods so that the problem will never occur.

The torque–speed characteristics of BLDC motors are different from those of DC motors. There are two torque parameters: (1) the stall torque,  $\tau_s$ ; and (2) the rated torque,  $\tau_r$ . During continuous operation the motor can be loaded up to the rated torque, which remains constant for a speed range up to the rated speed, as shown in Figure 3-26. The motor can be run faster than this, up to about 150% of the rated speed, but the torque drops. In addition, the motor can generate torques higher than the rated value, but these must be limited in duration; otherwise, the motor can overheat.

It is possible to control a BLDC without Hall sensor feedback. If the back EMF of each of the windings is monitored, it will transition through zero at about the same time that the Hall sensor state would change. This transition can be used to control the commutation as already discussed.

Though BLDC motors are much more complicated than normal brushed types, they are often used in demanding applications because of their improved reliability and compactness. Table 3-3 provides a comparison between the main features of BLDC and brushed DC motors.

### 3.2.3.1 Selecting a BLDC Motor

The parameters that govern motor selection for a particular application are as follows (Yedamale 2003):

- Peak torque required for the application
- Root mean square (RMS) torque required
- The range of operating speeds required

**Peak torque:** This can be thought of as the same as the stall torque and can be determined by summing the required load torque,  $\tau_l$ , the inertial torque,  $\tau_j$ , and the torque required to overcome friction,  $\tau_f$ . Other factors also contribute to the peak torque requirements. They include windage loss, which is contributed by the resistance of the air in the gap. These additional factors are difficult to calculate, so it is easier to add a safety factor of 0.2 when doing the calculation.

$$\tau_s = 1.2(\tau_l + \tau_j + \tau_f) \quad (3.38)$$

**TABLE 3-3** ■ Comparison Between a Brushless DC Motor and a Brushed DC Motor [Adapted from (Yedamale 2003)].

Feature	BLDC Motor	Brushed DC Motor
Commutation	Electronic commutation—Hall sensors	Mechanical commutation—brushes
Maintenance	Less required due to absence of brushes	Periodic maintenance
Life	Longer	Shorter
Speed/Torque	Flat—operation at all speeds with rated load	Moderately flat—at higher speeds brush friction increases, reducing output torque
Efficiency	High—no voltage drop across brushes	Moderate
Output Power/Frame Size	High—good thermal characteristics because BLDC has windings on the outside in contact with the frame	Moderate to low—heat dissipation across the air gap
Rotor Inertia	Low because the rotor is only a permanent magnet	Higher inertia for iron rotor motors, lower for ironless types
Speed range	Higher—no mechanical limitation imposed by brushes	Lower—mechanical limitations of brushes
EMI	Low	High—brush arcing causes EMI
Cost	Higher—Permanent magnet rotor expensive to manufacture	Low
Control	Complex and expensive	Simple and inexpensive
Control requirements.	A controller is always required to run the motor, though the same controller can also be used for speed control	No controller is required for fixed speed, a controller is required for variable speed operation

The inertial torque,  $\tau_j$ , is the torque required to accelerate the load from standstill to operational speed. It is the product of the load plus rotor inertias,  $J_{l+m}$  ( $\text{kg m}^2$ ), and the required acceleration,  $\alpha$  ( $\text{m/s}^2$ ).

$$\tau_j = J_{l+m}\alpha \quad (3.39)$$

The load and frictional torques are mostly determined by the mechanical system coupled to the motor shaft.

**Root mean square torque:** The RMS torque can be thought of as the average torque requirements during normal operation. This depends on, for example, the duty cycle of the system, the load torque, and required accelerations. In an application where the acceleration time is  $t_a$  (s), the run time is  $t_r$  (s), and the deceleration time is  $t_d$  (s), the RMS torque is

$$\tau_{rms} = \sqrt{\frac{\tau_s^2 t_a + (\tau_l + \tau_f)^2 t_r + (\tau_j - \tau_l - \tau_f)^2 t_d}{t_a + t_r + t_d}} \quad (3.40)$$

A good example of this is a powered prosthetic arm that may be called on to generate high peak torques for short periods but that has a low RMS value. It should therefore be possible to underspecify the motor to minimize its mass, with the knowledge that the duty cycle will be low.

**Speed range:** The motor speed required to drive the application is determined by that application. In an air pump for a CPAP application, speed variations are not very frequent, and the maximum speed required will be very close to the average. However, in the prosthetic arm example, the peak speed requirement will be much higher than the average.



**FIGURE 3-27 ■**  
Examples of some  
BLDC motors from  
various  
manufacturers.

### 3.2.3.2 BLDC Motor Types

Motors can, of course, be manufactured in all shapes and sizes to fulfill particular torque and speed requirements of specific applications. These include long, pencil-thin motors and flat pancake motors, as shown in Figure 3-27. Some of the more common BLDC motors are those used in hard-disk drives, CD/DVD drives, and electric fans.

A good example of one of the latest generations of high power-density motors is the EC16 from Maxon. It is 16 mm diameter  $\times$  56 mm long, can rotate at up to 60,000 rpm, and produces an output power of 40 W.

Power densities higher than this are available from the electric motors used to power radio control aircraft. These typically rotate at about 12,000 rpm and provide output powers of up to 300 W in a package less than 30 mm in diameter and only 40 mm long.

### 3.2.3.3 Biomechatronic Applications

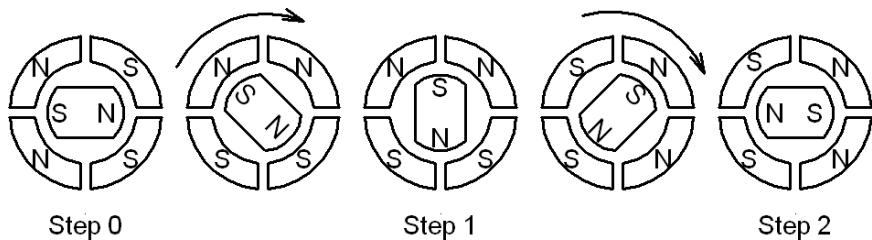
Brushless DC motors are the actuators of choice for a large number of biomechatronic applications. They are available in a wide range of sizes with associated gearheads for almost any speed-torque requirement, which makes them suitable for driving pulsatile artificial hearts, other pumps, and prosthetic limbs.

### 3.2.4 Stepper Motors

Stepper motors are DC motors comprising a permanent magnet or variable reluctance core that can rotate in precise angular increments in either direction and can sustain a strong holding torque at zero speed. These motors receive digitally controlled pulses correctly phased that cause them to step by an incremental angle that is typically between 1.8° and 30° per step. Microstepping circuitry can be designed to provide more than 10,000 steps per revolution (Alciatore and Histan, 2003). There are three basic types of stepper motors: (1) variable reluctance (VR); (2) hybrid; and (3) permanent magnet (PM).

Hybrid and PM steppers are bidirectional and can be operated over a wide speed range from incremental to speeds as high as 1800 rpm and from fixed to variable. The mechanical parameters including shaft and mounting configurations as well as gear boxes and pulley arrangements can be specified to meet particular requirements.

**FIGURE 3-28 ■**  
Stepper motor excitation sequence.



A commercial stepper motor has a large number of poles that provide many equilibrium positions on the rotor. In PM steppers, the stator consists of wound poles, and the rotor poles are permanent magnets. Exciting different stator winding combinations moves and holds the rotor at different positions. The VR motor has a ferromagnetic rotor that relies on the fact that the equilibrium points are in positions of minimum magnetic reluctance. These motors generally have a lower inertia than the magnetic variety, but they lack a holding torque unless energized. Hybrids are a combination of the two, with a core like a VR type that encases a permanent magnet.

Hybrid motors are typically chosen for applications needing a fine step angle output of  $1.8^\circ$ . They are also more efficient than PM steppers but also carry a higher cost. PM steppers are normally available in step increments from  $3.6^\circ$  to  $18^\circ$  and are inherently more adaptable. They are frequently less expensive and enjoy broad acceptance in diverse medical equipment applications.

PM stepper motors typically range in size from 15 to 70 mm in diameter and with gear reductions can generate output torques of 15 Nm or more.

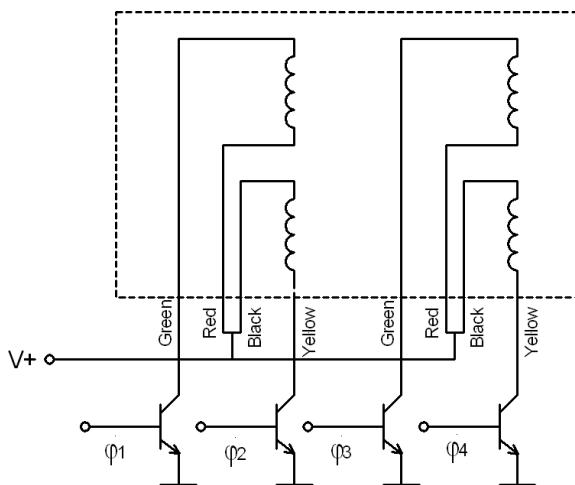
### 3.2.4.1 Operational Principle

To understand how a stepper motor works, consider the diagram shown in Figure 3-28. It consists of four stator poles and a permanent magnet rotor. The magnetic field on opposite poles of the stator can be reversed. At Step 0, the motor is in equilibrium with the opposing end permanent magnetic fields holding the rotor. If the polarity of one pair of stator windings is reversed, the rotor is no longer in equilibrium and will start to rotate as shown until it reaches equilibrium at Step 1. The polarity of the opposite pair of stator windings is then reversed, and the rotor rotates again until it reaches equilibrium at Step 2. This continues indefinitely.

The timing of the field reversals is important, particularly in multipole motors where the response can be underdamped, and it is easy to skip a step if the rate is too high. Damping can be increased, but even so the motor needs sufficient time to settle between steps.

### 3.2.4.2 Driving Stepper Motors

Stepper motors use two types of coil windings—unipolar or bipolar—which determine the kind of driver that is used. Unipolar coils incorporate bifilar windings with one center tap common to both coils. The center tap is normally tied to positive voltage. Unipolar circuitry is more cost-effective and is therefore used for high-volume applications. Bipolar coils contain one winding each with about the same number of turns found in unipolar coils, but with a larger wire gauge. This gives bipolar coils about 30% more torque, particularly at lower speeds. However, because it has no center tap, a bipolar coil requires additional circuitry to switch the direction of the current. This is accomplished cost-effectively with modern one-chip drivers.

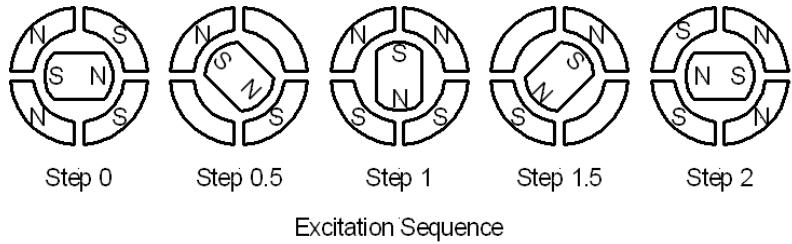


**FIGURE 3-29** ■ Schematic of a standard unipolar stepper motor drive configuration. [Adapted from (Alciatore et al., 2003)].

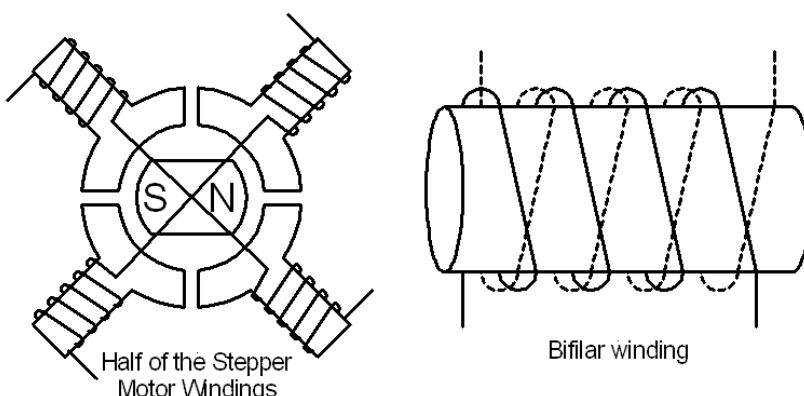
Unipolar configurations use a single voltage rail but can still drive in both directions because of the way that the windings are configured, as seen in Figure 3-29. The conventional color codes for these stepper motors are indicated in the diagram.

To implement the appropriate magnetic configurations of the stator shown in Figure 3-28, each pole pair must be wound with a pair of complementary windings (bifilar windings), as shown in Figure 3-30. A useful result of this configuration is the ability to half-step, where only one pair of windings is excited. The switch sequences for full stepping and half-stepping are listed in Table 3-4.

In the half-step mode, the resolution, or number of steps, is twice that of the full-step mode. In this case it increases from 4 steps ( $90^\circ$ ) per revolution to 8 steps ( $45^\circ$ ) per revolution.



**FIGURE 3-30** ■ Operation of a unipolar stepper motor with bifilar windings.



**TABLE 3-4** ■ Phase Sequences for Full-and Half-Step Drive

Step	$\varphi_1$	$\varphi_2$	$\varphi_3$	$\varphi_4$
1	On	Off	On	Off
2	On	Off	Off	On
3	Off	On	Off	On
4	Off	On	On	Off

Step	$\varphi_1$	$\varphi_2$	$\varphi_3$	$\varphi_4$
1	On	Off	On	Off
1.5	On	Off	Off	Off
2	On	Off	Off	On
2.5	Off	Off	Off	On
3	Off	On	Off	On
3.5	Off	On	Off	Off
4	Off	On	On	Off
4.5	Off	Off	On	Off

To improve the resolution a process called microstepping can be used. Instead of operating the windings in a digital on–off mode, the current flow is regulated to a number of different levels resulting in different magnetic equilibrium positions between the poles. In effect, discretized sine waves instead of square waves are applied to the poles with the result that a normal 200-step motor can be made to have 10,000 steps or more.

### 3.2.4.3 Real Stepper Motors

Figure 3-31 shows a photograph of a conventional stepper motor that has been dismantled. It is a hybrid configuration, as it combines the variable reluctance toothed rotor with a permanent magnet core. In this case the magnet is aligned along the axis of the motor so that one of the toothed sections is the south pole and the other is the north pole. The teeth in the two sections are misaligned by half a pitch.

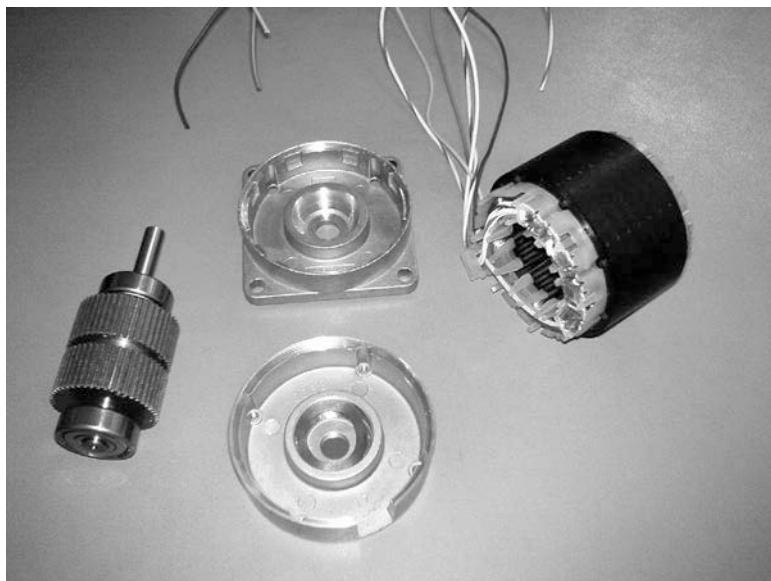
In the configuration shown in Figure 3-31, no effective torque is generated by the magnetic field of the coils alone, and only in combination with the fixed magnetic field of the permanent magnet is a rotational torque generated.

A simplified diagram showing the windings and the core of a typical stepper motor is shown in Figure 3-32.

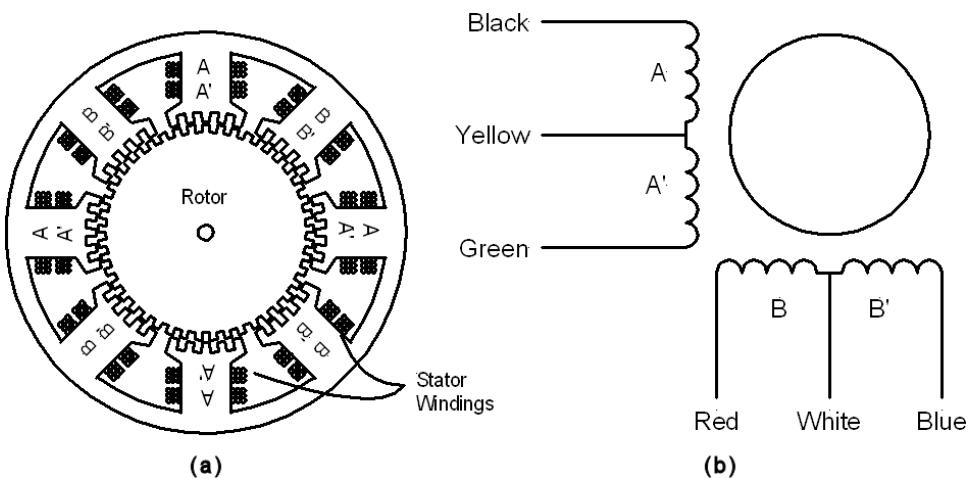
### 3.2.4.4 Stepper Motor Characteristics

**Static characteristics:** The stepper motor can be kept stationary by applying constant excitation to a single phase or a pair of phases. If an external torque is applied to the shaft, an angular displacement will occur. As the torque is increased the angle will increase, as will the resisting force until a maximum is reached. This relationship is known as the  $\tau/\theta$  characteristic of the motor. Once this is exceeded, the direction of the force will be reversed, and the rotor will be drawn to the next equilibrium point. The holding torque is defined as the maximum static torque that can be applied to the motor without causing continuous rotation.

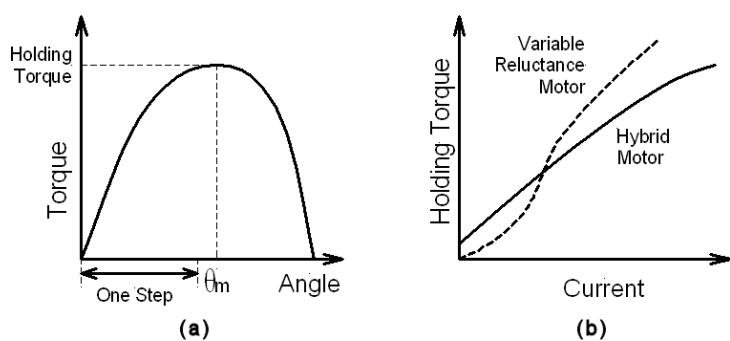
The holding torque increases with current, and the relationship is known as the  $\tau/I$  characteristic. This is typically very nonlinear for a variable reluctance motor but is reasonably linear for a hybrid, as shown in Figure 3-33.



**FIGURE 3-31** ■  
Photograph of a  
dismantled stepper  
motor.

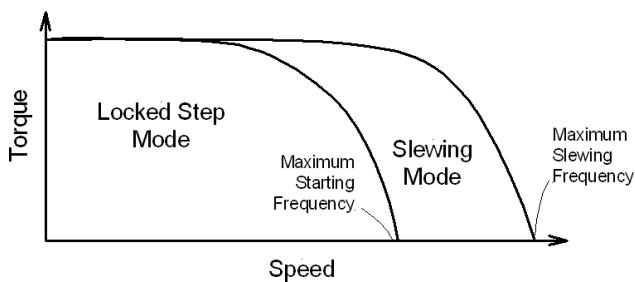


**FIGURE 3-32** ■  
Typical stepper  
motor construction  
(a) Rotor and stator  
configuration. (b)  
Wiring diagram.



**FIGURE 3-33** ■  
Static characteristics  
of a stepper motor.  
(a) Torque as a  
function of rotor  
angle. (b) Holding  
torque as a function  
of current.

**FIGURE 3-34** ■ Stepper motor torque–speed curves. [Adapted from (Alciatore et al., 2003)].



**Dynamic characteristics:** The torque–speed characteristics of stepper motors are nonlinear compared with DC motors. In addition, there are two different rotation modes, as illustrated in Figure 3-34. The normal operational mode is the locked step mode within which the motor can be stopped, started, and even reversed in the space of a single step. However, it is possible to continue to accelerate the motor by decreasing the step interval until the speed is too fast to allow for stopping and starting between steps. In this mode the motor can continue to travel in only one direction and has to be decelerated into the lock-step region before it can be stopped and reversed.

At high speeds, the performance of stepper motors can often be analyzed using the normal phasor analysis model that is used for synchronous AC motors.

**Step response:** At low speeds, each step is discernable, and the overall behavior is that of a series of step input transients. To determine their shape, the motor must be modeled as a set of nonlinear differential equations. For a variable reluctance motor, these equations are

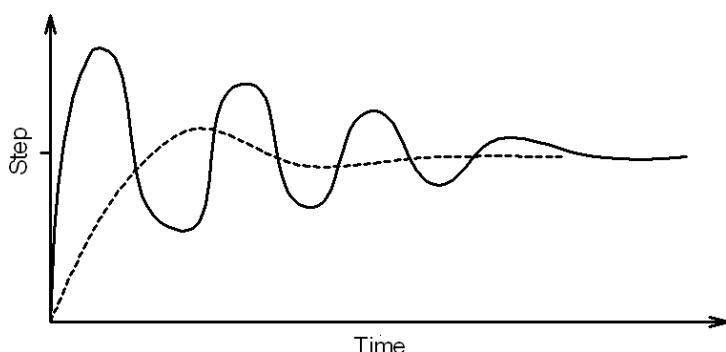
$$\frac{di}{dt} = -\frac{1}{L} \left[ R + \frac{dL(\theta)}{d\theta} \omega_r \right] i + \frac{1}{L} v \quad (3.41)$$

$$\frac{d\omega_r}{dt} = \frac{1}{J} \tau - \frac{1}{J} \tau_{load} \quad (3.42)$$

$$\frac{d\theta}{dt} = \omega_r \quad (3.43)$$

The motor response for a given step change on the input voltage can be obtained by solving these equations in conjunction with a number of initial conditions for the current, angle, and angular rate. For a typical motor in isolation, the response is strongly underdamped, as illustrated in Figure 3-35, and for that reason it is often expedient to introduce some form of mechanical damping in the load.

**FIGURE 3-35** ■ Step response without and with viscous inertial damping.





**FIGURE 3-36 ■**  
Sanyo Denki  
103-4902-0650  
stepper motor.

**Static position error:** When a stepper motor is employed to drive a load, the equilibrium position will always be away from the ideal step position because the torque produced must balance the load torque. The difference between the actual equilibrium position and the ideal is defined as the static position error. It can be approximated by

$$\theta_{err} = \frac{\sin^{-1}(-\tau_l/\tau_{pk})}{N_r} \quad (3.44)$$

where  $\tau_l$  (Nm) is the load torque,  $\tau_{pk}$  (Nm) is the holding torque, and  $N_r$  is the number of steps per revolution.

**Magnets:** Major torque improvements can be made with stronger magnets, and upgrading ceramic magnets to rare-earth materials such as neodymium increases motor performance significantly. The strength of rare-earth magnets can be adjusted by varying the percentage of fillers used, typically boron, cobalt, or iron. Greater availability of rare-earth materials has recently driven rare-earth magnet prices to competitive levels. However, rare-earth magnets are more susceptible to damage from heat than their ceramic counterparts.

### 3.2.4.5 Stepper Motor Types

The advent of digital electronics and strong rare-earth magnets has been the primary driving force behind the proliferation of stepper motor types in a wide variety of applications.

The specifications of the Sanyo Denki 103-4902-0650 stepper motor shown in Figure 3-36 are as follows:

- Holding torque: 0.032 Nm
- Step angle: 0.9°
- Voltage: 6.6 V
- Current/phase: 0.4 amp
- Ohm/phase: 16.5 Ω
- Inductance/phase: 3.5 mH
- Flange size: 39 mm<sup>2</sup> (1.54 in<sup>2</sup>)
- Connection: 6 Lead
- Rotor inertia: 0.009 kg.m<sup>2</sup> ( $\times 10^4$ )
- Mass: 0.135 kg

**FIGURE 3-37 ■**

Sanyo Denki  
103H8222-0941  
stepper motor.



Specifications of the larger Sanyo Denki 103H8222-0941 stepper motor shown in Figure 3-37, are as follows:

- Holding torque: 4.13 Nm
- Step angle: 1.8°
- Voltage: 3.88 V
- Current/phase: 4 amp
- Ohm/phase: 0.97 Ω
- Inductance/phase: 3.60 mH
- Flange size: 86 mm<sup>2</sup> (3.39 in<sup>2</sup>)
- Connection: 6 Lead
- Rotor inertia: 2.9 kg.m<sup>2</sup> ( $\times 10^4$ )
- Mass: 2.5 Kg

### 3.2.4.6 Biomechatronic Applications

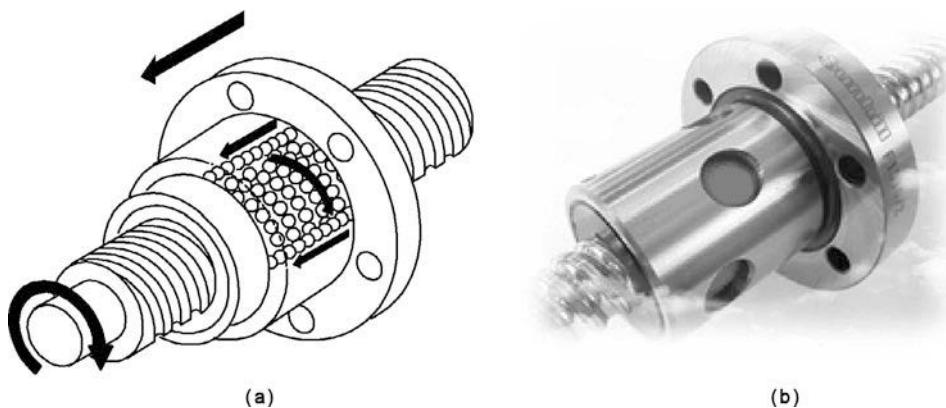
Peristaltic pumps provide medical facilities with accurate and repeatable pumping performance by the using microprocessor-controlled stepping motors. They are used to pump blood for heart-lung machines and kidney dialysis units. By varying the diameter of the tubing or the step rate of the motor, the pump volume and rate can be easily changed and controlled. Various fluids can be processed using the same pump by simply changing the tubes and reprogramming the pump parameters. Nothing but the tube touches the fluid, thus eliminating the risk of cross-contamination between the pump and the fluid.

### 3.2.5 Linear Actuators

There are many different methods of producing controlled linear motion, or linear positioning. Conventional means include air cylinders, ball screws, cables and pulleys, or even stacks of piezoelectric elements, but probably the most common is a stepping motor in which the shaft is replaced by a screw. The center part of the rotor is tapped to convert it to a “nut” that is threaded onto a screw. By preventing the screw turning (antirotation), it can move axially to give the linear motion required. The increment that is the linear progression corresponding to each step of the motor is given by the screw pitch divided by the number of motor steps. This is selected to suit the actuator characteristics required. The longer the pitch, the quicker the advance but the lower the force of the device. Conversely, a fine pitch limits the advance rate but increases the force available.



**FIGURE 3-38 ■**  
Noncaptive stepper  
motor driven linear  
actuator.



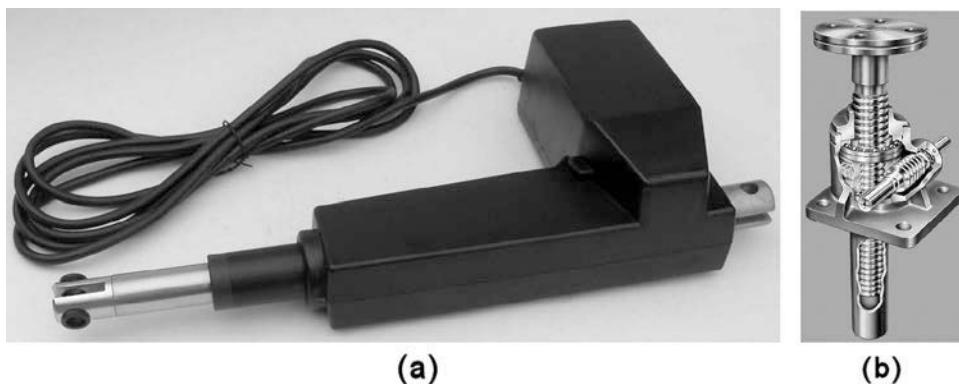
**FIGURE 3-39 ■**  
Ball-screw drive.  
(a) Schematic  
cutaway diagram.  
(b) Photograph.

Typical actuators of this type include the Portescap 26DAM series, which provides linear force up to 33 N, linear step resolution of 0.025 mm, 0.05 mm and 0.1 mm, and 3.4 W of power for fast response times and precise positioning capability. These linear actuators are available in captive and noncaptive versions (see Figure 3-38), with unipolar or bipolar coil construction and industry-standard frame sizes.

Torque and efficiency calculations for this actuator type can be found in the section on translation screws toward the end of this chapter.

For higher-performance applications that encompass longer life and smoother running, a ball-screw race usually replaces the simple nut, as shown in Figure 3-39. In this case, the threaded rod rotates in place producing translation in the ball-screw race.

Another common drive mechanism uses a worm gear to drive the linear actuator, as shown in Figure 3-40.



**FIGURE 3-40 ■**  
Worm gear drive  
actuator.  
(a) Photograph.  
(b) Cutaway  
diagram.

### 3.2.5.1 Piezoelectric Actuators

The piezoelectric effect is the generation of an electric charge by a crystalline material if it is stressed. It is caused by the stress-induced deformation of individual molecules within the lattice that exhibit changes in their charge distribution (also known as polarization; not to be confused with the polarization of light).

Natural materials such as quartz ( $\text{SiO}_2$ ) as well as some man-made ceramics and plastics exhibit this property. The latter are formed by poling (artificial polarization), a process of aligning the dipole fields of the molecules in the lattice in the presence of a strong electric field and then “freezing” them in place.

The piezoelectric effect is reciprocal insofar as applying a voltage will cause a strain in the crystal lattice and applying a strain to the lattice will generate a voltage across the faces. In each case, conductive electrodes on opposite faces are used to either apply the voltage or measure it.

The polarization vector,  $P$ , of a piezoelectric material is the sum of the three orthogonal polarization vectors (Fraden, 1996).

$$P = P_{xx} + P_{yy} + P_{zz} \quad (3.45)$$

Each of the individual vectors is related to the axial stress along the orthogonal axes

$$\begin{aligned} P_{xx} &= d_{11}\sigma_{xx} + d_{12}\sigma_{yy} + d_{13}\sigma_{zz} \\ P_{yy} &= d_{21}\sigma_{xx} + d_{22}\sigma_{yy} + d_{23}\sigma_{zz} \\ P_{zz} &= d_{31}\sigma_{xx} + d_{32}\sigma_{yy} + d_{33}\sigma_{zz} \end{aligned} \quad (3.46)$$

where the constants  $d_{mn}$  (C/N) are the piezoelectric coefficients along the orthogonal axes of the material.

The force,  $F_z$  (N), produced by a piezoelectric material is proportional to the charge,  $Q$  (C). In the z-direction this can be written as

$$F_z = \frac{Q_z}{d_{33}} \quad (3.47)$$

Because the crystal with its electrodes becomes a capacitor with capacitance  $C$  (F), the charge can be determined in terms of the applied voltage,  $V$ , as follows:

$$Q_z = CV \quad (3.48)$$

In turn, the capacitance is proportional to the electrode area,  $A$  ( $\text{m}^2$ ), and inversely proportional to the thickness of the material,  $t$  (m),

$$C = \varepsilon \varepsilon_o \frac{A}{t} \quad (3.49)$$

where  $\varepsilon$  is the dielectric constant of the material, and  $\varepsilon_o = 8.8542 \times 10^{-12} \text{ C}^2/\text{Nm}^2$  is the permittivity of free space.

Substituting (3.48) and (3.49) into (3.47) gives the force in the z-direction as a function of the applied voltage.

$$F_z = \frac{\varepsilon \varepsilon_o A}{d_{33} t} V \quad (3.50)$$

The displacement can be approximated from the generated force and the elastic modulus (Young's modulus),  $E$  ( $\text{N}/\text{m}^2$ ) of the material. The relationship between the stress ( $F/A$ )

**TABLE 3-5** ■ Properties of Some Piezoelectric Materials

	Piezoelectric Const $d$ (C/N)	Relative Dielectric Constant	Young's Modulus $E$ (N/m <sup>2</sup> )
BaTiO <sub>3</sub>	$d_{33} = 78 \times 10^{-12}$	1250 @ 20 °C	$110 \times 10^9$
PZT	$d_{31} = 180 \times 10^{-12}$ $d_{33} = 110 \times 10^{-12}$	1700	$83 \times 10^9$
PVDF	$d_{31} = 23 \times 10^{-12}$ $d_{33} = -33 \times 10^{-12}$	5–10	$2 \times 10^9$
Quartz	$2.3 \times 10^{-12}$	3.8	$77 \times 10^9$

Notes: PZT, lead zirconate titanate. PVDF, polyvinylidene fluoride.

and the strain ( $\Delta t/t$ ) is

$$\frac{F_z}{A} = E \frac{\Delta t}{t} \quad (3.51)$$

Therefore

$$\frac{\Delta t}{t} = \frac{\varepsilon \varepsilon_0}{E d_{33} t} V \quad (3.52)$$

The magnitudes of the piezoelectric constant and Young's modulus for a number of common materials are given in Table 3-5.

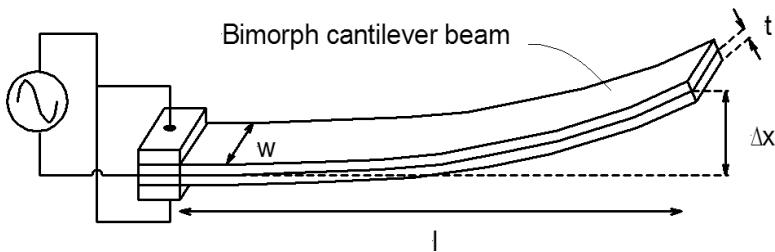
Piezoelectric actuators shown in Figure 3-41 are solid-state ceramic actuators that use this effect to convert electrical energy directly into linear displacement with an almost unlimited resolution. They rely on piezoelectric elements stacked in series with contacts sandwiched between each element. When a high voltage is applied to the elements, their individual lengths increase in the direction of the applied electric field by a small amount, resulting in a reasonably large total displacement (up to about 0.2% of the length of the actuator). Typical maximum displacements are between 5 and 200 with high force (up to 80 kN), high operational frequency, and a long lifetime. Unfortunately, these transducers do suffer from creep as well as hysteresis and are therefore mostly operated in conjunction with an accurate displacement transducer in a closed-loop configuration.

Other actuators operate on the bimorph principle. In this case a slab of piezoelectric material is bonded to a metal backing. When a voltage is applied to the material, it will expand forcing the composite structure to bend in exactly the same way as a bimetallic



**FIGURE 3-41** ■  
Piezoelectric  
actuators.  
(Physikinstrumente  
2008).

**FIGURE 3-42 ■**  
Configuration of a bimorph piezoelectric actuator.



strip bends. Alternatively, slabs of material can be bonded to either side of a thin metal film to form a sandwich with the potential applied to one slab being of the opposite polarity to the potential applied to the other. These actuators can provide larger displacements than stacks, but the maximum force is reduced. The principle is used by ultrasound transducers (Brooker, 2008).

The amount of deflection,  $\Delta x$  (m), at the tip of the cantilevered bimorph beam actuator shown in Figure 3-42 is

$$\Delta x = \frac{3Vd_{31}l^2}{4t^2} \quad (3.53)$$

where  $V$  (volts) is the applied voltage across the thickness of the material,  $d_{31}$  is the piezoelectric coefficient in the stretch (length) direction (see Chapter 2),  $l$  (m) is the length of the beam, and  $t$  (m) is its thickness.

Similarly, the force,  $F$  (N), at the tip of the beam is

$$F = \frac{3VEwd_{31}t}{2l} \quad (3.54)$$

where  $E$  (N/m<sup>2</sup>) is Young's modulus, and  $w$  (m) is the width of the beam.

### WORKED EXAMPLE

Consider a typical low-voltage actuator 50 mm high and 10 mm in diameter comprising 500 lead zirconate titanate (PZT) elements, each of which is 100  $\mu\text{m}$  high (including electrode thickness). A DC voltage of 20 V is applied to each element of the stack.

The force generated by each element can be determined from equation (3.50) and is proportional to the area. Therefore, calculate the area  $A$  (m<sup>2</sup>)

$$A = \frac{\pi d^2}{4} = \frac{\pi \times (10 \times 10^{-3})^2}{4} = 78.6 \times 10^{-6} \text{ m}^2$$

Now calculate the force per unit volt

$$\begin{aligned} F_z &= \frac{\varepsilon \varepsilon_0 A}{d_{33} t} \\ &= \frac{8.85 \times 10^{-12} \times 1700 \times 78.6 \times 10^{-6}}{110 \times 10^{-12} \times 100 \times 10^{-6}} \\ &= 107.5 \text{ N/V} \end{aligned}$$

So, for an applied voltage of 20 V the force is 2.15 kN per element, and because the elements are in series the total force is equal to the force of one element.

The strain can be determined from the force, the cross sectional area, and Young's modulus for the material

$$\begin{aligned}\frac{\Delta t}{t} &= \frac{F}{AE} \\ &= \frac{2.15 \times 10^3}{78.6 \times 10^{-6} \times 83 \times 10^9} \\ &= 329.6 \times 10^{-6}\end{aligned}$$

For a stack length of 50 mm, the change in length is  $50 \times 10^{-3} \times 329.6 \times 10^{-6} = 16.5 \mu\text{m}$ .

Note that the maximum applied voltage is governed by the electrical breakdown strength of the material, which is between 1 kV and 2 kV per mm. In this example the field strength is 20 V over 100  $\mu\text{m}$ , which is only 200 V/mm.

---

### WORKED EXAMPLE

Consider a bimorph cantilever 20 mm long and consisting of two strips of 9  $\mu\text{m}$  polyvinylidene fluoride (PVDF) film. Calculate the displacement at the tip of the beam if 100 V is applied across the film.

$$\begin{aligned}\Delta x &= \frac{3Vd_{31}l^2}{4t^2} \\ &= \frac{3 \times 100 \times 23 \times 10^{-12} \times (20 \times 10^{-3})^2}{4 \times (9 \times 10^{-6})^2} \\ &= 8.5 \text{ mm}\end{aligned}$$

What is the force at the tip of the beam if the width of the beam is 10 mm?

$$\begin{aligned}F &= \frac{3VEd_{31}t}{2l} \\ &= \frac{3 \times 100 \times 2 \times 10^9 \times 10 \times 10^{-3} \times 23 \times 10^{-12} \times 9 \times 10^{-6}}{2 \times 20 \times 10^{-3}} \\ &= 31 \mu\text{N}\end{aligned}$$


---

#### 3.2.5.2 Bellows Actuators

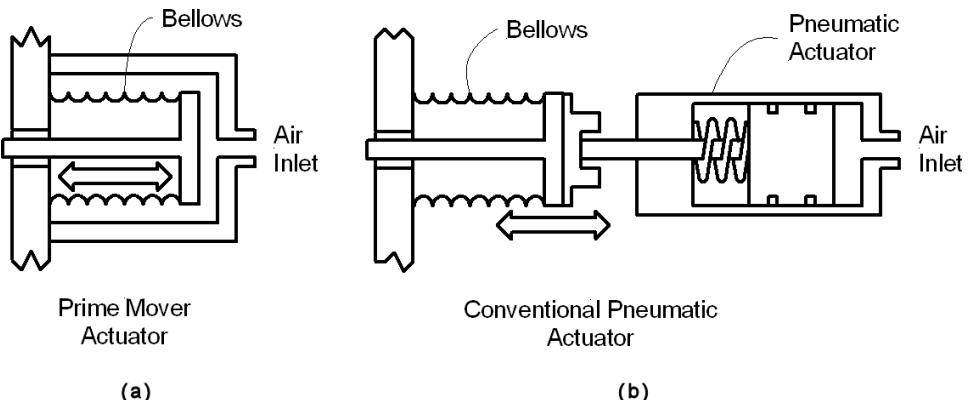
The metal bellows is essentially a hermetically sealed spring. When a pressure differential is applied across it, the bellows will compress and thus drive the actuator. When the pressure is removed, the bellows spring force will retract the actuator. In contrast to the conventional system that consists of the bellows and a pneumatic actuator, this construction provides a compact prime mover actuator, as shown in Figure 3-43.

The relationship among the applied pressure, the displacement, and the force for pneumatic and hydraulic actuators is described in more detail later in this chapter.

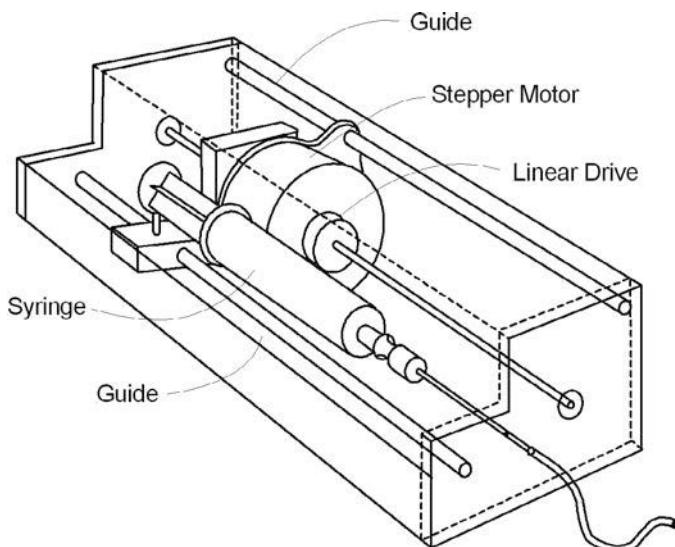
#### 3.2.5.3 Biomechatronic Applications

The stepper-motor-based syringe pump shown in Figure 3-44 automatically dispenses a precisely controlled amount of medication. Unlike an injection that is administered by a doctor in seconds, the microprocessor-controlled linear actuator can dispense medication

**FIGURE 3-43** ■ Pneumatic actuators. (a) Prime mover. (b) Conventional.



**FIGURE 3-44** ■ Precision syringe pump using a linear drive stepper motor.

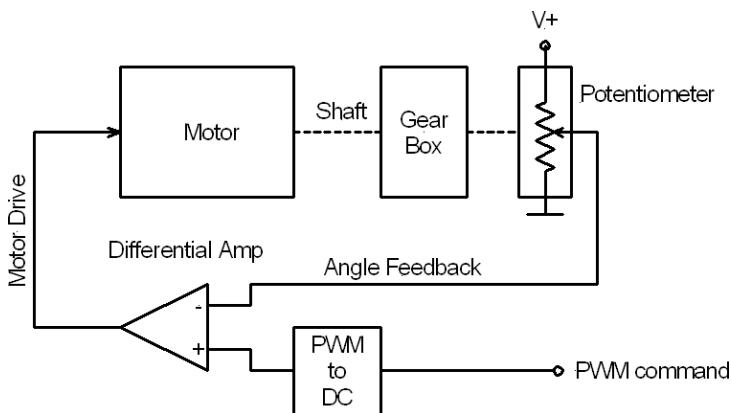


over long periods of time at precise rates and volumes. Additionally, unlike an intravenous feeder that relies on gravity to dispense liquids, the syringe pump controls flow using pressure. The two major factors—flow rate and dispensed volume—are controlled by a stepper motor.

Electronic pipettes are also linear-stroke devices that use linear actuators to control the amount of fluid being dispensed. The shaft of the actuator is coupled to a pipette fitted with a piston. A microprocessor controls the step rate of the actuator, which moves the shaft and piston in precise linear increments. Most modern electronic pipettes are portable handheld devices with built-in batteries and microcomputers.

### 3.2.6 Servo Motors

A servo mechanism, colloquially referred to as a *servo*, is a device using error-sensing feedback to control the performance of a mechanism. Servo mechanisms often use a servo motor, but this is not necessary as long as feedback is used to control some characteristic of the system. For example, an insulin pump driven by measured glycogen levels in the blood can be considered a servo mechanism.



**FIGURE 3-45** ■ RC type servo feedback mechanism.

This section is concerned only with servo systems that use a servo motor to control position. These are of particular importance to active prostheses, which are mostly powered by electric motors.

### 3.2.6.1 Radio Control Servos

The most common servo mechanisms are those originally developed for radio-controlled devices such as aircraft or cars but that are now used for a wide variety of precision position control applications. They consist of a DC motor mechanically linked to a potentiometer through reduction gears. Pulse-width modulated signals transmitted to the servo are translated into position commands by the internal electronics. The DC motor then drives the output shaft until feedback from the potentiometer reaches the commanded value. A schematic diagram illustrating the process is shown in Figure 3-45.

Due to their low cost, good reliability, and the ease with which they can be controlled by microprocessors, RC servos are used in many mechatronic applications. Power and control are provided through three wires usually colored as follows:

- Red: DC power (4.8–6 V)
- Black: Ground
- Brown: PWM position control

The servo is provided with a pulse every 20 ms (or less) with a width that varies from 1.25 to 1.75 ms. These correspond to servo angles of 0° and 180°, respectively. The neutral position that corresponds to an angle of 90° is provided by a pulse width of 1.5 ms. Physical limits to the angle of travel vary depending on the manufacturer, but they are generally larger than the electrical limits.

**Pulse-width-to-voltage converter:** The control pulse is fed to a pulse-width-to-voltage converter. This circuit charges a capacitor at a constant rate while the pulse is high. When the pulse goes low the charge on the capacitor is output via a suitable high impedance buffer amplifier. This produces a voltage related to the length of the applied pulse.

The circuit is tuned to produce a useful voltage over a 1 to 2 ms period. The output voltage is buffered to ensure that it does not decay significantly between control pulses, so the length of time between pulses is not critical. However, it should not exceed 20 ms for reliable operation.

**TABLE 3-6** ■ Specifications of a Number of Hitec Servos

Model	Features	Weight (g)	Speed 0–60° Step (s)	Torque (kg-cm)
HS50		6.1	0.09	0.6
HS-65MG	Metal gear, analog	11.2	0.11	2.2
HS-82MG	Metal gear, analog	19.0	0.10	3.4
HS-5645MG	Metal gear, analog	60.0	0.18	12.1
HS-5955TG	Titanium gear, digital	62.0	0.15	24

**Position sensor:** The angular position of the servo motor output shaft is read by a sensor. This is normally a potentiometer that produces a voltage related to the absolute angle of the output shaft. The position sensor feeds one input of the error amplifier that compares the current position with the commanded position from the pulse-width-to-voltage converter.

**Error amplifier:** The error amplifier is an operational amplifier (op amp) with negative feedback. It will always try to minimize the difference between the inverting (negative) and noninverting (positive) inputs by driving its output in the correct direction, as discussed in the section on op amps in Chapter 5.

The output of the error amplifier is either a negative or positive voltage representing the difference between its inputs. The greater the difference between the two inputs, the greater the output voltage.

The error amplifier output has sufficient drive capability to power the motor. If the output is positive the motor will turn in one direction, and if it is negative it will turn the other direction. This allows the error amplifier to reduce the difference between its inputs (thus closing the negative feedback loop) and cause the servo to turn to the commanded position. It is essential that some form of filtering be applied to the loop; this can be electrical or intrinsic to the mechanical configuration. Without it, the system response could be underdamped.

Different torques are provided by different RC servos as shown in Table 3-6. Depending on the application, these can range from about 0.6 kg-cm (58.8 mNm) to 24 kg-cm (2.35 Nm) for RC hobby types. They are determined by the motor size and its torque as well as the gear ratio of the gearbox. Examples of typical RC servos are shown in Figure 3-46.

As with most control systems, there is a trade-off in performance among a number of parameters. In the case of these RC servos, the trade-off is between the step response time (which should be fast) and the output torque (which should be high). Servo response is usually given as the time taken to step from 0° to 60°.

**FIGURE 3-46** ■  
Photographs of  
assorted Hitec  
servos.





**FIGURE 3-47 ■**  
Futaba RS403PR  
digital coreless  
servo specifications.

Futaba makes a number of high-speed, high-torque servos for robotic applications shown in Figure 3-47 and Figure 3-48. These operate at 6 V and have the following specifications:

Speed: 0.13 sec/60° @ 6 V  
Torque: 13.8 kg-cm @ 6 V  
Weight: 63 g  
Length: 40 mm  
Width: 20 mm

Height: 38 mm  
Communication: PWM  
Operation range: 180°  
Pulse cycle: 14.25 ms  
Pulse width: 920–2120  $\mu$ s

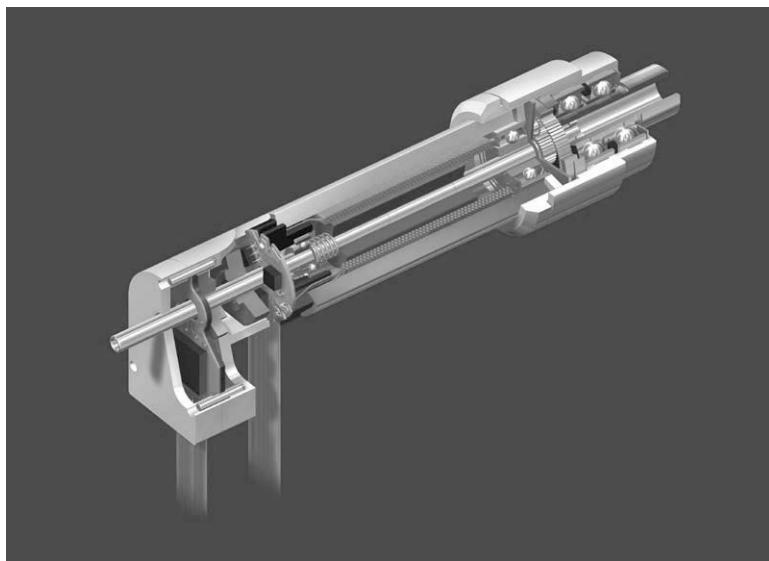
Speed: 0.20 sec/60° @ 6 V  
Torque: 8.9 kg-cm @ 6 V  
Weight: 47.5 g  
Length: 40 mm  
Width: 20 mm

Height: 39 mm  
Communication: PWM  
Operation range: 150°  
Pulse cycle: 14.25 ms  
Pulse width: 920–2120  $\mu$ s



**FIGURE 3-48 ■**  
Futaba RS404PD  
digital iron core  
servo.

**FIGURE 3-49** ■  
Maxon micro-drive  
servo motor.



### 3.2.6.2 Professional Servos

Hobby servos are obviously not really suitable for biomechatronic applications where good reliability and long life are essential. In these applications, high-quality motors such as those made by Maxon and Faulhaber should be used. In the past, a servo system would be assembled using a motor–gearbox combination with a separate encoder and power electronics, but now integrated systems such as the micro-drive system are starting to appear.

The micro-drive system, shown in Figure 3-49, consists of a 6 mm BLDC motor, magnetic encoder, and backlash-free harmonic drive gearhead. Supplying a constant 1.2 W, the motor can turn at up to 100,000 rpm and weighs only 2.8 g.

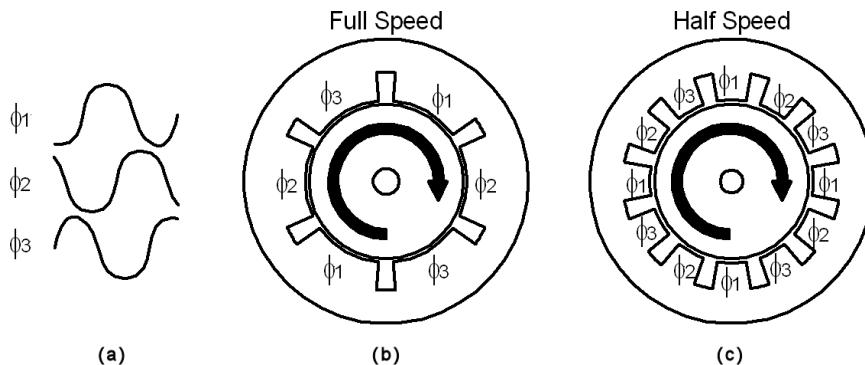
The backlash-free gearhead has a diameter of 8 mm and consists of a 1:160 reduction ratio harmonic drive and high-resolution 100-count incremental encoder. The positioning unit's shaft is hollow, and the drillhole diameter is 0.65 mm.

### 3.2.7 AC Motors

There are many different types of alternating current motors of which the most common are induction motors, hysteresis motors, and synchronous motors. This section outlines the operation of induction motors because they are probably encountered most often in the biomechatronics field.

#### 3.2.7.1 Induction Motors

In an induction motor, the stator's magnetic field induces an alternating current into the rotor squirrel-cage conductors, which constitute a shorted transformer secondary winding. This induced rotor current in turn creates a magnetic field. The rotating stator magnetic field interacts with this rotor field, which attempts to align with it. The result is rotation of the squirrel-cage rotor. If there were no load torque, no bearing, windage, or other losses, the rotor would rotate at the synchronous speed. However, the slip between the rotor and the synchronous speed stator field causes the magnetic flux to cut through the



**FIGURE 3-50** ■ Three-phase induction motor with one and two pole pairs per phase. (a) Waveform diagram. (b) half speed stator. (c) Full speed stator. [Adapted from (Kuphaldt 2008)].

rotor conductors, and this develops torque. Thus, a loaded motor will slip in proportion to the mechanical load. If the rotor were to run at synchronous speed, there would be no stator flux cutting the rotor conductors, no current induced in the rotor, and hence no torque.

**Motor speed:** The rotation rate of a stator rotating magnetic field is related to the number of pole pairs per stator phase. In the full-speed case shown in Figure 3-50, there are three pole pairs and three phases. The magnetic field rotates once per sine wave cycle. For 50 Hz power, it rotates at 50 rotations per second, or 3000 rpm; this is the synchronous speed of the motor. Though the rotor of an induction motor never achieves this speed, it is the upper limit of what is theoretically possible.

If the number of motor poles is doubled, the synchronous speed,  $N_s$  (rpm), is divided by two because the magnetic field rotates 180° in space for 360° of electrical sine wave. The relationship is

$$N_s = \frac{60f}{n_p} \quad (3.55)$$

where  $f$  (Hz) is the mains frequency, and  $n_p$  is the number of pole pairs per phase.

Therefore, in Figure 3-50 the rotation rate for the half-speed option is

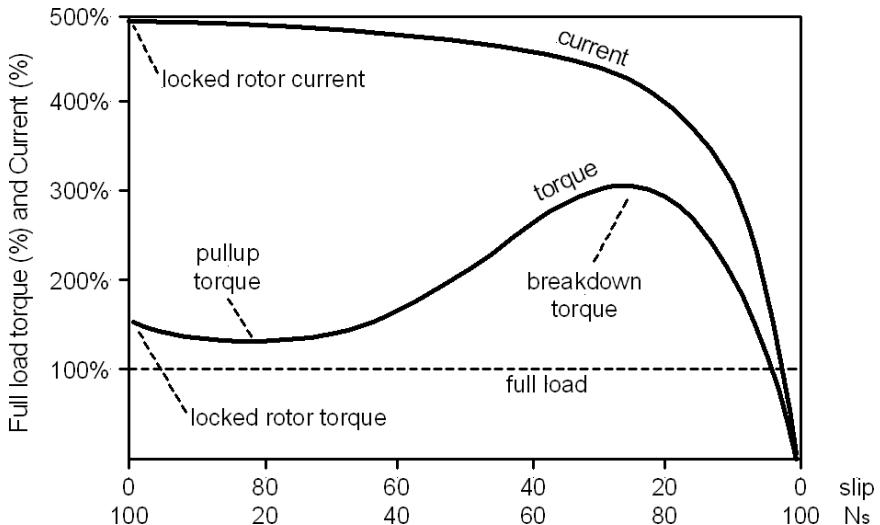
$$\begin{aligned} N_s &= \frac{60f}{n_p} \\ &= 60 \times 50/2 \\ &= 1,500 \text{ rpm} \end{aligned}$$

because there are two pole pairs per phase.

**Torque:** When power is first applied to the motor, the rotor is at rest, whereas the stator magnetic field rotates at the synchronous speed  $N_s$ . The stator field is cutting the rotor at the synchronous speed,  $N_s$ , and the current induced in the rotor shorted turns is a maximum, as is the frequency of the current. As the rotor speed increases, the rate at which stator flux cuts the rotor is the difference between synchronous speed,  $N_s$ , and actual rotor speed,  $N$ , or  $(N_s - N)$ . The ratio of actual flux cutting the rotor to synchronous speed is defined as slip,  $s$ ,

$$s = \frac{N_s - N}{N_s} \quad (3.56)$$

**FIGURE 3-51** ■  
Torque and current  
versus slip for an  
induction motor.  
[Adapted from  
(Kuphaldt 2008)]



The frequency of the current induced into the rotor conductors is only as high as the line frequency at motor start and decreases as the rotor approaches synchronous speed. This rotor frequency,  $f_r$  (Hz), is

$$f_r = sf \quad (3.57)$$

Slip at 100% torque is typically 5% or less in induction motors. Thus, for  $f = 50$  Hz line frequency, the frequency of the induced current in the rotor  $f_r = 0.05 \times 50 = 2.5$  Hz.

Figure 3-51 shows the torque, speed, and current relationships of a typical induction motor. It can be seen that the starting torque, known as locked rotor torque (LRT), is higher than 100% of the full-load torque (FLT), the safe continuous torque rating. The locked rotor torque is about 175% of FLT for the example shown in Figure 3-51.

Starting current, known as locked rotor current (LRC), is 500% of full-load current (FLC), the safe running current. The current is high because this is analogous to a shorted secondary on a transformer. As the rotor starts to rotate, the torque generally decreases slightly to a value known as the pull-up torque. This is the lowest value of torque ever encountered by the starting motor. As the rotor gains 80% of synchronous speed, torque increases from 175% up to 300% of the full-load torque. This breakdown torque is due to the larger than normal 20% slip. The current has decreased only slightly at this point but decreases rapidly beyond it. As the rotor accelerates to within a few percent of synchronous speed, both torque and current decrease sharply.

Slip will be only a few percent during normal operation. For a running motor, any portion of the torque curve below 100% rated torque is normal. The motor load determines the operating point on the torque curve. While the motor torque and current may exceed 100% for a few seconds during starting, continuous operation above 100% can damage the motor. Any motor torque load above the breakdown torque will stall the motor. The torque, slip, and current will approach zero for a *no mechanical torque* load condition (Kuphaldt, 2008).

**Efficiency:** Large three-phase motors are more efficient than smaller three-phase motors and almost all single-phase motors. Large induction motor efficiency can be as high as

95% at full load. Efficiency for a lightly loaded or unloaded induction motor is poor because most of the current is involved with maintaining magnetizing flux. As the torque load is increased, more current is consumed in generating torque while current associated with magnetizing remains fixed. Efficiency at 75% FLT can be slightly higher than that at 100% FLT.

### 3.3 HYDRAULIC ACTUATORS

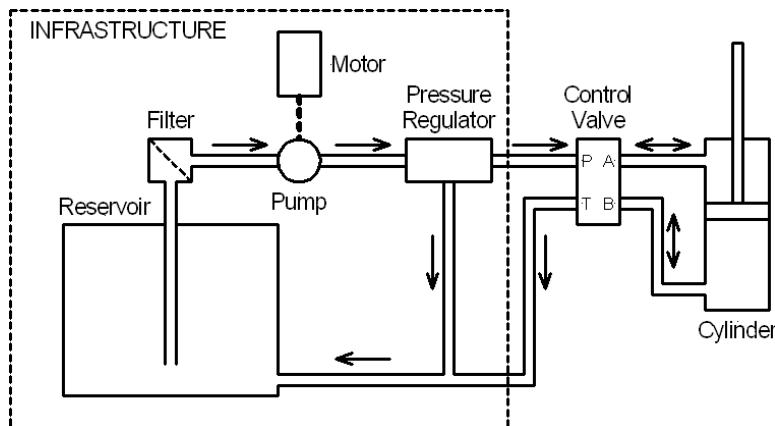
Hydraulic systems are designed to handle large loads by introducing high-pressure fluid into cylinders or sometimes rotary actuators. As shown in Figure 3-52, a typical hydraulic system comprises a pump to deliver high-pressure fluid, a pressure regulator to maintain a constant pressure in the rest of the system, valves to control flow rate, and a distribution system consisting of a number of pipes. This infrastructure is required to drive an actuator cylinder.

Hydraulic pumps are mostly powered by large AC electric motors. For large systems, pressures produced can range from 5 to 20 MPa. The hydraulic fluid must be incompressible, must have good lubrication qualities, and must be resistive to corrosion.

Most hydraulic pumps are defined as positive displacement because they provide a fixed volume of fluid with every cycle. These can include gear, vane, and piston pumps. Because of the fixed volume flow rate, it is necessary to include a pressure relief valve (called a pressure regulator) to prevent the pressure from exceeding design limits for the components. The simplest pressure regulators are spring-ball types, shown schematically in Figure 3-53. The threshold pressure is altered by adjusting the compression on the spring.

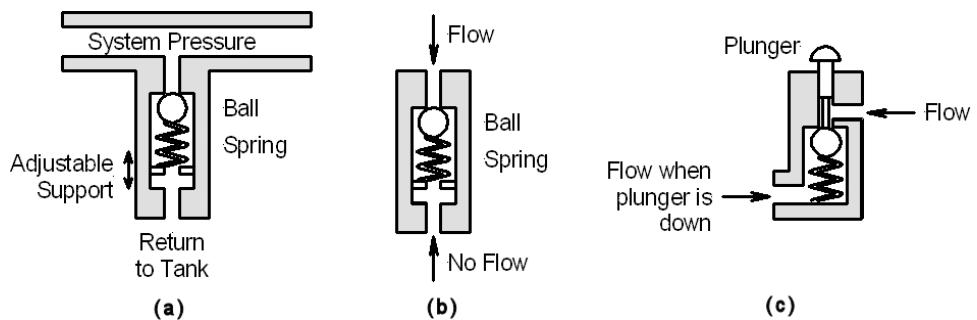
Hydraulic control valves can be either proportional, which allows any position between open and closed, or two positions, open–closed only. The latter are usually described in terms of the number of ports and positions (similar to electrical switches). A four-port, three-position valve is described as a 4/3 valve. Check valves allow flow in one direction only, while poppet valves are check valves that can be forced to allow reverse flow.

Spool valves are very common for controlling actuators as they can control multiple flows, and because the static hydraulic forces are balanced actuators need to overcome only the hydrodynamic forces associated with flow momentum. A linear force is applied

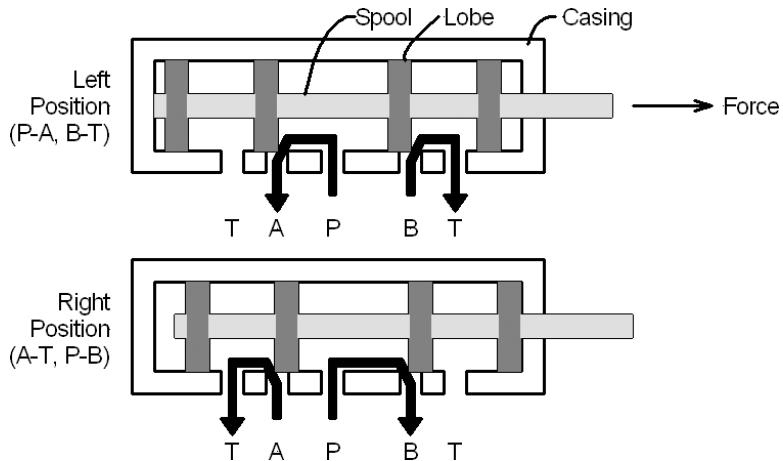


**FIGURE 3-52** ■ Components of a hydraulic system. [Adapted from (Alciatore et al., 2003)].

**FIGURE 3-53 ■**  
Hydraulic valves.  
(a) Pressure regulator. (b) Check valve. (c) Poppet valve. [Adapted from (Alciatore et al., 2003)].



**FIGURE 3-54 ■**  
Hydraulic spool valve.



by hand or using an electrical solenoid that switches inputs and outputs, as illustrated in Figure 3-54.

Spool valves can be made proportional by moving the spool a distance proportional to some input command. Linear electrical actuators are usually used to control these valves. If the hydrodynamic pressures are very high, then direct electrical actuation becomes impractical, and a smaller pilot hydraulic valve is used to control the larger valve.

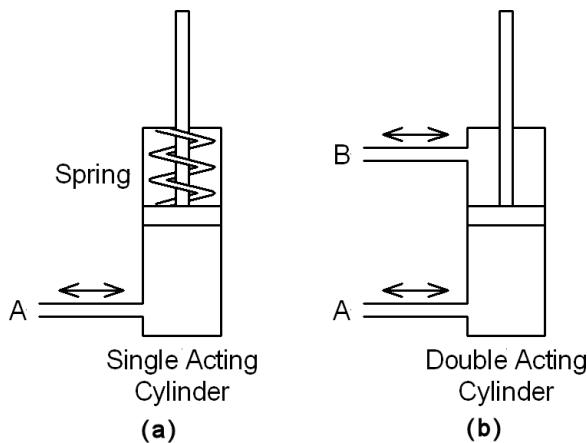
The most common hydraulic actuator is a simple cylinder with a piston driven by the pressurized fluid. These can be single acting, in which the cylinder is held in one position by the fluid pressure acting against a spring, or double acting, in which hydraulic pressure is used to drive the piston in both directions, as illustrated in Figure 3-55.

The force available from a single-acting cylinder is the product of the fluid pressure,  $P_A$  ( $\text{Pa} = \text{N/m}^2$ ), and the cross sectional area of the piston,  $A_A$  ( $\text{m}^2$ ), minus the correcting force applied by the spring. The spring force is equal to the product of the spring constant,  $k$  ( $\text{N/m}$ ), and the displacement from the neutral position,  $d$  ( $\text{m}$ )

$$F = P_A A_A - kd \quad (3.58)$$

In the double-acting cylinder, the force is determined by the difference in the pressure between the  $A$  and  $B$  inlets and the area of the pistons. It can be seen that the piston area on the output side is reduced by the shaft area.

$$F = P_A A_A - P_B A_B \quad (3.59)$$



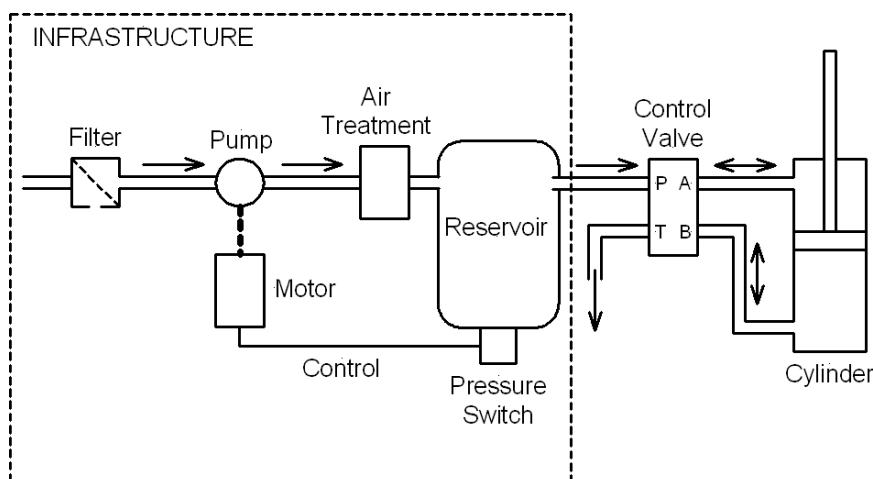
**FIGURE 3-55** ■  
Hydraulic actuators.  
(a) Single acting.  
(b) Double acting.

Hydraulic systems generate large forces from very compact actuators. They can also provide precise control at low speeds. However, their main drawbacks from a biomechanical perspective include the large amount of infrastructure required, noisy operation, and particularly the potential for fluid leaks. Hydraulic fluid is not very compatible with biological organisms, and, because of the extremely high-working pressure, fluid jets from pinhole-sized breaks are easily injected through the skin into the tissue below.

## 3.4 | PNEUMATIC ACTUATORS

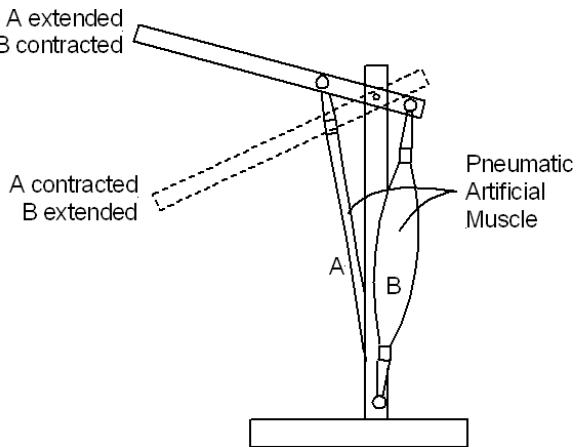
Pneumatic actuators are similar in principle to hydraulic systems, but they use compressed air as a working fluid. The components of a pneumatic system are similar to those of hydraulic systems insofar as they contain a compressor and a reservoir and this is fed to actuators through control valves, as shown in Figure 3-56. Working pressures are limited to between 450 kPa and 1 MPa, and because these are much lower than those used by hydraulic systems the available forces are much lower.

Filtered air is compressed by a motor-driven compressor, after which it is cooled and dried before entering the reservoir. Unlike hydraulic systems, the reservoir is maintained



**FIGURE 3-56** ■  
Components of a  
pneumatic system.

**FIGURE 3-57 ■**  
Pneumatic artificial muscles.



at the working pressure and is regulated by a pressure switch that controls the compressor motor.

Control valves and actuators operate in much the same way as hydraulic systems, but instead of returning the fluid to a tank it is vented into the air. Working with air makes them cleaner than hydraulics, but some form of lubrication is required to minimize wear on working surfaces.

Pneumatic actuators are ideal for providing low-force linear motion between two well-defined end points. However, because air is compressible, pneumatic systems are not good at providing accurate motion between the end points—particularly if the load is varying.

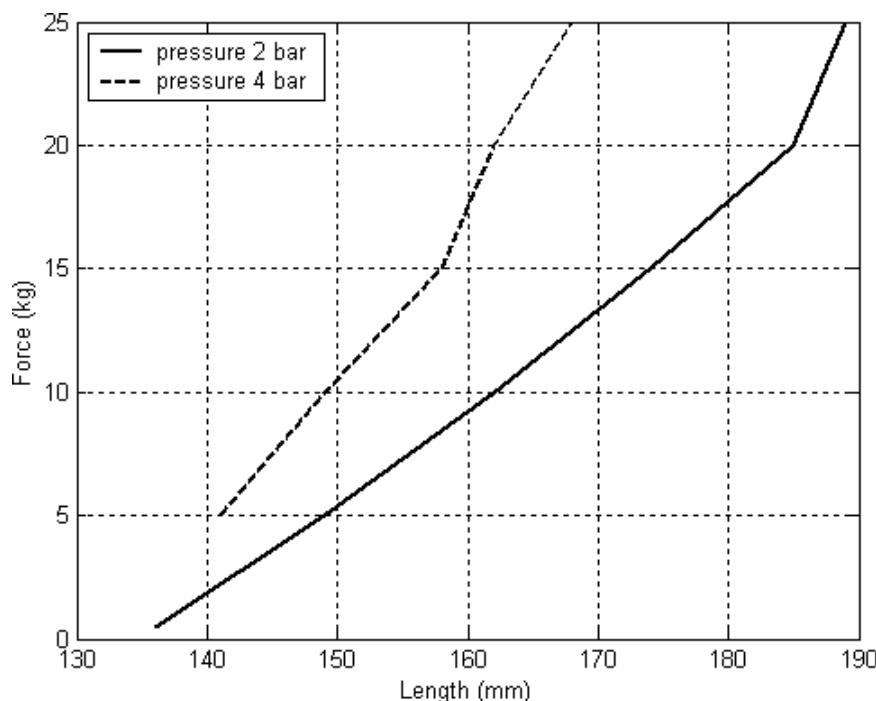
### 3.4.1 Pneumatic Muscles

Pneumatic muscles, also known as pneumatic artificial muscles (PAMs), are contractile or extensional devices that use pressurized air and rubber bladders surrounded by a scissor weave mesh instead of a piston–cylinder construction. When the bladder is pressurized, its diameter increases, which forces the mesh to contract longitudinally with significant force. Because they apply force only in contraction, these actuators are generally grouped into pairs of agonists and antagonists similar to human muscles, as shown in Figure 3-57.

PAMs were first developed under the name of McKibben artificial muscles in the 1950s for use in artificial limbs. These were commercialized in the 1980s as Rubbertuators (Wikipedia, 2008).

PAMs are lightweight because their main element is a thin membrane. This allows them to be directly connected to the structure they power, which is an advantage when considering the replacement of a defective muscle or as part of a prosthetic limb. Another advantage of PAMs is their inherent compliant behavior: When a force is exerted on the PAM, it allows some movement without increasing the force in the actuation. This is an important feature when the PAM is used as an actuator in a prosthesis that interacts with a human or when delicate operations have to be carried out.

As a PAM contracts under constant pressure, the pulling force generated decreases. Therefore, the maximum possible force at a given pressure is obtained when the PAM is extended as far as possible. If the actuator is not taut when the air pressure is increased, it will not yield its full force. For a constant pressure, the relationship between the force and the length of the PAM is as shown in Figure 3-58.



**FIGURE 3-58** ■ Relationship between force and length for a 210 mm long, 20 mm diameter PAM at constant pressure.

Data for a number of different PAMS manufactured by the Shadow Robot Company are summarized in Table 3-7 in terms of the percentage change in their lengths for various force and pressure combinations.

**TABLE 3-7** ■ Relationship Among Load, Pressure, and Percentage Decrease in Length for Various PAMs

6 mm PAM		Load (kg)				
Pressure (bar)	0.5	1	2	3.2	4.6	
0	n/a	3%	2%	2%	1%	
2	12%	10%	7%	5%	3%	
4		20%	28%	17%	11%	
20 mm PAM		Load (kg)				
Pressure (bar)	0.5	5	10	15	20	25
0	n/a	8%	6%	2%	1%	0%
2	35%	29%	23%	17%	12%	10%
4		33%	29%	25%	23%	20%
30 mm PAM		Load (kg)				
Pressure (bar)	0.5	10	20	35	50	70
0	n/a	2%	1%	0%	0%	0%
2	35%	33%	30%	27%	24%	21%
4		35%	33%	31%	29%	27%

Source: Courtesy Shadow Robot Company, <http://www.shadowrobot.com/>, with permission.

**TABLE 3-8** ■ Specifications of Some Commercially Available PAMs

Product	Diameter (mm)	Extended Length (mm)	Pull at 3.5 bar (kg)	Maximum Pull (kg)
	6	150	3	7
	20	210	12	20
	30	290	35	70

Source: Courtesy Shadow Robot Company, <http://www.shadowrobot.com/>, with permission.

At a constant extension, the relationship between force and pressure is linear; therefore, some control of displacement against a spring-type load can be achieved by regulating the system pressure. The relationship between force and extension in PAMs mirrors what is seen in the length-tension relationship in biological muscle systems.

In summary, PAMs can be made with extremely high force-to-weight ratios, as seen in Table 3-8.

Finally, because the air used to inflate these actuators is compressible, a PAM that uses long tubes must have a control system that can deal with a delay between the movement control signal and the effective muscle action. A PAM actuator system needs the infrastructure shown in Figure 3-56.

According to Ku and Bradbeer (2008), the response time of PAMs increases with applied load. In tests conducted using a 20 mm PAM, it took 0.7 s, 0.9 s, and 1 s for 5, 10, and 12 kg of load, respectively, from the application of 4 bar of pressure to reach full contraction. These time constants are consistent with human muscle responses.

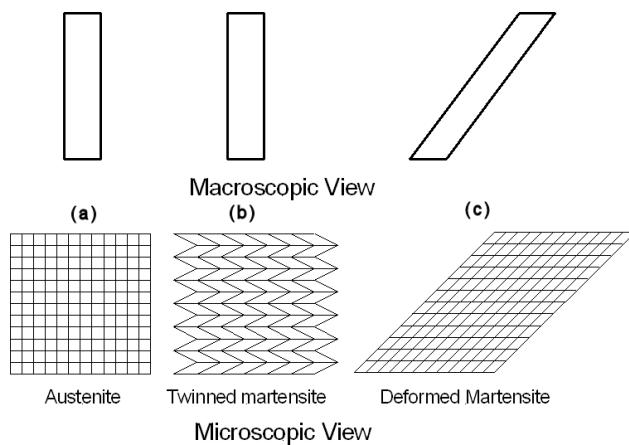
## 3.5 | SHAPE MEMORY ALLOY

Shape memory alloys (SMAs) are metals that exhibit two unique properties: (1) pseudo-elasticity; and (2) the shape–memory effect. Though the effect was discovered in 1938, serious research did not begin until the 1960s. The most effective and widely used SMAs include nickel titanium (NiTi), copper zinc aluminum (CuZnAl), and copper aluminum nickel (CuAlNi).

### 3.5.1 Principle of Operation

The properties described for SMAs are made possible through a solid-state change in phase. These are molecular rearrangements that occur in the shape memory alloy. A solid-state phase change is similar to a solid–liquid phase change in that a molecular rearrangement occurs, but the molecules remain closely packed so that the substance remains a solid. In most shape memory alloys, a temperature change of only about 10 °C is necessary to initiate this phase change. The two phases that occur in shape memory alloys are martensite and austenite.

Martensite is the relatively soft and easily deformed phase of shape memory alloys that exists at low temperatures. The molecular structure in this phase is twinned, as seen

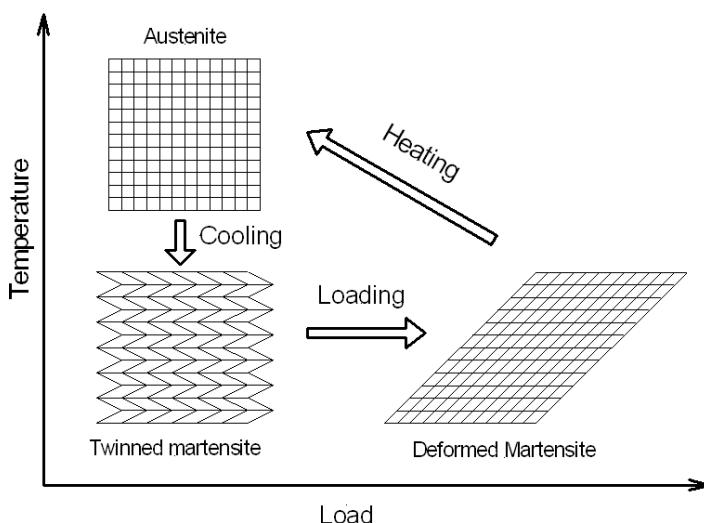


**FIGURE 3-59** ■ Macroscopic and microscopic views of the two phases of shape memory alloys showing (a) Austenitic Phase, (b) Martensitic Phase, (c) Deformed Martensitic Phase. [Adapted from SMA/MEMS Research Group, 2001.]

in Figure 3-59b. Upon deformation this phase takes on the deformed form (Figure 3-59c). Austenite, the stronger phase of shape memory alloys, occurs at a higher temperature. The shape of the austenite structure is cubic, as illustrated in Figure 3-59a. The undeformed martensite phase is the same size and shape as the cubic austenite phase on a macroscopic scale, so that no change in size or shape is visible in shape memory alloys until the martensite is deformed (SMA/MEMS Research Group, 2008).

The temperatures at which each of these phases begins and finishes forming are a function of the applied load and the composition of the wire.

As illustrated in Figure 3-60, the shape memory effect is observed when the temperature of a piece of shape memory alloy is cooled to below the temperature  $M_f$ . At this stage the alloy is completely composed of martensite, which can be easily deformed. After distorting the SMA, the original shape can be recovered simply by heating the wire above the temperature  $A_f$ . The heat transferred to the wire is the power driving the molecular rearrangement of the alloy, similar to heat melting ice into water but the alloy remains solid. The deformed martensite is now transformed to the cubic austenite phase, which is configured in the original shape of the wire.



**FIGURE 3-60** ■ Diagram of the microscopic shape memory effect [Adapted from SMA/MEMS Research Group, 2001]

**FIGURE 3-61 ■**  
**DM01 SMA actuator**  
 made by MIGA.  
 (Courtesy of MIGA  
 Motor Company.)



Actuators made from SMAs are manufactured by a number of companies. One example is the MIGA Motor Company, which makes a range of devices that can provide forces from 5 N from a unit weighing only 2.5 g up to 70 N for a unit weighing 30 g. A typical actuator is the DM01-15, shown in Figure 3-61, with a stroke of 12 mm and a maximum force of 20 N. It is only 80 mm long and weighs 20 g. Both push and pull versions are available.

The SMA wire segments within the actuator are electrically heated by passing a current through them. When the temperature reaches 75 °C, the wires begin to contract and are fully contracted by the time they reach 110 °C. If the temperature reaches 150 °C, permanent damage to the SMA wires results. When power is removed, the wire must return to 60 °C before the actuator returns to the neutral position.

The time taken to reach full contraction depends on the applied current, and in the DM01-15, which has a resistance of 3 Ω, this takes 1 s at 3 A, 0.5 s at 4 A, and 0.1 s at 9.3 A. The 1-second rule applies, and currents of less than 3 A can be applied continuously without overheating the SMA wires. However, currents higher than this must be interrupted to avoid damaging the actuator. This is generally achieved using some form of pulse width modulation.

Open-loop position control is not possible as the actuator wires creep with age. Therefore, if accurate positioning is required then so are a separate sensor and closed-loop control.

Some of the main advantages of shape memory alloys include the following:

- Biocompatibility
- Diverse fields of application
- Good mechanical properties (e.g., strong, corrosion resistant)
- Strong contraction force at start of stroke
- No inrush current or back EMF

There are still some problems with shape memory alloys that must be overcome before they can live up to their full potential. They are relatively expensive to manufacture compared with other materials such as steel and aluminium. Most SMAs have poor fatigue properties, with an equivalent steel component able to survive at least 100 times as many cycles as an SMA would under the same cyclical loading conditions.

### 3.5.2 Biomechatronic Applications

Because of their high force-to-weight ratio as well as the ease with which they can be configured for a specific space, SMAs have wide appeal as actuators in biomedical applications. These include actuators for finger joints, heart-assist devices, and vascular stents, some of which are discussed later in this book.

## 3.6 | MECHANICAL AMPLIFICATION

As mentioned earlier in this chapter, individual actuators are often not well matched to the load they have to drive in terms of displacement or, in some cases, speed or even torque. Mechanical methods of performing this matching process can be considered as mechanical amplification or mechanical impedance transformation. In an ideal situation, the actuator is matched specifically for the application, but in reality off-the-shelf devices such as voice-coil actuators and motors are often used to drive biomechanical systems. In this process the overall system efficiency is reduced somewhat, and less power is coupled to the output than could otherwise be obtained.

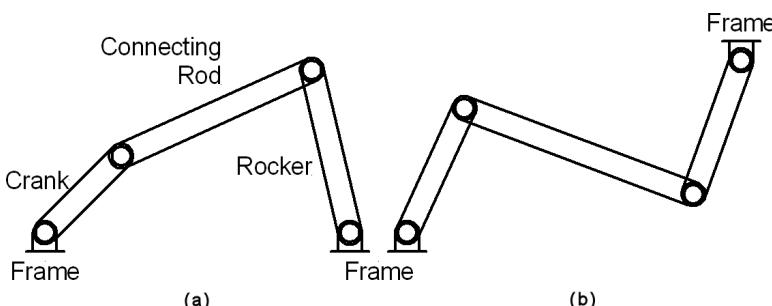
### 3.6.1 Linkages and Levers

A link is a rigid body having two or more elements paired together and connected to other bodies for the purpose of transmitting force or motion (Ham, Crane, and Rogers, 1958). The term *linkage* refers, therefore, to a mechanism made up from links such as cranks, rods, and levers that comprise elements moving in a linear fashion and in rotation.

The primary function of a link mechanism is to produce rotation, oscillation, or reciprocating motion from the rotation of a crank or the reverse. They may be used to convert the following:

- Continuous rotation into continuous rotation with a constant or variable angular velocity ratio
- Continuous rotation into oscillation or reciprocation (or the reverse)
- Oscillation into oscillation
- Reciprocation into reciprocation

Many mechanisms can be constructed from the simple constrained four-link structure consisting of four bar-shaped links and four turning pairs in an open or crossed configuration, as shown in Figure 3-62.

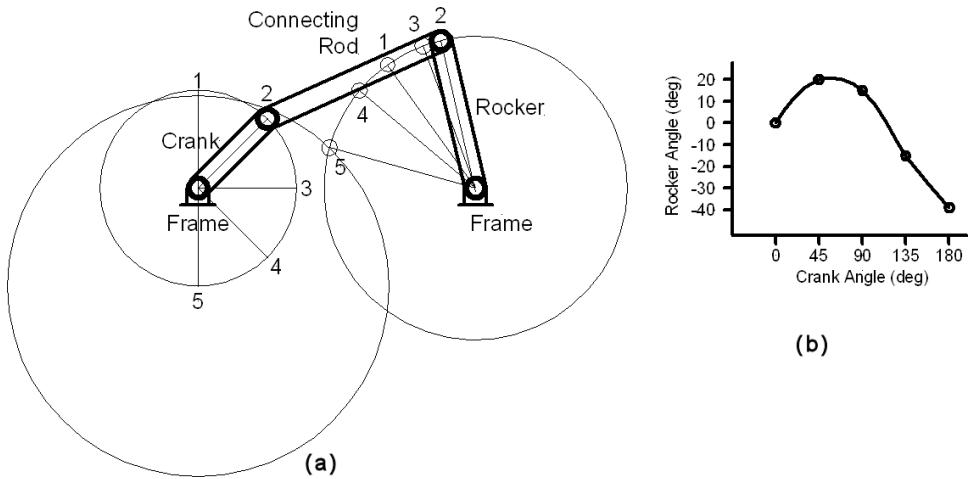


**FIGURE 3-62** ■  
Four-bar mechanisms.  
(a) Open. (b)  
Crossed.

**FIGURE 3-63**

Determining the angle of the rocker arm.

- (a) Construction.
- (b) Relationship between crank angle and rocker angle.



The fixed link may be an actual bar, but in most cases it represents the frame of the machine. Generally the other three links are referred to as the crank, the connecting rod (con rod), and rocker or follower. Depending on the relative lengths of the links, the crank may make a complete rotation or may oscillate, whereas the follower may oscillate or in some cases may also rotate.

Considering the mechanism shown in Figure 3-62a, the relative angle through which the rocker moves can be determined easily by finding the intersection of the circle from the tip of the crank arm and the locus of the tip of the rocker arm, as shown in Figure 3-63. In this example, the motion of the crank arm is limited to 180°, resulting in the irregular motion of the rocker arm through about 60°, as shown in the graph.

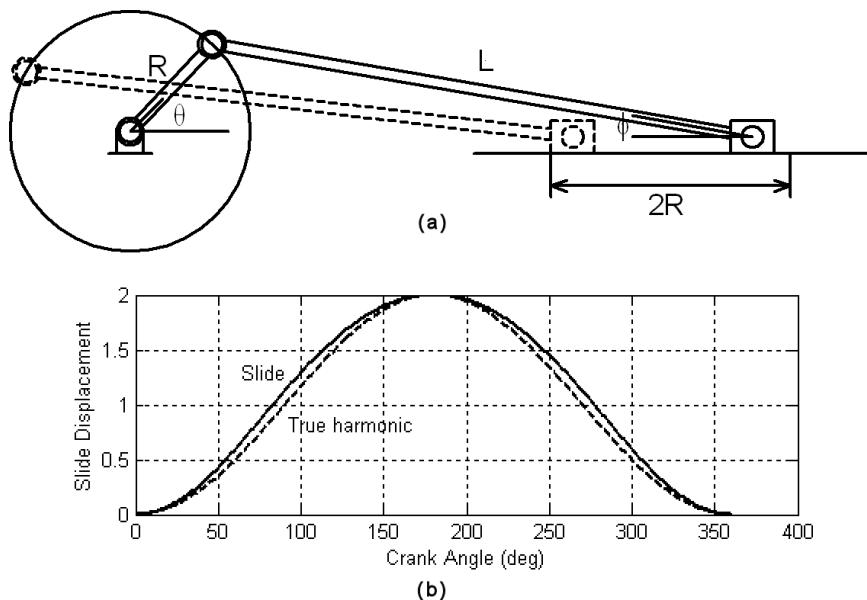
Care should be taken when designing these mechanisms that the lengths of the various elements are never such that the connecting rod and the rocker fall in a straight line. This is known as a dead point because any driving force transmitted to the rocker will be radial and hence cannot affect its rotation.

One of the most common applications of the slider-crank mechanism is to convert linear reciprocating motion of a piston into rotary motion or vice versa. If the connecting rod has a finite length, the motion of the slider is not a simple harmonic motion because the connecting rod operates at an angle to the slider, as shown in Figure 3-64.

Considering Figure 3-64, the displacement of the slider can be written in terms of the length of the crank,  $R$  (m), the length of the connecting rod  $L$  (m), and the two angles.

$$\begin{aligned}
 x &= R + L - R \cos \theta - L \cos \phi \\
 &= R(1 - \cos \theta) + L(1 - \cos \phi) \\
 &= R(1 - \cos \theta) + L \left[ 1 - \sqrt{1 - \sin^2 \phi} \right] \\
 &= R(1 - \cos \theta) + L \left[ 1 - \sqrt{1 - \left( \frac{R}{L} \right)^2 \sin^2 \theta} \right]
 \end{aligned} \tag{3.60}$$

It is obvious that the second term of this expression describes the deviation of the displacement from simple harmonic motion and that as the ratio  $R/L$  decreases—that is, the length of the crank arm tends toward infinite—the difference reduces to zero.



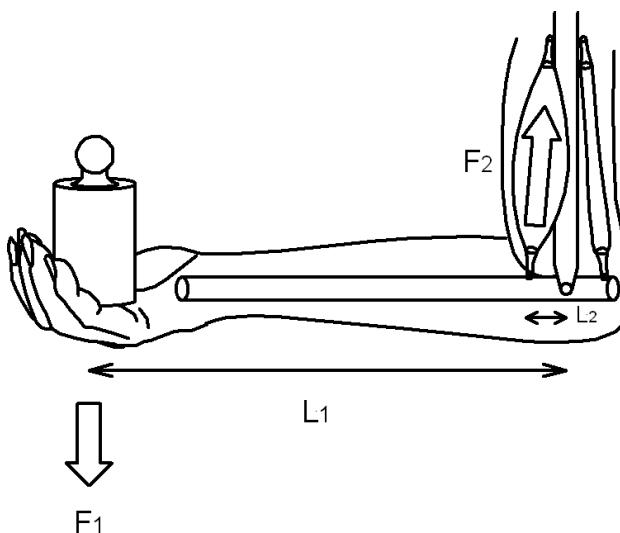
**FIGURE 3-64** ■ Displacement of a slider-crank mechanism with  $R = 1$  and  $L = 4$ .  
 (a) Construction.  
 (b) relationship between the crank angle and the slide displacement.

Linkages can be used to facilitate quite complex motion of prostheses, particularly fingers, where a limited number of actuators must provide as much dexterity as possible. Some of these possibilities are considered in Chapter 11.

A lever, as illustrated in Figure 3-65, consists of a single bar with a fulcrum or pivot that attempts to match the force and displacement of the actuator with the requirements of the system. The equilibrium position of the mechanism can be determined by balancing the moments around the pivot point. However, in the case of the PAMs shown in the diagram in Figure 3-65, the analysis is complicated by the nonlinear force-to-length characteristics of the two actuator elements.

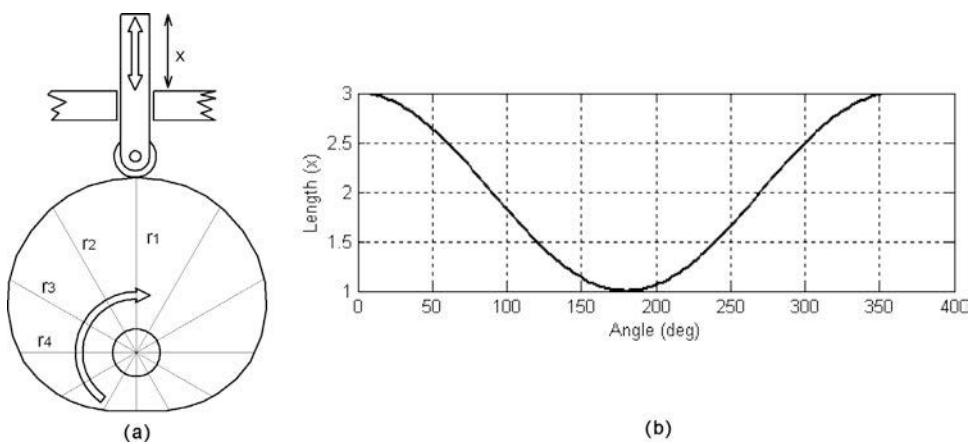
As a first approximation, it can be assumed that the triceps PAM is not contributing to the equation; therefore, in equilibrium the force relationship is

$$F_1 L_1 - F_2 L_2 = 0 \quad (3.61)$$



**FIGURE 3-65** ■ Prosthetic elbow actuated using PAMs.

**FIGURE 3-66** ■  
Cam mechanism with a roller follower.  
(a) Schematic diagram. (b)  
Relationship between the cam  
angle and the follower length.



As the angle between the humerus and the ulna decreases from  $90^\circ$ , the moment is reduced proportional to the sine of the angle. For accurate analysis, a complete free-body diagram should be constructed for each activity, and these can be used to determine the required strength of the individual actuators.

Levers are commonly used in prosthetic limbs as they allow the high-force linear actuators such as those based on pneumatic, hydraulic, shape memory alloy, and piezoelectric materials to be incorporated within the normal form of the limb. They are also very quiet compared with electric motor and gearbox-driven joints. Applications are discussed later in this chapter and more comprehensively in Chapter 10.

### 3.6.2 Cams

Like linkages, cams are convenient, compact, and simple mechanisms for converting rotary motion into other forms of motion. Cams consist of an element with a curved outline that rotates or oscillates to generate a predetermined motion to a second element called a follower. A common example of an ovoid cam with a roller follower is shown in Figure 3-66. In this example, the linear displacement of the follower is determined by the radius of the cam.

In the example in Figure 3-66, the cam mechanism has been developed to produce simple harmonic motion where the radius of the cam is

$$r = 2 + \cos \theta \quad (3.62)$$

The base curve in Figure 3-66 shows the displacement of the follower. Its derivative and second derivative provide the velocity and acceleration, respectively. In this example, these are well behaved because of the nature of the curve, but it is possible to produce cams where extremely high velocities and accelerations occur. This is generally undesirable, and care must be taken to ensure that the magnitude of the acceleration is not excessive. Another consideration is the pressure angle, which is the deviation of the direction of the force of the cam compared with the motion vector of the follower. If the angle is large, the side thrust applied to the follower is also large, and the follower can jam in its guides. This is limited by increasing the size of the cam while providing the same total displacement.

In addition to cams to produce simple harmonic motion, other profiles include straight lines and parabolic curves. The parabolic curve provides the smoothest motion (constant acceleration and therefore a constant force to produce that acceleration) and the least power for operation. The end sections of these profiles are usually modified to limit the

maximum acceleration and, hence, the force required to move the follower (Ham et al., 1958).

Cam mechanisms are fairly common in biomechatronic applications. They can be used to generate specific motion profiles for prosthetic limbs or to drive pump diaphragms in pulsatile artificial hearts.

### 3.6.3 Gears and Belt Drives

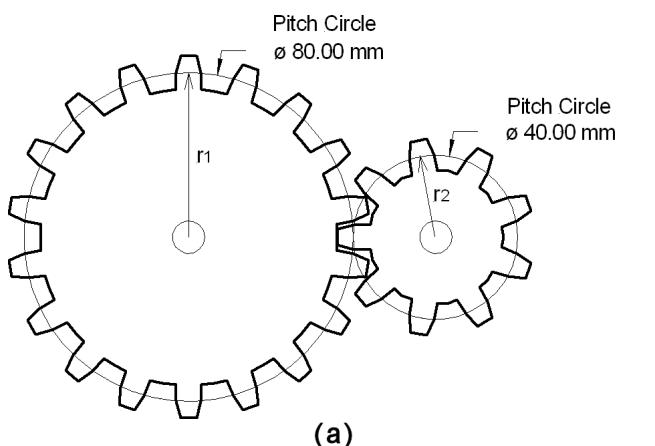
Gears and belt drives provide a means of matching the characteristics of a motor to the load using circular elements of different diameters. Considering the case where all of the power from the motor is transmitted to the load, the available load torque,  $\tau_L$ , (Nm) is

$$\tau_L = \frac{\omega_m}{\omega_L} \tau_m = \frac{r_L}{r_m} \tau_m \quad (3.63)$$

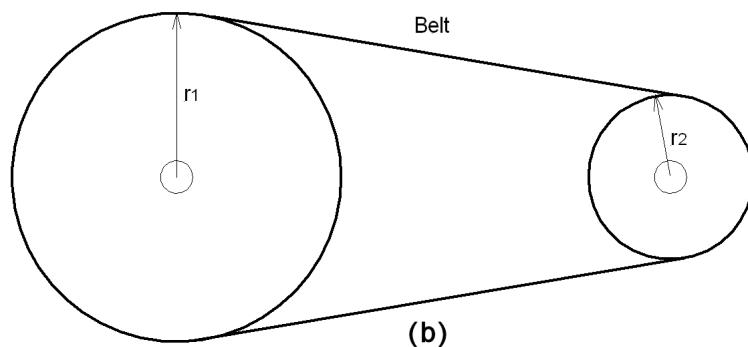
where  $\omega_m$  (rad/s) is the speed of the motor,  $\omega_L$  (rad/s) is the speed of the load, and  $\tau_m$  (Nm) is the motor torque.

The gear (or belt drive) ratio  $\omega_L/\omega_m$  is determined by the ratio of the two pitch circle radii  $r_m/r_L$ , which in turn determines the load speed as a function of the motor speed. The reciprocal of this gives the load torque as a function of the motor torque, therefore decreasing the output speed results in increasing available torque and vice versa.

In Figure 3-67, where  $r_1 = 40$  mm is the pitch circle radius of the driven element and the load is driven by the element with radius  $r_2 = 20$  mm, the output speed will be double that of the input and the available torque will be halved.

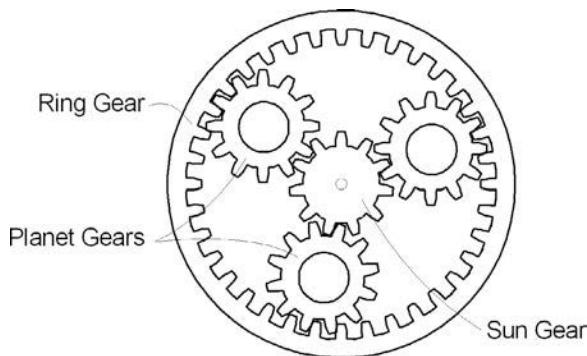


(a)



**FIGURE 3-67 ■**  
Torque and speed  
matching methods.  
(a) Spur gears. (b)  
Belt drives.

**FIGURE 3-68 ■**  
Gear configuration  
for a planetary  
gearhead.



The design of gears must consider strength under static and dynamic loads, quietness and smoothness of operation, backlash, lubrication, temperature and heat dissipation, as well as overall lifetime (Black and Adams, 1968). Most motor manufacturers provide a wide range of gearheads designed for their motors, so it is seldom necessary to design gearing systems from scratch.

Two common gearhead designs are spur and planetary. In general, spur gearheads are simpler and less expensive than planetary units and perform well in low-torque applications. Torque capacity of spur types is limited because each gear in the train bears the entire torsional load. In contrast, planetary gearheads share the load over multiple planet gears. As shown in Figure 3-68, the input shaft drives a central sun gear that, in turn, drives the planet gears. Each of the planet gears simultaneously delivers torque to a rotating carrier plate (not shown) coupled to a geared output shaft.

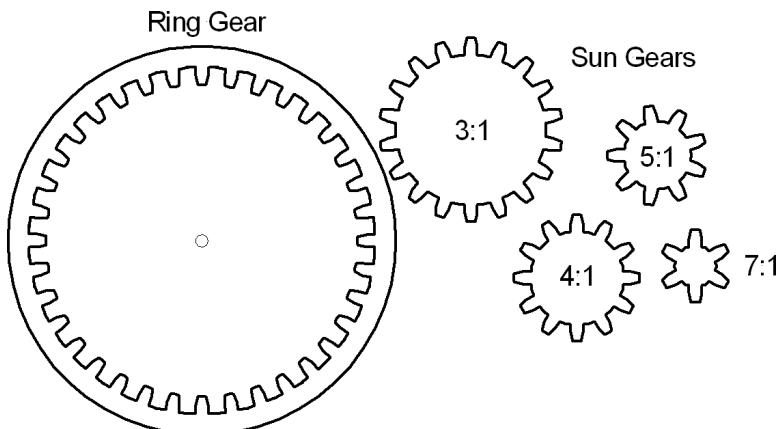
The gear ratio,  $N$ , for a stationary ring gear is determined by the internal diameter of the ring gear,  $\phi_r$  (mm), and the outer diameter of the sun gear,  $\phi_s$  (mm),

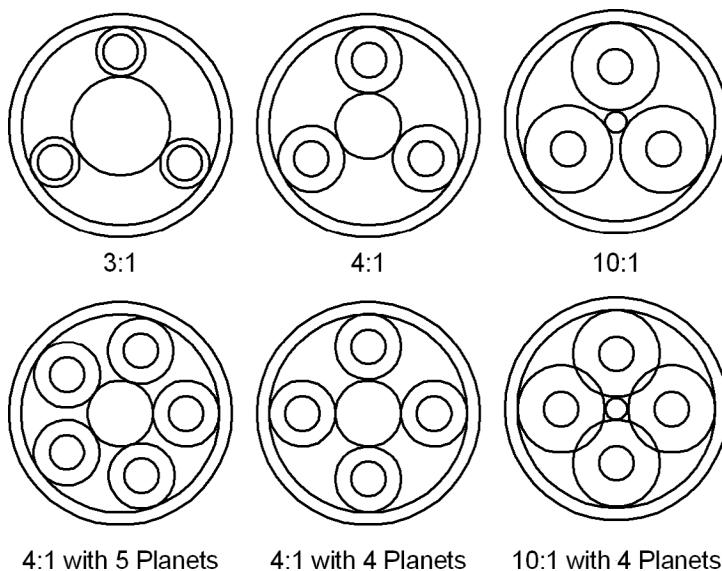
$$N = \frac{\phi_r}{\phi_s} + 1 \quad (3.64)$$

The formula holds if the respective diameters are replaced by the number of teeth in each case.

It is obvious that a 2:1 ratio cannot be achieved as it would require that the diameter of the sun and the ring gears be the same, and that is not possible. Typical ratios are from 3:1 to 7:1, as shown in Figure 3-69.

**FIGURE 3-69 ■**  
Size of sun gear  
compared with ring  
gear for different  
planetary gear  
ratios.





**FIGURE 3-70** ■  
Different configurations for a planetary gearhead.

Because the size of the gears determines their load carrying capacity, it is important to maintain a good balance between the diameters. Figure 3-70 shows the gears in a 3:1, 4:1, and 10:1 system. For a 3:1 ratio, the sun gear is large and the planets are small. In this case the planets have thin walls and therefore limit the space for bearings and carrier pins and thus the load torque. The 4:1 ratio is well balanced with sun and planets about the same size, while the 5:1 and 6:1 ratios (not shown) still yield fairly well-balanced gear sizes. With higher ratios approaching 10:1, the small sun gear becomes a strong limiting factor for the transferable torque.

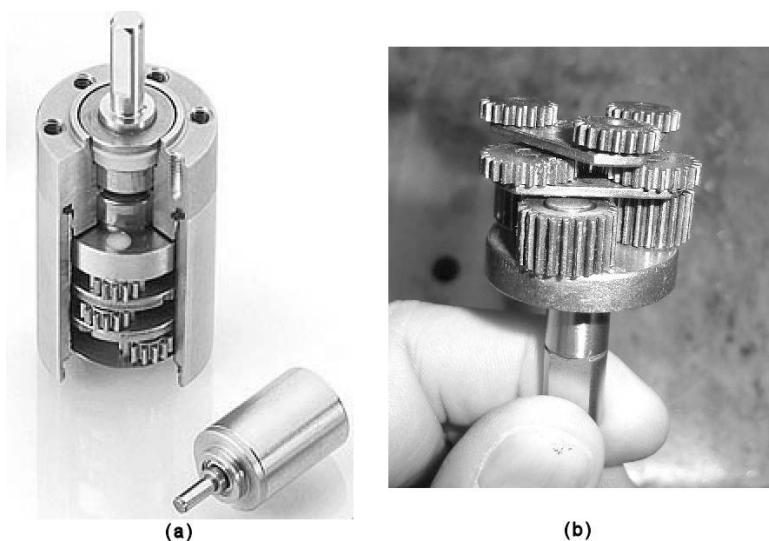
Adding planets increases the torque transfer capacity of the arrangement. Figure 3-70 shows that with lower ratios additional planet gears can be used but that with higher ratios such as 10:1 multiple gears would cause interference (Anthony, 2008).

The material from which the gears are built is also a consideration in regard to torque capability. Sintered nickel steel is often used because the sintering process produces gears that are able to run at closer tolerances. Additionally, because the material is porous, these gears hold lubricant better than steel units do. However, cut-steel gears tend to be more durable and are therefore a better choice for higher-torque applications. That notwithstanding, good lubrication is important regardless of gear material, especially at high speeds and loads. Planetary gearheads have an advantage because oil flying outward from the sun gear is captured by the planet gears and carrier plate. Spur gearheads, on the other hand, fling lubricant off and away from the gears. This is one reason planetary gearheads have higher speed ratings.

Backlash is a measure of positional accuracy usually specified in arc-minutes. For example, a typical spur gearhead has about 10 arc-min of backlash, whereas its planetary counterpart may be rated better than half of that.

Reduction ratios for both spur and planetary gearheads range from near unity up to several thousand to one. Spur gearheads, with a single geared input shaft coupled to a geared output shaft (single stage), provide up to about a 6:1 reduction. As discussed previously, planetary units can reach 10:1 in a single stage. For higher ratios and proportionally greater output torque, multiple stages or gear sets are stacked together axially, as shown

**FIGURE 3-71 ■**  
**Stacked planetary gearhead.**  
 (a) Cutaway of motor. (b) detail of gearhead. (Courtesy of MicroMo Electronics.)



in Figure 3-71. Increasing the number of stages increases the reduction ratio and output torque as well as overall length and also lowers the mechanical efficiency. A typical single-stage spur gearhead is about 90% efficient, whereas a two-stage device is about 85% efficient. Planetary gearheads are slightly more efficient at approximately 97% and 94%, respectively, for one- and two-stage units. Obviously, the efficiency decreases further as additional stages are included.

MicroMo Electronics offers a wide range of gearheads with diameters of from 6 mm to 44 mm (to suit motor diameters). Reduction ratios are available from 3.7:1 to 23.104:1. Maximum input speeds range from 3,500 rpm to 8,000 rpm, and continuous load torques are available from 25 mNm to 16,000 mNm.

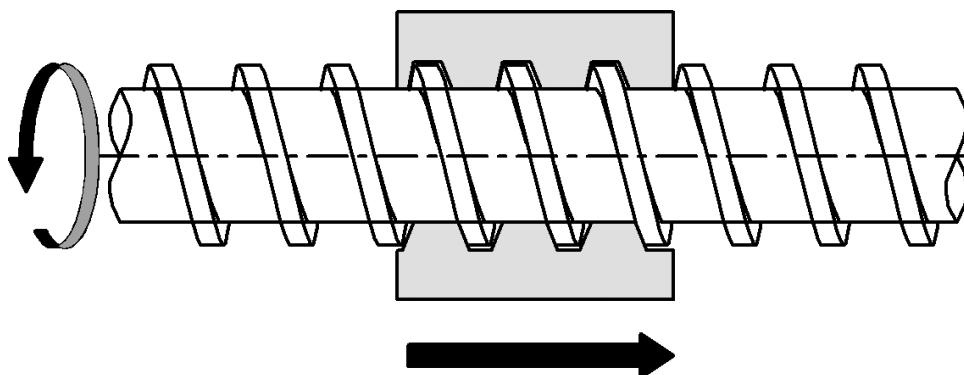
As can be seen in Table 3-9, planetary gearheads are superior to spurs on all counts except mechanical noise and cost. Selection of the appropriate gearhead for a mechatronic application should take all of these factors into account.

Belt drives are commonly used over longer center distances than those used for gears. Conventional flat or V belts slip slightly so the driven speed is less than the ratio of pulley diameters. However, there is no slip in toothed belts.

The maximum amount of power than can be transmitted by a belt drive is determined by belt stretch or excessive slippage on the smaller pulley. This is in turn is determined by the angle of contact, the belt tension, and the coefficient of friction between the belt and the pulley.

**TABLE 3-9 ■** Comparison of planetary and spur gear characteristics

Factor	Planetary	Spur
Backlash	Low	High
Efficiency	High	Low
Load capacity	Good	Adequate
Operating speed	Fast	Slower
Mechanical noise	Noisy	Quiet
Size	Small	Large
Cost	High	Lower



**FIGURE 3-72** ■ Translation screw operation.

In a toothed belt, the sheaves have axial grooves that engage teeth on the belt. The belts usually include a number of thin steel cables that carry tension under load, which permits a light drive to operate at high speed (up to 100 m/s). These drives are compact, light, quiet, and low maintenance. However, they are more sensitive to misalignment than flat or V belts (Black and Adams, 1968).

### 3.6.4 Translation Screw Devices

Translation screws are commonly used compact devices to convert rotary motion produced by a motor into linear motion against large forces. In most translation screws, the screw rotates in its bearings, and the nut moves axially, as shown in Figure 3-72. However, in some devices the nut rotates while the screw moves axially with no rotation.

Consider that the nut of a screw is moved against an axial load by the rotation of the screw, illustrated in Figure 3-73. The load on the nut will be transferred to the screw as a distributed load on the surface of the threads in contact. To facilitate analysis, it can be assumed that the load is concentrated at point *o* on the mean circumference,  $\pi d$  (m), of the thread.

Under static conditions the direction of the load on the thread must be normal to the thread surface, along the vector *ao* as shown in Figure 3-73. As the screw rotates so that the nut is moved against an external force, *Q* (N), the line of action *ao* will be rotated through the angle of friction,  $\phi$  (rad), to *bo*.

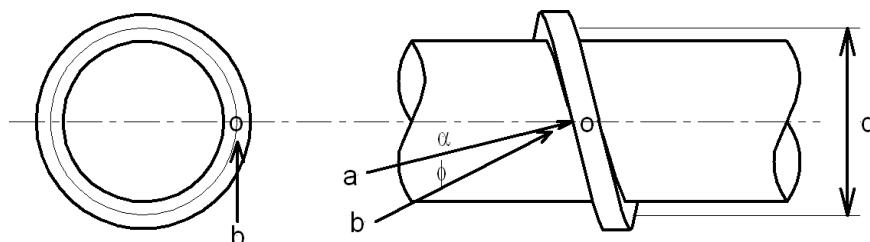
For equilibrium of forces, the component of *bo* parallel to the axis of the screw is

$$Q = bo \cos(\alpha + \phi) \quad (3.65)$$

where  $\alpha$  (rad) is the lead angle of the screw, and  $\phi$  (rad) is the friction angle.

The component of *bo* at right angles to the axis of the screw is

$$F = bo \sin(\alpha + \phi) \quad (3.66)$$



**FIGURE 3-73** ■ Forces on a square threaded translation screw.

Equating  $F$  in terms of  $Q$

$$F = Q \tan(\alpha + \phi) \quad (3.67)$$

The torque,  $\tau$  (Nm), is

$$\tau = F \frac{d}{2} = \frac{Qd}{2} \tan(\alpha + \phi) \quad (3.68)$$

To determine the efficiency of the screw, consider the required torque,  $\tau_o$  (Nm), in the absence of any friction. Equation (3.68) can be rewritten as

$$\tau_o = \frac{Qd}{2} \tan \alpha \quad (3.69)$$

The efficiency is

$$e = \frac{\tau_o}{\tau} = \frac{\tan \alpha}{\tan(\alpha + \phi)} \quad (3.70)$$

The coefficient of thread friction  $f = \tan \phi$  is about 0.1 for well-machined, run-in, and well-lubricated threads. This increases to about 0.125 for average workmanship and quality. Static friction can be approximated as 1.33 times the dynamic friction (Black and Adams, 1968).

Application of this technology in linear actuators was discussed briefly earlier in this chapter.

To improve the overall efficiency of translation screw devices, steel balls can be introduced between properly formed threads on the screw and nut. This ball-screw mechanism substitutes rolling friction for sliding friction. When the nut is translated by rotation of the screw, the balls run out of the thread of the nut and are returned to the other end via a recirculating groove. Efficiencies of higher than 90% can be achieved.

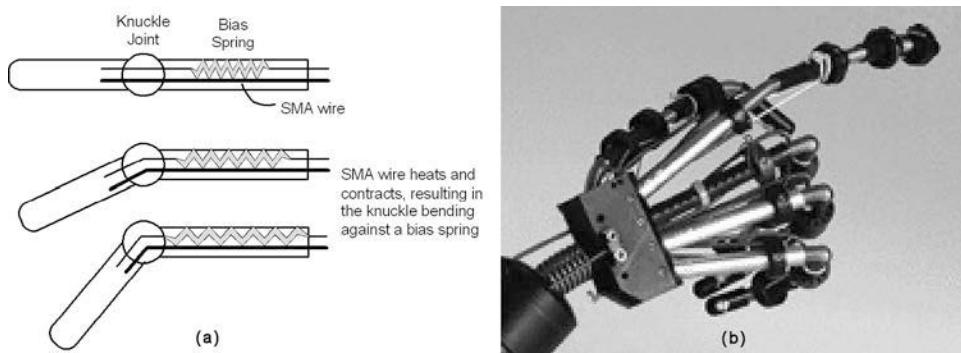
## 3.7 | PROSTHETIC HAND ACTUATION

---

Many attempts have been made to recreate human anatomy through mechanical means; most of these have been met with limited success because the complexity of the human body makes it very difficult to duplicate even simple functions effectively. Most prosthetic hands in industrial production have been restricted to rugged two- or three-finger grippers and are used to execute relatively simple movements. Prosthetic hands for more delicate tasks have proven expensive and difficult to manufacture in the past due to the lack of available technology. However, in the past decade, miniaturization of electronics and mechanical systems has allowed for good progress, as discussed in Chapter 10.

To reproduce human extremities, a number of aspects must be considered:

- The gripping force required to manipulate different objects (e.g., eggs, pens, tools)
- The motion capabilities of each joint of the hand
- Minimization of actuator noise
- The ability to feel or touch objects (tactile senses)
- The method of controlling movement within the limb
- Emulation of real human movement (i.e., smoothness and speed of response)



**FIGURE 3-74** ■  
Using shape memory alloy to actuate a prosthetic hand. (a) Schematic diagram. (b) Photograph of hand. (Courtesy Robotics and Mechatronics Laboratory, Rutgers University.)

Many different solutions have been proposed for these problems, including using “muscles” controlled by air pressure, piezoelectric materials, or SMAs.

### 3.7.1 Shape Memory Alloys

Shape memory alloys mimic human muscles and tendons quite well. They are strong and compact so that large groups of them can be used for prosthetic applications. In addition, the motion with which they contract and expand is very smooth, producing a lifelike movement unavailable in other systems.

Creating human motion using SMA wires is a complex task, but the basics are straightforward, as illustrated in Figure 3-74. For example, to create a single direction of movement, a bias spring shown in the upper portion of the finger holds the finger straight, stretching the SMA wire. When the wire on the bottom portion of the finger is heated, it shortens, and the joint will be bent downward. The heating takes place by passing an electric current through the wire.

Some challenges must still be overcome before prosthetic hands can become more commonplace: (1) generating the computer software used to control the artificial muscle systems within the robotic limbs; (2) creating sufficient movements to emulate human flexibility (i.e., being able to bend the joints as far as human beings can); and (3) reproducing the speed and accuracy of human reflexes.

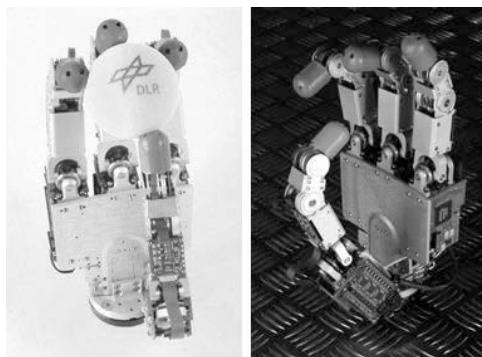
### 3.7.2 Electric Motors

A combination of microelectronics and micromechanics has provided the means to produce prosthetic hands with separately controllable fingers and joints based on human hands. One example is the device developed by the German Aerospace Centre (DLR), in cooperation with the Harbin Institute of Technology (HIT). This prototype prosthetic hand, shown in Figure 3-75, uses miniature actuators and high-speed bus technology.

Constructing a prosthetic hand with the strength and dexterity of a human hand requires at least four fingers: three fingers to allow the hand to grip conically shaped parts; and a thumb in opposition. Consequently, the DLR hand consists of three fingers, each containing four joints with three degrees of freedom. The fourth finger, designed as a thumb, has four degrees of freedom.

Tiny but powerful motors and shaft encoders are fitted directly in the finger. Each finger joint also houses a contactless angle sensor and a torque sensor. Since both sensors provide extremely high-resolution feedback, a high-speed, three-wire serial connection

**FIGURE 3-75 ■**  
DLR/HIT prosthetic hand based on miniature electric motors. (Courtesy MicroMo.)



conveys sensed data back to field-programmable gate array (FPGA) based controllers. Presently these are on a plug-in peripheral component interconnect (PCI) card integrated in a standard personal computer (PC), but in the future it will be possible to integrate all of the processing into the hand.

The complexity of this prosthetic hand has its price with each finger requiring several separately controllable actuators. In this case, there are 12 16 mm diameter Faulhaber analog Hall sensor-controlled brushless DC motors per hand. These are commercially available, high-performance motors. They can be connected with gear systems of the same diameter to form a single integrated unit. The motors produce an output power of 11 W and a maximum continuous torque of up to 2.6 mNm without gearing. However, for this application the motor's no-load speed of 29,900 rpm is reduced to 188 rpm through all-metal planetary gearheads with a ratio of 159:1. These simultaneously increase the torque to a maximum value of 450 mNm. Hall sensors provide rotor position to the controller and deliver the requisite feedback information with a resolution of at least eight bits. The Hall sensors and motor form a compact 31 g unit with a length of only 28 mm and an outer diameter of 16 mm.

### 3.7.3 Pneumatic Artificial Muscles

The Shadow Hand is the closest robot hand to the human hand available. It provides 24 movements, allowing a direct mapping from a human hand to the prosthesis. In addition, it has integrated sensing and position control, allowing precise computer control.

The hand contains an integrated bank of 40 air muscles for actuation, some of which are visible in Figure 3-76. The muscles are compliant, which allows the hand to be used

**FIGURE 3-76 ■** The Shadow Hand in action. (Courtesy Shadow Robot Company <http://www.shadowrobot.com/>, with permission.)



around soft or fragile objects, and because it can be fitted with touch sensitive pads on its fingertips it can be made sufficiently sensitive to detect an object the size of a small coin.

## 3.8 REFERENCES

---

- Alciatore, D. and M. Histan. (2003). *Introduction to Mechatronics and Measurement Systems*, 2d ed. Boston: McGraw Hill.
- Anthony, G. (2008). "The Best Balanced Planetary Ratio from a Torque Density Point of View." Retrieved December 2008 from <http://www.motioncontrol.com/products/index.cfm/Balance-Planetary-Ratio-Torque-Density>
- Black, P. and O. Adams. (1968). *Machine Design*. Tokyo: McGraw-Hill Kogakusha, Ltd.
- Brooker, G. (2008). *Sensors for Ranging and Imaging*. Raleigh, NC: Scitech.
- Brown, W. (2002). AN857: Brushless DC Motor Control Made Easy. *Microchip*. Retrieved June 2008 from <http://www.microchip.com/>
- Fraden, J. (1996). *Handbook of Modern Sensors*. New York: AIP Press, Springer-Verlag.
- Ham, C., Crane, E., and Rogers, W. (1958). *Mechanics of Machinery*, 4th ed. New York: McGraw Hill Kogakusha.
- Ku, K., and Bradbeer, R. (2008). Modelling Pneumatic Muscles as Hydraulic Muscles for Use as an Underwater Actuator. In *Mechatronics and Machine Vision in Practice*, J. Billingsley and R. Bradbeer (Eds.). Berlin: Springer.
- Kuphaldt, T. (2008). "Lessons in Electric Circuits: Volume II—AC." Retrieved July 2008 from <http://www.ibiblio.org/kuphaldt/electricCircuits/AC/index.html>
- MicroMo. (2008a). "How to Select a DC Motor." Retrieved July 2008 from <http://www.faulhaber-group.com/n390290/n.html>
- MicroMo. (2008b). "Micro-actuators Move Sensitive 3-Finger Hand." Retrieved June 2009 from <http://www.micromo.com/n280126/n.html>
- Physikinstrumente. (2008). "Piezo Actuators Overview." Retrieved September 2009 from <http://www.physikinstrumente.com/en/products/piezo/index.php>
- Portescap. (2009). "Ironless DC Motor." Retrieved June 2009 from <http://www.portescap.com/>
- SMA/MEMS Research Group. (2001). "Shape Memory Alloys." Retrieved July, 2008 from [http://www.cs.ualberta.ca/~database/MEMS/sma\\_memes/sma.html](http://www.cs.ualberta.ca/~database/MEMS/sma_memes/sma.html)
- Wikipedia. (2008). "Pneumatic Artificial Muscles." Retrieved July 2008 from [http://en.wikipedia.org/wiki/Pneumatic\\_artificial\\_muscles](http://en.wikipedia.org/wiki/Pneumatic_artificial_muscles)
- Yedamale, P. (2003). AN885: Brushless DC (DLDC) Motor Fundamentals. *Microchip*. Retrieved June 2008 from <http://www.microchip.com/>



# Feedback and Control Systems

## Chapter Outline

4.1	Introduction .....	159
4.2	Biological Feedback Mechanisms.....	160
4.3	Biomechatronic Feedback Mechanisms .....	160
4.4	System Representation .....	162
4.5	System Models .....	164
4.6	System Response .....	174
4.7	System Stability .....	181
4.8	Controllers.....	188
4.9	Controller Implementation .....	201
4.10	References .....	205

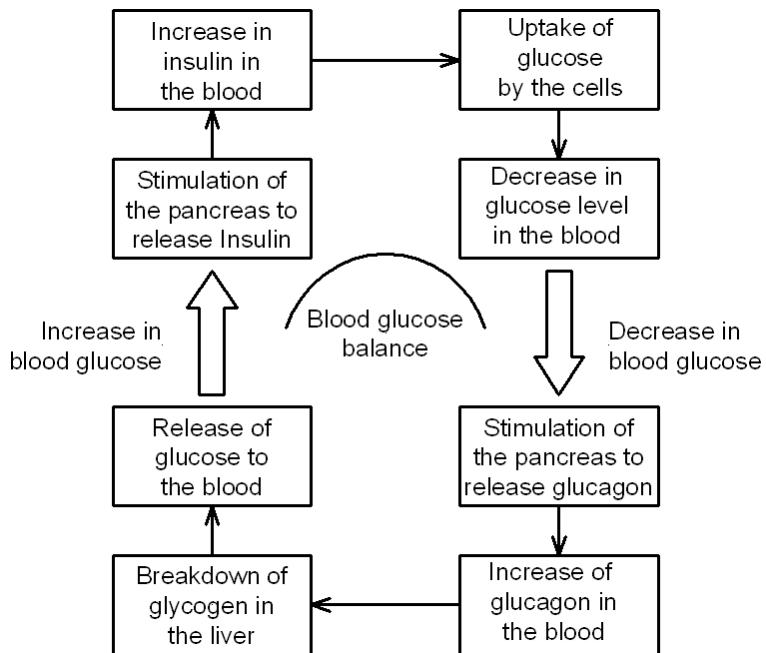
## 4.1 INTRODUCTION

A *system* can be defined as the artificial boundary surrounding a collection of interacting components, which are considered a “black box” consisting of a limited number of inputs and outputs. In the natural world, this could be a complete ecosystem, a single human being, or a physiological system (e.g., cardiovascular, respiratory) within a human being. From a mechatronic perspective, systems can be as complex as atomic power stations or as simple as a pneumatic actuator or a single electronic component.

From a biomechatronic perspective, it is often important that a system’s characteristics be well enough understood to be controlled. The control can occur in an open-loop manner, which implies that the system is provided with an input, and this will determine the output state. For example, a skin patch that provides a constant supply of insulin to a diabetic is an open-loop system. Unfortunately, open-loop control does not cater to changes in the system’s characteristics, so if the diabetic drank a large glass of fruit juice, his blood sugar level would increase substantially, with potentially serious consequences. If, however, the skin patch was replaced with a more sophisticated device that measured the blood sugar level and compensated for changes by adjusting the rate at which insulin was infused, the blood sugar level would remain reasonably stable. This is known as closed-loop control, and it relies on feedback of an output parameter to help regulate the system.

This controlled process may be applied on a macroscopic scale (i.e., feedback from the output to the input of a complex system) or on a much smaller scale, such as with a simple filter that applies a proportion of the output voltage to the input.

**FIGURE 4-1** ■  
Feedback mechanisms used to control blood glucose levels.



## 4.2 | BIOLOGICAL FEEDBACK MECHANISMS

The human body is, of course, a riot of cross-coupled feedback processes, some of which are understood but many of which are still beyond our comprehension. Figure 4-1 illustrates the control aspects that govern the aforementioned interaction between insulin and glucose levels. It is one of the simpler internal processes, yet it involves adjusting the levels of both insulin and glucagon to control glucose production and uptake.

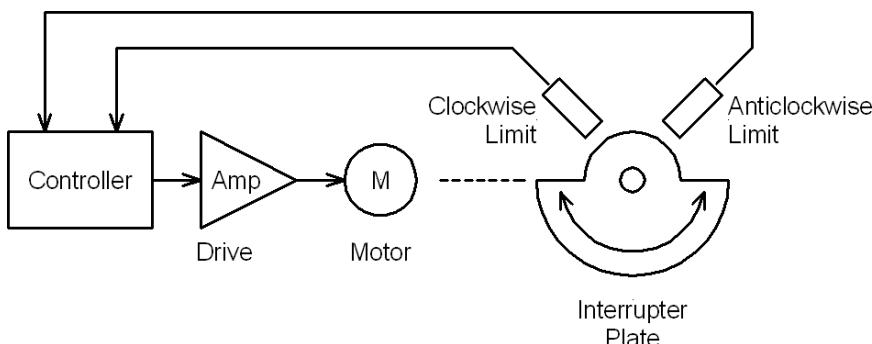
Insulin is synthesised in the pancreas, and its secretion from  $\beta$ -cells into the blood is controlled by the concentration of glucose in the blood. As the level of glucose rises (e.g., after a meal), the insulin level follows suit. Most cells in the body have receptors to which insulin binds, and when this occurs the cell is stimulated to absorb glucose from the bloodstream.

When glucose levels fall below a certain threshold,  $\alpha$ -cells in the pancreas are stimulated to secrete glucagon; this is carried directly to the liver where it stimulates the breakdown of glycogen to glucose, which is released into the bloodstream.

Other biological feedback mechanisms are discussed briefly in later chapters. One example, discussed in Chapter 9, is the short- and long-term control of blood pressure; another is the less well-understood neural feedback mechanisms in the visual cortex that allow us to interpret visual signals from the retina.

## 4.3 | BIOMECHATRONIC FEEDBACK MECHANISMS

Almost all biomechatronic systems incorporate some form of feedback for control purposes. Some examples discussed in later chapters include the control of the output level of a hearing prosthesis and the control of blood flow in artificial hearts and ventricular



**FIGURE 4-2 ■**  
Application of limit switches as feedback elements.

assist devices. The latter example is interesting because the parameter being controlled is not, in fact, measured and must be estimated from other available data like the current drawn by the pump motor. However, in most cases it is possible to measure the value of the parameter to be controlled and to apply feedback to control it. This chapter examines the basic principles of modeling systems and applying feedback to improve their control.

### 4.3.1 Limit Switches

Probably the simplest of all of the feedback mechanisms is the limit switch. These devices can be mechanical or electrical, as discussed in Chapter 2, and their function is to produce an output when a mechanism has reached the end of its allowable range of travel. An example of such a system is shown in Figure 4-2.

In this example, an electric motor drives an interrupter plate that is designed to break the optical path of a pair of optical switches at the clockwise and counterclockwise limits of travel. These outputs are fed into a controller that can be used to cut power to the motor or to reverse the direction of travel. This form of feedback can be applied to a pulsatile ventricular assist device to produce a continuous reciprocating pumping action. In most cases, however, limit switches are used only as warning devices in association with more complex mechanisms to ensure that mechanical damage does not occur by driving a device beyond its allowable limits.

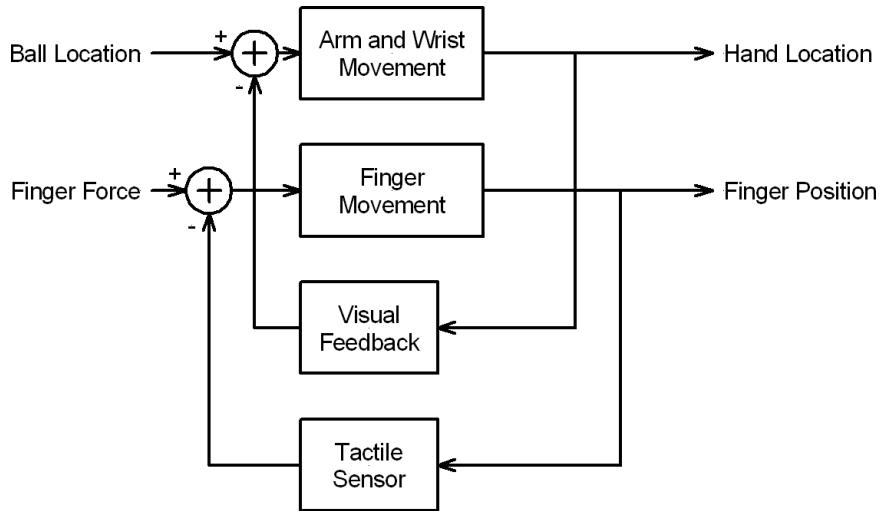
### 4.3.2 Proportional and Higher-Order Controllers

In a proportional control system, the set point is compared with the measured output of the system, and the difference between the two is applied as a correction input to the system. In most cases, the sense of the correction signal drives the output to minimize, and sometimes even to eliminate, the error completely. Under some conditions, simple proportional controllers are either not sufficiently fast or are not stable and need to be augmented by additional states such as the derivative or the integral of the error signal.

In addition, as discussed, systems often comprise multiple inputs and outputs that are often cross-coupled, which makes efficient control that much more difficult. For example, consider the sophisticated prosthetic arm shown in Figure 4-3, which is commanded to pick up a ball.

In this case, both the position of the hand used to reach the ball and the positions of the fingers used to grasp it need to be controlled. In the case of the position of the hand, it is assumed that visual feedback is used to control this function and that when it is in

**FIGURE 4-3 ■**  
Block diagram for multi-input, multi-output control.



place the hand can be commanded to close and tactile feedback from force sensors in the fingertips is used to control the grip strength. What is not shown in the diagram is the cross-coupling between the actual finger positions and the position of the arm, which ensures that the ball is grasped accurately and not bumped out of reach.

## 4.4 | SYSTEM REPRESENTATION

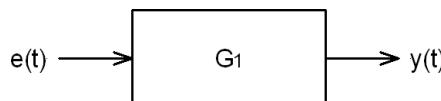
The classical representation of a system is a rectangular box that represents the system transfer function. A single box may be used, or multiple boxes may be linked in series, in which case the overall system transfer function is the product of the individual transfer functions. An open-loop system can be represented by a box with an input and an output, as shown in Figure 4-4.

In this case, the input,  $e(t)$ , is referred to as the actuating signal, and  $y(t)$  is the controlled variable, where  $G_1$  is the control element, so

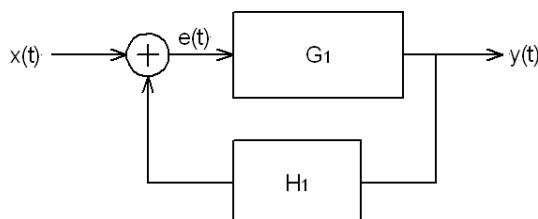
$$y(t) = G_1 e(t) \quad (4.1)$$

With the application of feedback, as shown in Figure 4-5, the block diagram starts to resemble those in the previous examples.

**FIGURE 4-4 ■**  
Open-loop system.



**FIGURE 4-5 ■**  
Closed-loop system.



In this case,  $x(t)$  is the set point, and  $e(t)$  is the error signal, which is the difference between the set point and the feedback signal.

$$e(t) = x(t) - b(t) \quad (4.2)$$

where  $b(t)$  is the feedback signal and is related to the output signal, or controlled variable,  $y(t)$ , by the feedback element  $H_1$

$$b(t) = H_1 y(t) \quad (4.3)$$

If equation (4.2) is written in terms of (4.1) and (4.3)

$$\begin{aligned} \frac{y(t)}{G_1} &= x(t) - H_1 y(t) \\ y(t) \left[ \frac{1}{G_1} + H_1 \right] &= x(t) \\ y(t) \left[ \frac{1 + G_1 H_1}{G_1} \right] &= x(t) \end{aligned} \quad (4.4)$$

The system transfer function is the ratio of the output over the input

$$T_1 = \frac{y(t)}{x(t)} = \frac{G_1}{1 + G_1 H_1} \quad (4.5)$$

The main advantages of using closed-loop feedback are that the system becomes less sensitive to disturbances and small changes to the characteristics of the control element  $G_1$ . It can also result in faster response to changes in the set point. However, the overall gain is reduced from  $G_1$  to  $T_1$ , and the system can become unstable under some circumstances.

### WORKED EXAMPLE

---

#### Motor Control

In an open-loop controller, the transfer function of a motor is  $G = 250 \text{ rpm/V}$ . Determine the steady-state speed if 12 V is applied to the input.

In this case, the open-loop model shown in Figure 4-4 is used, where  $G_1 = 250$  and  $e(t) = 12 \text{ V}$

$$\begin{aligned} y(t) &= Ge(t) \\ &= 250 \times 12 \\ &= 3000 \text{ rpm} \end{aligned}$$

In the closed-loop case, consider that a tachogenerator is coupled to the motor shaft and its transfer function is  $H_1 = 3 \text{ mV per rpm}$ . What is the new transfer function?

From equation (4.5)

$$\begin{aligned} T_1 &= \frac{G_1}{1 + G_1 H_1} \\ &= \frac{250}{1 + 250 \times 3 \times 10^{-3}} \\ &= 142.86 \text{ rpm/V} \end{aligned}$$

The application of 12 V will produce an output speed of  $12 \times 142.86 = 1715 \text{ rpm}$ .

At this speed, the tachogenerator will produce an output of  $1714 \times 3 \times 10^{-3} = 5.14$  V, producing an error voltage of  $12 - 5.14 = 6.86$  V, which in turn produces an output speed of

$$\begin{aligned}y(t) &= e(t)G_1 \\&= 6.86 \times 250 \\&= 1715 \text{ rpm}\end{aligned}$$

as expected.

The control elements  $G_1$  and  $H_1$  can be simple gains, as shown in this example, but in general they are more complex and are modeled as differential equations.

---

## 4.5 | SYSTEM MODELS

One way of analyzing a system is to produce a lumped parameter model of each of the elements and then to combine them to form a system. Fortunately, there are strong similarities among the behaviors of most types of elements of interest in biomechatronics, including mechanical, electrical, fluidic, and thermal systems, and this makes their analysis reasonably straightforward.

### 4.5.1 Mechanical Elements

Mechanical systems consist of springs, dashpots, and masses, where the springs are stiffness, dashpots are the forces opposing motion (damping and friction), and masses are the inertial components or resistance to acceleration.

The stiffness of a spring is described by the linear relationship between the compressive or expansive force,  $F$  (Newton), and the displacement,  $x$  (m)

$$F = kx \quad (4.6)$$

where  $k$  (N/m) is the spring constant.

A dashpot consists of a piston moving within a closed, fluid-filled cylinder. As the piston moves, fluid must pass through a small hole through the piston from one side of the cylinder to the other. This flow produces a resistive force,  $F$  (N), which is proportional to the velocity,  $v$  (m/s), of the piston

$$F = cv \quad (4.7)$$

where  $c$  (N per m/s) is the constant of proportionality.

This can be rewritten in the form of a derivative, because velocity is the rate of change of position

$$F = c \frac{dx}{dt} \quad (4.8)$$

Finally, in the case of the mass,  $m$  (kg), for a given acceleration,  $a$  (m/s<sup>2</sup>), the greater the mass, the higher the required force,  $F$  (N)

$$F = ma \quad (4.9)$$

where the constant of proportionality between the acceleration and the force is the mass.

Because acceleration is the rate of change of velocity, which is in turn equal to the rate of change of displacement, equation (4.9) can be rewritten as

$$F = m \frac{dv}{dt} = m \frac{d^2x}{dt^2} \quad (4.10)$$

It is interesting to note that potential energy is stored in the compression of the spring and kinetic energy is stored in the velocity of the mass, but energy is dissipated in the dashpot. In this case, the potential and kinetic energy (Joules) is

$$\begin{aligned} E_{spring} &= \frac{1}{2} k x^2 \\ E_{mass} &= \frac{1}{2} m v^2 \end{aligned} \quad (4.11)$$

and the power,  $P_{dash}$  (Watt), dissipated in the dashpot is proportional to the square of the velocity

$$P_{dash} = cv^2 \quad (4.12)$$

Similar mechanical models can be derived for the rotary equivalents of these linear equations.

For a torsion spring

$$T = k\theta \quad (4.13)$$

where  $T$  (Nm) is the torque,  $\theta$  (rad) is the angle of rotation, and  $k$  (Nm per rad) is the constant of proportionality.

In the case of a rotary damper, often modeled as a fan in a fluid bath, the equation is

$$T = c\omega = c \frac{d\theta}{dt} \quad (4.14)$$

where  $c$  (Nm per rad/s) is the damping constant, and  $\omega$  (rad/s) is the angular rate.

The moment of inertia,  $I$  ( $\text{kgm}^2$ ), is the constant of proportionality between the torque and the angular acceleration  $\alpha$  ( $\text{rad/s}^2$ )

$$T = I\alpha = I \frac{d\omega}{dt} = I \frac{d^2\theta}{dt^2} \quad (4.15)$$

As before, potential energy is stored in the torsion spring

$$E_{spring} = \frac{1}{2} k \theta^2 \quad (4.16)$$

and kinetic energy in the rotating inertia

$$E_I = \frac{1}{2} I \omega^2 \quad (4.17)$$

but power is dissipated by the rotary damper

$$P_{damp} = c\omega^2 \quad (4.18)$$

These equations are summarized in Table 4-1.

**TABLE 4-1** ■ Description of Mechanical Blocks

Block	Describing Equation	Energy/Power
Spring	$F = kx$	$E_{spring} = \frac{1}{2}kx^2$
Torsion spring	$T = k\theta$	$E_{spring} = \frac{1}{2}k\theta^2$
Mass	$F = m \frac{dv}{dt} = m \frac{d^2x}{dt^2}$	$E_{mass} = \frac{1}{2}mv^2$
Moment of inertia	$T = I\alpha = I \frac{d\omega}{dt} = I \frac{d^2\theta}{dt^2}$	$E_I = \frac{1}{2}I\omega^2$
Dashpot	$F = cv = c \frac{dx}{dt}$	$P_{dash} = cv^2$
Rotational damper	$T = c\omega = c \frac{d\theta}{dt}$	$P_{damp} = c\omega^2$

### 4.5.2 Mechanical Model

Many models used in biomechatronic systems can be modeled using the building blocks described in the previous system. For example, consider the actuator that is driving the incus in a middle ear implantable hearing device (MEIHD), shown in Figure 4-6.

As a first approximation, the driven mass consists of the sum of the mass of the actuator pin and the ossicles. The spring component results from the springiness of the flexible linkages between the ossicles and their attachment to the eardrum and the round window, while the damping component is also determined by the energy absorbed in these linkages.

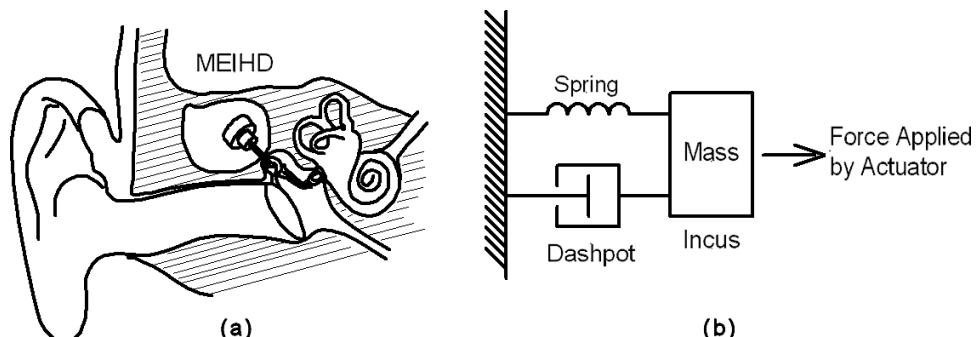
In this model, it is assumed that the applied forces all operate in the direction of movement of the ossicles. Therefore, the net force applied to the mass  $m$  (kg) is  $F - kx - cv$ , and it results in an acceleration  $a$  ( $\text{m/s}^2$ ). Written as an equation

$$ma = F - kx - cv \quad (4.19)$$

This can be written as a differential equation

$$m \frac{d^2x}{dt^2} = F - kx - c \frac{dx}{dt} \quad (4.20)$$

**FIGURE 4-6** ■ Diagram of (a) A middle ear implantable hearing device and (b) Simplified free-body diagram of the system.



which can be rearranged; thus,

$$m \frac{d^2x}{dt^2} + c \frac{dx}{dt} + kx = F \quad (4.21)$$

This is the classic form for a second-order differential equation that can be rewritten in terms of its natural oscillation frequency,  $\omega_n$ , and damping ratio,  $\xi$ , where

$$\omega_n = \sqrt{\frac{k}{m}} \quad (4.22)$$

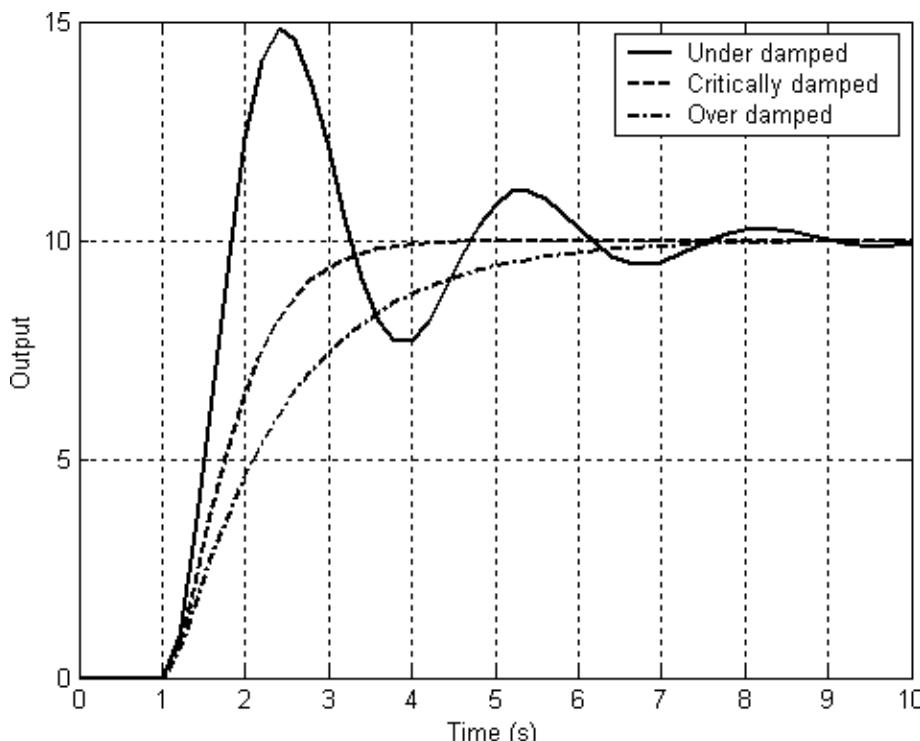
and

$$\xi = \frac{c}{2\sqrt{mk}} \quad (4.23)$$

Rewriting equation (4.21) in terms of these two parameters

$$\frac{1}{\omega_n^2} \frac{d^2x}{dt^2} + \frac{2\xi}{\omega_n} \frac{dx}{dt} + x = \frac{F}{k} \quad (4.24)$$

The response of the system depends on the natural frequency and the damping ratio. In this application, the actuator characteristics would be adjusted so that the natural frequency of the system is outside the normal audio range to ensure that audio signals do not excite any resonances, and the damping ratio would be adjusted to ensure that the response is critically damped or even overdamped so that no ringing occurs. The step responses shown in Figure 4-7 are typical for a second-order differential equation.



**FIGURE 4-7** ■ Response of a second-order differential equation to a step.

### 4.5.3 Electrical Elements

The basic building blocks for electrical circuits are inductors, capacitors, and resistors, as discussed in Chapter 2. These elements are described in terms of the relationship between the applied voltage and the current flow.

For an inductor, the potential difference,  $V$  (volts), across it is proportional to the rate of change of current,  $i$  (amps), through it

$$V = L \frac{di}{dt} \quad (4.25)$$

where  $L$  (Henry) is the inductance, and the polarity of the induced potential difference is opposite to the polarity of the potential difference used to drive the current.

This relationship can also be described in terms of an integral

$$i = \frac{1}{L} \int V dt \quad (4.26)$$

For a capacitor,  $C$  (farads), the potential difference,  $v$  (volts), depends on the charge,  $q$  (coulomb),

$$V = \frac{q}{C} \quad (4.27)$$

From the definition, current is the rate of change of charge

$$i = \frac{dq}{dt} \quad (4.28)$$

which can be rewritten as an integral

$$q = \int i dt \quad (4.29)$$

Substituting into equation (4.27)

$$V = \frac{1}{C} \int i dt \quad (4.30)$$

or

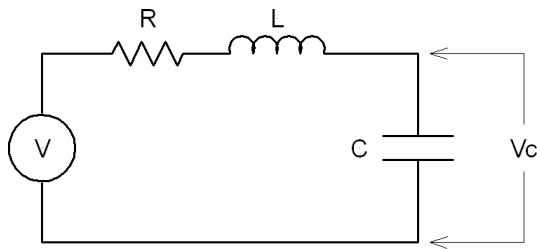
$$i = C \frac{dV}{dt} \quad (4.31)$$

The final electrical element is the resistor,  $R$  (ohms), and in this case the potential difference,  $V$  (volts), across it is proportional to the current,  $i$  (amps), flowing through it

$$V = Ri \quad (4.32)$$

As with the mechanical model, in the electrical case, both the inductor and the capacitor store energy, where

$$\begin{aligned} E_{ind} &= \frac{1}{2} L i^2 \\ E_{cap} &= \frac{1}{2} C V^2 \end{aligned} \quad (4.33)$$



**FIGURE 4-8** ■ Simple RLC circuit.

while power is dissipated by the resistor

$$P_{res} = \frac{1}{R} V^2 \quad (4.34)$$

#### 4.5.4 Electrical Model

The equations that describe how electrical building blocks are combined are known as Kirchhoff's laws and can be summarized as follows:

- The sum of all of the currents flowing into a junction (or a node) is zero. That is, the same amount of current that flows into a junction must flow out again.
- In a circuit, the algebraic sum of the potential differences around the circuit is equal to the applied electromotor function (EMF).

Consider a simple circuit that contains a resistor, an inductor, and a capacitor, as shown in Figure 4-8, and determine the equations that describe the potential difference across the capacitor as a function of time.

Applying Kirchhoff's second law

$$V = V_R + V_L + V_C \quad (4.35)$$

Because there is a single loop, the same current must flow through each of the components; therefore,

$$\begin{aligned} V_R &= iR \\ V_L &= L \frac{di}{dt} \end{aligned}$$

Substituting into equation (4.35)

$$V = iR + L \frac{di}{dt} + V_C \quad (4.36)$$

and from Table 4-2 we know that

$$i = C \frac{dV_C}{dt} \quad (4.37)$$

Therefore, substituting (4.37) into (4.36) gives

$$V_C = V - RC \frac{dV_C}{dt} - LC \frac{d^2V}{dt^2} \quad (4.38)$$

This is a second-order differential equation and can therefore be expressed in the same way as its mechanical counterpart.

**TABLE 4-2** ■ Description of Electrical Blocks

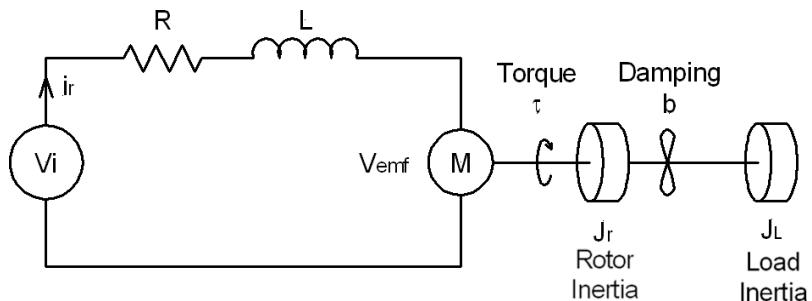
Block	Describing Equation		Energy/Power
Inductor	$V = L \frac{di}{dt}$	$i = \frac{1}{L} \int V dt$	$E_{ind} = \frac{1}{2} Li^2$
Capacitor	$V = \frac{1}{C} \int i dt$	$i = C \frac{dV}{dt}$	$E_{cap} = \frac{1}{2} CV^2$
Resistor	$V = Ri$	$i = \frac{V}{R}$	$P_{res} = \frac{1}{R} V^2$

**WORKED EXAMPLE****Direct Current Motor Model**

A simple model for an electric motor combines electrical and mechanical elements, as shown in Figure 4-9 and described in the following section.

**FIGURE 4-9** ■

Block diagram of elements of an electric motor.



The torque,  $\tau(t)$  (Nm), generated by the motor is proportional to the current,  $i_r(t)$  (A), in the rotor

$$\tau(t) = K_m i_r(t) \quad (4.39)$$

where  $K_m$  (Nm/A) is the motor torque constant and is related to the physical properties of the motor, including the number of turns and the strength of the magnetic field.

The back EMF,  $V_{emf}(t)$  (volts), is proportional to the motor speed,  $\omega(t)$  (rad/s)

$$V_{emf}(t) = K_e \omega(t) = K_e \frac{d\theta(t)}{dt} \quad (4.40)$$

where  $K_e$  (V/rpm) is the back EMF constant, also known as the speed constant.

The equations that relate the motor output to the input current are described equations for the electrical and for the mechanical sections of the circuit. A second-order differential equation describes the mechanical section

$$(J_r + J_L) \frac{d^2\theta(t)}{dt^2} + b \frac{d\theta(t)}{dt} = K_m i_r(t) \quad (4.41)$$

where  $J_r$  ( $\text{kgm}^2$ ) and  $J_L$  ( $\text{kgm}^2$ ) are the moments of inertia of the rotor and the load respectively, and  $b$  is the damping coefficient.

A first-order differential equation describes the electrical portion of the motor

$$L \frac{di_r(t)}{dt} + Ri_r(t) = V_i(t) - K_e \frac{d\theta(t)}{dt} \quad (4.42)$$

Alternatively, equations (4.41) and (4.42) can be written in terms of the speed of the motor

$$(J_r + J_L) \frac{d\omega(t)}{dt} + b\omega(t) = K_m i_r(t) \quad (4.43)$$

$$L \frac{di_r(t)}{dt} + Ri_r(t) = V_i(t) - K_e \omega(t) \quad (4.44)$$

### 4.5.5 Similarities of the Two Models

SPICE software used to analyze electrical circuits is both accurate and easily available; therefore, mechanical models are often converted to their electrical counterparts to make use of this software. For example, the electrical equivalent of the standard mass-spring-damper block diagram is the parallel capacitor-inductor-resistor circuit shown in Figure 4-10.

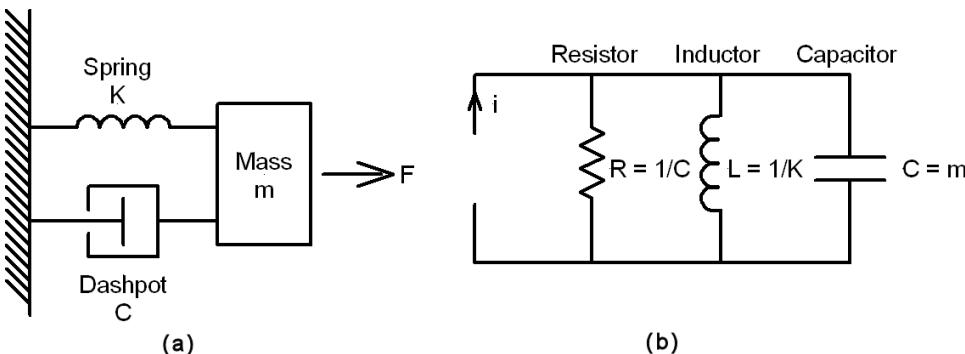
Table 4-3 lists the equations that describe the equations for the linear and torsional mechanical elements along with their electrical equivalents.

### 4.5.6 Fluid Flow Elements

Similar models can be derived for fluid flow components. However, in this instance there are two different formulations: one for incompressible hydraulics; and another for compressible pneumatics, where changes in pressure result in changes in density. This book examines only the incompressible hydraulic model.

The relationship between the volume rate of flow,  $q$  ( $\text{m}^3/\text{s}$ ), and the pressure difference,  $\Delta p = p_1 - p_2$ , is determined by the hydraulic resistance,  $R$ . This is the hydraulic equivalent to Ohm's law, where the hydraulic resistance is analogous to electrical resistance, the volume flow rate is equivalent to current, and the pressure difference is equivalent to the potential difference.

$$\Delta p = p_1 - p_2 = qR \quad (4.45)$$



**FIGURE 4-10** ■  
Equivalent 2nd  
order models  
(a) Mechanical.  
(b) Electrical.

**TABLE 4-3** ■ Comparison Between Electrical and Mechanical Elements

Block	Describing Equation	Energy/Power
Inductor	$i = \frac{1}{L} \int V dt$	$E_{ind} = \frac{1}{2} Li^2$
Spring	$F = kx = k \int v dt$	$E_{spring} = \frac{1}{2} kx^2$
Torsion spring	$T = k\theta = k \int \omega dt$	$E_{spring} = \frac{1}{2} k\theta^2$
Capacitor	$i = C \frac{dV}{dt}$	$E_{cap} = \frac{1}{2} CV^2$
Mass	$F = m \frac{dv}{dt} = m \frac{d^2x}{dt^2}$	$E_{mass} = \frac{1}{2} mv^2$
Moment of inertia	$T = I\alpha = I \frac{d\omega}{dt} = I \frac{d^2\theta}{dt^2}$	$E_I = \frac{1}{2} I\omega^2$
Resistor	$i = \frac{V}{R}$	$P_{res} = \frac{1}{R} V^2$
Dashpot	$F = cv = c \frac{dx}{dt}$	$P_{dash} = cv^2$
Rotational damper	$T = c\omega = c \frac{d\theta}{dt}$	$P_{damp} = c\omega^2$

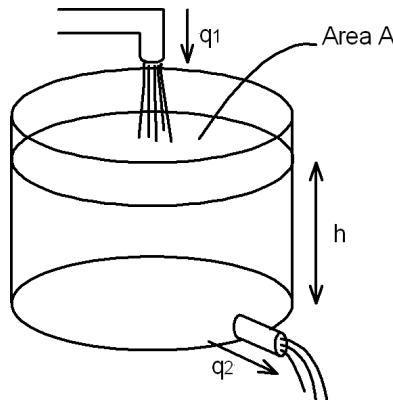
Hydraulic capacitance is the term that describes the potential energy storage by a fluid in terms of its pressure head. Consider the diagram shown in Figure 4-11, where the rate of change of volume,  $V$  ( $\text{m}^3$ ), in the reservoir is equal to the difference in the volume flow rates into and out of the reservoir.

$$q_1 - q_2 = \frac{dV}{dt} \quad (4.46)$$

However, if the volume is written in terms of the cross sectional area,  $A$  ( $\text{m}^2$ ), and the height of the fluid,  $h$  (m), in the reservoir, then

$$q_1 - q_2 = A \frac{dh}{dt} \quad (4.47)$$

**FIGURE 4-11** ■ Potential energy of fluid in a reservoir.



The pressure difference between the input and the output is a function of the density,  $\rho$  ( $\text{kg/m}^3$ ), and the acceleration due to gravity,  $g$  ( $\text{m/s}^2$ ),

$$P = \rho gh \quad (4.48)$$

Rewriting equation (4.47) in terms of the pressure gives

$$q_1 - q_2 = \frac{A}{\rho g} \frac{dP}{dt} \quad (4.49)$$

Where the hydraulic capacitance,  $C$ , is defined as

$$C = \frac{A}{\rho g} \quad (4.50)$$

Equation (4.49) can then be written as a differential equation in terms of the hydraulic capacitance

$$q_1 - q_2 = C \frac{dP}{dt} \quad (4.51)$$

or in terms of an integral as

$$P = \frac{1}{C} \int (q_1 - q_2) dt \quad (4.52)$$

Finally, hydraulic inertiance is equivalent to electrical inductance. Consider the cylindrical section of fluid shown in Figure 4-12. The difference between the two forces can be described in terms of the difference between the two pressures ( $P_1 - P_2$ ) across the cylinder

$$F_1 - F_2 = P_1 A - P_2 A = (P_1 - P_2) A \quad (4.53)$$

The net force,  $F$  (N), that causes the mass of fluid to accelerate at  $a$  ( $\text{m/s}^2$ ) is

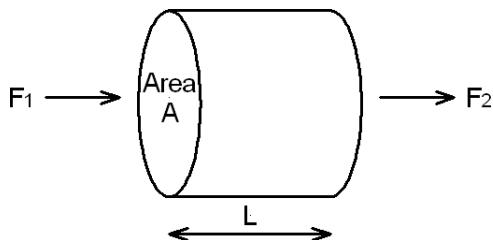
$$F = ma \quad (4.54)$$

Therefore,

$$(P_1 - P_2)A = ma \quad (4.55)$$

Equation (4.55) can be rewritten in terms of the rate of change of velocity

$$(P_1 - P_2)A = m \frac{dv}{dt} \quad (4.56)$$



**FIGURE 4-12** ■ Forces applied to a cylinder of fluid.

**TABLE 4-4** ■ Comparison Between Electrical and Hydraulic Elements

Block	Describing Equation	Energy/Power
Inductor	$i = \frac{1}{L} \int V dt$	$E_{ind} = \frac{1}{2} Li^2$
Hydraulic inertiance	$q = \frac{1}{I} \int (P_1 - P_2) dt$	$E_{HI} = \frac{1}{2} I q^2$
Capacitor	$i = C \frac{dV}{dt}$	$E_{cap} = \frac{1}{2} CV^2$
Hydraulic capacitance	$q = C \frac{d(P_1 - P_2)}{dt}$	$E_{mass} = \frac{1}{2} C(P_1 - P_2)^2$
Resistor	$i = \frac{V}{R}$	$P_{res} = \frac{1}{R} V^2$
Hydraulic resistance	$q = \frac{1}{R}(P_1 - P_2)$	$P_{HR} = \frac{1}{R}(P_1 - P_2)^2$

The mass of the fluid cylinder is equal to the product of the volume of the fluid and the density

$$m = \rho V = \rho A L \quad (4.57)$$

Therefore,

$$(P_1 - P_2)A = \rho A L \frac{dv}{dt} \quad (4.58)$$

But the volume flow rate is  $q = Av$ ; therefore,

$$(P_1 - P_2)A = \rho L \frac{dq}{dt} \quad (4.59)$$

The pressure difference can be written as

$$(P_1 - P_2) = \frac{\rho L}{A} \frac{dq}{dt} = I \frac{dq}{dt} \quad (4.60)$$

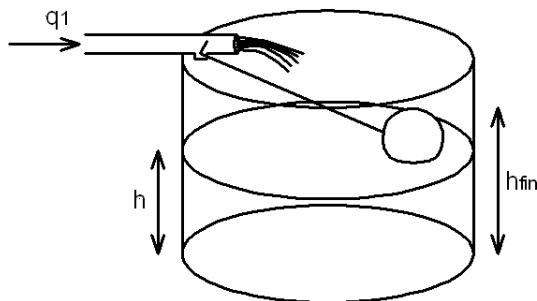
where the hydraulic inertiance,  $I$  is defined as

$$I = \frac{\rho L}{A} \quad (4.61)$$

As with the mechanical elements discussed earlier, hydraulic elements can be thought of as equivalent to their electrical counterparts, which allows SPICE models to be used to simulate fluid flow. The equivalence is shown in Table 4-4.

## 4.6 | SYSTEM RESPONSE

If the input to a system changes, the output will change in two different stages. The first will be a transient which will settle into the steady-state response. The solution to the differential equations discussed earlier is used to determine analytically what these responses will be.



**FIGURE 4-13 ■**  
Float controller reservoir for a heart-lung machine.

In a real system, the input may change in an arbitrary way, but to quantify its behavior responses are usually determined for impulses, steps, ramps, and sinusoidal inputs. If the system is linear and time invariant (LTI), the output will be the algebraic sum of any number of responses to any of these functions.

It is, in fact, possible to construct any of the inputs described from the sum of a sequence of impulses, and therefore the impulse response provides a common method to characterize a system in the time domain. However, the manipulation of sums of impulse responses is cumbersome, so another method of analysis, the Laplace transform, is easier.

Consider the simple example of a float-valve controller that could be used to maintain the level of blood in an open reservoir in a heart-lung machine, as shown in Figure 4-13. In this system, the rate at which blood enters the tank is dependent on the difference between the current blood depth,  $h$  (m), and the final depth,  $h_{fn}$  (m).

A simple first-order differential equation can be written

$$\frac{du}{dt} = k(h_{fn} - h) \quad (4.62)$$

The solution of this form of differential equation can be determined by inspection

$$h = h_{fn}(1 - e^{-kt}) \quad (4.63)$$

In the case of a second-order differential equation such as the mass–spring–damper system discussed earlier, the response is described by the natural frequency of the system and the damping ratio. Depending on the magnitudes of these parameters, the response to a step input could be an undamped sinusoid, a damped sinusoid, or an exponential, as shown in Figure 4-7.

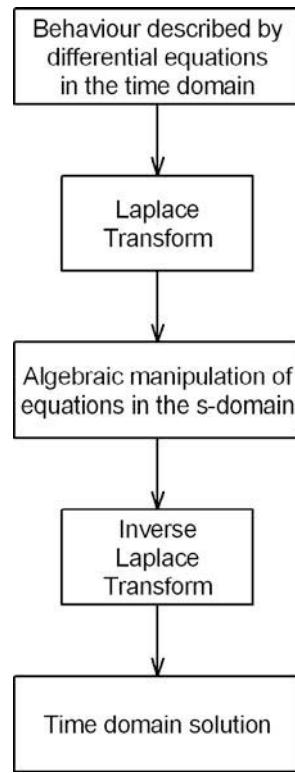
It is often more difficult to fit the response of a system to one of the standard solutions, particularly if the order is high and there are feedback components. This is another reason for describing systems in terms of their Laplace transforms, which make them far easier to manipulate and solve. The process is shown in Figure 4-14.

For a function,  $f(t)$ , in the time domain, the Laplace transform is

$$F(s) = \int_0^{\infty} f(t)e^{-st} dt \quad (4.64)$$

Under normal circumstances it is not necessary to evaluate this integral because tables have been compiled for the most commonly occurring functions, which, in conjunction with a few rules for handling combinations of transforms, allow for the solution of most problems encountered in biomechatronics and other disciplines. Some of the rules are listed in Table 4-5.

**FIGURE 4-14** ■ Process of using Laplace transforms to determine the time-domain responses of complex systems.



**TABLE 4-5** ■ Rules for Handling Laplace Transforms (Bolton 1992)

Time Domain	S (Laplace) Domain
$f_1(t) + f_2(t)$	$F_1(s) + F_2(s)$
$f_1(t) - f_2(t)$	$F_1(s) - F_2(s)$
$af(t)$	$aF(s)$
$f(t - T)$	$e^{-Ts} F(s)$
$\frac{d}{dt} f(t)$	$sF(s) - f(o)$
$\frac{d^2}{dt^2} f(t)$	$s^2 F(s) - sf(0) - \frac{df(0)}{dt}$
$\frac{d^n}{dt^n} f(t)$	$s^n F(s) - s^{n-1} f(0) \cdots - \frac{d^{n-1} f(0)}{dt^{n-1}}$
$\int_0^t f(t) dt$	$\frac{1}{s} F(s)$

To solve differential equations using the Laplace transform method, perform the following steps:

- Transform each term in the differential equation into its Laplace transform.
- Perform the algebraic manipulation for a specific input (e.g., step, impulse, sine wave).

**TABLE 4-6** ■ Laplace Transforms

S (Laplace) Domain $F(s)$	Time Domain $f(t), t \geq 0$
1	$\delta(t)$
$\frac{1}{s}$	$1(t)$
$\frac{1}{s^2}$	$t$
$\frac{2!}{s^3}$	$t^2$
$\frac{3!}{s^4}$	$t^3$
$\frac{m!}{s^{m+1}}$	$t^m$
$\frac{1}{s + a}$	$e^{-at}$
$\frac{1}{(s + a)^2}$	$te^{-at}$
$\frac{1}{(s + a)^m}$	$\frac{1}{(m - 1)!} t^{m-1} e^{-at}$
$\frac{a}{s(s + a)}$	$1 - e^{-at}$
$\frac{a}{s^2(s + a)}$	$\frac{1}{a}(at - 1 + e^{-at})$
$\frac{b - a}{(s + a)(s + b)}$	$e^{-at} - e^{-bt}$
$\frac{s}{(s + a)^2}$	$(1 - at)e^{-at}$
$\frac{a^2}{s(s + a)^2}$	$1 - (1 + at)e^{-at}$
$\frac{(b - a)s}{(s + a)(s + b)}$	$be^{-bt} - ae^{-at}$
$\frac{a}{s^2 + a^2}$	$\sin at$
$\frac{s}{s^2 + a^2}$	$\cos at$
$\frac{s + a}{(s + a)^2 + b^2}$	$e^{-at} \cos bt$
$\frac{b}{(s + a)^2 + b^2}$	$e^{-at} \sin bt$

- Convert the results into the standard forms using the partial fraction expansion process.
- Invert the results back into the time domain.

### WORKED EXAMPLE

---

#### Response of a RC Circuit

Consider the step change to the input voltage of a series RC circuit, and determine the voltage,  $V_C(t)$ , across the capacitor, as a function of time.

The differential equation can be written as

$$V = RC \frac{dV_C}{dt} + V_C \quad (4.65)$$

with  $V_C = 0$  at  $t = 0$ .

Take the Laplace transform of each of the terms

$$\frac{V}{s} = RC s V_C(s) + V_C(s) \quad (4.66)$$

Rewrite equation (4.66) with  $V_C(s)$  the subject of the formula

$$\begin{aligned} V_C(s)[1 + RCs] &= \frac{V}{s} \\ V_C(s) &= \frac{V}{s(1 + RCs)} \end{aligned} \quad (4.67)$$

Manipulate so that equation (4.67) conforms to one of the standard forms found in tables for which there is a solution. In this case the obvious choice is

$$\frac{a}{s(s + a)}$$

for which the time-domain solution is  $1 - e^{-at}$

$$\begin{aligned} V_C(s) &= \frac{V/RC}{s(1/RC + s)} \\ &= V \frac{1/RC}{s(s + 1/RC)} \end{aligned}$$

Therefore,  $a = 1/RC$

Taking the inverse Laplace transform

$$V_C(t) = V(1 - e^{-t/RC}) \quad (4.68)$$

In this case  $1/RC$  is the time constant for the time-domain response.

---

#### 4.6.1 Partial Fraction Expansion

To simplify more complex algebraic functions into a form that can be solved using tables, it is necessary to use the partial fraction expansion method. To enable this, the denominator

must be factorizable, and the power of the highest term in the numerator is at least one lower than that of the denominator.

Convert the expression  $F(s) = \frac{s+3}{s^2+3s+2}$  so that it can be written as a partial fraction expansion.

The denominator can be factorized into two roots,  $(s+1)(s+2)$ , so the expression can be rewritten in the following form

$$\begin{aligned} F(s) &= \frac{A}{s+1} + \frac{B}{s+2} \\ &= \frac{A(s+2) + B(s+1)}{(s+1)(s+2)} \end{aligned}$$

For the two expressions to be equal requires that the numerators be equal

$$A(s+2) + B(s+1) = s+3$$

This can be solved using simultaneous equations by selecting the value for  $s$  in a pair of equations in which  $A = 0$  in the one and  $B = 0$  in the other

$$\begin{aligned} s &= -2 \\ B(-2+1) &= -2+3 \\ -B &= 1 \\ B &= -1 \\ s &= -1 \\ A(-1+2) &= -1+3 \\ A &= 2 \end{aligned}$$

Therefore,

$$F(s) = \frac{2}{s+1} + \frac{-1}{s+2}$$

## 4.6.2 Analyzing Complex Models

The ability to manipulate time-domain equations in the Laplace domain makes it practical to generate and analyze complex models of biomechatronic devices.

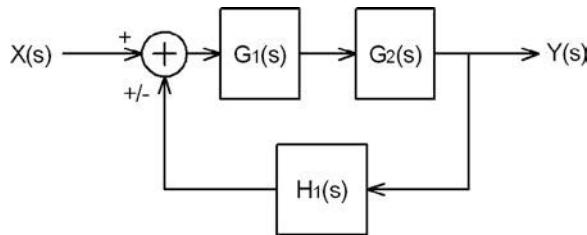
In a typical system with two feed-forward blocks,  $G_1(s)$  and  $G_2(s)$ , and one feedback block,  $H_1(s)$ , shown in Figure 4-15, the feed-forward blocks can be combined

$$G_3(s) = G_1(s)G_2(s) \quad (4.69)$$

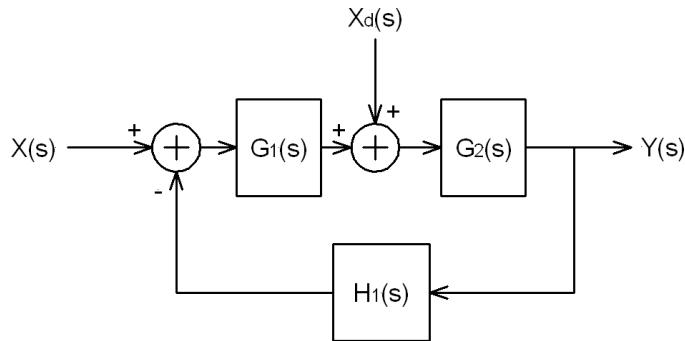
Second, the inclusion of the feedback element can be achieved using the relationship described in equation (4) for negative feedback

$$T(s) = \frac{G_3(s)}{1 + G_3(s)H_1(s)} \quad (4.70)$$

**FIGURE 4-15** ■  
System block diagram in the s domain.



**FIGURE 4-16** ■  
System block diagram with a disturbance.



If the feedback is positive, then the sign changes

$$T(s) = \frac{G_3(s)}{1 - G_3(s)H_1(s)} \quad (4.71)$$

If the controller has multiple inputs, it is best to deal with them individually by setting all of the other inputs to zero and performing the analysis for each in turn. Because the systems are linear and time invariant, the output is obtained by performing the algebraic sum of all of the results. This method is often used to determine the effects of disturbance inputs.

Consider the previous example with the addition of a disturbance between the two feed-forward blocks,  $G_1(s)$  and  $G_2(s)$ , as shown in Figure 4-16.

Start by setting  $X_d(s) = 0$  to obtain the system transfer function

$$\frac{Y(s)}{X(s)} = \frac{G_1(s)G_2(s)}{1 - G_1(s)G_2(s)H_1(s)} \quad (4.72)$$

Now rearrange the block diagram with  $X(s) = 0$ , so that  $X_d(s)$  is the input as shown in Figure 4-17

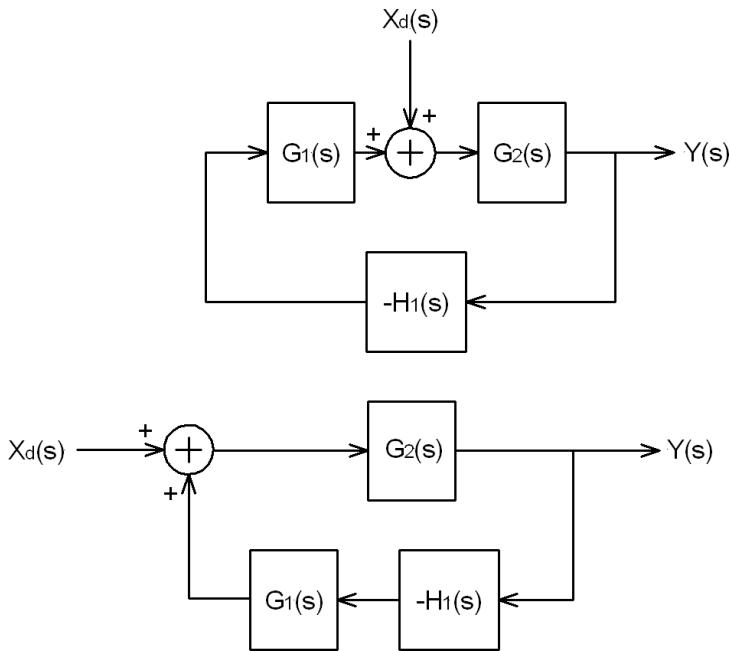
By inspection, the new transfer function can be written as

$$\frac{Y(s)}{X_d(s)} = \frac{G_2(s)}{1 - G_1(s)G_2(s)H_1(s)} \quad (4.73)$$

Combining equations (4.72) and (4.73) to produce the output of the system in terms of the two inputs

$$Y(s) = \frac{G_1(s)G_2(s)X(s) + G_2(s)X_d(s)}{1 - G_1(s)G_2(s)H_1(s)} \quad (4.74)$$

This equation can then be used to determine the relative effect of the disturbance, or interference signal, compared with the effect of the input signal.



**FIGURE 4-17** ■  
System block diagram for the disturbance input.

## 4.7 | SYSTEM STABILITY

A system is considered to be stable if every bounded input produces a bounded output. An alternative definition is that a system is stable if the impulse response dies away to zero.

A transfer function,  $G(s)$ , can be written as the ratio of two polynomials that can be factorized into their respective roots, where the roots can be real or can occur in complex pairs,

$$G(s) = \frac{K(s + z_1)(s + z_2) \cdots (s + z_m)}{(s + p_1)(s + p_2) \cdots (s + p_n)} \quad (4.75)$$

The roots of the numerator  $z_1, z_2, \dots, z_m$ , are called zeros, and the roots of the denominator  $p_1, p_2, \dots, p_n$ , are called poles.

The zeros are the values of  $s$  for which the transfer function becomes zero, while the poles are the values of  $s$  for which the transfer function becomes infinite.

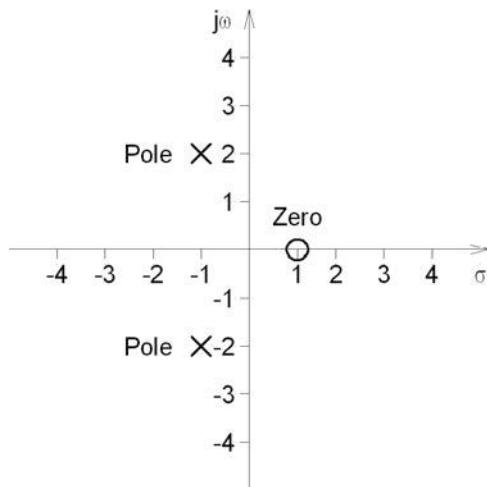
Consider a simple transfer function

$$G(s) = \frac{s - 1}{s^2 + 4s + 5} \quad (4.76)$$

The denominator can be factorized to produce a complex pole pair and plotted on the  $s$ -plane in Figure 4-18.

$$G(s) = \frac{s - 1}{(s - [1 + j2])(s - [1 - j2])} \quad (4.77)$$

**FIGURE 4-18 ■**  
S-plane plot  
showing the poles  
and zeros.



The stability can be determined by considering how the output changes with time after an impulse. For example, consider a first-order system with a pole at  $-3$

$$G(s) = \frac{1}{s + 3}$$

The output and input are related by

$$Y(s) = X(s)G(s)$$

For a unit impulse,  $X(s) = 1$ ; therefore,

$$Y(s) = G(s) = \frac{1}{s + 3}$$

Taking the inverse Laplace transform gives

$$y(t) = e^{-3t}$$

which decreases to zero as time increases.

Now consider a similar system with the pole at  $+3$

$$G(s) = \frac{1}{s - 3}$$

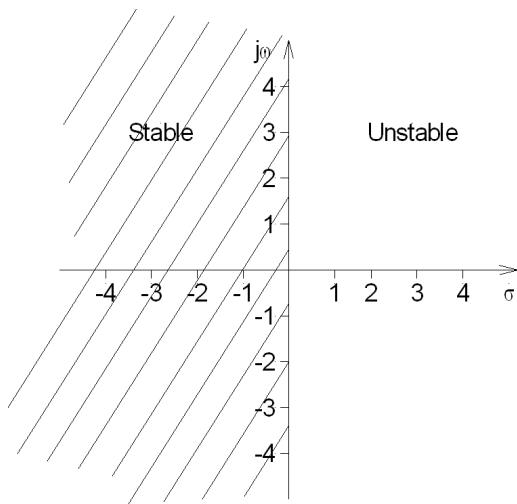
Repeating the previous analysis results in the following response:

$$y(t) = e^{+3t}$$

This increases with time; therefore, the system is unstable.

In general terms, the output of a system excited by an impulse will be the sum of a number of exponential terms. If any of those exponentials increases with time, the system is unstable.

In terms of the poles, this translates into the following, as illustrated in Figure 4-19: If all of the poles are in the left-hand half of the plane (shown shaded), then the system is stable. If any lie on the y-axis, the system is said to be critically stable, and if any are in the right-hand half of the plane the system will be unstable.



**FIGURE 4-19 ■**  
S-plane showing the criterion for system stability.

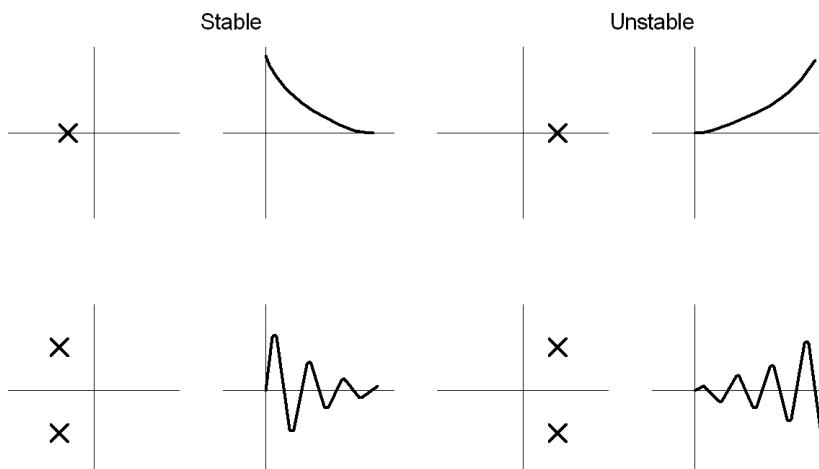
If it is difficult to factorize the denominator, then it is possible to use the Routh-Hurwitz criteria to determine the system stability (Raven, 1978). For the denominator written as a polynomial

$$a_n s^n + a_{n-1} s^{n-1} + \cdots + a_1 s + a_0 \quad (4.78)$$

- If any of the coefficients are negative, then the system will be unstable.
- If any coefficient is missing then the system is, at best, critically stable.

Other criteria exist for stability, and it is suggested that the reader investigate these further if they are of concern.

As illustrated in Figure 4-20, the actual time-domain response that occurs for an impulse is a function of the position of the poles and whether they are real or complex.



**FIGURE 4-20 ■**  
Relationship between pole position and system response.

### 4.7.1 Root Locus

The stability of a system based on the positions of the poles as the controller gains are changed is known as the root locus method. It is one way to design controllers and to determine the limits in the gains that can be applied by a controller before instability results.

The simplest controller that can be implemented is one that includes a gain,  $K$ , in the forward path, as shown in Figure 4-21.

The system transfer function is

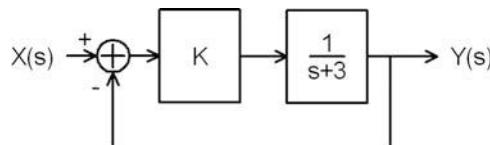
$$\begin{aligned} T(s) &= \frac{G(s)}{1 + G(s)} \\ &= \frac{K/(s+3)}{1 + K/(s+3)} \\ &= \frac{K}{s+3+K} \end{aligned}$$

The characteristic equation is  $s + 3 + K = 0$ , so the open-loop pole occurs when  $K = 0$ , and this will be at  $s = -3$ . However, when  $K > 0$ , then the pole will occur at  $s = -(3+K)$ , so as  $K$  increases, the pole will become more negative, as shown in Figure 4-22.

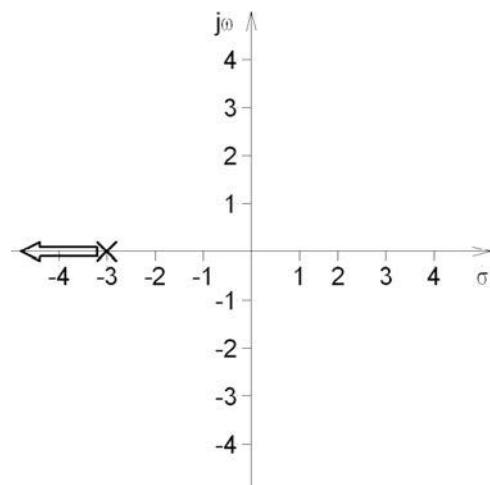
The system response can be solved for an impulse or a step response, and it will be found that the speed of response increases as  $K$  increases.

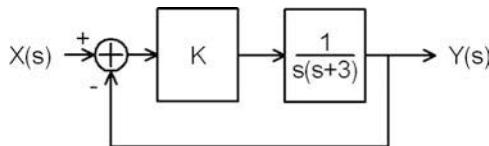
In a second-order system, the response is completely different. Consider the proportional controller shown in Figure 4-23, also with gain  $K$ .

**FIGURE 4-21** ■  
Proportional controller for a first-order system with unity feedback.



**FIGURE 4-22** ■  
Root locus plot for a simple controller.





**FIGURE 4-23** ■ Proportional controller for a second-order system with unity feedback.

In this case the system transfer function is

$$\begin{aligned} T(s) &= \frac{K/s(s+3)}{1+K/s(s+3)} \\ &= \frac{K}{s^2 + 3s + K} \end{aligned}$$

The characteristic equation is now a quadratic,  $s^2 + 3s + K = 0$ , and the roots are

$$\begin{aligned} s &= \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \\ &= -\frac{3}{2} \pm \frac{1}{2}\sqrt{9 - 4K} \end{aligned}$$

When  $K = 0$ , the roots are at

$$\begin{aligned} s &= -\frac{3}{2} \pm \frac{3}{2} \\ &= 0 \text{ or } 3 \end{aligned}$$

This is as expected.

As  $K$  increases from 0 to  $9/4$ , the root at 0 becomes more negative, moving toward  $-3/2$ , whereas the root at  $-3$  becomes more positive, also moving toward  $-3/2$ .

As  $K$  continues to increase from  $9/4$ , the roots become complex and split. By the time  $K = 3$ , the roots are

$$s = -\frac{3}{2} - j\sqrt{\frac{3}{4}}$$

and

$$s = -\frac{3}{2} + j\sqrt{\frac{3}{4}}$$

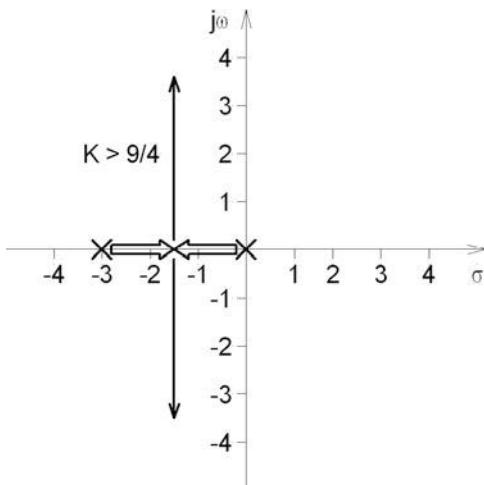
The root locus plot can easily be plotted, as shown in Figure 4-24.

In this case, the response moves from being overdamped to critical damping to being underdamped for a step input, as shown in Figure 4-25.

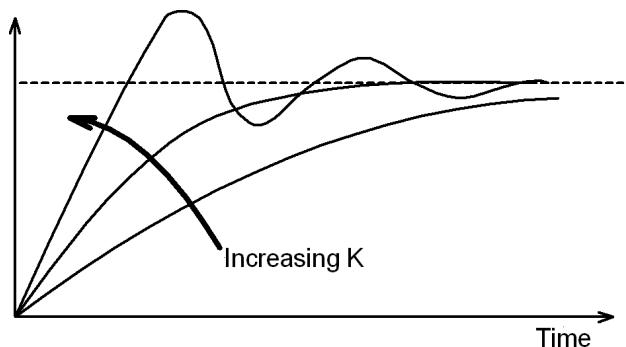
As systems become more complicated, it becomes less convenient to plot root loci manually. Fortunately, MATLAB includes scripts that automate the process, and the previous example is shown in Figure 4-26.

Note that the open-loop transfer function is represented by the coefficients of the polynomials describing the numerator and the denominator. The root locus command assumes that the system includes unity gain feedback with a proportional controller gain

**FIGURE 4-24** ■ Root locus plot for the quadratic system with proportional controller.



**FIGURE 4-25** ■ Time-domain response for quadratic system with proportional controller.



implicit in the feed-forward path.

```
%rl01.m
% simple root locus
% G = 1/s(s+3)
num = [1];
den = [1 3 0];
rlocus(num,den);
```

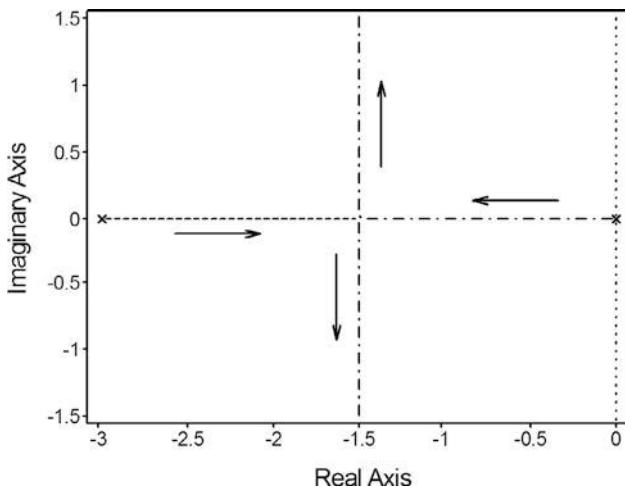
Consider a more complex function with an open-loop transfer function

$$G_o(s) = \frac{K}{(s + 1)(s + 2)(s + 3)}$$

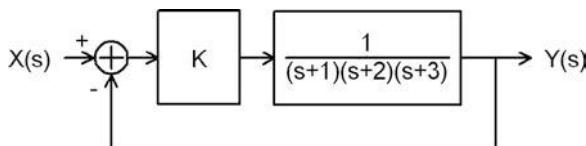
The block diagram is shown in Figure 4-27.

The numerator and denominator are expanded into their polynomial forms for representation by MATLAB (Figure 4-28).

$$G_o(s) = \frac{K}{s^3 + 6s^2 + 11s + 6}$$



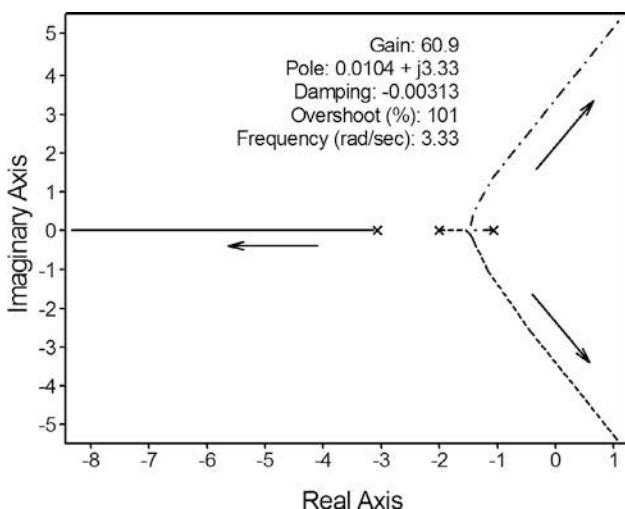
**FIGURE 4-26** ■ MATLAB version of root locus plot for quadratic system with proportional controller.



**FIGURE 4-27** ■ Proportional controller for a third-order system with unity gain feedback.

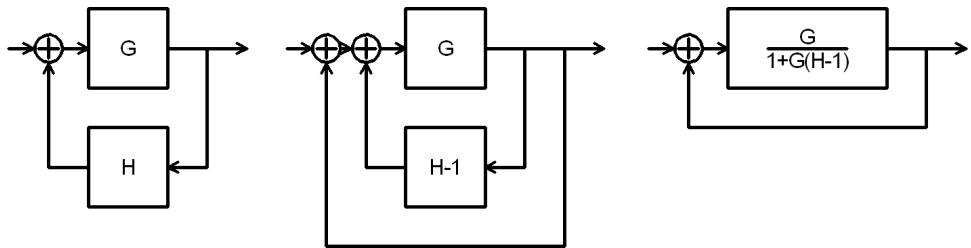
```
%rl02.m
% simple root locus
% G = 1/(s+1)(s+2)(s+3)
num = [1];
den = [1 6 11 6];
rlocus(num,den);
```

In this case it can be seen that two of the roots become complex and then move towards the right where, for  $K > 60$  they cross into the right-hand half-plane and the system will become unstable.



**FIGURE 4-28** ■ MATLAB version of root locus plot for third-order system with proportional controller.

**FIGURE 4-29** ■  
Conversion to unity gain feedback.



### 4.7.2 Steady-State Error

The steady-state error is the difference between the commanded output and the actual output,

$$E(s) = X(s) - Y(s) \quad (4.79)$$

The error depends not only on the system but also on the form of the input  $X(s)$ . To determine the magnitude of this error, it is necessary to convert a system with feedback to one with unity gain feedback, as shown in Figure 4-29.

The forward path transfer function for the unity gain feedback system

$$G_o(s) = \frac{1}{1 + G(H + 1)} \quad (4.80)$$

is then used in the following limit:

$$e_{ss} = \lim_{s \rightarrow 0} \left[ s \frac{1}{1 + G_o(s)} X(s) \right] \quad (4.81)$$

These limits are considered further in the section on system control.

## 4.8 | CONTROLLERS

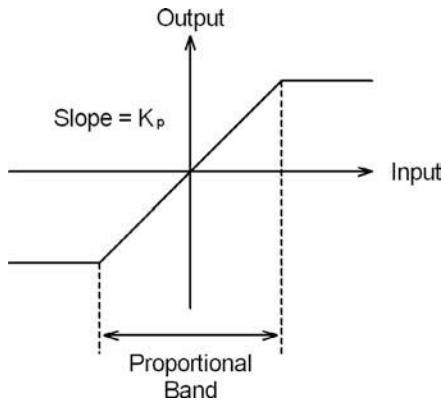
### 4.8.1 Proportional Controller

The controllers discussed in the previous sections comprise a simple gain stage, which results in the output being directly proportional to the input, with  $K_a$  being the proportional gain. In reality these controllers are proportional only over a limited band until the output saturates, as illustrated in Figure 4-30.

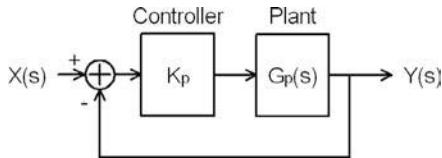
In general terms, for a unity gain feedback system and a plant with a transfer function  $G_P(s)$ , the configuration is shown in Figure 4-31.

$$T(s) = \frac{K_p G_p(s)}{1 + K_p G_p(s)} \quad (4.82)$$

A number of factors determine how well the controller works: (1) the ultimate stability of the system; (2) the steady-state error for a specific input; and (3) the response time until the error is within specific bounds. One other consideration is the ability of the controller to reject disturbances within the loop. Methods to determine all of these are discussed earlier in this chapter.



**FIGURE 4-30** ■  
Proportional controller with saturation.



**FIGURE 4-31** ■  
Classic configuration of a proportional controller.

### WORKED EXAMPLE

#### Motor Speed and Position Control

Consider the direct current (DC) motor model developed earlier and described by equations (4.43) and (4.44) and reproduced here:

$$(J_r + J_L) \frac{d\omega(t)}{dt} + b\omega(t) = K_m i_r(t) \quad (4.43)$$

$$L \frac{di_r(t)}{dt} + R i_r(t) = V_i(t) - K_e \omega(t) \quad (4.44)$$

To develop a controller, it is convenient to convert these to their Laplace transform form with the moments of inertia lumped together so  $J = J_r + J_L$ :

$$J s \Omega(s) + b \Omega(s) = K_m I_r(s) \quad (4.83)$$

$$L s I_r(s) + R I_r(s) + K_e \Omega(s) = V_i(s) \quad (4.84)$$

From equation (4.83), the current can be written as

$$I_r(s) = \frac{\Omega(s)}{K_m} (J s + b) \quad (4.85)$$

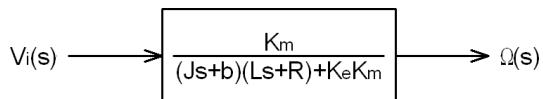
Substituting (4.85) into (4.84)

$$\frac{\Omega(s)}{K_m} (J s + b) (L s + R) + K_e \Omega(s) = V_i(s)$$

Simplifying,

$$\frac{\Omega(s)}{K_m} [(J s + b) (L s + R) + K_e K_m] = V_i(s)$$

**FIGURE 4-32** ■ DC motor block diagram.



The transfer function that relates the input voltage to the output speed, as shown in Figure 4-32, can be written as

$$T(s) = \frac{\Omega(s)}{V_i(s)} = \frac{K_m}{(Js + b)(Ls + R) + K_e K_m} \quad (4.86)$$

An alternative is to consider the two governing equations separately, with the back EMF applied as feedback.

The current drawn by the motor can be obtained by rewriting equation (4.84)

$$I_r(s)(Ls + R) = V_i(s) - K_e \Omega(s)$$

$$I_r(s) = \frac{1}{Ls + R}(V_i(s) - K_e \Omega(s)) \quad (4.87)$$

The equation for the mechanical portion can be obtained by rewriting equation (4.85) in terms of the input current:

$$\begin{aligned} \Omega(s)(Js + b) &= K_m I_r(s) \\ \Omega(s) &= \frac{K_m}{Js + b} I_r(s) \end{aligned} \quad (4.88)$$

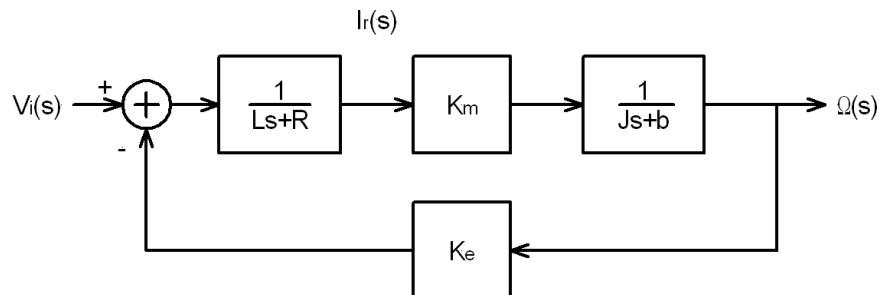
A block diagram can be constructed that is equivalent to that shown in Figure 4-32 but includes the two individual transfer functions as well the back EMF feedback. This is shown in Figure 4-33.

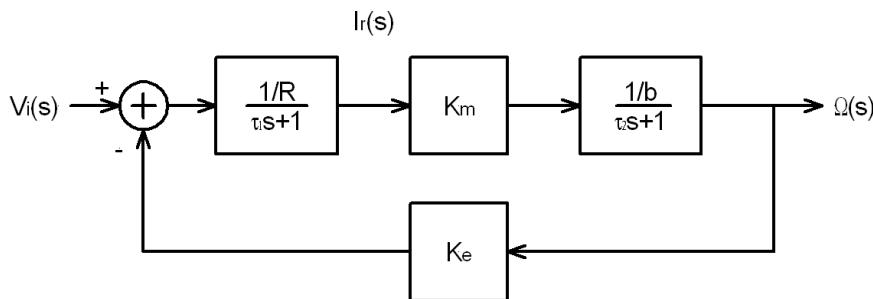
The advantage of using this form to describe the open-loop motor model is that each of the first-order systems can be rewritten in a form that is easy to interpret in the time domain. In addition, it gives access to the torque node (just after the  $K_m$  block), which then allows for the introduction of an external torque requirement or disturbance.

From the table of Laplace transforms (Table 4-6), it would be convenient to write the first-order blocks in the form

$$F(s) = \frac{1}{\tau s + 1}$$

**FIGURE 4-33** ■ Alternative form for DC motor block diagram.





**FIGURE 4-34** ■ DC motor block diagram in time-constant form.

which has the tabled solution  $f(t) = e^{-t/\tau}$ . Therefore, the electrical model of the motor can be modified to conform as follows:

$$\frac{1}{Ls + R} = \frac{1/R}{\frac{L}{R}s + 1} = \frac{1/R}{\tau_1 + 1}$$

so  $\tau_1 = L/R$  is the electrical time constant of the motor.

Similarly,  $\tau_2 = J/b$  is the mechanical time constant for the motor and load.

The block diagram can be redrawn with the time constants for the electrical and mechanical portions of the motor as shown in Figure 4-34.

Consider the Maxon motor selected to drive the turbine of a continuous positive airway pressure (CPAP) air pump discussed in Chapter 3, whose specifications are listed in Table 4-7.

A number of factors must be determined or scaled so that the model shown in Figure 4-34 can be used. First, the electrical time constant  $\tau_1 = L/R = 40.9 \mu s$ . The mechanical time constant is given as  $\tau_2 = 5.28 \text{ ms}$ ; therefore, the damping,  $b$ , is

$$\begin{aligned} b &= \frac{J_r}{\tau_2} \\ &= \frac{1.22 \times 10^{-7}}{5.28 \times 10^{-3}} \\ &= 2.31 \times 10^{-5} \end{aligned}$$

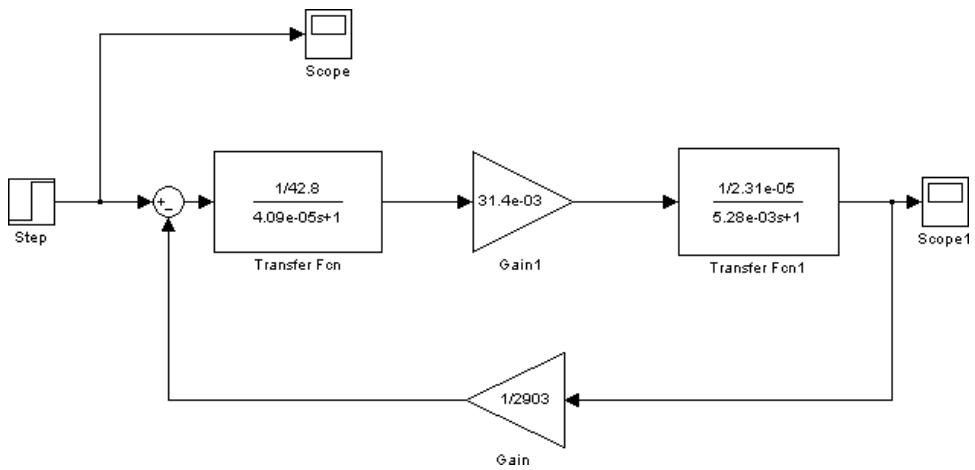
Note that  $J_r$  has been converted from  $\text{g.cm}^2$  to  $\text{kg.m}^2$  for this calculation.

$K_m$  is converted to  $31.4 \times 10^{-3} \text{ Nm/A}$ , and  $K_e$  is converted to  $2903 \text{ rad/s per V}$ . It is also important to remember that in the feedback path the conversion is from speed to voltage; therefore, the reciprocal of the constant is used in the model.

**TABLE 4-7** ■ Maxon RE 16 Motor Specifications

No-load speed @ 24 V $n_o$ (rpm)	7250
No-load current $I_o$ (mA)	3.11
Terminal resistance $R$ ( $\Omega$ )	42.8
Terminal inductance $L$ (mH)	1.75
Torque constant $K_m$ (mNm/A)	31.4
Speed constant $K_e$ (rpm/V)	304
Output power $P_o$ (W)	3.2
Mechanical time constant $\tau_2$ (ms)	5.28
Rotor inertia $J_r$ ( $\text{gcm}^2$ )	1.22

**FIGURE 4-35** ■  
Simulink model  
of the Maxon RE  
16 motor.

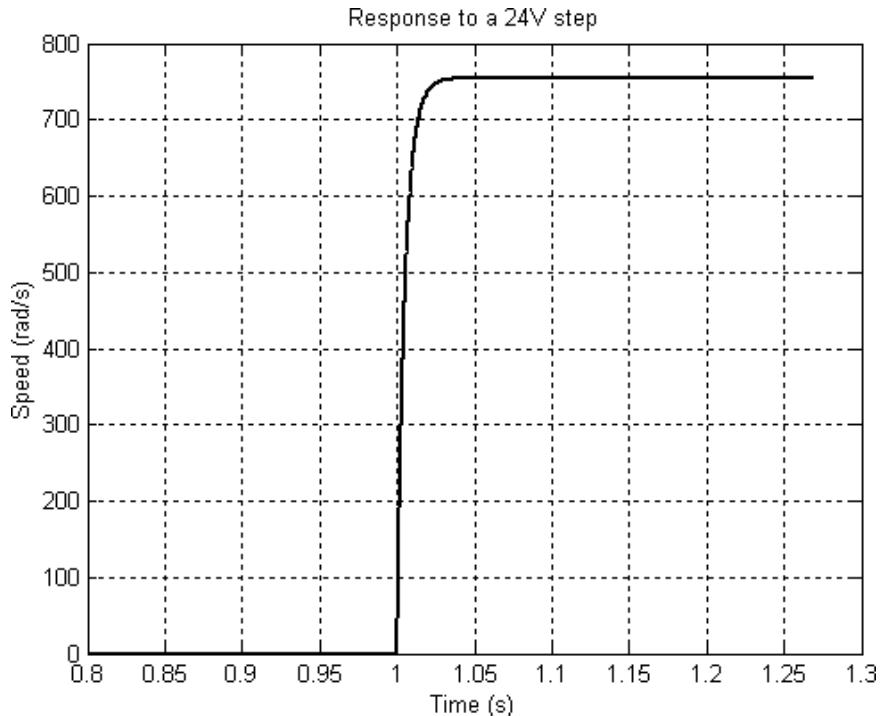


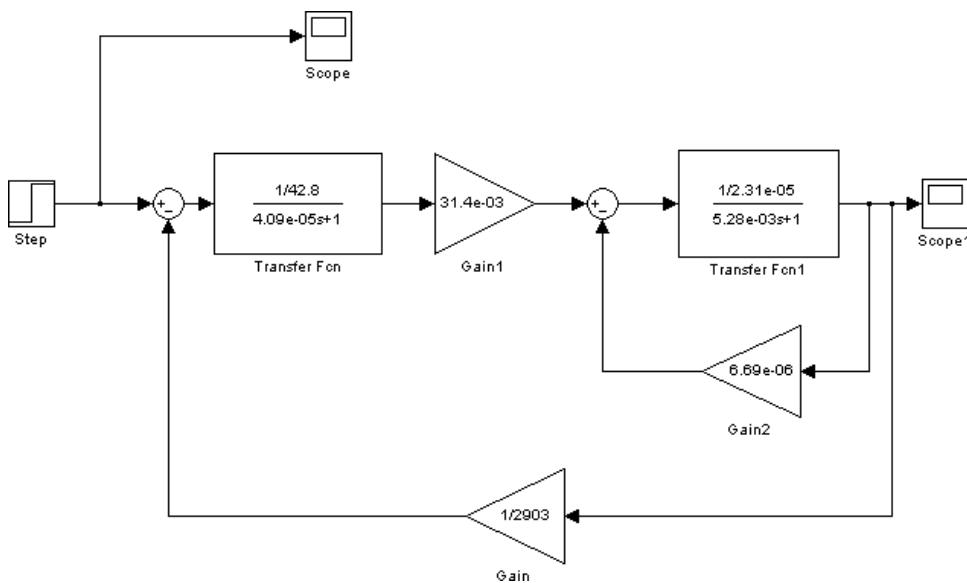
The Simulink model for the motor is shown in Figure 4-35.

The step response shown in Figure 4-36 is determined by a combination of the electrical and mechanical time constants of the motor. The mechanical time constant is by far the slower and therefore has the dominant effect on the settling time. The open-loop speed is 754 rad/s (7200 rpm), which is close to the expected no-load speed of 7250 rpm, as expected.

The nominal load for the CPAP turbine is 3.5 mNm at 5000 rpm. At a first approximation, the flow rate is directly proportional to the turbine speed, and therefore the load

**FIGURE 4-36** ■  
Open-loop step  
response of the  
motor for a 24 V  
step.





**FIGURE 4-37** ■  
Simulink model of  
the Maxon RE 16  
motor driving a load.

torque will be directly proportional to the rotation speed.

$$T_L(s) = K_L \Omega(s) \quad (4.89)$$

The constant of proportionality,  $K_L$  (Nm per rad/s), is

$$K_L = \frac{3.5 \times 10^{-3}}{5000 \times 2\pi/60} = 6.69 \times 10^{-5}$$

This load torque is introduced into the loop as shown in Figure 4-37.

The step response in this case, shown in Figure 4-38, exhibits a similar time constant, but because of the linear relationship between the torque and speed discussed in Chapter 3 the speed is reduced proportionally.

Obviously, the inertia of the turbine (load) must also be accommodated in the model, and this is achieved by adding  $J_L$  to the rotor inertia. It may also alter the mechanical time constant of the system.

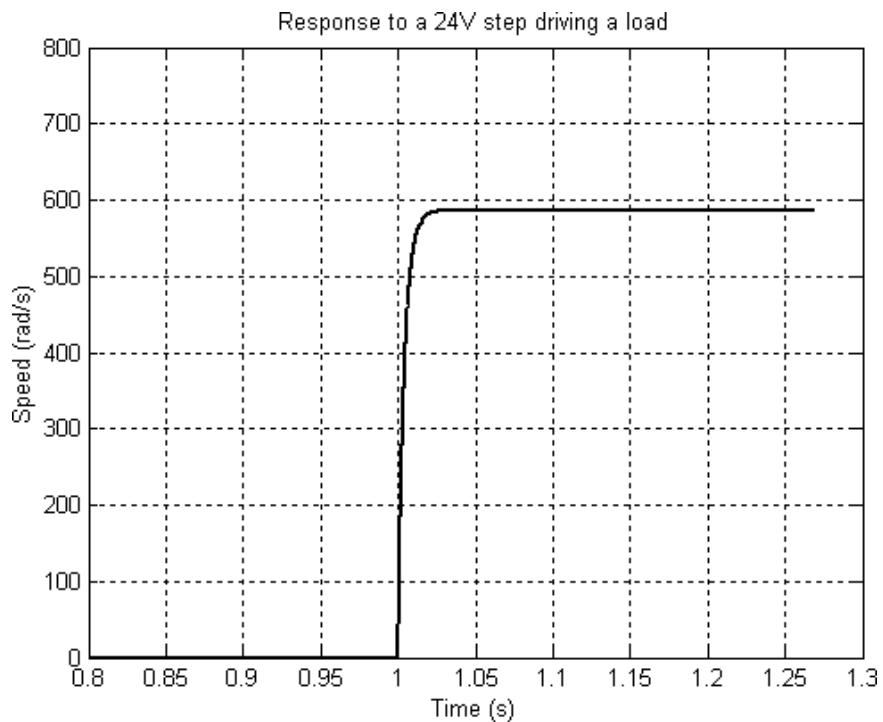
Assuming that the turbine inertia  $J_L = 100 \text{ gcm}^2$  (about 100 times that of the motor) and that the damping,  $b$ , remains unchanged, then the new mechanical time constant is

$$\begin{aligned} \tau_2 &= \frac{(J_r + J_L)}{b} \\ &= \frac{(1.22 \times 10^{-7} + 100 \times 10^{-7})}{2.31 \times 10^{-5}} \\ &= 0.44 \text{ s} \end{aligned}$$

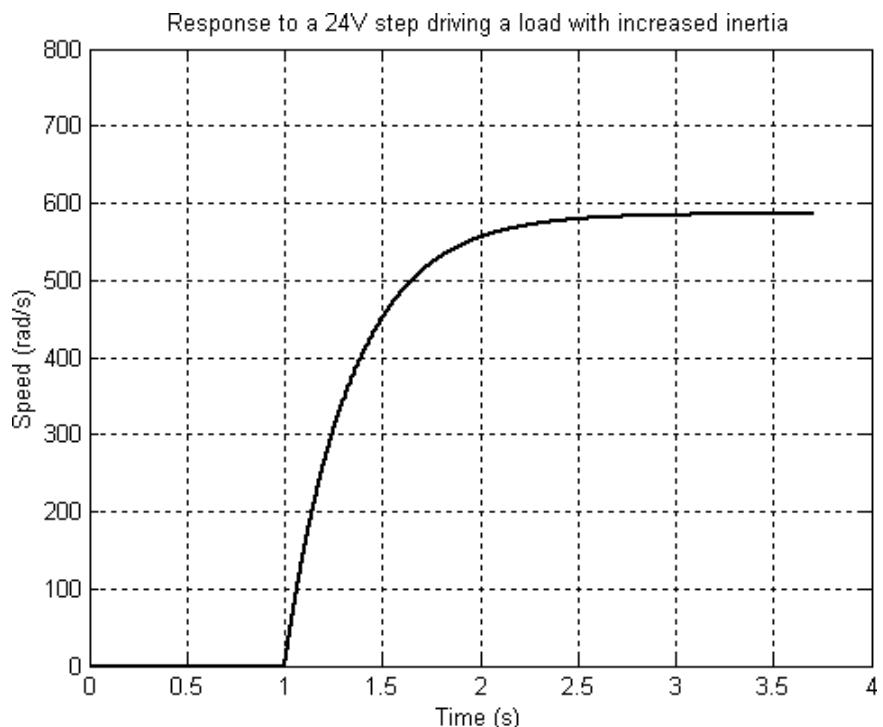
The result of these changes is that the final speed remains the same as before, but the motor takes much longer to reach steady state, as shown in Figure 4-39.

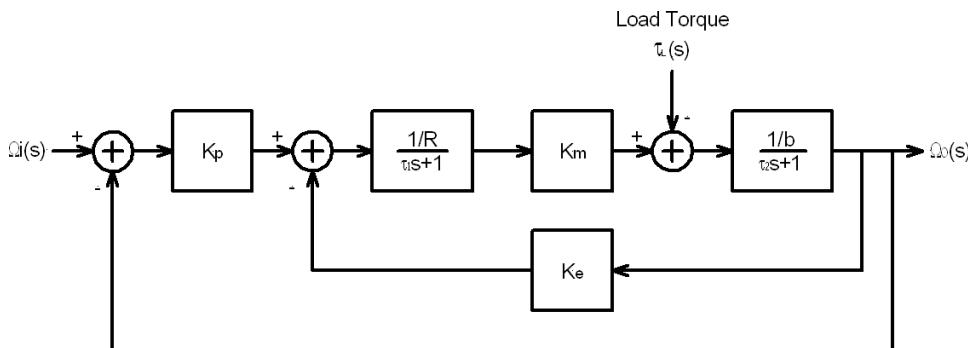
In Chapter 3 it was shown that motor speed decreases with increasing load torque. Therefore, if a speed controller is required for a motor, an additional loop and a controller have to be implemented. A proportional controller to perform this function is shown in Figure 4-40.

**FIGURE 4-38** ■  
Open-loop step  
response for the  
motor driving a load.

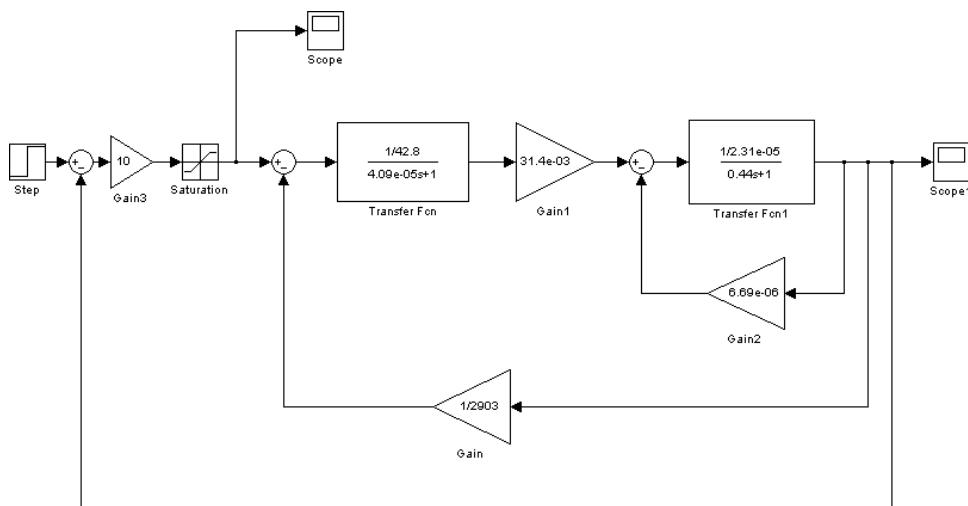


**FIGURE 4-39** ■  
Open-loop step  
response for the  
motor driving a load  
with 100 times the  
inertia.





**FIGURE 4-40** ■  
Proportional controller of motor speed.



**FIGURE 4-41** ■  
Simulink model of a proportional speed controller for a motor driving a load.

It is important to remember that the output of the controller will be limited to the maximum allowable voltage for the motor, so the complete Simulink model is shown in Figure 4-41.

The step response to a commanded speed of 600 rad/s in Figure 4-42 shows a residual speed error of 15 rad/s because the commanded speed is beyond the capability of the motor given the allowable voltage and the load torque. However, if the commanded speed is reduced to 500 rad/s, the controller performs perfectly.

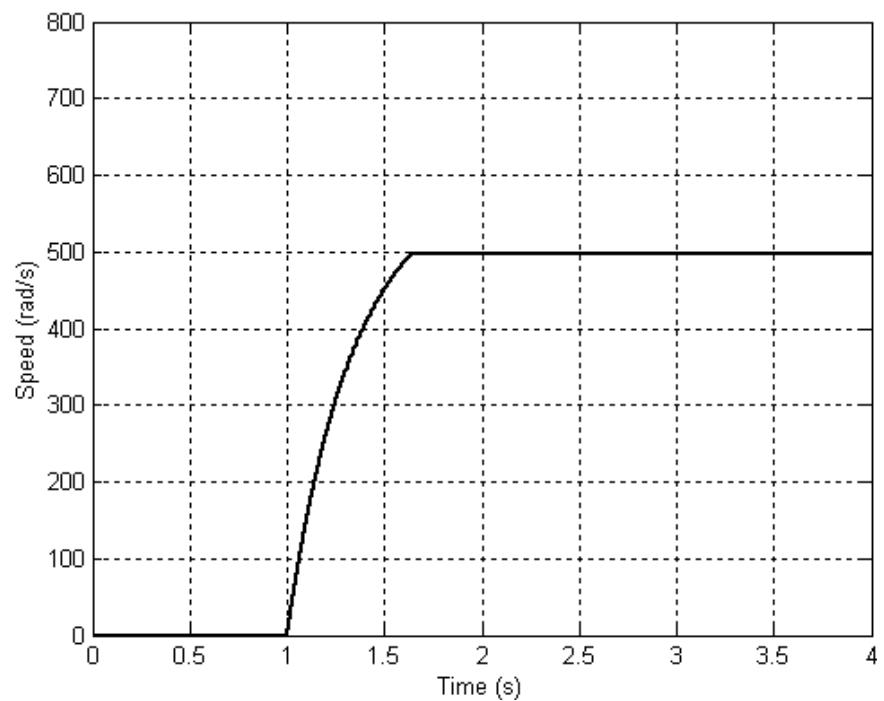
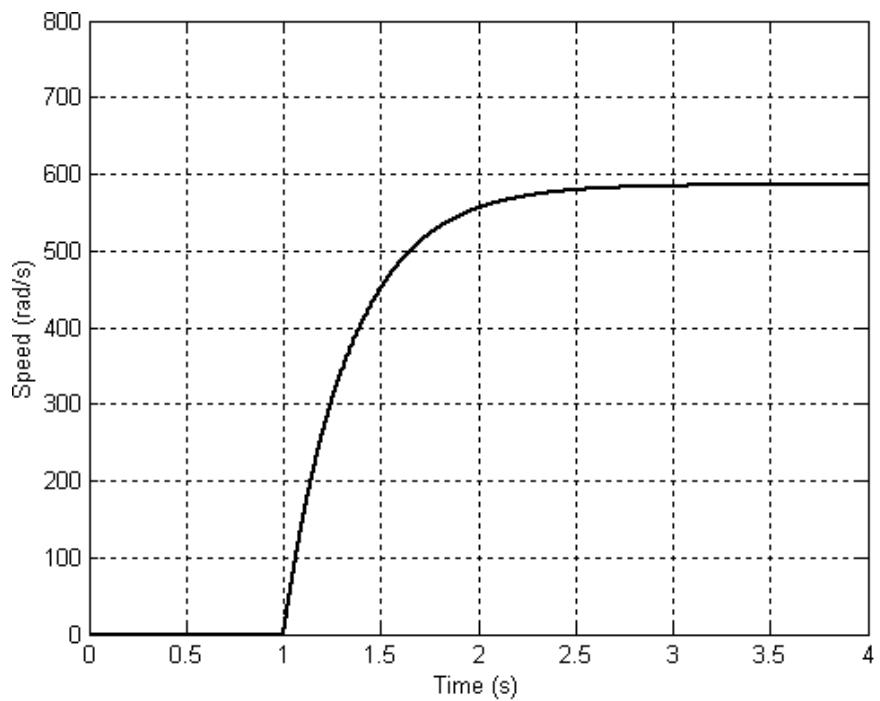
In many biomechatronic applications, it is not the motor speed but the angle that needs to be controlled. This requires that an additional block be incorporated into the output of the system to integrate the angular rate and to give angular position.

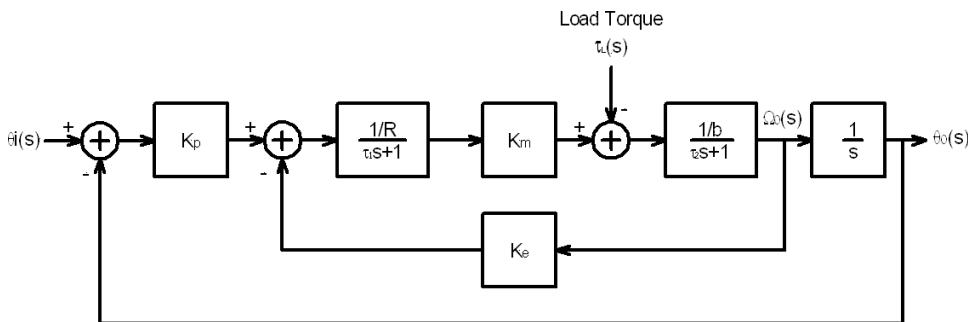
$$\theta(s) = \frac{1}{s} \Omega(s) \quad (4.90)$$

Therefore, the closed-loop position controller is as shown in Figure 4-43 and the Simulink model in Figure 4-44.

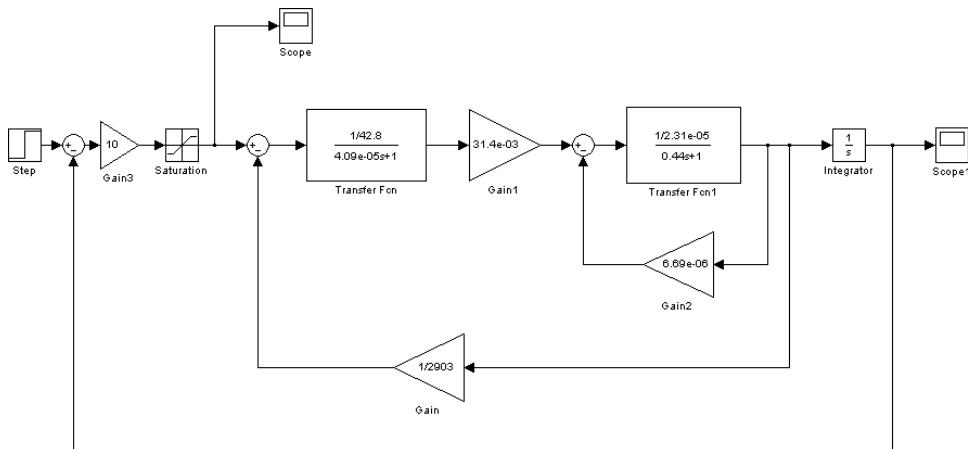
In this case, the proportional controller is highly underdamped, as can be seen in Figure 4-45. Increasing the gain doesn't improve the response time at all, but it does increase the natural frequency of the system with the result that the oscillation frequency increases.

**FIGURE 4-42** ■  
Response to a  
commanded speed  
of 600 rad/s and  
500 rad/s.

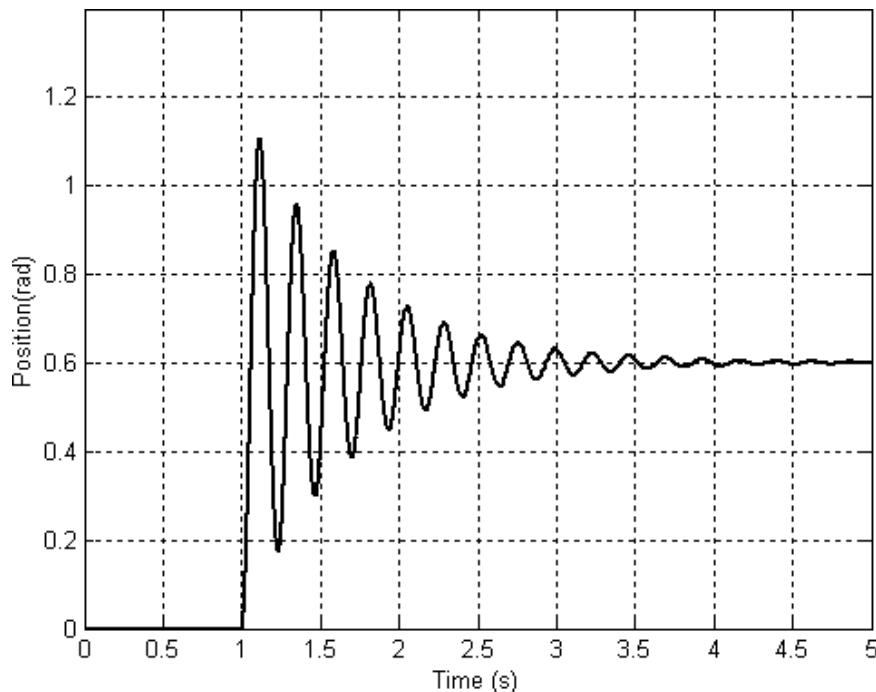




**FIGURE 4-43** ■  
Proportional controller for motor angular position.

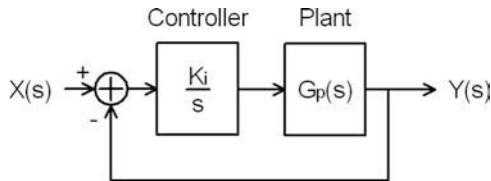


**FIGURE 4-44** ■  
Simulink model of a proportional position controller for a motor driving a load.



**FIGURE 4-45** ■  
Response to a commanded position of 0.6 rad.

**FIGURE 4-46** ■  
Classic  
configuration of an  
integral controller.



In general, a gearbox is incorporated at the output of the motor to better match the motor torque to the load. This can be represented by a gain that is representative of the gear ratio.

#### 4.8.1.1 Disadvantages of Proportional Controllers

The main disadvantage of this form of controller is that the transfer function of the plant with the controller is still the same order, so if the plant exhibits a steady-state error the controller does not eliminate it completely.

#### 4.8.2 Integral Controller

With integral control, the output of the controller is equal to the integral of the error. This is shown in a unity gain configuration in Figure 4-46.

The forward path transfer function for this control system is

$$G_o(s) = \frac{K_i}{s} G_p(s) \quad (4.91)$$

It can be seen that the controller introduces an additional  $s$  into the denominator, with the result that the order of the system increases. A system that exhibited a steady-state error for a step input will now converge with no error. The main disadvantage of the system is that it introduces a pole at the origin ( $s = 0$ ) and that can result in a less stable system.

If an integral controller is used to control the motor position, the system becomes unstable for any gain, as is shown in the step response in Figure 4-47.

#### 4.8.3 Proportional Plus Integral Controller

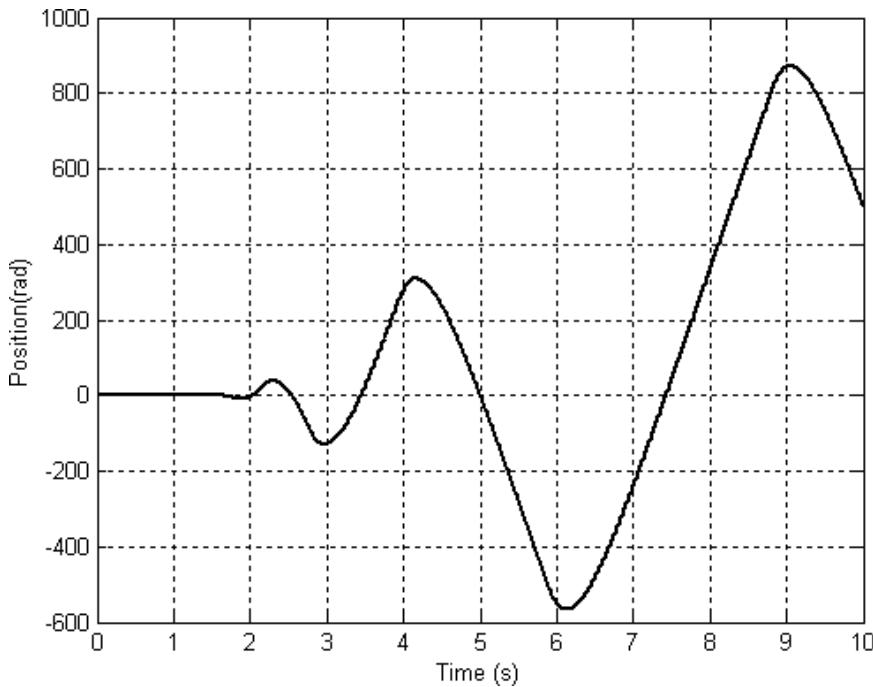
The reduction of the relative controllability of a plant as a result of using integral control can sometimes be overcome by introducing a proportional term to make the control function

$$G_c(s) = K_p + \frac{K_i}{s} \quad (4.92)$$

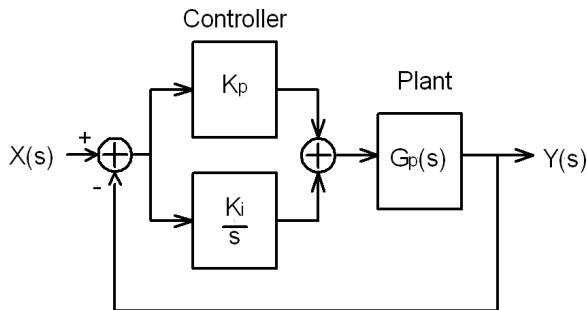
Implementation of this controller results in a system block diagram, with unity gain feedback, as shown in Figure 4-48.

The forward path transfer function for this control system is

$$G_o(s) = \left[ K_p + \frac{K_i}{s} \right] G_p(s) \quad (4.93)$$



**FIGURE 4-47** ■ Response to a step input for an integral controller of the motor position.



**FIGURE 4-48** ■ Classic configuration of a proportional plus integral controller.

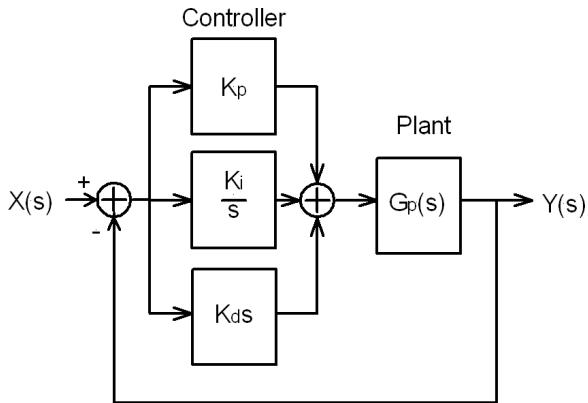
This can be rewritten as

$$\begin{aligned}
 G_o(s) &= \left[ K_p + \frac{K_i}{s} \right] G_p(s) \\
 &= \frac{sK_p + Ki}{s} G_p(s) \\
 &= \frac{K_p (s + K_i/K_p)}{s} G_p(s) \\
 &= \frac{K_p (s + 1/\tau_i)}{s} G_p(s)
 \end{aligned}$$

where  $K_p/K_i = \tau_i$  is the integral time constant.

It is obvious that a zero at  $-1/\tau_i$  and a pole at the origin have been added to the system. This maintains the same difference in the number of poles and zeros as the plant had and increases the system order. This serves to maintain the improvement in steady-state error as well as to improve the relative stability somewhat.

**FIGURE 4-49** ■  
Classic configuration  
of a PID controller.



In terms of motor control, the proportional plus integral controller returns the system to stability, but it is still highly underdamped.

#### 4.8.4 Proportional–Integral–Derivative Controller

Proportional–integral–derivative (PID) controllers are also known as three-terminal controllers and have the transfer function

$$G_c(s) = K_p + \frac{K_i}{s} + K_d s \quad (4.94)$$

The system block diagram is shown in Figure 4-49.

As with the proportional plus integral (PI) controller, equation (4.94) can be rewritten in terms of the integral time constant,  $\tau_i = K_p/K_i$ , and a derivative time constant,  $\tau_d = K_d/K_p$ ,

$$G_c(s) = K_p \left( 1 + \frac{1}{\tau_i s} + \tau_d s \right) \quad (4.95)$$

In this case the forward transfer function is

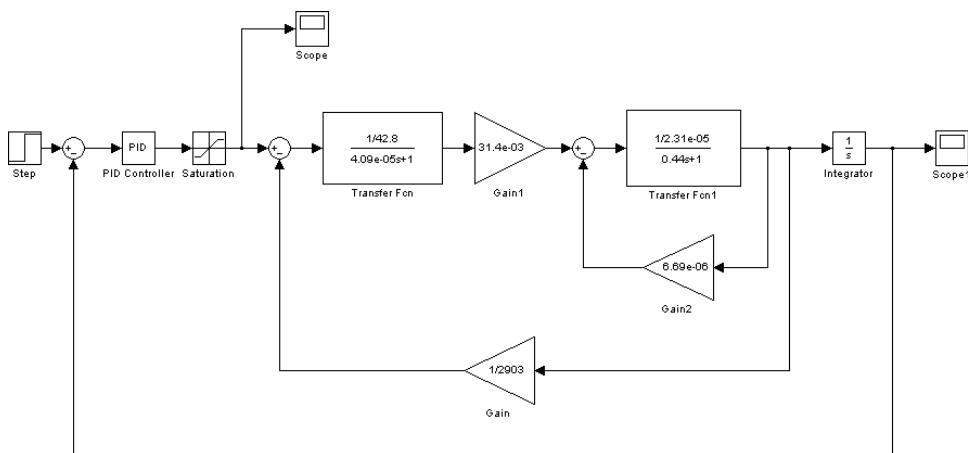
$$G_o(s) = \frac{K_p(\tau_i s + 1 + \tau_i \tau_p s^2)}{\tau_i s} G_p(s) \quad (4.96)$$

It can be seen that this has increased the number of zeros by two and the number of poles by one as well as increasing the overall system order by one.

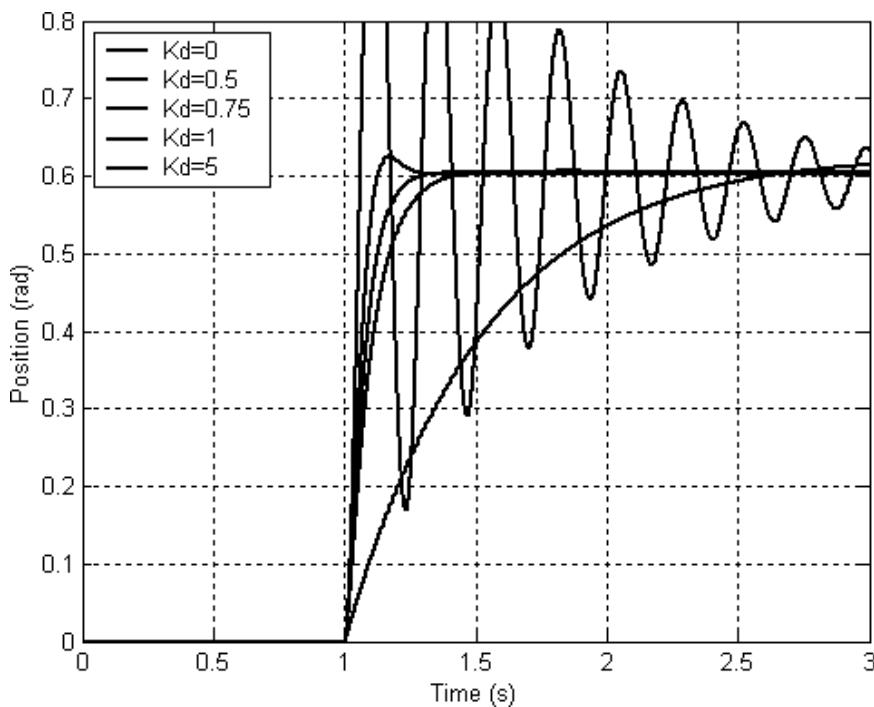
In the motor control example, a PID controller block replaces the proportional gain block, as shown in Figure 4-50.

In this case the proportional gain,  $K_p = 10$ , the integral gain,  $K_i = 1$ , and the derivative gain,  $K_d$ , are set to a range of values. The step response for the PID controlled position controller is shown in Figure 4-51.

It can be seen that the response is very similar to that for a proportional controller when  $K_d = 0$ . However, as the derivative gain increases, damping is increased and the ringing decreases. The critically damped response occurs for  $K_d = 0.75$ , and above that the response becomes overdamped.



**FIGURE 4-50** ■  
Simulink model of a  
PID position  
controller for a  
motor driving a load.



**FIGURE 4-51** ■  
Response to a step  
input for a PID  
controller of the  
motor position.

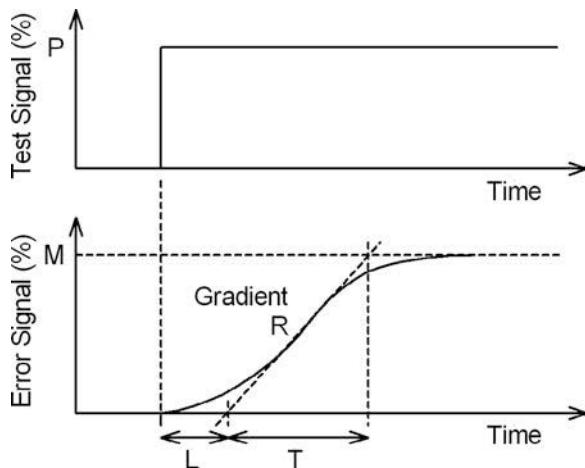
## 4.9 | CONTROLLER IMPLEMENTATION

### 4.9.1 Selection of Controller Gains

Selection of the gains for the controllers can be done by trial and error, but a number of techniques have been developed to help with the process, one of which is the process reaction curve method (Bolton, 1992). It is performed as follows:

- Open the control loop so that no control action occurs (break between the controller and the correction unit).
- Apply a small test signal step,  $P$ , expressed as a percentage change in the correction.

**FIGURE 4-52** ■ Response of the open-loop controller.



**TABLE 4-8** ■ Gain relationship for Optimum Controllers per Zeigler and Nichols

Control Mode	$K_p$	$K_i$	$K_d$
Proportional	$P/RL$		
PI	$0.9P/RL$	$1/1.33L$	
PID	$1.2P/RL$	$1/2L$	$L/2$

- Measure the error signal as a function of time.
- Plot as shown in Figure 4-52.
- Draw a tangent to the maximum gradient of the error signal curve. The gradient  $R = M/T$ .
- Measure the lag,  $L$ , which is defined as the period from the start of the test to the intersection of the tangent with the x-axis.
- Zeigler and Nichols (cited in Franklin, Powell, and Emami-Naeini, 1991) suggest adjusting the controller gains based on the measured values of  $P$ ,  $R$ , and  $L$ , as shown in Table 4-8.

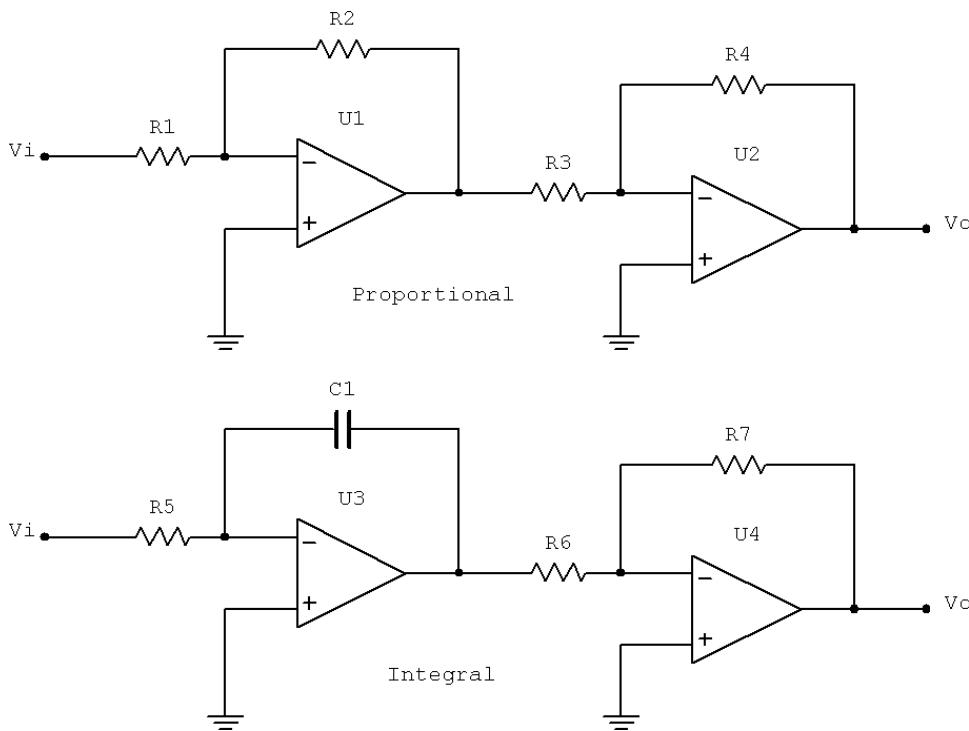
Other optimization techniques based on state-space representations are also available (Bryson and Ho, 1975) but are beyond the scope of this book.

### 4.9.2 Controller Hardware

Most modern control systems are implemented digitally since this gives the most scope for optimization of the controller structure and the gains. However, for simple applications it is easy to implement proportional, integral, PI, and PID controllers using the operational amplifier (op amp) building blocks discussed in Chapter 5.

Figure 4-53 shows the electronic circuits used to implement proportional and integral controllers.

In these examples, the purpose of the first op amp is to perform the required function, and that of the second op amp is to operate as an inverter to restore the correct sign to the output; therefore,  $R_3 = R_4$  and  $R_6 = R_7$ .



**FIGURE 4-53** ■ Op amp circuits used to implement proportional and integral control strategies.

The controller transfer function for the proportional case is

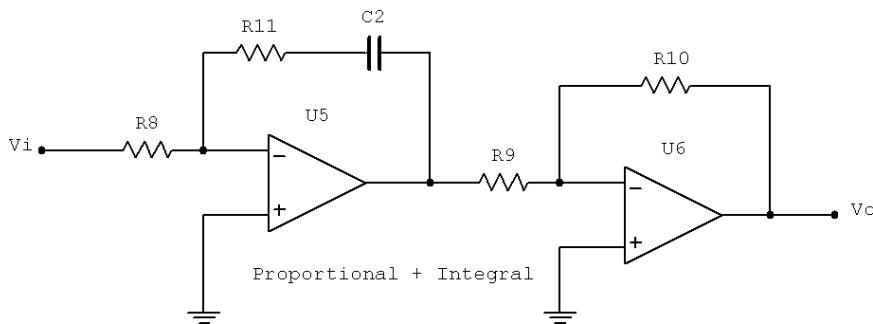
$$G_c(s) = K_p = \frac{R_2}{R_1} \quad (4.97)$$

and, for the integral case,

$$G_c(s) = \frac{K_i}{s} = \frac{1/R_5C_1}{s} \quad (4.98)$$

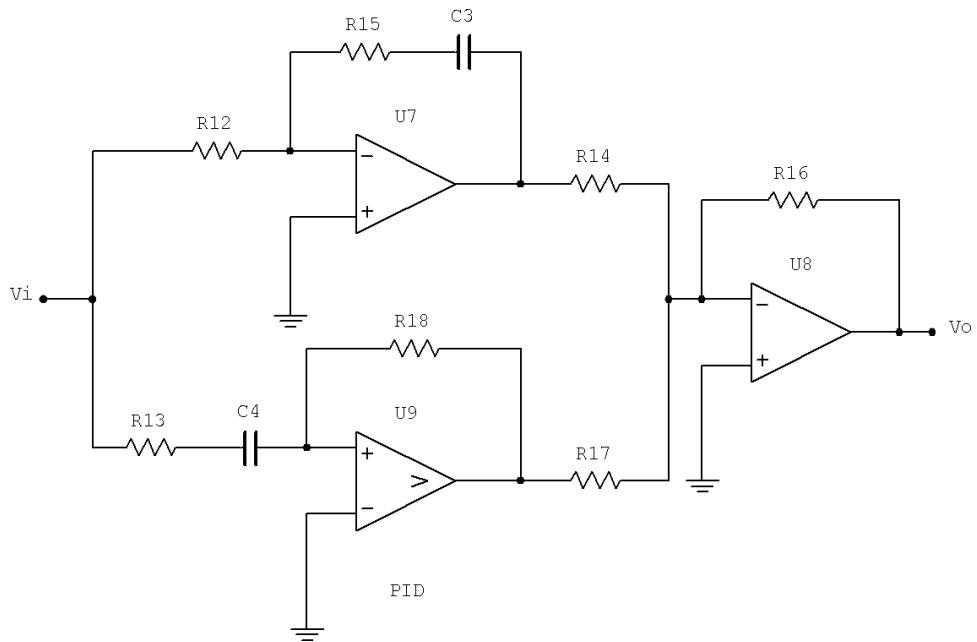
The PI controller can be implemented by combining the feedback elements of the previous circuits, as shown in Figure 4-54. In this case the transfer function is

$$G_c(s) = \frac{R_{11}}{R_8} + \frac{1/R_8C_2}{s} \quad (4.99)$$



**FIGURE 4-54** ■ Op amp circuit used to implement PI control strategy.

**FIGURE 4-55** ■ Op amp circuit used to implement PID control strategy.



The final circuit to implement a PID controller requires an extra op amp to implement the derivative function. However, the pure derivative is seldom used because it adds excessive amounts of high-frequency noise, so a lead compensator is selected instead.

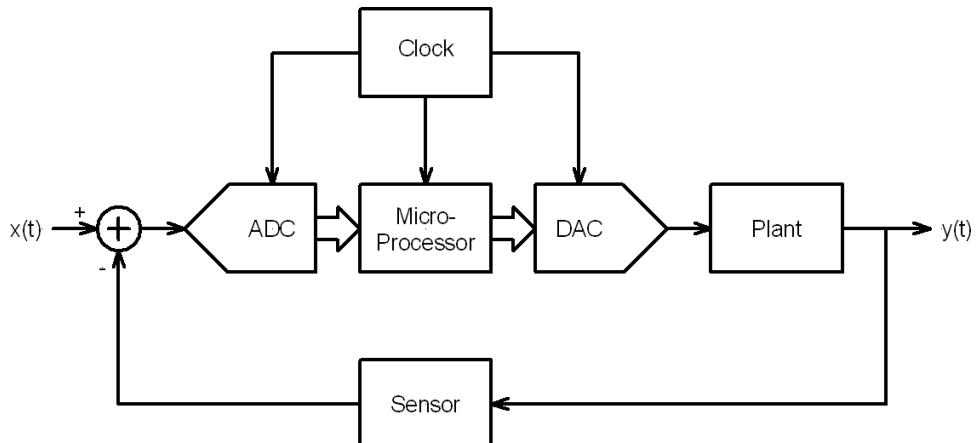
This complete circuit shown in Figure 4-55 has the transfer function

$$G_c(s) = \frac{R_{15}}{R_{12}} + \frac{1/R_{12}C_3}{s} + \frac{R_{18}C_4s}{R_{13}C_4s + 1} \quad (4.100)$$

In a digital controller, the error signal is digitized and then run through the appropriate control algorithms in a microprocessor before being converted back to analog and introduced into the plant, as shown in Figure 4-56.

The main difference between analog and digital controllers is that the latter operate in discrete time steps. Conversion of an analog signal to its digital representation takes a finite

**FIGURE 4-56** ■ Digital controller implementation.



amount of time, as does the reverse process. Additionally, implementing the controller algorithm in a microprocessor also takes time. A clock is generally used to synchronize the various processes, some of which are discussed in Chapter 5.

## 4.10 | REFERENCES

---

- Bolton, W. (1992). *Control Engineering*. Harlow, UK: Longman Scientific & Technical.
- Bryson, A. and Y.-C. Ho. (1975). *Applied Optimal Control—Optimization, Estimation and Control*. New York: Hemisphere Publishing Corporation.
- Franklin, G., J. Powell, and A. Emami-Naeini. (1991). *Feedback Control of Dynamic Systems*. Boston: Addison-Wesley Publishing Company.
- Raven, F. (1978). *Automatic Control Engineering*, 3d ed. New York: McGraw Hill.



# Signal Processing

## Chapter Outline

5.1	Introduction.....	207
5.2	Biomedical Signals.....	207
5.3	Signal Acquisition.....	211
5.4	Analog Signal Processing.....	224
5.5	Digital Signal Processing.....	241
5.6	Statistical Techniques and Machine Learning.....	264
5.7	Isolation Barriers.....	270
5.8	References .....	274

## 5.1 INTRODUCTION

This chapter deals with two separate aspects of biomechatronic signal acquisition and processing. The first is concerned with signals obtained directly from the organism, including electrical, chemical, and pressure. The second is concerned with the remaining signals generated as part of a biomechatronic process—including tactile signals from a prosthetic hand, outputs of potentiometers, and rate gyros—that are used for control or feedback.

These signals are manipulated in various ways using analog circuitry or digital algorithms to produce outputs that can be used to actuate or control a mechatronic device such as a prosthetic limb. Alternatively, the output can be used as an indicator of the state of the organism, such as an electrocardiograph trace that is indicative of the condition of the heart.

## 5.2 BIOMEDICAL SIGNALS

Biomedical signals originate from a number of sources including the following (Bronzino, 2006):

- **Bioelectric signals:** This is a generic term for all of the electrical signals generated by nerve and muscle cells. The source is the membrane potential that, under certain circumstances, may generate an action potential. In single-cell measurements, the action potential is the bioelectric signal. However, in most cases an embedded or surface electrode measures the sum of the action potentials of a large number of cells, the latter of which include myoelectric and electrocardiogram (ECG) signals.

- **Bioimpedance signals:** The electrical impedance of tissue contains important information concerning, for example, its makeup, blood volume, and endocrine activity. These signals are normally obtained by injecting low-current (<20 mA) sinusoidal electrical signals into the body at frequencies between 50 kHz and 1 MHz and monitoring the relationship between the current and the voltage.
- **Bioacoustic signals:** Many biological phenomena generate acoustic outputs, and these are indicative of the function being performed. A good example of this is the “lub-dub” sound produced by the pumping heart. Other sounds are generated by blood flow, air flow, and the transit of solids, liquids, and gas through the digestive system.
- **Biomagnetic signals:** All of the organs in which electrical activity occurs generate magnetic fields as a result of this activity. These organs include the brain, heart, and lungs as well as the skeletal muscles. Unfortunately, the amplitude of these signals is very small, and they are difficult to measure.
- **Biomechanical signals:** These signals include motion and displacement as well as pressure, tension, and flow within the organism. Measurement of these requires the use of sensors and transducers discussed in Chapter 2. Unlike electrical and magnetic signals, these signals generally do not propagate (with the exception of pressure) and so are mostly measured at the source.
- **Biochemical signals:** These result from chemical activity within the organism that can alter its chemical composition in both subtle and gross ways. Chemical composition is measured using the sensors discussed in Chapter 2 and can include gases such as CO<sub>2</sub> and O<sub>2</sub> as well as dissolved solids like glucose and various salts.
- **Bio-optical signals:** These are signals related to the optical reflectance or transmission of the tissue. For example, blood oxygen levels may be determined from the relationship between the reflectance of infrared (IR) and visible wavelengths. Other information may be gathered using fluorescence characteristics based on injected dye materials.

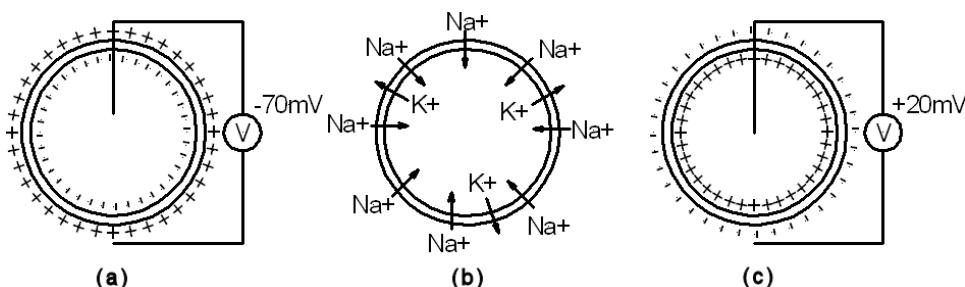
The signals can be classified with regard to their source or application or in terms of the signal characteristics. Biological signals are classified as either continuous or discrete. Continuous signals include temperature, pressure, and chemical concentration, and discrete signals include electrical impulses generated by individual nerve cells.

These signals can be divided into broad classes depending on the rate and nature of the variations (Carr, 1997).

### 5.2.1 Bioelectric Signals

The idea that electricity is generated by the body was first introduced by Luigi Galvani in 1786, but it was not until 1903 when William Einthoven was able to measure these potentials using an improved string galvanometer, that their true potential as a diagnostic tool became apparent. This potential was further enhanced by the invention of the vacuum tube amplifier a few years later.

Bioelectric potentials are actually ionic voltages produced as a result of electrochemical activity in some types of specialist cells. Conductive solutions consisting of dissolved salts surround the cells in the body. The principal ions in solution are sodium (Na<sup>+</sup>), potassium (K<sup>+</sup>), and chloride (Cl<sup>-</sup>). The semipermeable membranes of nerve and muscle cells



**FIGURE 5-1** ■ Process of triggering cell action potential. (a) Equilibrium state. (b) Depolarization process as sodium ions flow across cell membrane. (c) Depolarized state.

allow the entry of the potassium and chloride ions but block the sodium ions. Because the ions seek to balance both potential and concentration across the membrane, the restriction on the diffusion of sodium ions results in an imbalance in the concentration of sodium ions, with fewer within the cell and more in the intercellular fluid. In an attempt to balance the charge, additional potassium ions enter the cell. At equilibrium, a potential difference of  $-70\text{ mV}$  exists across the cell membrane, with the interior being negative with respect to the exterior. This potential difference, which has been measured at between  $-60$  and  $-100\text{ mV}$ , is called the resting potential of the cell, and cells at this potential are referred to as polarized.

When a section of the cell membrane is excited by the flow of ionic current or by some other form of excitation energy, the membrane becomes permeable to sodium ions and begins to flow across the boundary, as illustrated in Figure 5-1. This ionic current increases the permeability further with the result that the current flow increases exponentially (the avalanche effect), and sodium ions rush into the cell to try to reach equilibrium. At the same time, some potassium ions begin to move out of the cell for the same reasons. Potassium ions are slower than sodium ions, and as a result the cell is left with a slight potassium imbalance, which results in a positive potential difference of about  $+20\text{ mV}$  across the cell membrane. This is known as the action potential of the cell and is considered depolarized when in this state.

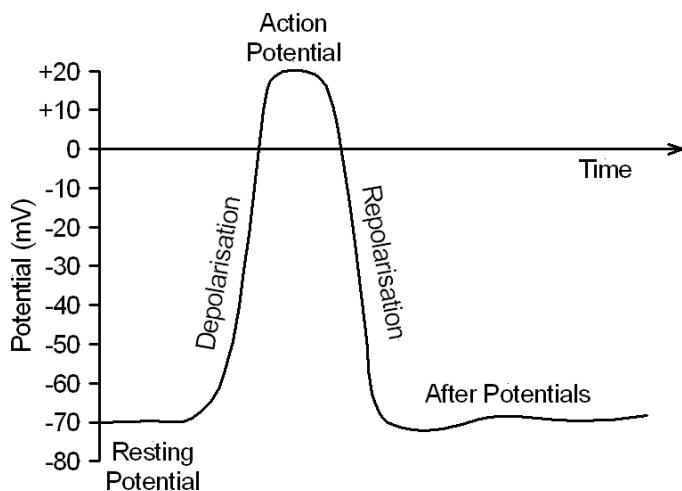
At this time, the cell reverts to its semipermeable state, and an active process called a sodium pump quickly transports sodium ions back out of the cell during the process known as repolarization.

The time periods involved in this process vary with different cell types. For example, in nerve and muscle cells the repolarization process is so quick that the action potential appears as a sharp spike as little as  $1\text{ ms}$  wide. Heart muscle repolarizes much more slowly, with the action potential lasting from  $150$  to  $300\text{ ms}$ . However, as shown in Figure 5-2, regardless of the duration the resting and action potentials are always the same.

When a cell is excited and generates an action potential, ionic currents flow in the intercellular fluid or in adjacent areas of the same cell and can excite neighboring cells. In the case of nerve cells with a long axon fiber, the action potential is generated in a very small segment of the fiber but propagates rapidly in both directions from the trigger point. Under normal conditions, nerve fibers are excited only near their input end and thus propagate in one direction only.

As an action potential travels down the fiber, it cannot reexcite the portion of fiber immediately prior because of a refractory period that follows the action potential. The rate at which the potential travels down the fiber is called the propagation rate, nerve

**FIGURE 5-2** ■  
Waveform showing the potential difference across a cell membrane as a function of time.



conduction rate, or conduction velocity. This rate varies widely depending on the type and diameter of the nerve fiber, but it is usually something between 20 and 140 m/s in nerves. Propagation through heart muscle is much slower, with an average rate of only 0.2 to 0.4 m/s, and some time-delay fibers between the atria and ventricles propagate even more slowly at between 0.03 and 0.05 m/s (Cromwell, Weibell et al., 1973).

### 5.2.2 Signals Characterized by Source

As can be seen from Table 5-1, a range of potentials from less than  $1 \mu\text{V}$  right up to 100 mV and frequencies right up to 10 kHz must be accommodated. One of the major difficulties arises where a small amplitude signal must be examined in the presence of noise or another much larger signal.

### 5.2.3 Signals Characterized by Type

#### 5.2.3.1 Static and Quasi-Static Signals

Static signals are by definition unchanging over a long period of time. Such signals are essentially direct current (DC) levels and in isolation convey very little information. Quasi-static signals are those that change very slowly with time, such as the long-term drift on a sensor or the decreasing voltage on a slowly discharging battery.

#### 5.2.3.2 Periodic and Repetitive Signals

Periodic signals repeat themselves on a regular basis, though the timescale for repetition can be from femtoseconds up to days or even years. These include sine, square, and sawtooth waves, and their defining nature is that each waveform is identical. Repetitive signals are periodic in nature, but the exact shape may change slightly with time. Ultimately, very few biological signals are truly periodic, so ECG signals and estrogen levels in the blood are good examples of repetitive signals.

#### 5.2.3.3 Transient and Quasi-Transient Signals

Transient signals are, by definition, one time only, whereas quasi-transient signals are periodic but with a duration that is very short compared with the period of the waveform.

**TABLE 5-1** ■ Characteristics of Some Bioelectric Signals

Classification	Acquisition	Frequency Range (Hz)	Dynamic Range	Description
Action potential	Micro electrode	100–2000	10 $\mu$ V–100 mV	Cell membrane potential
Electroneurogram (ENG)	Needle electrode	100–1000	5 $\mu$ V–10 mV	Potential of a nerve bundle
Electroretinogram (ERG)	Micro electrode	0.2–200	0.5 $\mu$ V–1 mV	Evoked flash potential
Electroencephalogram (EEG)				
Surface	Surface electrode	0.5–100	2–100 $\mu$ V	Multichannel scalp potential
<i>Delta range</i>		0.5–4		Children, deep sleep
<i>Theta range</i>		4–8		Temporal and central—alert state
<i>Alpha range</i>		8–13		Awake, relaxed, closed eyes
<i>Beta range</i>		13–22		
<i>Sleep spindles</i>		6–15	50–100 $\mu$ V	Bursts of 0.2–0.6 s
<i>K-complexes</i>		12–14	100–200 $\mu$ V	Bursts during deep sleep
Evoked potentials (EP)	Surface electrode		0.1–20 $\mu$ V	Brain response to stimulus
<i>Visual (VEP)</i>		1–300	1–20 $\mu$ V	Occipital lobe 200 ms duration
<i>Somatosensory (SEP)</i>		2–3000		Sensory cortex
<i>Auditory (AEP)</i>		100–3000	0.5–10 $\mu$ V	Vertex
Electrocorticogram	Needle electrode	100–5000		Exposed brain surface potentials
Electromyogram (EMG)				
<i>Single fiber</i>	Needle electrode	500–10000	1–10 $\mu$ V	Action potential—single fiber
<i>Motor unit action pot.</i>	Needle electrode	5–10000	100 $\mu$ V–2 mV	
Surface EMG (SEMG)	Surface electrode			
<i>Skeletal muscle</i>		2–500	50 $\mu$ V–5 mV	
<i>Smooth muscle</i>		0.01–1		
Electrocardiogram (ECG)	Surface electrode	0.05–100	1–10 mV	

Once again, this classification is rather arbitrary, with the differences between quasi-transient and periodic signal types being rather vague.

#### 5.2.3.4 Stochastic Signals

Stochastic signals are generated by a stochastic biological process that produces sample functions, each of which differs from the others from a temporal perspective but shares the same distribution characteristics. These signals cannot be classified exactly and are described by the statistics of the whole collection of sample functions (referred to as an ensemble). These classifications can include the probability density function (PDF), which describes the amplitude characteristics of the signal, and the autocorrelation function or its associated power spectral density (PSD). They are often also characterized in terms of their mean and variance.

## 5.3 | SIGNAL ACQUISITION

The acquisition of electrical signals generated by the organism can be achieved using one of myriad electrode types discussed in Chapter 2. Any of the other signal types are converted to electrical signals by one of the sensors discussed in that chapter. These signals are generally acquired in a continuous manner, after which they are processed using some analog circuitry before sampling and digitization. This is followed by storage or display, as for electroencephalograms (EEGs), or as inputs into some form of actuation, such as an insulin dispenser if the blood glycogen levels are incorrect.

### 5.3.1 Noise

In all of the examples, the ultimate electrical signal will be accompanied by some form of noise. This noise could come from other physiological activities, from external interference, or, for the smaller signals, from thermal noise generated by the acquisition electronics. It may be random, or it may be repetitive or even periodic, depending on its source. There is therefore no single signal processing technique that can be used to minimize the noise and hence to maximize the signal-to-noise ratio.

#### 5.3.1.1 Thermal Noise

Thermal noise is electrical noise generated by random fluctuations of the voltage or current due to the thermal agitation of electrons within a conductor. It is therefore common to any electrical circuit associated with biomechatronic sensors.

More formally, if  $v(t)$  is the thermal noise voltage across the terminals of a resistor,  $R$ , then if this voltage is measured at regular intervals over a long period the mean value,  $\bar{v}$ , is

$$\bar{v} = \frac{1}{m} \sum_{j=1}^m v_j \quad (5.1)$$

Measurements show that in the limit as  $m \rightarrow \infty$  the mean value approaches zero. This result can be justified by considering the random motion of large numbers of electrons that produce fluctuations in the potential. These must average out to zero in the long-term; otherwise, they would result in the flow of a current.

The time-averaged squared signal  $\bar{v}^2$  is determined in a similar way:

$$\bar{v}^2 = \frac{1}{m} \sum_{j=1}^m v_j^2 \quad (5.2)$$

In the limit as  $m \rightarrow \infty$  this mean squared value,  $\bar{v}^2$  ( $V^2$ ), can be shown to approach

$$\bar{v}^2 = 4kT R \Delta f \quad (5.3)$$

where  $k$  is Boltzmann's constant ( $1.38 \times 10^{-23}$  J/K),  $T$  (Kelvin) is the absolute temperature,  $R$  (ohms) is the resistance value, and  $\Delta f$  (Hz) is the bandwidth (Young, 1990).

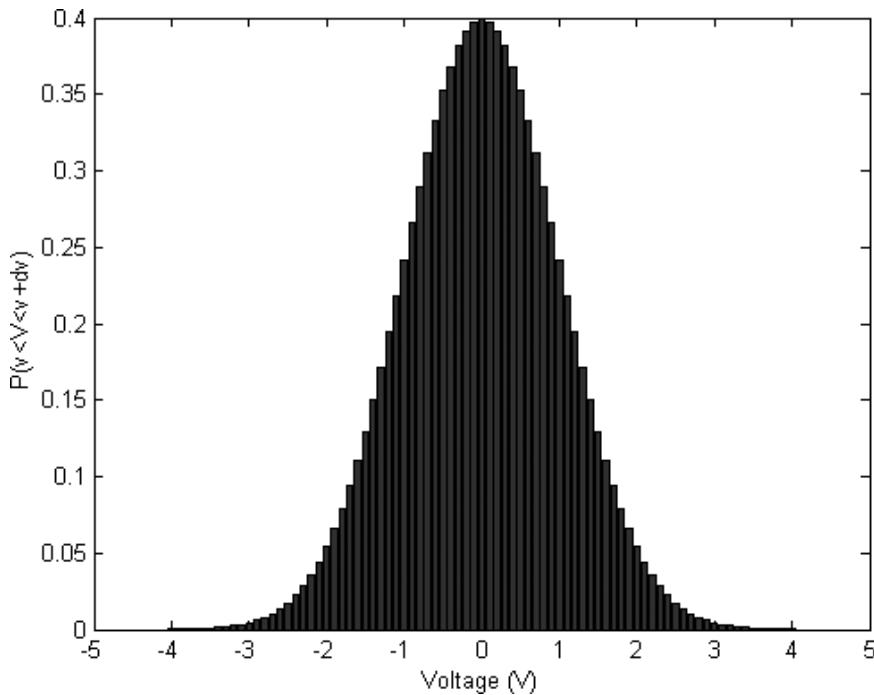
If samples of the noise voltage are taken over a long period and the results are plotted as a histogram with bin widths,  $dV$ , a distribution of the form shown in Figure 5-3 is produced (Brooker, 2008).

The probability  $p\{V\}dV$  that any future measurements will fall in the range  $V \rightarrow V + dV$  is given by this plot, which is known as the PDF. This function approximates the normal or Gaussian distribution (Walpole and Myers, 1978), which can be described by

$$p\{V\}dV = \frac{1}{\sigma\sqrt{2\pi}} e^{-V^2/2\sigma^2} \quad (5.4)$$

The time-averaged squared value,  $\bar{v}^2$ , equates to the variance,  $\sigma^2$ , because the distribution is Gaussian. Its value is a measure of how wide the distribution is; hence, it is a useful indicator of the amount of noise present. However, in practice it is more common to specify the noise level in terms of the root mean square (RMS) quantity, where  $v_{rms}$  is

$$v_{rms} = \sqrt{\bar{v}^2} \quad (5.5)$$



**FIGURE 5-3** ■ Histogram showing the PDF for thermal noise with unity variance

**Noise power spectrum for thermal noise:** In theory this is completely flat spectrally and is known as *white noise* as an analogy to white light, which comprises a uniform mix of all the colors. Strictly speaking, however, it is not possible to produce a power spectrum that is truly white over an infinite frequency range, as the total power integrated over this bandwidth would be infinite. In reality, all noise-generating processes are subject to some band-limiting mechanism that produces a finite noise bandwidth. In addition, the measurement process is also band-limited and restricts the measured value for the total noise power still further.

It is often convenient to remove the bandwidth dependence on the power spectrum; this is referred to as the PSD or voltage variance per hertz,  $V^2/\text{Hz}$ . Electronic noise levels are often quoted as volts per root hertz ( $\text{V}/\sqrt{\text{Hz}}$ ) as this serves as a reminder that the RMS voltage increases with the square root of the noise bandwidth.

### WORKED EXAMPLE

---

Determine the noise power spectral density of a  $100 \text{ k}\Omega$  resistor at a temperature of  $25^\circ\text{C}$

$$\begin{aligned} \bar{v}^2 &= 4kTR (\text{V}^2/\text{Hz}) \\ &= 4 \times 1.38 \times 10^{-23} \times (273 + 25) \times 100 \times 10^3 \\ &= 1.65 \times 10^{-15} \text{ V}^2/\text{Hz} \end{aligned}$$

By taking the square root, it is possible to obtain the PSD in the more common form

$$v_{rms} = 40.56 \text{ nV}/\sqrt{\text{Hz}}$$

A data acquisition system has a  $20 \text{ kHz}$  bandwidth and a very high input impedance compared

with the resistor value, so it would measure the RMS voltage across the resistor to be

$$\begin{aligned} v_{rms} &= 40.56 \times 10^{-9} \times \sqrt{20 \times 10^3} \\ &= 5.7 \mu\text{V} \end{aligned}$$

### 5.3.1.2 Shot Noise

Shot or Schottky noise is typically generated by the current flowing across a barrier such as the action potential developed across a cell membrane. The noise is generated by the migration of individual charge elements across the barrier at random intervals, so even though on average the current flow may be constant, fluctuations around the average take the form of a Poisson distribution (Walpole and Myers, 1978)

$$p(n, \gamma) = \frac{e^{-\gamma} \gamma^n}{n!} \quad (5.6)$$

where  $e$  is the base of the natural log (2.71828),  $n$  is the actual number of occurrences, and  $\gamma$  is the expected number of occurrences during a given time period.

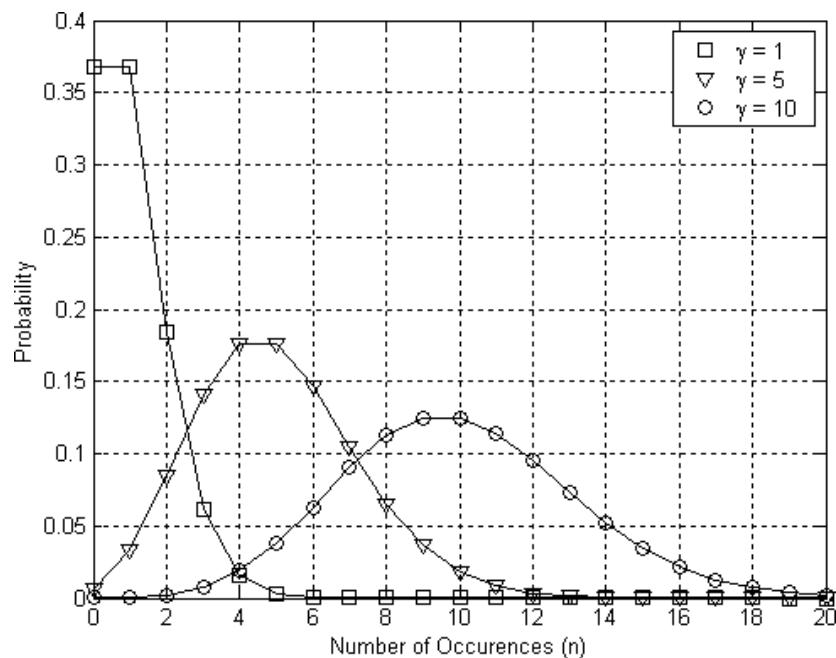
Figure 5-4 shows a number of examples of the distribution for different values of  $\gamma$ . Note that the occurrences must be discrete integers, so the lines joining these points are for illustration only.

The Poisson distribution has a number of interesting characteristics, one of which is that the mean and the variance are both equal to  $\gamma$ . In addition the figure shows that, as the expected number of occurrences in a given time interval increases, the distribution becomes more normal. For  $\gamma > 1000$ , a Gaussian distribution with both mean and variance equal to  $\gamma$  is an excellent approximation of the Poisson distribution.

One of the results of this relationship is that as the mean current,  $I_{dc}$  (A), along a cluster of nerve fibers increases, the RMS noise current,  $i_{rms}$  (A), increases proportionally:

$$i_{rms} = \sqrt{2qI_{dc}\Delta f} \quad (5.7)$$

**FIGURE 5-4 ■**  
Poisson distributions for differing occurrence expectations  
(Brooker 2008.)



where  $q$  is the electron charge ( $1.6 \times 10^{-19}$  C), and  $\Delta f$  (Hz) is the bandwidth. As with the thermal noise case, the magnitude of shot noise is also proportional to the measurement bandwidth (Young, 1990).

If this current is measured using an electrode, is buffered, and flows through a load resistor,  $R_{load}$  ( $\Omega$ ), then the RMS noise voltage,  $v_{rms}$  (V), will be

$$v_{rms} = i_{rms} R_{load} \quad (5.8)$$

**Noise power spectrum for shot noise:** This can be determined by examining the noise-generation process. As each charge element flows across a cell membrane, it generates a current pulse. Because the duration of each pulse is relatively short, its effective bandwidth is reasonably wide (by biometric standards). It can therefore be concluded that shot noise is *white*. A good analogy of the process is the sound of rain on a corrugated iron roof.

### WORKED EXAMPLE

Determine the noise power spectral density,  $i_{rms}$  ( $A\sqrt{\text{Hz}}$ ) of the shot noise for the spinal cord carrying a current of 0.5 mA.

$$\begin{aligned} i_{rms} &= \sqrt{2qI_{dc}} \\ &= \sqrt{2 \times 1.6 \times 10^{-19} \times 0.5 \times 10^{-3}} \\ &= 12.6 \text{ pA}/\sqrt{\text{Hz}} \end{aligned}$$

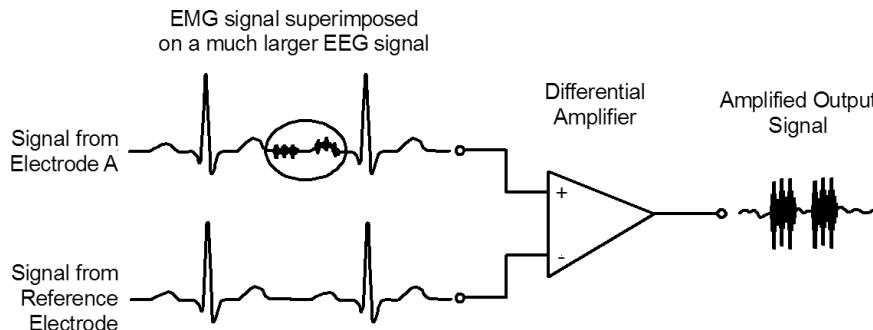
If this current flows through a load resistor with  $R_{load} = 1 \text{ k}\Omega$ , the RMS voltage spectral density will be  $v_{rms} = 12.6 \text{ nV}/\sqrt{\text{Hz}}$ .

The RMS noise voltage measured by the data acquisition system with a 20 kHz bandwidth will be

$$v_{rms} = 12.6 \times 10^{-9} \times \sqrt{20 \times 10^3} = 1.79 \mu\text{V}$$

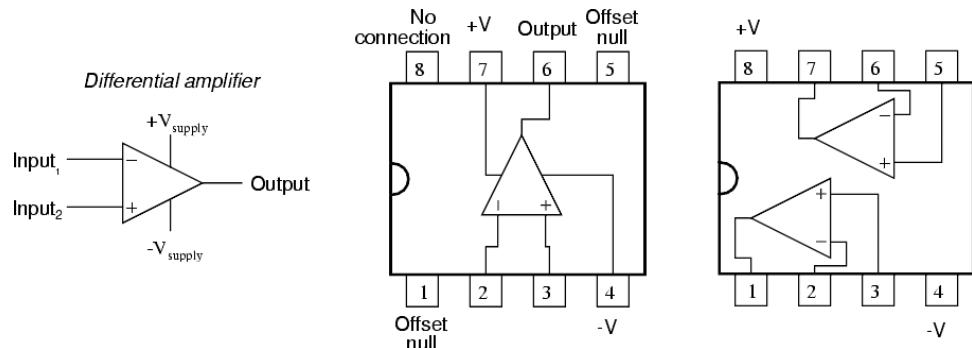
#### 5.3.1.3 Common-Mode and Differential Mode Noise

Measurement of a single-fiber electromyogram with a magnitude of only 10  $\mu\text{V}$  in the presence of an ECG signal 1000 times as large would seem to be a lost cause. However, if the electrodes are placed with care, the vast majority of the ECG signal will be common-mode noise, which can be eliminated using a combination of common-mode filtering and an amplifier with a differential input. This process is illustrated in Figure 5-5.



**FIGURE 5-5** ■ Common-mode noise rejection using a differential amplifier.

**FIGURE 5-6** ■ Op amp symbol and common pin configurations.



Differential-mode noise generated by electrode movement or any other mechanism is indistinguishable from the electromyogram (EMG) signal and cannot be removed by this technique. It is therefore output along with the EMG signal and would have to be removed by the signal processor.

### 5.3.2 Amplifiers

Most analog signal processing is performed using combinations of operational amplifiers shown schematically in Figure 5-6. Known as op amps, these versatile integrated circuits consist of hundreds of transistors, resistors, and capacitors to produce an extremely high-gain, wide-bandwidth amplifier with a differential input.

The ideal op amp is considered to have infinite gain at DC, infinite input impedance, and zero output impedance. As a consequence of the infinite input impedance, no current is drawn by the inputs, and as a result of the infinite gain any difference between the two inputs results in an infinite output voltage. The zero output impedance results in an output voltage that is independent of the output current.

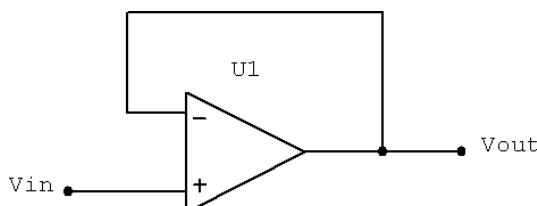
These characteristics may seem to be illogical, but they provide a good approximation of the actual performance of an op amp in the case where negative feedback is used to reduce the actual gain.

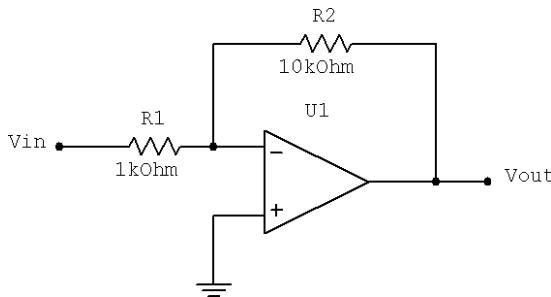
#### 5.3.2.1 Negative Feedback

If the output of an op amp is connected to its inverting input and a voltage signal is applied to the noninverting input, as shown in Figure 5-7, the output voltage of the op amp closely follows that input voltage.

As  $V_{in}$  increases,  $V_{out}$  will increase in accordance with the differential gain. However, as  $V_{out}$  increases that output voltage is fed back to the inverting input, thereby acting to decrease the voltage difference between inputs. This acts to reduce the output. What happens for any given voltage input is that the op amp will output a voltage very nearly equal to  $V_{in}$  but just low enough for the voltage difference left between  $V_{in}$  and the negative (–) input to be amplified to generate the output voltage.

**FIGURE 5-7** ■ Op amp circuit with negative feedback.





**FIGURE 5-8** ■ Op amp circuit for an inverting amplifier.

The circuit will quickly reach equilibrium at a point where the output voltage is exactly the right value to maintain the correct amount of differential, which in turn produces the right amount of output voltage. This technique is known as negative feedback, and it is the key to having a self-stabilizing system.<sup>1</sup> This stability gives the op amp the capacity to work in its linear region.

For an op amp with a gain  $A_v$ , the output voltage,  $V_{out}$ , is the product of the voltage gain and the differential input voltage

$$V_{out} = (V_{in} - V_{out}) A_V \quad (5.9)$$

This can be rewritten as

$$V_{out} = V_{in} \frac{A_V}{1 + A_V} \quad (5.10)$$

Typical op amps have DC gains of  $A_v = 10^6$  or more; therefore, to a good approximation, the output voltage will be equal to the input voltage, and the differential voltage will be zero (Kuphaldt, 2003).

### 5.3.2.2 Inverting Amplifier

The relationship between the input and output voltages can be easily be determined if it is remembered that the input impedance to the amplifier is extremely high, and therefore the current into node A of the circuit shown in Figure 5-8 must equal the current exiting it (Kirchhoff's law). Therefore,

$$-\frac{V_{in}}{R_1} = \frac{V_{out}}{R_2} \quad (5.11)$$

from which it is simple to determine that the output voltage will be

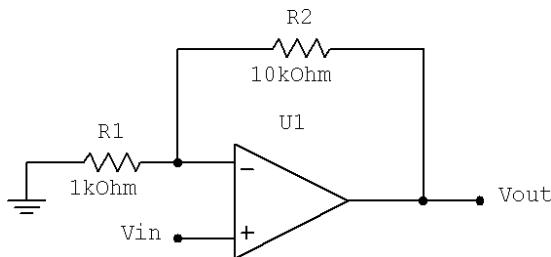
$$V_{out} = -V_{in} \frac{R_2}{R_1} \quad (5.12)$$

In the example shown in Figure 5-8, the gain  $A_V = -R_2/R_1 = -10$ ; therefore, the magnitude of output voltage will be 10 times as large as the input voltage.

In most modern op amps it is possible to produce a gain of less than one and still maintain the stability of the circuit. However, in most cases these op amp circuits are used to increase the level of the signal and  $R_2 \gg R_1$ .

<sup>1</sup>This is true not only of op amps but also of any dynamic system in general.

**FIGURE 5-9** ■ Op amp circuit for a noninverting amplifier.



One further consideration is that because the  $-ve$  input to the op amp is a virtual earth, the input impedance of the inverting amplifier is equal to  $R_1$ , which is usually less than 100 k $\Omega$  or so (depending on the actual input impedance of the op amp and the bias current required). This can result in this amplifier circuit loading the input, with unforeseen consequences (Kuphaldt, 2003).

### 5.3.2.3 Noninverting Amplifier

The relationship between the input and output voltage can easily be derived from Figure 5-9. Consider that the input impedance of the op amp is extremely high; therefore, the current flowing into node A must equal the current flowing out.

$$\frac{V_{in}}{R_1} = \frac{V_{out} - V_{in}}{R_2} \quad (5.13)$$

from which it is easy to determine that

$$V_{out} = V_{in} \frac{R_1 + R_2}{R_1} \quad (5.14)$$

For the component values shown in Figure 5-9, the gain is  $A_V = 11$ , making the output 1.1 V for an input of 100 mV.

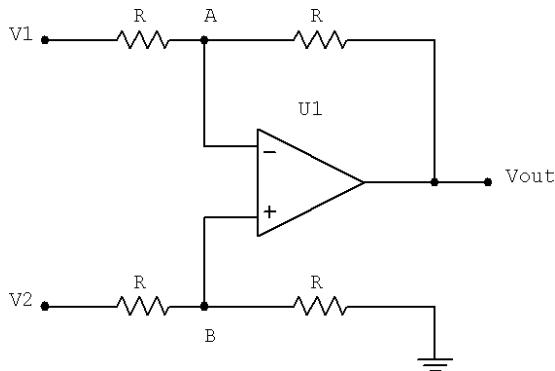
If the resistance of  $R_2$  is reduced to zero, the circuit reduces to that in Figure 5-7, and the gain reduces to unity.

In this configuration, the input impedance of the amplifier is equal to the input impedance of the op amp, which is extremely high, and hence no loading to any previous stages will occur.

### 5.3.2.4 Differential Amplifier

An op amp with no feedback already provides differential gain, amplifying the voltage difference between the two inputs. However, its gain cannot be controlled, and it is generally too high to be of any use except as a comparator. In the previous examples, the application of negative feedback to op amps has resulting in the practical loss of one of the inputs, with the resulting amplifier good only for amplifying a single voltage signal input. However, an op amp circuit maintaining both voltage inputs with a controlled gain set by external resistors can be constructed as shown in Figure 5-10.

If all the resistor values are equal, this amplifier will have a differential voltage gain of 1. The analysis of this circuit is similar to that of an inverting amplifier, except that the noninverting input (+) of the op amp is at a voltage equal to a fraction of  $V_2$  rather than being connected directly to ground. In this configuration,  $V_2$  functions as the noninverting



**FIGURE 5-10** ■ Op amp circuit for a differential amplifier.

input, and  $V_1$  functions as the inverting input of the final amplifier circuit. The voltage,  $V_B$ , at node  $B$  is

$$V_B = V_2 \frac{R}{2R} = \frac{V_2}{2}$$

Using Kirchhoff's law for currents into node  $A$

$$\frac{V_A - V_1}{R} = \frac{V_{out} - V_A}{R}$$

Because the gain is large, the voltage at node  $A$  is equal to that at node  $B$ ,  $V_A = V_B$ .

Substituting for  $V_A$  and simplifying gives

$$V_{out} = V_2 - V_1 \quad (5.15)$$

If a differential gain of anything other than unity is required, then the resistances of both the upper and lower voltage dividers would have to be adjusted simultaneously for balanced operation, and that is not practical. As with the inverting amplifier, the gain would be determined by the ratio of the resistors,  $A_v = R_2/R_1$ , and the output voltage would be

$$V_{out} = (V_2 - V_1) \frac{R_2}{R_1} \quad (5.16)$$

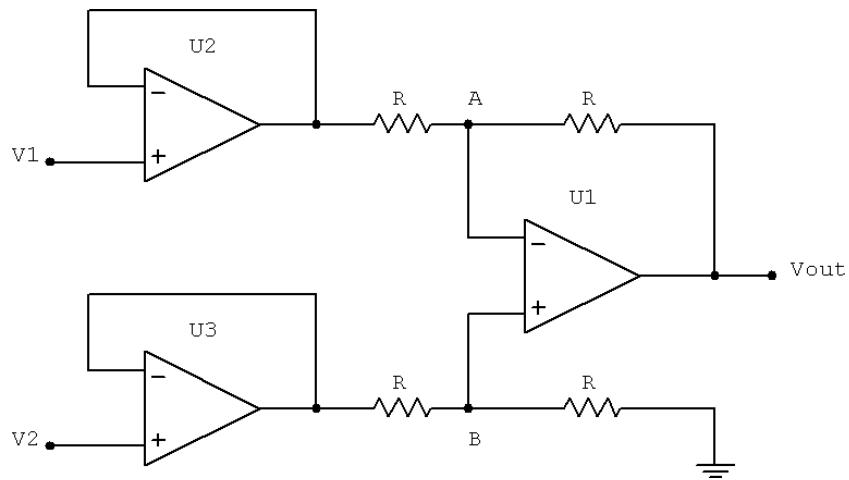
Another limitation of this amplifier design is the fact that its input impedances are rather low compared with that of some other op amp configurations. Each input voltage source has to drive current through a resistance to a virtual ground, which constitutes a far lower impedance than the extremely high input impedance of the op amp alone. This problem is solved by buffering the inputs as shown in Figure 5-11.

Now the  $V_1$  and  $V_2$  input lines are connected straight to the inputs of two voltage-follower op amps, giving very high impedance. The two op amps on the left now handle the driving of current through the resistors instead of letting the input voltage sources do it (Kuphaldt, 2003).

### 5.3.2.5 Instrumentation Amplifier

To construct a practical instrumentation amplifier, it is necessary to modify the previous circuit to simplify the gain control. This is achieved using three new resistors linking the

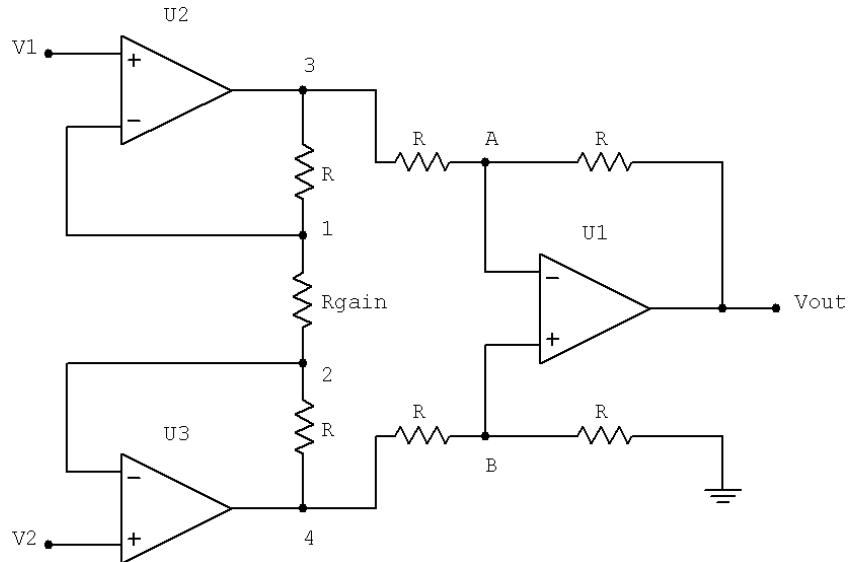
**FIGURE 5-11** ■ Op amp circuit for a buffered input differential amplifier.



two buffer circuits together, as shown in Figure 5-12. Consider all resistors to be of equal value except for  $R_{gain}$ . The negative feedback of the upper-left op amp causes the voltage at node 1 to be equal to  $V_1$ . Likewise, the voltage at node 2 is held to a value equal to  $V_2$ . This establishes a voltage drop across  $R_{gain}$  equal to the voltage difference between  $V_1$  and  $V_2$ . That voltage drop causes a current through  $R_{gain}$ , and since the feedback loops of the two input op amps draw no current an identical current to that flowing through  $R_{gain}$  must pass through the two  $R$  resistors above and below it. This produces a voltage drop between nodes 3 and 4 equal to

$$V_{3-4} = (V_2 - V_1) \left( 1 + \frac{2R}{R_{gain}} \right) \quad (5.17)$$

**FIGURE 5-12** ■ Practical differential input instrumentation amplifier.



The differential amplifier on the right-hand side of the circuit then takes this voltage drop between nodes 3 and 4 and amplifies it by a gain of 1 (assuming again that all  $R$  resistors are of equal value).

Though this may appear to be a cumbersome way to build a differential amplifier, it has the distinct advantages of possessing extremely high input impedances on the  $V_1$  and  $V_2$  inputs, and adjustable gain that can be set by a single resistor. The voltage gain of the instrumentation op amp is

$$A_V = \left( 1 + \frac{2R}{R_{gain}} \right) \quad (5.18)$$

The overall gain can still be adjusted by changing the values of the other resistors as well as  $R_{gain}$ , but this would necessitate balanced resistor value changes for the circuit to remain symmetrical.

Note that the lowest gain possible with the circuit shown in Figure 5-12 is obtained with  $R_{gain}$  completely open (infinite resistance) and that the gain value is 1 (Kuphaldt, 2003).

One of the major problems with this circuit implementation using discrete components is that, even with precision resistors, imbalances remain and the common-mode rejection ratio (CMRR) remains poor. Fortunately, instrumentation amplifiers are available as commercial integrated circuits with laser-trimmed internal resistors to maximize the CMRR. These include the AD524 and AD624 from Analog Devices as well as the LM623 from National Semiconductor. The CMRR for the AD524 increases from 90 dB at unity gain up to 120 dB at a gain of 1000.

### 5.3.2.6 Charge Amplifier

A charge amplifier has as its input a capacitance that provides an extremely high input impedance at low frequencies. Contrary to what the name suggests, charge amplifiers do not amplify electric charge but convert the input charge into a voltage and present it as a low impedance output. It should therefore be called a charge-to-voltage converter. Common applications include capacitive accelerometers and piezoelectric sensors.

In effect, a charge amplifier consists of a high-gain inverting voltage amplifier with a metal–oxide–semiconductor field-effect transistor (MOSFET) or junction gate field-effect transistor (JFET) at its input to achieve a sufficiently high input impedance (see Figure 5.13).

In this schematic,  $C_t$  is the sensor capacitance,  $C_c$  is the cable capacitance,  $C_r$  is the feedback capacitor,  $R_t$  is the time constant resistor (or insulation resistance of range capacitor),  $R_i$  is the insulation resistance of input circuit (cable and sensor),  $q$  is the charge generated by sensor, and  $A_V$  is the open-loop gain of the op amp.

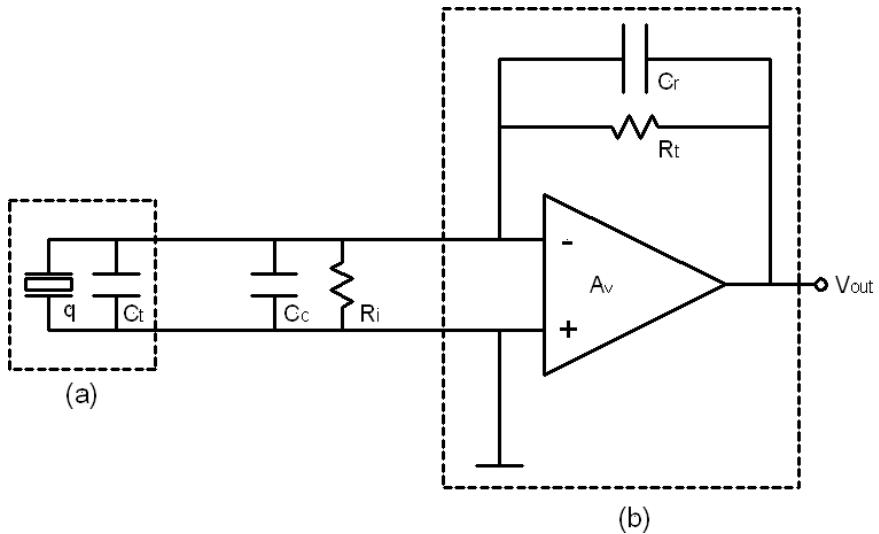
Neglecting the effects of  $R_t$  and  $R_i$ , the resulting output voltage,  $V_{out}$ , becomes

$$V_{out} = \frac{-q}{C_r} \frac{1}{1 + \frac{1}{A_V C_r} (C_t + C_r + C_c)} \quad (5.19)$$

For sufficiently high open-loop gain, the cable and sensor capacitance can be neglected, leaving the output voltage dependent only on the input charge and the range capacitance

$$V_{out} = -\frac{q}{C_r} \quad (5.20)$$

**FIGURE 5-13** ■  
Schematic showing  
(a) piezoelectric  
accelerometer  
connected to (b) a  
charge amplifier.



In short, the amplifier acts as a charge integrator that balances the sensor's electrical charge with a charge of equal magnitude and opposite polarity. This ultimately produces a voltage across the range capacitor. In effect, the purpose of the charge amplifier is to convert the high impedance charge input,  $q$ , into a usable output voltage,  $V_{out}$ .

Two of the more important considerations in the practical use of charge amplifiers are time constant and drift. The time constant,  $\tau_c$ , is defined as the discharge time of an alternating current (AC)-coupled circuit. In a period of time equivalent to one time constant, a step input will decay to 37% of its original value. The time constant of a charge amplifier is determined by the product of the range capacitor,  $C_r$ , and the time constant resistor,  $R_t$ :

$$\tau_c = R_t C_r \quad (5.21)$$

Drift is defined as an undesirable change in output signal over time and is independent of the input signal. In a charge amplifier it can be caused by low insulation resistance at the input,  $R_i$ , or by leakage current of the input MOSFET or JFET in the op amp.

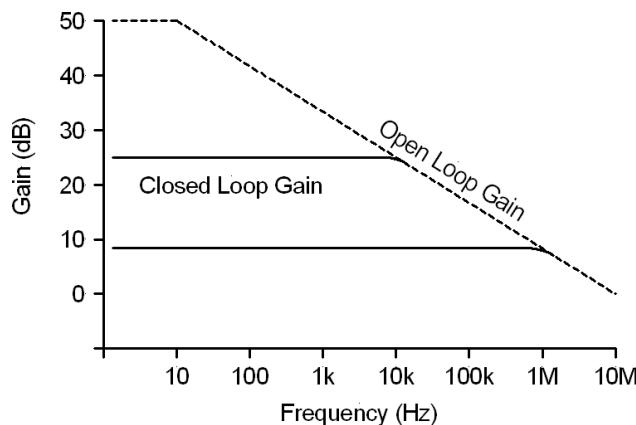
Drift and time constant simultaneously affect a charge amplifier's output. One or the other will be dominant. Either the charge amplifier output will drift toward saturation (power supply rail) at the drift rate, or it will decay toward zero at the time constant rate.

### 5.3.3 Practical Considerations

It must be remembered that op amps do not really have infinite gain or infinite input impedances, among other problems.

In addition, real op amps exhibit an imbalance caused by a small mismatch between the input transistors, which results in unequal bias currents flowing through the input terminals and offsets the output voltage. This offset can be balanced out by introducing a compensating bias voltage.

As mentioned earlier, one of the main uses of differential amplifiers in biomechatronic applications is to eliminate common-mode voltages. In reality, this rejection is not perfect



**FIGURE 5-14** ■ Bandwidth and closed-loop gain of an op amp.

because of circuit imbalances, and some of the common-mode signal leaks through to the output. Op amp specifications include their performance in this regard. The CMRR is defined as the difference between the differential mode gain and the common-mode gain of the amplifier.

Because of internal compensation capacitors within the op amp that keep it stable, gain rolls off with frequency, as shown in Figure 5-14. This limits the highest frequency that can be amplified by an op amp and the maximum effective closed-loop gain.

### 5.3.4 Op Amp Specifications

There are literally hundreds of op amp models to choose from. Many sell for less than \$1 a piece, while special-purpose instrumentation and radio frequency (RF) op amps may be quite a bit more expensive. Table 5-2 lists the performance of a range of general purpose and specialist op amps.

**TABLE 5-2** ■ Some Common Op Amp Specifications

Model	Package	Supply Voltage (min/max)	Bandwidth (MHz)	Bias Current (nA)	Slew Rate (V/ $\mu$ s)	Output Current (mA)
TL082	dual	12/36	4	8	13	17
LM301A	single	10/36	1	250	0.5	25
LM318	single	10/40	15	500	70	20
LM324	quad	3/32	1	45	0.25	20
LF353	dual	12/36	4	8	12	20
LF356	single	10/36	5	8	12	25
LF411	single	10/36	4	20	15	25
LM741C	single	10/36	1	500	0.5	25
LM833	dual	10/36	15	1050	7	40
LM1458	dual	6/36	1	800	10	45
CA3130	single	5/16	15	0.05	10	20
CLC404	single	10/14	232	44000	2600	70
CLC425	single	5/14	1900	40000	350	90
LM12CL	single	15/80	0.7	1000	9	13000
LM7171	single	5.5/36	200	12000	4100	100

There is substantial variation in performance between some of these units. Take, for instance, the parameter of input bias current: The CA3130 wins the prize for lowest at 0.05 nA (or 50 pA), and the LM833 has the highest at slightly over 1  $\mu$ A. The model CA3130 achieves its incredibly low bias current through the use of MOSFET transistors in its input stage. One manufacturer advertises the CA3130's input impedance as 1.5 terra-ohms, or  $1.5 \times 10^{12} \Omega$ . Other op amps shown here with low bias current figures use JFET input transistors, whereas the high bias-current models use bipolar input transistors.

While the 741 is specified in many electronic circuit schematics, its performance has long been surpassed by other devices. Even some designs originally based on the 741 have been improved over the years to beat the original design specifications. Op amps with JFET and MOSFET input transistors far exceed the 741's performance in terms of bias current and generally manage to beat the 741 in terms of bandwidth and slew rate as well.

When low bias current is a priority (such as in low-speed integrator circuits), choose the CA3130. For general-purpose DC amplifier work, the LM1458 offers good performance. For an upgrade in performance, choose the LF353, as it is a pin-compatible replacement for the LM1458. The LF353 is designed with JFET input circuitry for very low bias current and has a bandwidth four times greater than the LM1458, although its output current limit is lower (but still short-circuit protected).

If low power supply voltage is a requirement, the model LM324 is suitable, as it functions on as low as 3 volts DC. Its input bias current requirements are also low, and it provides four op amps in a single 14-pin package. Its major weakness is speed, limited to 1 MHz bandwidth and an output slew rate of only 0.25 volts per  $\mu$ s. For high-frequency AC amplifier circuits, the LM318 is a good general-purpose device.

Special-purpose op amps are available for modest cost, which provide better performance specifications. Many of these are tailored for a specific type of performance advantage, such as maximum bandwidth or minimum bias current. The CLC402 and CLC425 have bandwidths of 250 MHz and 1.9 GHz, respectively. In both cases, high speed is achieved at the expense of high bias currents and restrictive power supply voltage ranges.

The last two op amps in the table provide high output current capabilities for driving low impedance loads.

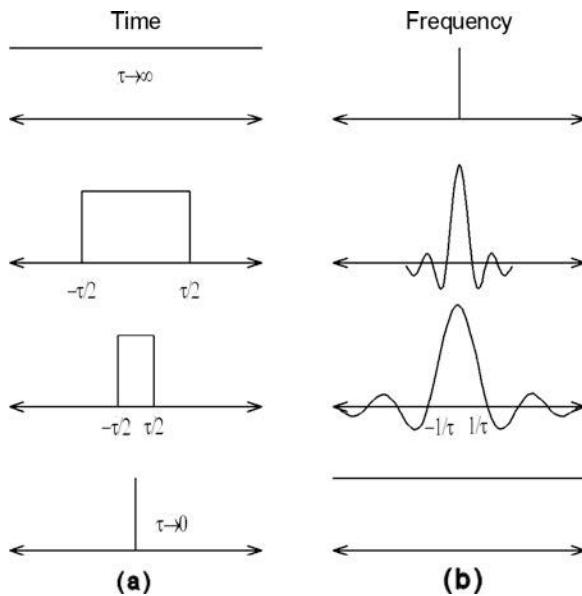
## 5.4 | ANALOG SIGNAL PROCESSING

---

### 5.4.1 Frequency Content of a Signal

In the frequency domain, a continuous sinusoidal signal of infinite duration can be represented in terms of its position on the frequency continuum and its amplitude only. However, most practical signals are not of infinite duration, so there is some uncertainty in the measured frequency, which is represented in the frequency domain by a finite spectral width.

From a mathematical perspective, this is equivalent to windowing the continuous sinusoidal signal using a rectangular pulse of duration  $\tau(s)$ . Because windowing, or multiplication, in the time domain becomes convolution in the frequency domain, the continuous signal spectrum must be convolved by the spectrum of a rectangular pulse to obtain the spectrum of the windowed signal.



**FIGURE 5-15** ■ Mapping the relationship between the duration of a pulse and its spectrum (a) Time domain. (b) Frequency domain.

The spectrum of a rectangular pulse is the sinc function (Carlson, 1998)

$$F(\omega) = \tau \frac{\sin(\omega\tau/2)}{\omega\tau/2} \quad (5.22)$$

and the spectrum of a continuous sinusoidal signal is an impulse,  $\delta(\omega)$ , so the resultant convolution is just the sinc function.

It can be seen from equation (5.22) that as the duration of the signal decreases,  $\tau \rightarrow 0$ , its spectral width increases until, in the limit, when the signal can be represented by an impulse  $\delta(t)$ , the spectral width is infinite. This relationship is shown graphically in Figure 5-15.

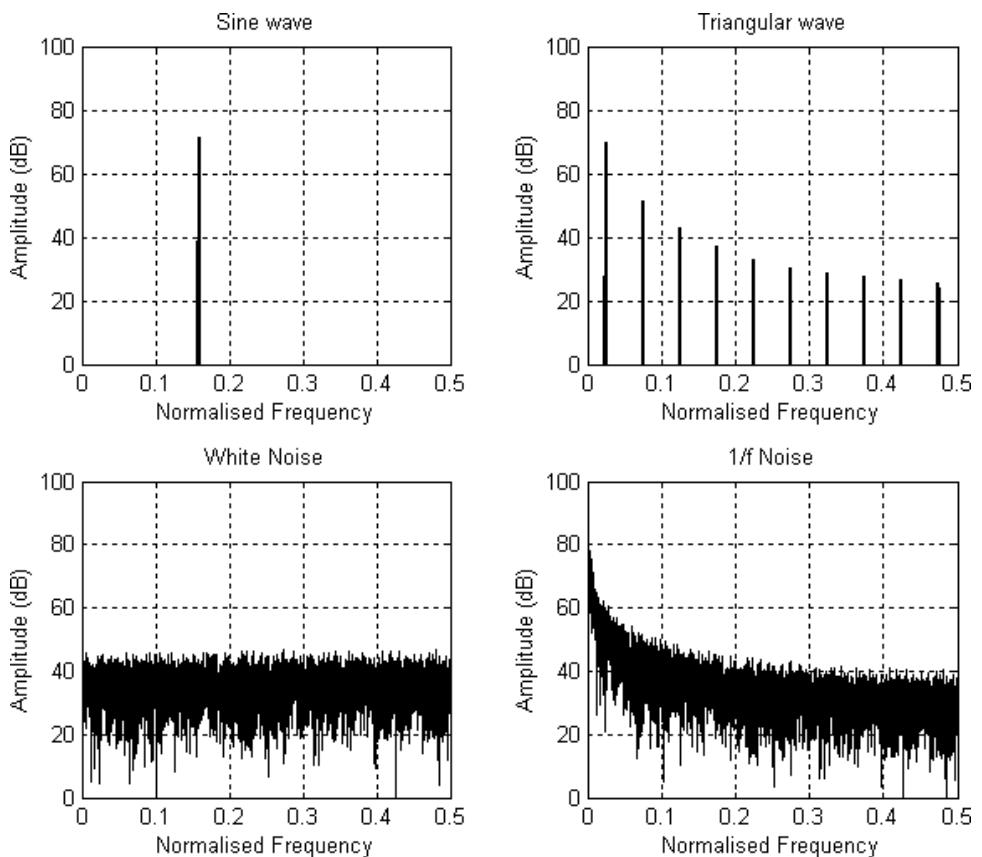
More complex signals can usually be made up of a number of sinusoidal components of varying amplitudes. These can be calculated using the Fourier series, so it is often easier to identify the spectrum of a time-domain signal by processing it through a Fourier transform and then examining the amplitudes of the resultant components. Some examples of this process are shown in Figure 5-16.

As can be seen from the harmonics for the triangular wave signal shown in Figure 5-16, even though the series is infinite the coefficients decrease in amplitude and eventually become so small that their contribution is considered to be negligible. For example, the electrocardiogram trace shown in Figure 5-17, with a fundamental frequency of about 1.2 Hz, can be reproduced with 70 to 80 harmonics, which equates to a bandwidth of about 100 Hz (Carr, 1997).

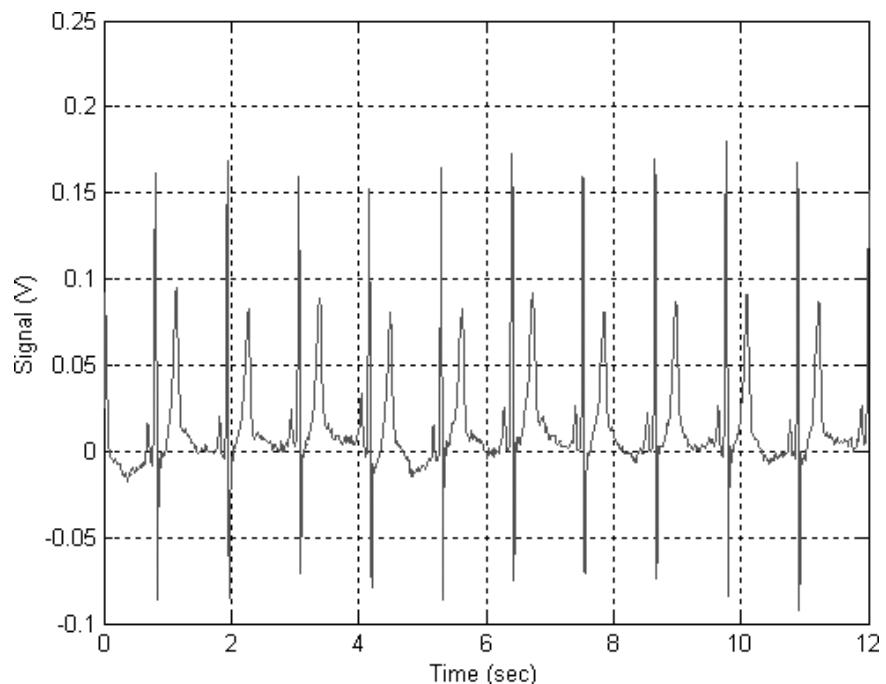
## 5.4.2 Analog Filters

The most common filter responses are the Butterworth, Chebyshev, Bessel, and elliptical types. Many other types are available, but most applications can be solved with one of these four. Butterworth ensures a flat response in the passband and an adequate rate of roll-off. Good all around, the Butterworth filter is simple to understand and suitable for applications such as audio processing.

**FIGURE 5-16 ■**  
Spectra of various types of signal and noise.



**FIGURE 5-17 ■**  
Typical electrocardiogram trace.



The Chebyshev gives a much steeper roll-off, but passband ripple makes it unsuitable for audio systems. It is superior for applications in which the passband includes only one frequency of interest (e.g., the derivation of a sine wave from a square wave, by filtering out the harmonics).

The Bessel filter gives a constant propagation delay across the input frequency spectrum. Therefore, applying a square wave (consisting of a fundamental and many harmonics) to the input of a Bessel filter yields an output square wave with no overshoot (all the frequencies are delayed by the same amount). Other filters delay the harmonics by different amounts, resulting in an overshoot on the output waveform.

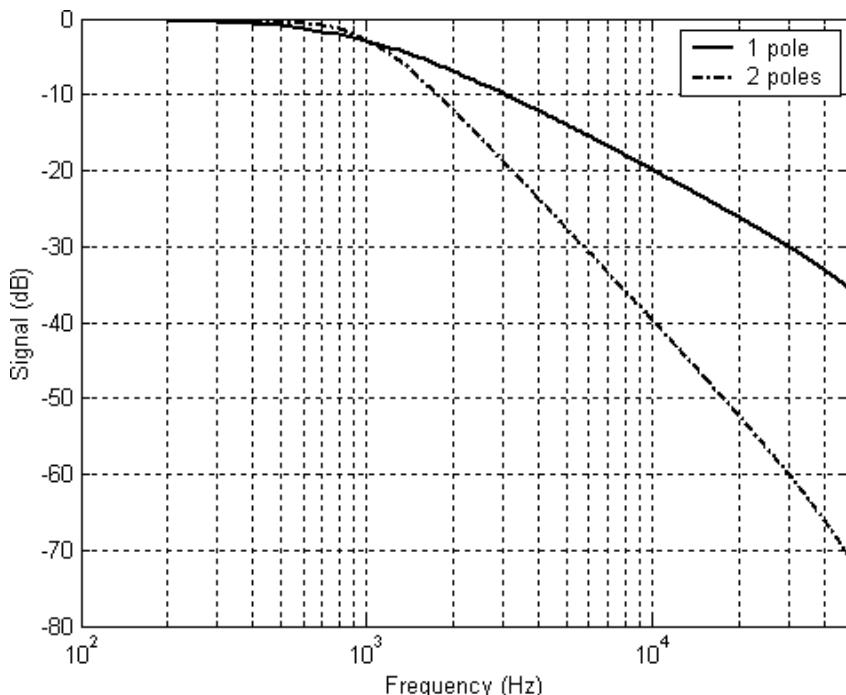
The elliptical filter is more complicated than the other types, but because it has the steepest roll-off it is often used in antialiasing filters.

#### 5.4.2.1 Low-Pass Filter

A low-pass filter is a filter that passes low frequencies and attenuates high frequencies, as shown in Figure 5-18. The amplitude response of a low-pass filter is flat from DC or near DC to a point where it begins to roll off. A standard reference point for this roll-off is the point where the amplitude has decreased by 3 dB, to 70.7% of its original amplitude (volts).

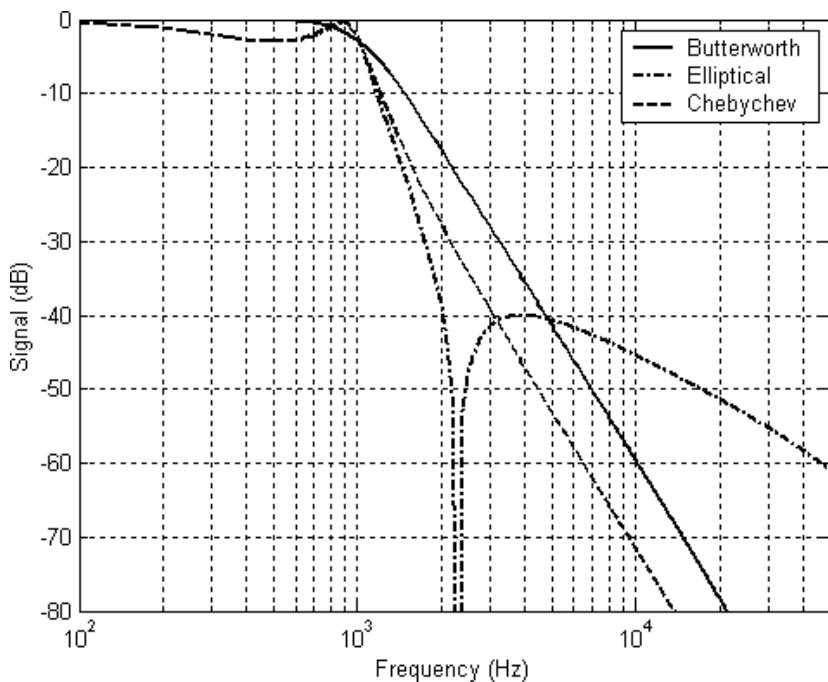
The region from around DC to the point where the amplitude is down 3 dB is defined as the passband of the filter. The range of frequencies from the 3 dB point to infinity is defined as the stopband of the filter.

The amplitude of the filter response at ten times the 3 dB frequency is attenuated a total of 20 dB for a single-pole filter and a total of 40 dB for a two-pole Butterworth filter. At higher frequencies, the amplitude continues to roll off in a linear fashion, where the slope of the line is  $-20$  dB per decade (10 times frequency) for a single-pole filter and  $-40$  dB per decade for a two-pole filter.



**FIGURE 5-18** ■ Frequency response of Butterworth low-pass filters.

**FIGURE 5-19** ■ Comparison among the frequency response of Butterworth, Chebyshev, and elliptical low-pass filters.



The term Butterworth refers to a type of filter response, not a type of filter. It is sometimes called the maximally flat approximation, because for a response of order  $n$  the first  $(2n - 1)$  derivatives of the gain with respect to frequency are zero at frequency = 0. There is no ripple in the passband, and DC gain is maximally flat.

The term Chebyshev also refers to a type of filter response, not a type of filter. It is sometimes referred to as an equal-ripple approximation. It features superior attenuation in the stopband at the expense of ripple in the passband, as shown in Figure 5-19. Generally, the designer will choose a ripple depth of between 0.1 and 3 dB. Chebyshev filter response, therefore, is not limited to a single value of response.

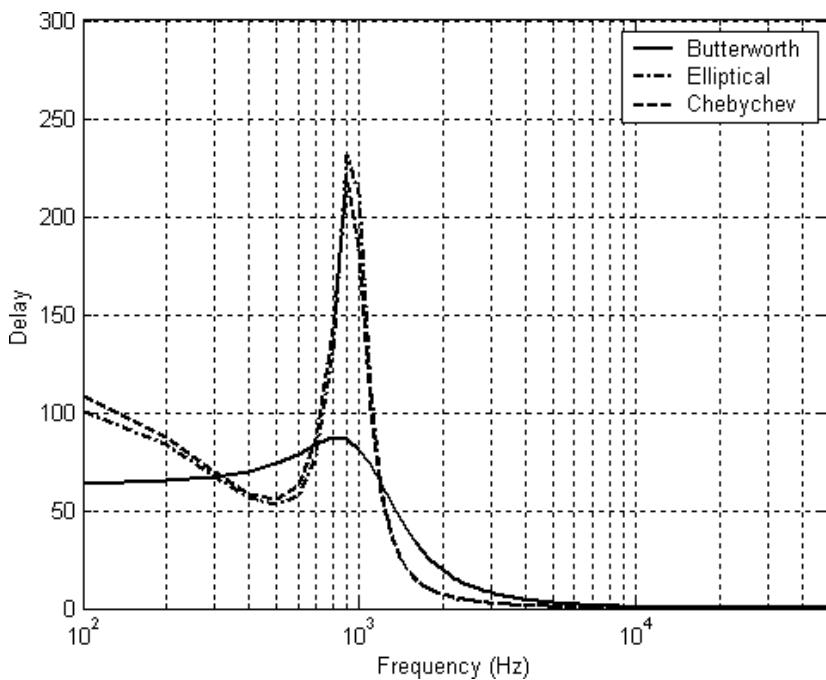
Elliptical filters have the best roll-off characteristics even for low-order filters, as shown in Figure 5-19. This characteristic makes them very useful for antialiasing filters prior to digitization or to remove clock spurs in direct digital synthesis (DDS) systems. In these applications, the filter zeros are generally placed at multiples of the clock frequency.

Unfortunately, all of these filter types suffer from group delay problems, as shown in Figure 5-20. However, very few filters are designed with square waves in mind because most of the time the signals filtered are sine waves, or close enough that the effect of harmonics can be ignored. If a waveform with high harmonic content is filtered, such as a square wave, the harmonics can be delayed with respect to the fundamental frequency, and distortion will result.

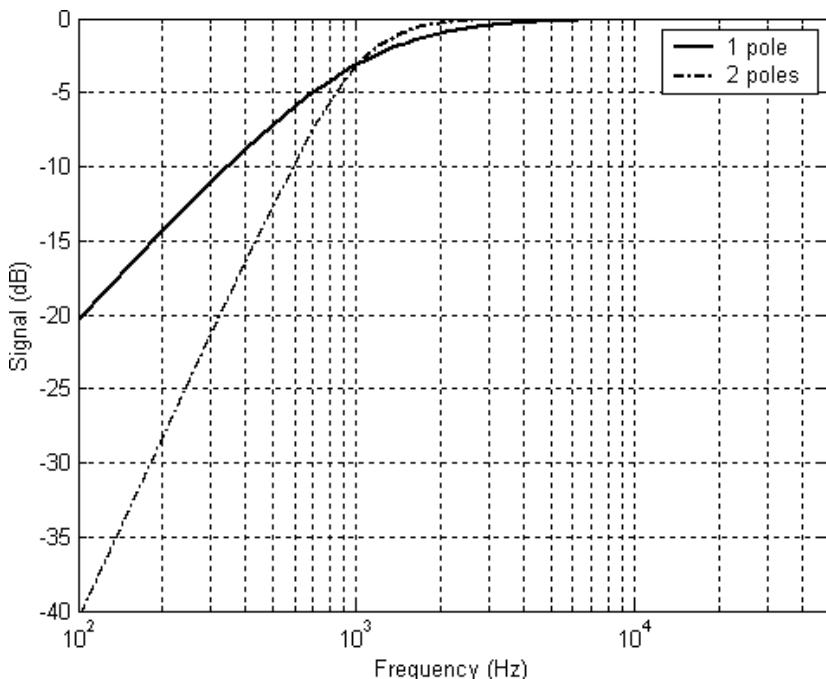
To counteract this problem, filters with a Bessel response are used. This response features flat group delay in the passband, the characteristic of Bessel filters that makes them valuable to digital designers.

#### 5.4.2.2 High-Pass Filters

A high-pass filter is a filter that passes high frequencies and attenuates low frequencies, as illustrated in Figure 5-21. The amplitude response of a high-pass filter is flat from infinity



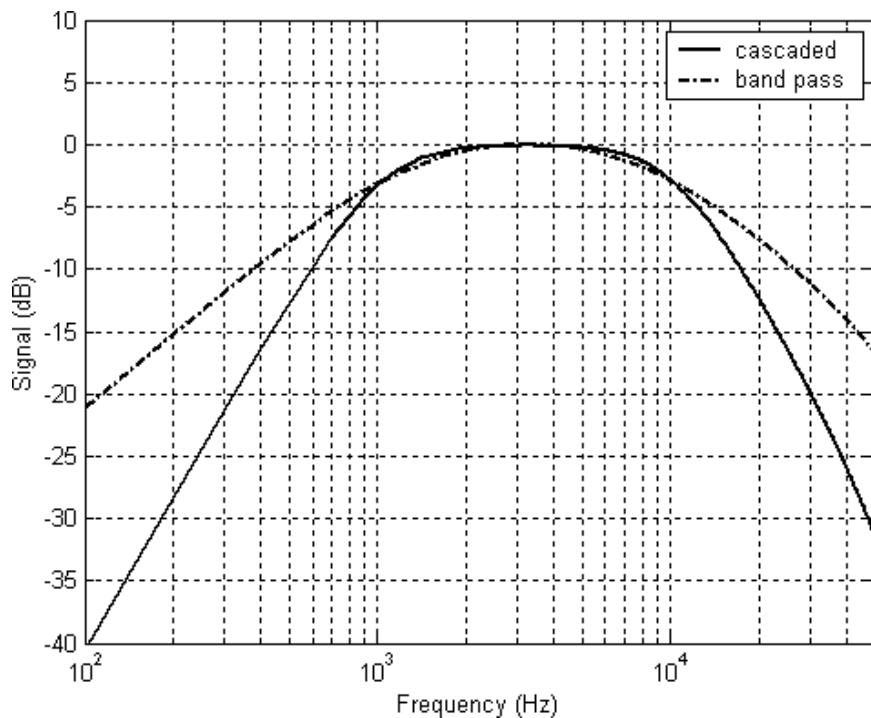
**FIGURE 5-20** ■ Comparison of the group delay of Butterworth, Chebyshev, and Bessel low-pass filters.



**FIGURE 5-21** ■ Frequency response of Butterworth high-pass filters.

down to a point where it begins to roll off. A standard reference point for this roll-off is the point where the amplitude has decreased by 3 dB, to 70.7% of its original amplitude. The region from infinity to the point where the amplitude is down 3 dB is defined as the passband of the filter. The range of frequencies from the 3 dB point down to zero (or near zero) is defined as the stopband of the filter.

**FIGURE 5-22** ■ Band-pass filter frequency response—second-order band-pass compared with cascaded second-order high- and low-pass stages.



The amplitude of the filter at  $1/10$  the 3 dB frequency is attenuated a total of 20 dB for a one-pole filter and a total of 40 dB for a two-pole Butterworth filter. At lower frequencies, the amplitude continues to roll off in a linear fashion, where the slope of the line is  $-20$  dB per decade (10 times frequency) for a single-pole filter and  $-40$  dB per decade for a two-pole filter.

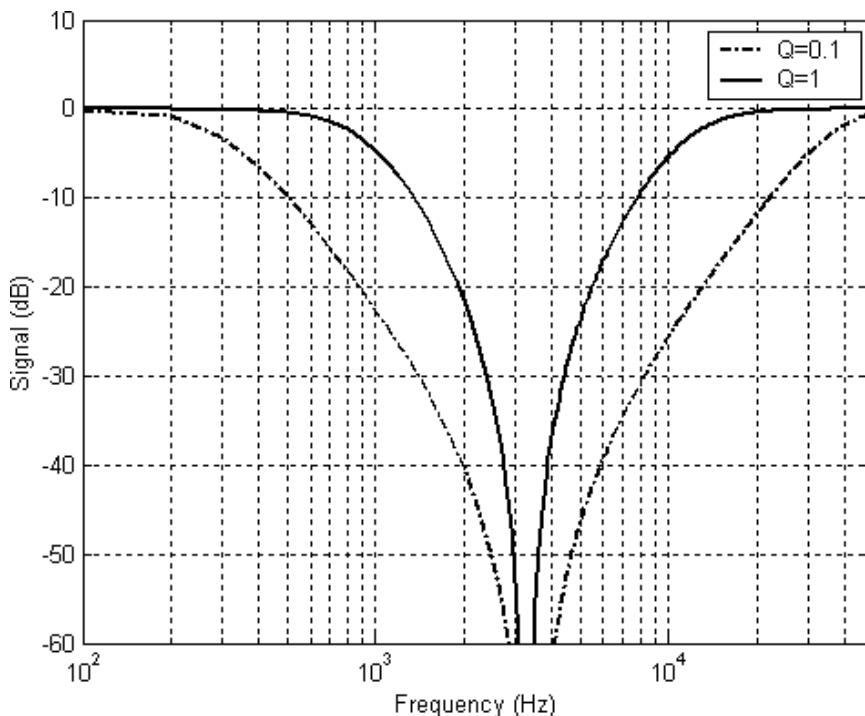
#### 5.4.2.3 Band-Pass Filters

Band-pass filters are those that pass a range of frequencies above a minimum and below a maximum, as shown in Figure 5-22. These filters can be made using cascaded band-pass stages or by cascading high- and low-pass sections:

- The figure shows the effect of cascading double-pole Butterworth filters.
- Both the high-pass and low-pass filters must have the same gain.
- The high-pass stage must come first, followed by the low-pass. In this way, high-frequency noise from the high-pass filter will be attenuated by the low-pass filter.
- Cascaded high-pass and low-pass filters probably take the same number of op amps as cascading band-pass filters, yet the response is clearly sharper, giving a double-pole characteristic at the low- and high-frequency 3 dB points instead of 6 dB of roll-off caused by cascading two band-pass filters, each with single-pole response.

#### 5.4.2.4 Notch and Band Reject Filters

A notch filter passes all frequencies except those in a stopband centered on a specific frequency. High Q notch filters eliminate a single frequency or narrow band of frequencies, while a band reject filter eliminates a wider range of frequencies.



**FIGURE 5-23** ■  
Notch filter frequency responses.

The amplitude response of a notch filter is flat at all frequencies except at the stopband on either side of the center frequency. The standard reference points for the roll-offs on each side of the stopband are the points where the amplitude has decreased by 3 dB, to 70.7% of its original amplitude.

The  $-3$  dB points and  $-20$  dB points are determined by the size of the stopband in relation to the center frequency—in other words, the  $Q$  of the filter. The  $Q$  is the center frequency divided by the bandwidth.

In Figure 5-23, the following points should be noted:

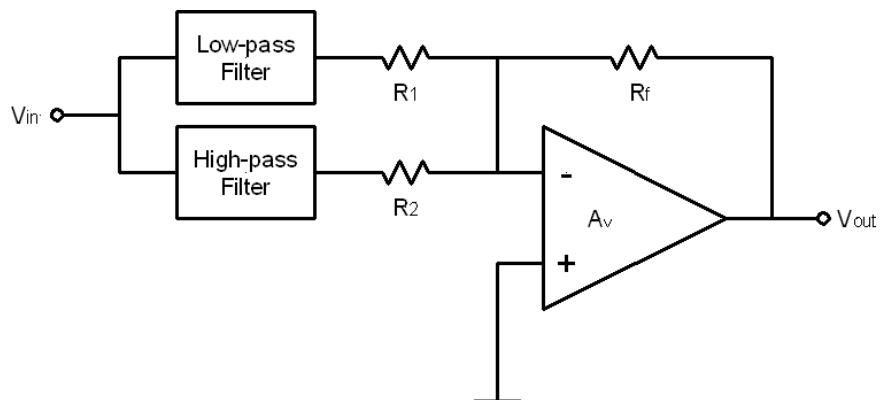
- The  $-3$  dB points are at about 300 Hz and 30 kHz for a  $Q$  of 0.1 and a center frequency of 3.1 kHz. At 1/10 the bandwidth, the amplitude is down 20 dB.
- The response of the band-pass filter with a  $Q = 1$  is also shown. The  $-3$  dB points are at about the same frequencies as the  $-20$  dB levels for a  $Q$  of 0.1.
- At very high  $Q$  values the response of the circuit will begin to have overshoot and undershoot, which will destroy the integrity of the notch. The frequency that was supposed to be rejected may actually be amplified.

A band-reject filter is constructed by summing the outputs of two parallel low-pass and high-pass filters, as illustrated in Figure 5-24.

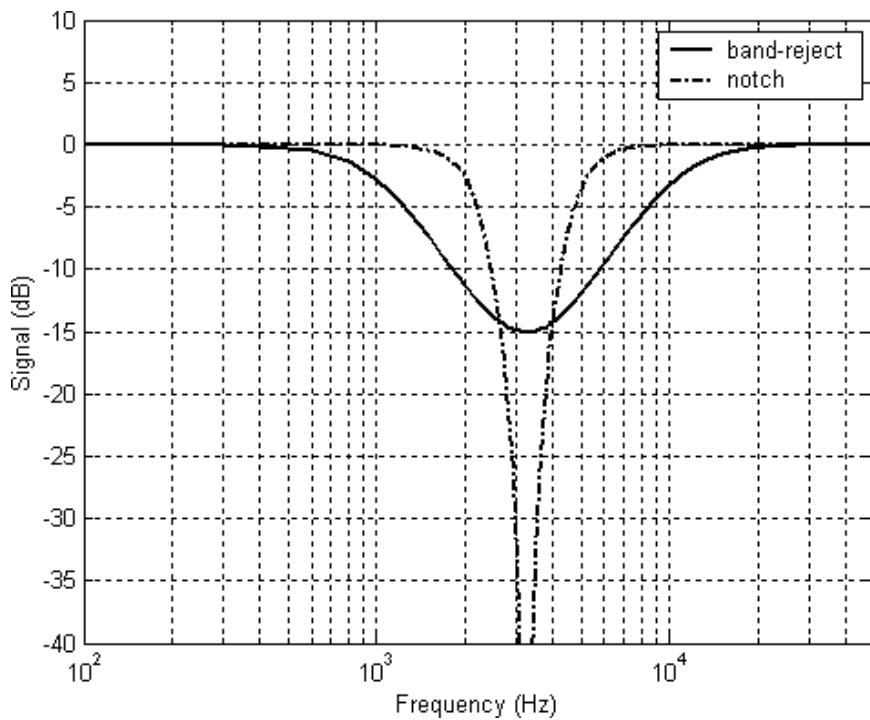
Some points are worth noting in regard to the relative merits of notch and band-reject filters, as shown in Figure 5-25:

- The performance increase that comes with summing low-pass and high-pass filter outputs comes at the expense of an additional op amp, the op amp that performs the summing function.

**FIGURE 5-24** ■ Construction of a band-reject filter.



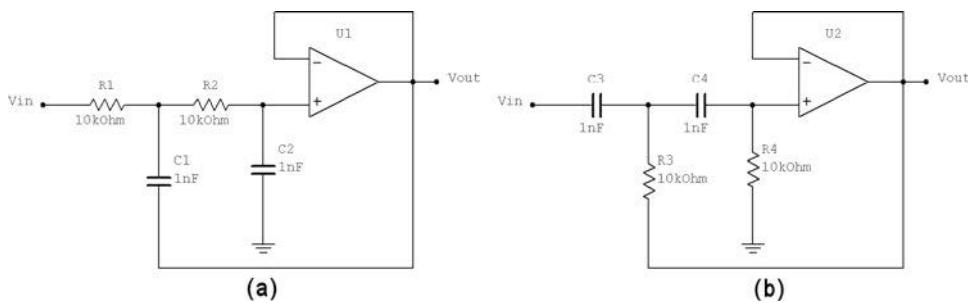
**FIGURE 5-25** ■ Comparison between a band-reject and notch filter.



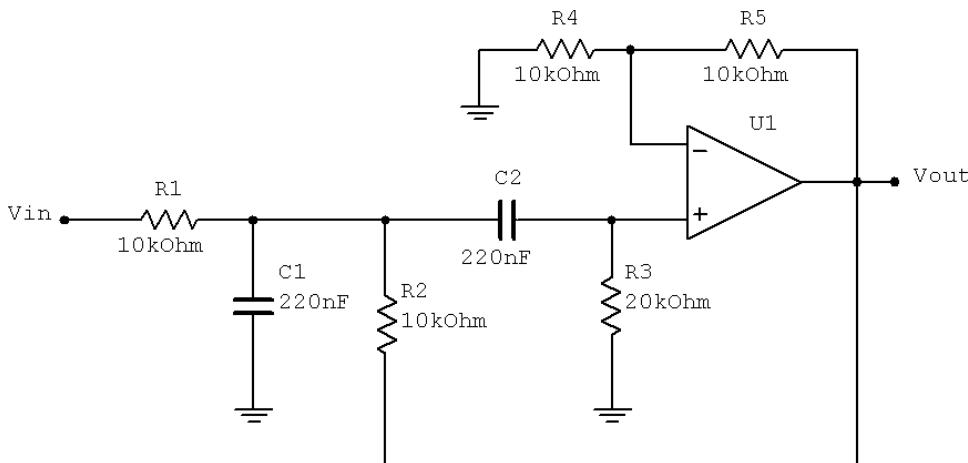
- Higher-order low-pass and high-pass filters will improve the performance of the band-reject filter.
- The farther apart the passbands are, the better the performance of the band-reject filter.

#### 5.4.2.5 Active Filter Implementation

A number of different topologies can be used to implement the filters discussed in the previous section. Some of these use the standard op amp feedback mechanisms. Practical realizations of analog filters are usually built from cascading second-order sections based on a complex conjugate pole pair or a pair of real poles. A first-order section is then added if an odd-order filter is required.



**FIGURE 5-26** ■  
Sallen Key filter topology. (a) Low pass. (b) High pass.



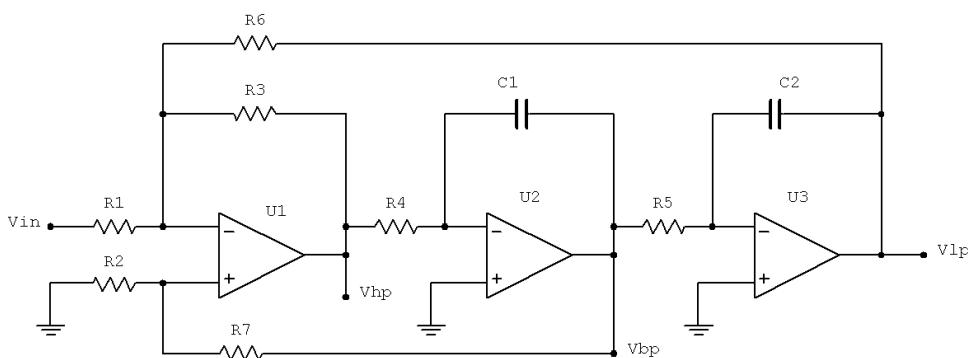
**FIGURE 5-27** ■  
Sallen Key band-pass filter.

One is the Sallen Key filter topology, which is particularly valued for its simplicity. The circuit produces a second-order (12 dB/octave) low-pass or high-pass response using only two capacitors and two resistors and a unity gain buffer, as shown in Figure 5-26.

The band-pass configuration is slightly more complex, and it is shown in Figure 5-27 for a nonunity gain configuration.

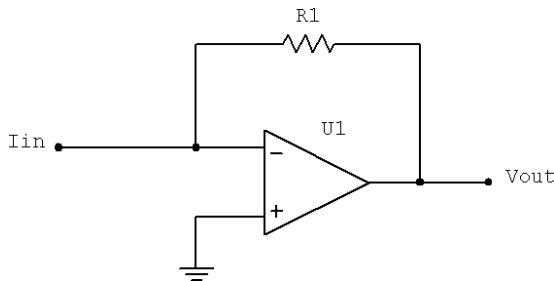
Higher-order filters can be obtained by cascading these building blocks.

One of the more interesting topologies is that of the state-variable filter. As shown in Figure 5-28, it consists of a number of cascaded integrators fed back into a summing amplifier. This configuration emulates the state-space model of a linear time invariant



**FIGURE 5-28** ■  
State-variable filter topology.

**FIGURE 5-29** ■ Current-to-voltage converter.



(LTI) system exactly. The outputs of the various integrators therefore correspond to the state-space model's state variables and become the low-pass, high-pass, and band-pass filter outputs.

Design of these filters is mostly done using any of a large number of web-based packages or applets supplied with electronic design automation (EDA) software.

### 5.4.3 Other Analog Circuits

#### 5.4.3.1 Current-to-Voltage Converter

Many sensor types output a current rather than a voltage and include photomultiplier- and photodiode-based devices. However, it is inconvenient to process currents using the analog techniques discussed in this section or to digitize them for further processing. The circuit shown in Figure 5-29, based on the inverting amplifier, is a simple method to convert a current to a voltage

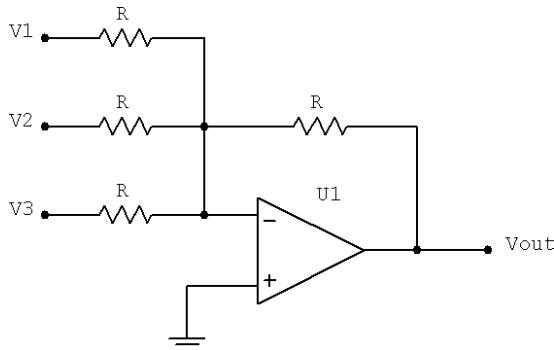
$$V_{out} = I_{in} R \quad (5.23)$$

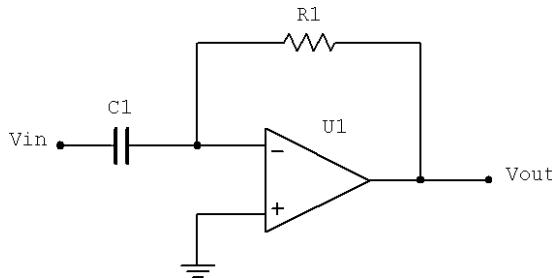
The advantage of using this circuit rather than a simple resistor is that the current flows into a virtual ground, eliminating the effects of the internal resistance of the current source.

#### 5.4.3.2 Summing Amplifier

The circuit shown in Figure 5-30 operates in a similar manner to the standard inverting amplifier except that the current at node A comprises the currents from the three inputs and that fed back from the output. With all resistor values equal, the currents through each of the three resistors will be proportional to their respective input voltages. Since those

**FIGURE 5-30** ■ Summing amplifier.





**FIGURE 5-31** ■ Differentiator.

three currents input add at the virtual ground node, the algebraic sum of those currents through the feedback resistor will produce a voltage at  $V_{out}$  equal to

$$V_{out} = -(V_1 + V_2 + V_3) \quad (5.24)$$

### 5.4.3.3 Integrator and Differentiator

By introducing electrical reactance into the feedback loops of op amp amplifier circuits, it is possible to cause the output to respond to changes in the input voltage over time. Drawing their names from their respective calculus functions, an integrator produces a voltage output proportional to the product of the input voltage and time, and a differentiator produces a voltage output proportional to the input voltage's rate of change.

An op amp circuit that determines the change in voltage by measuring current through a capacitor and outputs a voltage proportional to that current can be constructed, as shown in Figure 5-31.

The right-hand side of the capacitor is held to a voltage of 0 volts, due to the “virtual” ground. Therefore, current flowing through the capacitor is solely due to *change* in the input voltage. A steady input voltage won’t cause a current through C, but a changing input voltage will.

From a more rigorous perspective, consider first the current that flows through the feedback resistor to the virtual ground to produce a voltage drop across it. This will be identical to the output voltage. Therefore,

$$V_{out} = iR$$

From the charge relationship and the definition of capacitance, it can be shown that the voltage developed across a capacitor is equal to the integral of the current flowing into it

$$V_{cap} = \frac{1}{C} \int i dt$$

From Kirchhoff’s law, the two currents that flow into the negative input node must balance. Therefore, the current through the capacitor must equal the current through the resistor,  $i_{cap} = -i_{res}$ , which makes it

$$i_{cap} = -\frac{V_{out}}{R}$$

Substituting for this current into the integral gives

$$V_{cap} = -\frac{1}{RC} \int V_{out} dt$$

Taking the derivative of both the left-hand side and the right-hand side gives

$$\frac{dV_{cap}}{dt} = -\frac{1}{RC} V_{out}$$

Finally, because one terminal of the capacitor is the virtual earth  $V_{in} = V_{cap}$

$$V_{out} = -RC \frac{dV_{in}}{dt}. \quad (5.25)$$

A positive rate of change of input voltage will result in a steady negative voltage at the output of the op amp if the rate of change is linear. Conversely, a linear, negative rate of change of input voltage will result in a steady positive voltage at the output of the op amp. This polarity inversion from input to output is because the input signal is applied to the inverting input of the op amp so it acts like the inverting amplifier analyzed previously. The faster the rate of change of input voltage, the greater the voltage at the output.

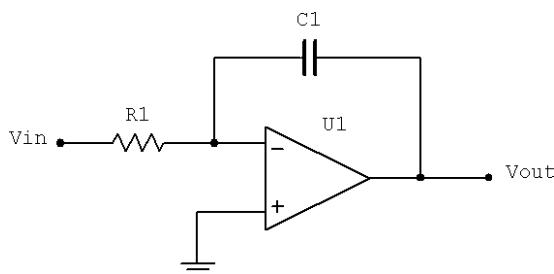
Applications for this besides representing the derivative function in an analog computer include rate-of-change indicators for process instrumentation. One such rate-of-change signal application might be for monitoring (or controlling) the rate of temperature change in an incubator. In a prosthetic arm, it could be used to limit the slew rate of the elbow joint.

To produce an integrator, the op amp circuit must generate an output voltage proportional to the magnitude and duration of the input voltage signal. Stated differently, a constant input signal will generate a certain rate of change in the output voltage. To achieve this, the capacitor and resistor in the previous circuit are simply swapped, as shown in Figure 5-32.

The negative feedback of the op amp ensures that the inverting input will be held at 0 volts (the virtual ground). If the input voltage is exactly 0 volts, there will be no current through the resistor; therefore, the capacitor will not be charged, and the output voltage will not change. However, the output voltage may not be zero if the capacitor has been previously charged.

If a constant positive voltage is applied to the input, though, the op amp output will fall negative at a linear rate in an attempt to produce the changing voltage across the capacitor necessary to maintain the current established by the voltage difference across the resistor. Conversely, a constant, negative voltage at the input results in a linear, rising (positive) voltage at the output. The output voltage rate of change will be proportional to the value of the input voltage.

**FIGURE 5-32 ■**  
Integrator.



The equation describing this relationship is

$$V_{out} = -\frac{1}{RC} \int V_{in} dt + V_c \quad (5.26)$$

where  $V_c$  is the initial charge on the capacitor at  $t = 0$ .

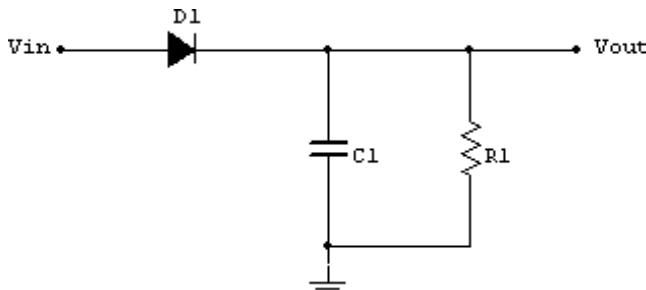
Real integrator circuits generally have a switch to short out the capacitor at the start of an integration cycle and sometimes also a large value resistor across it to drain away any residual voltage that may have accumulated due to leakage.

#### 5.4.3.4 Envelope Detection

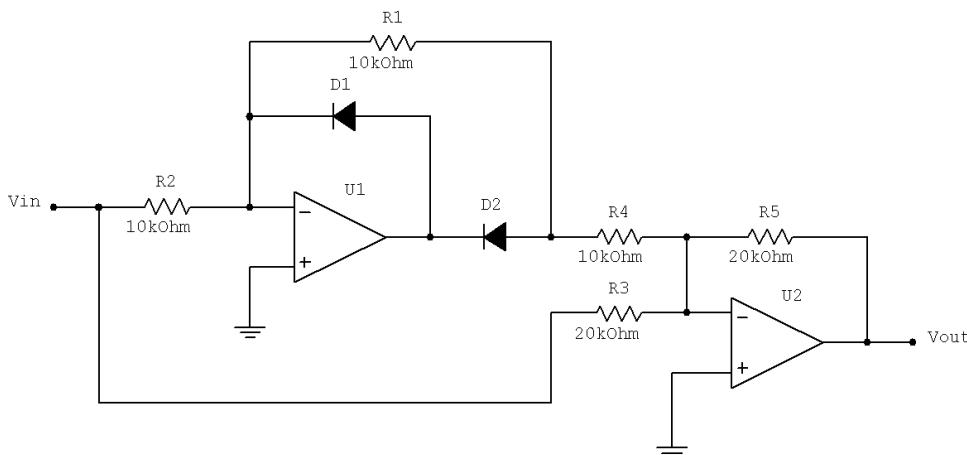
For an amplitude modulated signal, the envelope is a construct that joins the positive peaks of the signal to recreate the original modulation waveform. The process of envelope detection starts with rectification of the bipolar signal. If the signal amplitude is sufficiently large, a conventional half- or full-wave rectifier can be used—notwithstanding the volt drop across the diodes. A conventional envelope detector consisting of a diode and a low-pass filter is shown in Figure 5-33.

If the signal is small, though, then a precision full-wave rectifier such as the one shown in Figure 5-34 must be used.

This circuit is very common and has been around for many years. The tolerance of R2, 3, 4, and 5 is critical for good performance, and all four resistors should be 1% or



**FIGURE 5-33** ■ Conventional envelope detector.



**FIGURE 5-34** ■ Precision full-wave rectifier.

better. Note that the diodes have been reversed to obtain a positive rectified signal. The second stage inverts the signal polarity. To obtain improved high-frequency response, the resistor values should be reduced.

This circuit is sensitive to source impedance, so it is important to ensure that it is driven from a low impedance, such as an op amp buffer stage. The circuit has good linearity down to a couple of mV at low frequencies but has a limited high-frequency response. Use of high-speed diodes, lower resistance values, and faster op amps is recommended if greater sensitivity or higher-frequency response is required.

The output of the rectifier is passed through a low-pass filter to produce the envelope.

#### 5.4.3.5 Myoelectric Signal Processing

Developing a differential amplifier for EMG poses few real problems if the appropriate precautions are taken. When a muscle is caused to contract, the distribution of electrolytes within the tissue changes, which induces small voltages on the surface of the skin that can be picked up using the appropriate electrodes. From the specifications in Table 5-1, the voltage levels vary from  $50 \mu\text{V}$  up to a maximum of  $5 \text{ mV}$ . However, this signal will probably be contaminated by other much larger biopotential signals and certainly by induced *mains hum*.

The problem is to sense and isolate this signal so that it can be used to control the movement of a prosthetic device by driving a pneumatic artificial muscle (PAM), discussed in Chapter 3.

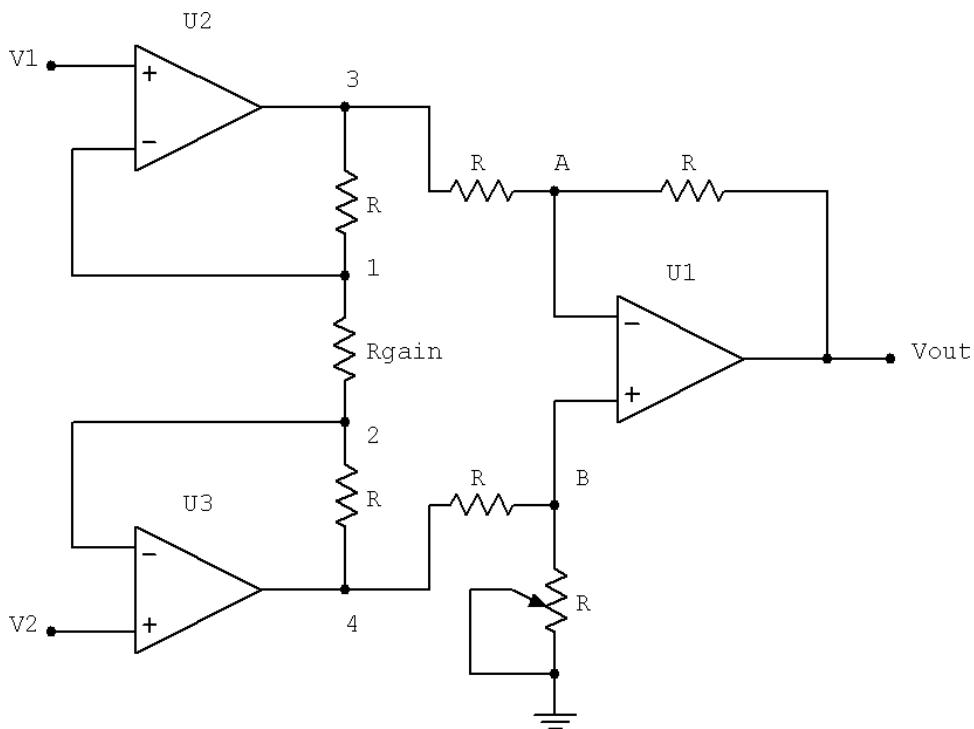
The first requirement is to select an appropriate set of electrodes to detect this small voltage with the addition of as few measurement artifacts as possible. Electrode types were discussed in Chapter 2, and from this discussion it is obvious that a silver–silver chloride surface electrode and a conductive gel are best. These electrodes can be applied directly over the muscle complex of interest. The electrode interface is reasonably high impedance and also picks up  $50 \text{ Hz}$  mains, ECG signals, and myriad other electrical noise generated by the body and from extraneous sources.

An instrumentation amplifier is required to provide a high input impedance, a good CMRR, and the appropriate gain to amplify the small signal to a usable level for further processing. This amplifier must be provided with a pair of differential inputs to sense the myoelectric signal and a ground reference, as shown in Figure 5-35.

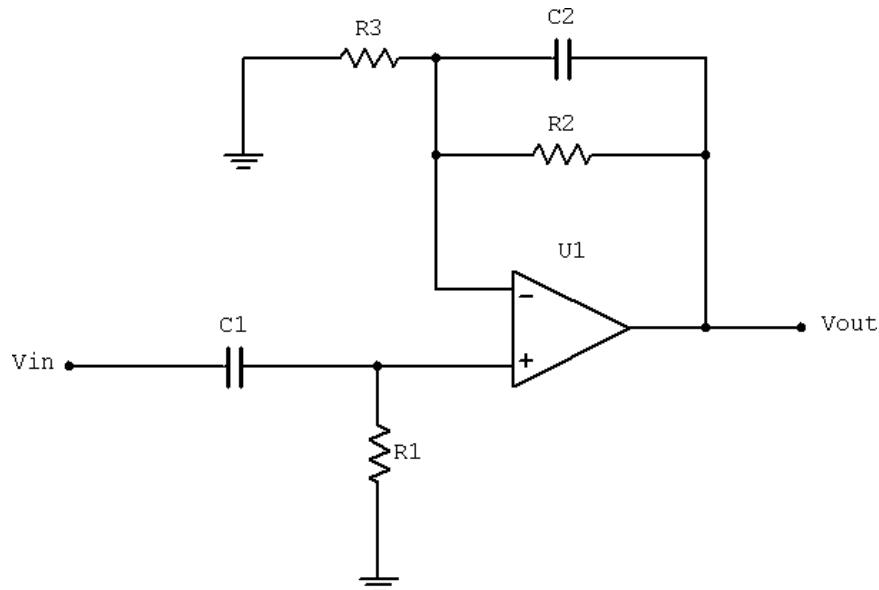
This instrumentation configuration can easily achieve a CMRR in excess of  $60 \text{ dB}$  to eliminate most of the common-mode  $50 \text{ Hz}$  while still providing a voltage gain of 100. This will provide an output of between  $5 \text{ mV}$  and  $500 \text{ mV}$ .

The next problem to be solved is that of electrode movement. As discussed in Chapter 2, electrode movement alters the cell potential of the conductive gel, which results in the generation of large electrical signals called motion artifacts. These can saturate any further stages of amplification and need to be removed.

Fortunately, motion artifacts are found at the lower end of the spectral response of the EMG signal and can therefore be removed by high-pass filtering without removing too much of the real signal. A first-order high-pass filter with a  $5 \text{ Hz}$  cutoff frequency realized by  $C_1$  and  $R_1$ , followed by an amplifier with a voltage gain of 10, can easily be designed to implement this function, as shown in Figure 5-36. Note that a small capacitor,  $C_2$ , is included in the feedback path to reduce the high-frequency gain of the amplifier. This eliminates high-frequency noise above the maximum bandwidth of the EMG signal.



**FIGURE 5-35** ■  
Instrumentation op amp with ground reference electrode.

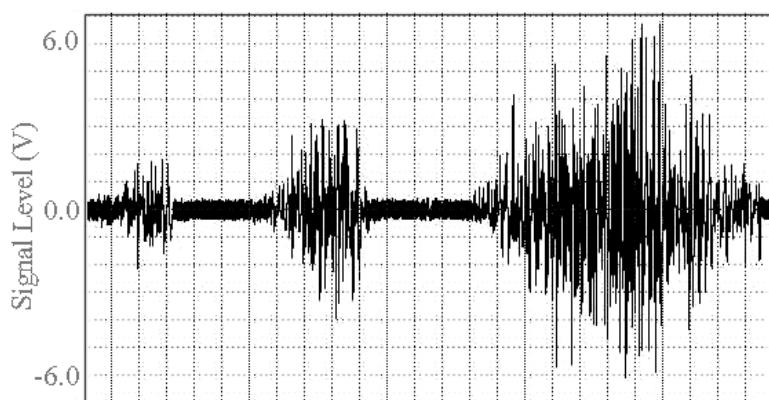


**FIGURE 5-36** ■  
First-order high-pass filter followed by further amplification.

At this stage, the EMG signal should have the features shown in Figure 5-37. Its frequency content should vary between 5 Hz and 300 Hz, and its amplitude should vary between 5 mV and a maximum of about 5 V.

The next step in the process is to extract the envelope of the EMG signal. A simple rectifier diode requires a turn-on voltage of 0.7 V, which is larger than a good proportion of

**FIGURE 5-37** ■  
EMG signal after  
amplification and  
filtering.



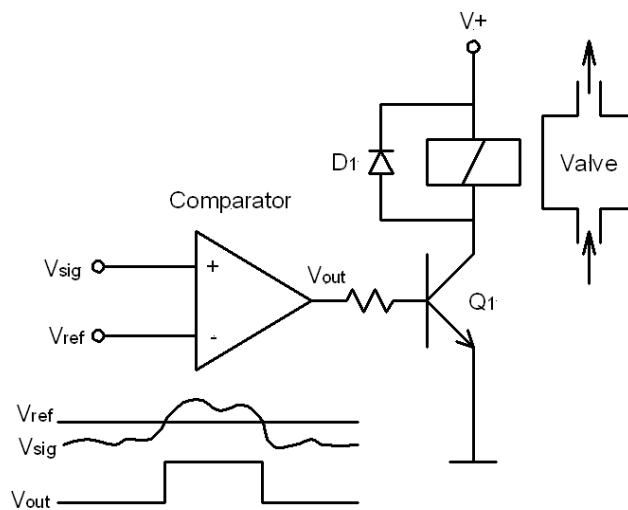
**FIGURE 5-38** ■  
Outputs from the  
rectifier and  
envelope detector.



the signal shown. Therefore, a precision full-wave rectifier shown in the previous section should be used. This is followed by a low-pass filter to produce only the signal envelope, as shown in Figure 5-38.

Various techniques are available for using this envelope to actuate pneumatic artificial muscles, and these are addressed in detail in Chapter 10. The simplest method is to open the pneumatic solenoid valve to the actuator if the envelope exceeds a threshold voltage and to keep it closed if the level is below this. A comparator, discussed in the following section, followed by a power transistor can be used to control the solenoid, as shown in Figure 5-39.

**FIGURE 5-39** ■  
Threshold  
voltage-based  
control of a  
pneumatic solenoid  
valve.



## 5.5 | DIGITAL SIGNAL PROCESSING

### 5.5.1 The Comparator

The comparator is the simplest analog-to-digital converter (ADC), as it converts an analog signal with an infinite number of amplitudes to a single-bit digital output. If the analog signal is above a given threshold the output is high, and if it is lower the output is low, as shown in Figure 5-40.

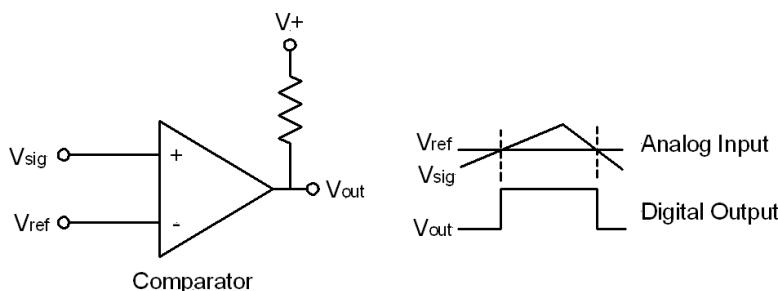
In reality, the analog signal almost always includes a small amount of noise, so as it nears the threshold level multiple triggering may occur. This can be eliminated by including some positive feedback from the comparator output to add hysteresis to the threshold level. This circuit configuration is referred to as a Schmitt trigger.

### 5.5.2 Signal Acquisition and Processing Overview

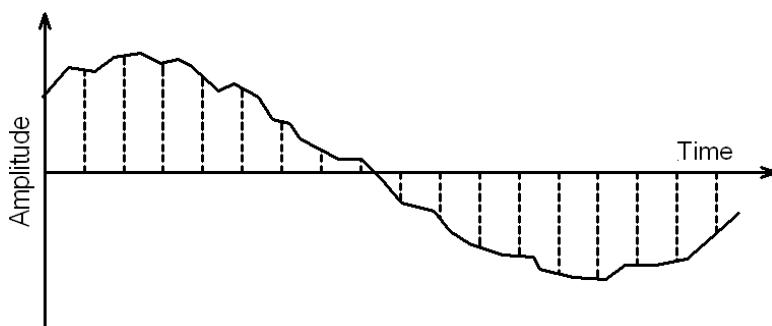
Processing signals within a computer (digital signal processing) requires that they be sampled periodically and then converted to a digital representation using an ADC. As a result, the continuous signal is reduced to being defined only at discrete points, as shown in Figure 5-41. The sampling theorem (also known as Shannon's sampling theorem) states that to ensure accurate representation the signal must be sampled at the Nyquist rate, which is defined as at least double the highest significant frequency component of the signal.

$$f_s \leq 2f_{max} \quad (5.27)$$

In addition, the number of discrete levels to which the signal is quantized must also be sufficient to represent variations in the amplitude to the required accuracy. Most ADCs quantize to 12 or 16 bits, which represent  $2^{12} = 4096$  or  $2^{16} = 65536$  discrete levels, respectively.

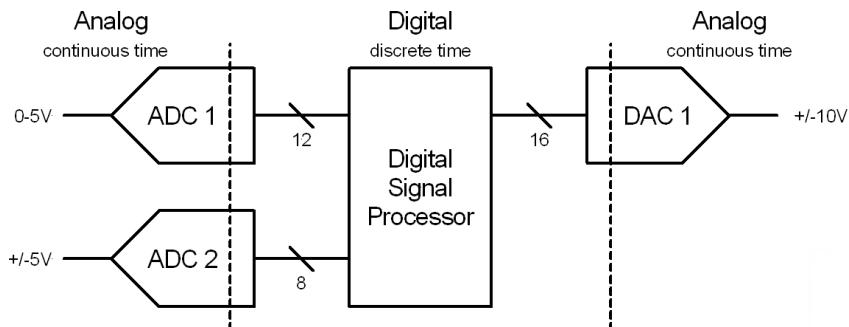


**FIGURE 5-40** ■  
Comparator circuit.



**FIGURE 5-41** ■  
Digitizing a signal.

**FIGURE 5-42** ■ Typical configuration for a DSP application.  
[Adapted from (Brooker 2008).]

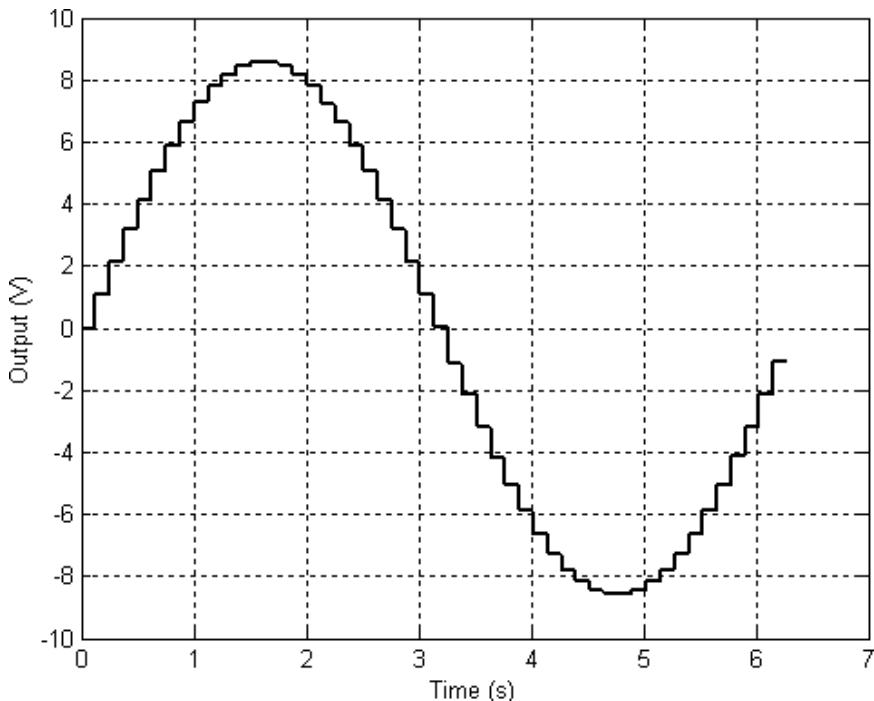


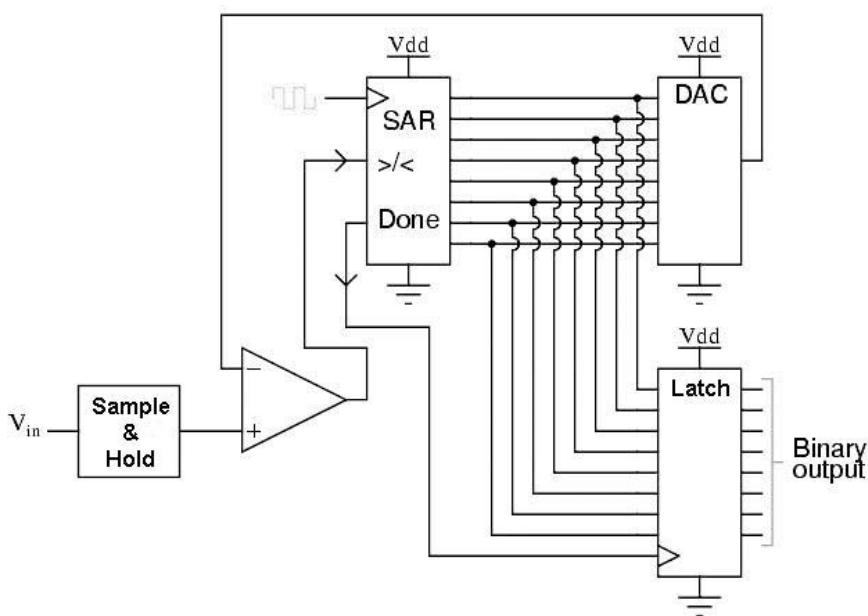
The signal can then be manipulated within a computer in various ways. This process often produces intermediate results that are much larger than the input values, requiring that a larger word size be used if fixed-point values are used; more commonly, a floating-point representation is used. Once the signal has been processed, the final result is often passed through a digital-to-analog converter (DAC) to change it from the internal binary representation back to a voltage. It is therefore possible to represent a complete digital signal processor (DSP), as shown in Figure 5-42, with ADCs a processor and DACs.

The DAC outputs new values only at discrete times, so a continuous signal needs to be reconstructed. This generally involves holding the signal constant (zero-order hold) during the period between samples, as shown in Figure 5-43. This signal is then cleaned up by being passed through a low-pass filter to remove high-frequency components generated by the sampling process.

Other common methods of interfacing to the outside world include manipulating a single-bit or a complete digital word to drive some peripheral device like a video display

**FIGURE 5-43** ■ Analog reconstruction of a sampled signal using a zero-order hold.





**FIGURE 5-44** ■  
Successive  
approximation ADC.

unit. Alternatively, it is possible to output clock signal at a specific frequency or with a particular duty cycle to represent a number.

### 5.5.3 ADCs and DACs

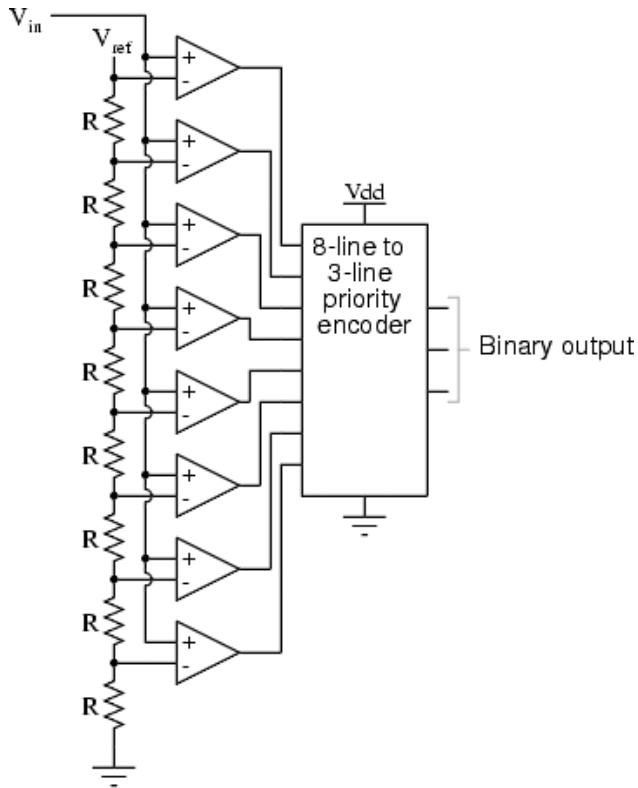
There are many different ways of converting an analog signal to its digital equivalent. These are selected depending on the number of digital bits required and the rate of conversion. Two of the most common, used in commercial designs, are the successive approximation and flash conversion, though others including single- and dual-slope integration, switched capacitor, and delta-sigma conversion are implemented for specialist applications.

The successive approximation method is used in bioelectrical applications, for relatively low conversion rates, because it is low cost. As can be seen in the block diagram shown in Figure 5-44, it consists of a number of modules including a DAC in the feedback path.

The process of conversion is as follows:

- At the start of the conversion, the analog signal is latched (captured and held) using a sample and hold (S&H).
- The successive approximation register (SAR) sets its most significant bit (MSB), which causes the DAC to output an analog value voltage equal to  $V_{max}/2$ .
- The DAC value is compared with the analog input value using a comparator.
- If the analog signal is still larger than this value, then the output of the comparator remains high and the bit remains set. If it is smaller, then the output of the comparator goes low and the bit is reset.
- The process is then repeated with the next most significant bit and then the one after, until at last the least significant bit (LSB) has been output.
- The successive approximation register then contains a digital equivalent of the analog input signal, which is clocked into the output register and latched so that the whole process can be repeated.

**FIGURE 5-45** ■ Flash ADC.



An  $n$ -bit ADC converter requires  $n \Delta T$  seconds to perform one conversion, where  $\Delta T$  is the cycle time for the DAC and control unit. Typical conversion times for 8- to 12-bit successive approximation ADCs range from 1 to 100  $\mu\text{s}$  (Alciatore and Histand, 2003).

The fastest ADCs are known as flash converters. As shown in Figure 5-45, these consist of a bank of comparators acting in parallel to identify the signal level. A resistive potential divider chain forms the references for each of the comparators, so that the moment an analog input is presented the respective comparator outputs will have encoded for it. A priority encoder then converts the comparator output format to a binary word.

As the name implies, DACs convert a digital word into an analog output. The simplest kind use an R–2R resistor ladder network connected to an inverting summer op amp circuit, as shown in Figure 5-46.

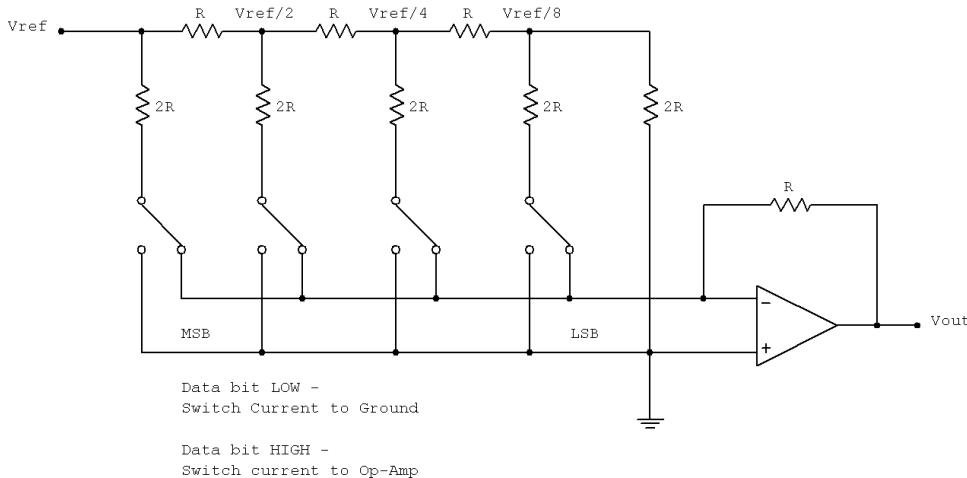
This configuration is convenient and accurate because it uses only two precision resistor values,  $R$  and  $2R$ , to produce a series of reference voltages, each half the size of the previous one with the MSB value being  $V_{REF}$  and the next  $V_{REF}/2$ , down to  $V_{REF}/8$  in this example.

The analog output is generated by switching the  $2R$  resistor, either to the negative input of the op amp (virtual ground) if the bit is high or to the true ground if the bit is low.

Consider the case where only the LSB is high, in which case  $D_o$  switches the voltage  $V_{REF}/8$  to the negative input of the op amp through  $2R$  and the remaining resistors are grounded.

If the gain of the op amp is  $A_V$

$$A_V = -\frac{R}{2R} = -\frac{1}{2} \quad (5.28)$$



**FIGURE 5-46 ■**  
Ladder resistor DAC.

then the output will be

$$\begin{aligned} V_{out0} &= \frac{V_{REF}}{8} A_V \\ &= -\frac{V_{REF}}{16} \end{aligned} \quad (5.29)$$

In a similar manner, if  $D_3$  is high, the MSB switches  $V_{REF}$  to the negative input, and the output of the op amp goes to  $V_{out3} = -V_{REF}/2$ . If both are switched, the output will be the sum of the two values, as discussed earlier in this chapter.

In the general case for a four-bit DAC with a binary input word  $b_3 b_2 b_1 b_0$

$$V_{out} = b_3 V_{out3} + b_2 V_{out2} + b_1 V_{out1} + b_0 V_{out0} \quad (5.30)$$

If  $V_{REF} = 10$  V, then  $V_{out}$  can vary from 0 V up to a maximum of

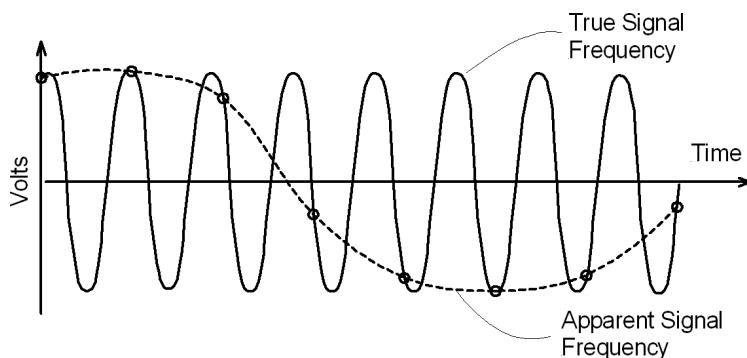
$$\begin{aligned} V_{out} &= -V_{REF} \left( \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} \right) \\ &= -\frac{15}{16} V_{REF} \end{aligned}$$

#### 5.5.4 Signal Aliasing

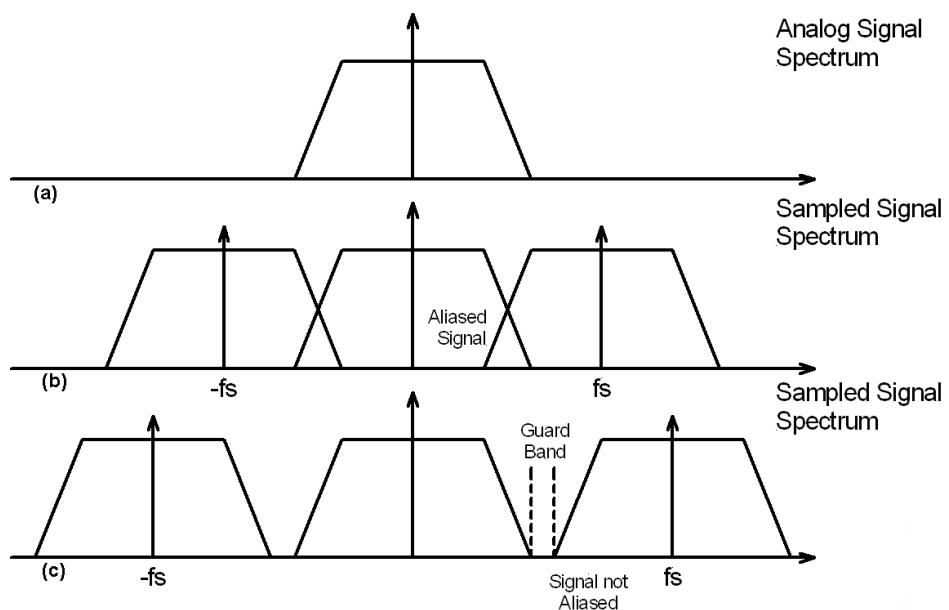
If the Nyquist criterion is not satisfied and the analog signal is not sampled at least twice the frequency of the highest-frequency component, then these high-frequency signals are folded, or *aliased*, down to a lower frequency, as illustrated in the time-domain representation shown in Figure 5-47.

In the frequency domain, a generic analog signal may be represented in terms of its amplitude and total bandwidth, as shown in Figure 5-48a. A sampled version of the same signal can be represented by a repeated sequence spaced at the sample frequency (generally denoted  $f_s$ ), as shown in Figures 5-48b and Figure 5-48c. If the sample rate is not sufficiently high, then the sequences will overlap and high-frequency components will appear at a lower frequency (albeit with reduced amplitude).

**FIGURE 5-47** ■ Interpretation of aliasing effects in the time domain.



**FIGURE 5-48** ■ Interpretation of aliasing effects in the frequency domain.  
[Adapted from (Brooker 2008).]

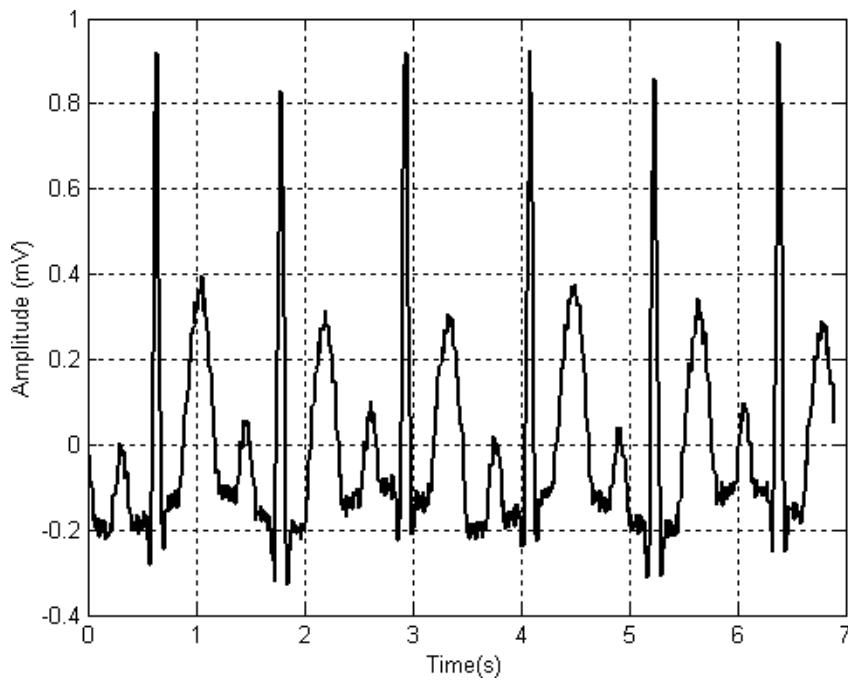


In reality, the finite roll-off of the filter response requires that a guard band be maintained between the spectra. This is achieved by selecting an antialiasing filter with a cutoff frequency that is less than 0.4 times the sample frequency. Using this ratio as a rule of thumb, a typical third-order low-pass filter will attenuate these unwanted signals by between 40 and 60 dB (1/100 to 1/1000) in voltage.

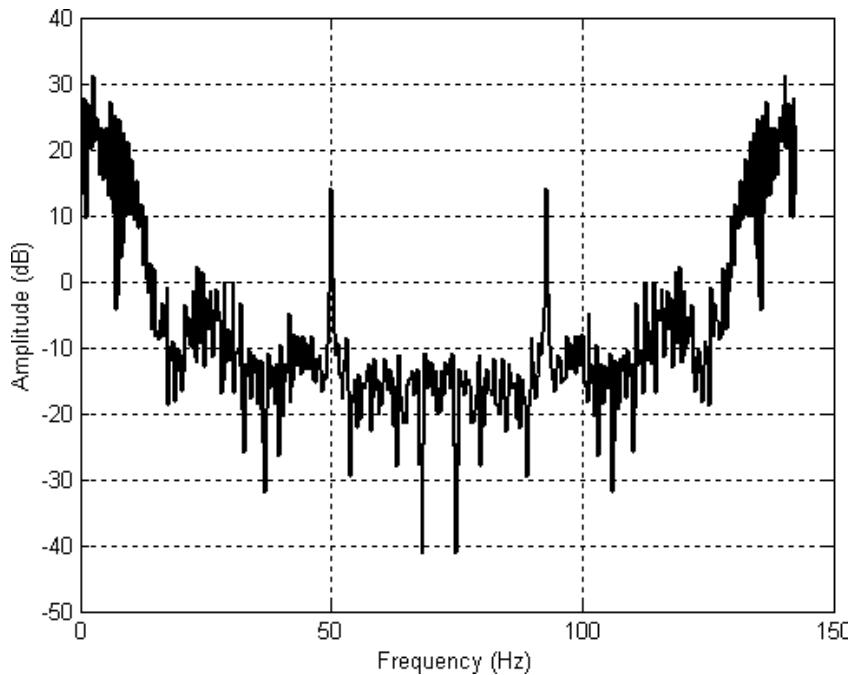
#### 5.5.4.1 Worked Example: Electrocardiogram

Consider the filtered electrocardiogram signal shown in Figure 5-49. At a glance it appears to be reasonably noise free.

If the signal is sampled at about 143 samples per second and the spectrum is examined, Figure 5-50 shows that there is a significant frequency component at 50 Hz. This is caused by the residual mains hum that was not removed by the CMRR of the differential amplifier. Because most of the signal content of the actual ECG resides at frequencies below 30 Hz, it seems reasonable to sample this signal at 75 to 85 Hz.



**FIGURE 5-49** ■  
Measured  
electrocardiogram.

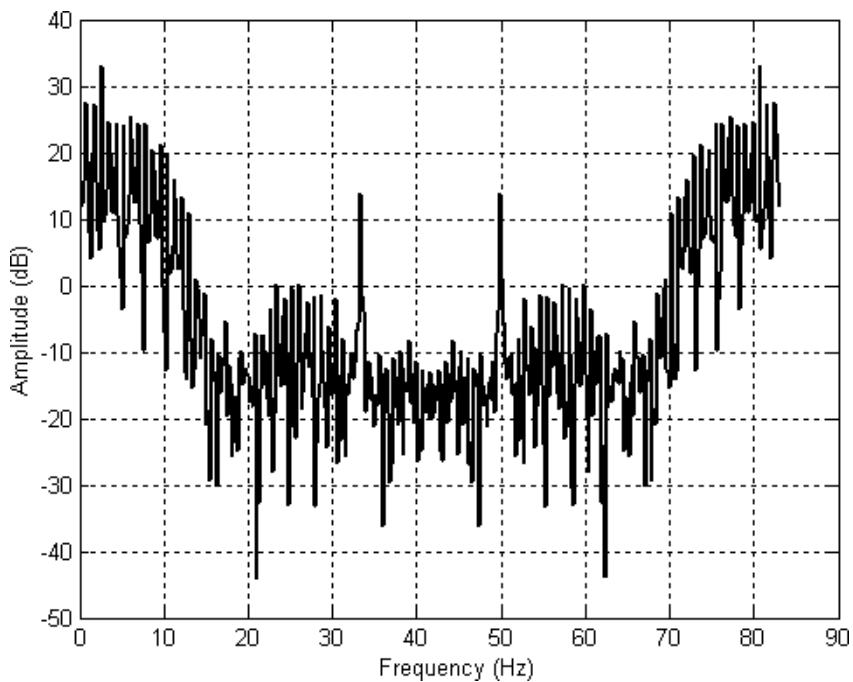


**FIGURE 5-50** ■  
ECG spectrum for a  
142.9 Hz sample  
rate.

If this same ECG signal is sampled at 83 samples per second and the spectrum is generated, the 50 Hz component has been aliased down to about 33 Hz (83–50 Hz), as shown in Figure 5-51. In this instance, there is no way of knowing that the 33 Hz artifact is not a genuine component of the ECG signal.

The only way to eliminate this aliasing is to ensure that the analog signal is filtered to less than  $f_s/2$  prior to sampling and digitization.

**FIGURE 5-51** ■  
ECG spectrum for an  
83.3 Hz sample rate.



### 5.5.5 Digital Filters

Digital filters perform the same functions as their analog counterparts; they can be configured as low- or high-pass filters or have any of the other functions discussed earlier, as well as a whole lot more. However, instead of manipulating actual voltages, digital filters manipulate numbers—the digital representations of the sampled analog signals.

#### 5.5.5.1 Difference Equations and Transfer Functions

Any filter can be written in terms of a difference equation, which is the weighted sum of a sequence of input values and output values that together generate the next value of the output sequence. This is shown in general terms (Stanley, 1975):

$$y(n) = \sum_{k=0}^M (b_k x(n-k)) - \sum_{k=1}^N (a_k y(n-k)) \quad (5.31)$$

where  $y(n - k)$  represents the outputs, and  $x(n - k)$  represents the inputs. The value of  $N$  gives the order of the difference equation and corresponds to the depth of the memory required to implement it. Because this equation relies on past values of the output, it is recursive, and these initial values (initial conditions) must be known. It is in this form that the actual numbers are manipulated by the DSP to perform the required function, but in general the filters are described by their transfer functions.

The  $z$ -transform is useful for the manipulation of discrete data sequences and has acquired a new significance in the formulation and analysis of discrete-time systems since the advent of digital computers. It is used extensively in the areas of applied mathematics and digital signal processing. In these cases discrete models are solved with difference equations in a manner that is analogous to solving continuous models with differential

equations. The role played by the  $z$ -transform in the solution of difference equations corresponds to that played by the Laplace transforms in the solution of differential equations.

The transfer function  $H(z)$  is easily obtained from the difference equation. The first step involves taking the Fourier transform before using the linearity property to move the transform to within the summation. The time-shifting property of the  $z$ -transform is used to change the time shifting terms to exponentials.

For  $a_0 = 1$ , this process results in

$$Y(z) = \sum_{k=0}^M (b_k X(z) z^{-k}) - \sum_{k=1}^N (a_k Y(z) z^{-k}) \quad (5.32)$$

The transfer function  $H(z)$  is then

$$H(z) = \frac{Y(z)}{X(z)} = \frac{\sum_{k=0}^M (b_k z^{-k})}{1 + \sum_{k=1}^N (a_k z^{-k})} \quad (5.33)$$

The frequency response is then obtained by replacing every instance of  $z$  with  $e^{j\omega}$ .

$$H(\omega) = \frac{\sum_{k=0}^M (b_k e^{-j\omega k})}{\sum_{k=0}^N (a_k e^{-j\omega k})} \quad (5.34)$$

Fortunately, this process of obtaining the transfer function is easily achieved in MATLAB.

### WORKED EXAMPLE

#### Obtaining a Filter Transfer Function Using MATLAB

The difference equation for a Butterworth low-pass filter can be written in terms of

$$\begin{aligned} y(n) = & 0.0029x(n) + 0.0087x(n-1) + 0.0087x(n-2) + 0.0029x(n-3) \\ & - [2.3741y(n-1) - 1.9294y(n-2) + 0.5321y(n-3)] \end{aligned}$$

The coefficients of the difference equation can be written in as arrays, where  $A$  includes the coefficients of  $y$ , and  $B$  includes the coefficients of  $x$

$$\begin{aligned} A &= (a_0, a_1, a_2, a_3) \\ B &= (b_0, b_1, b_2, b_3) \end{aligned}$$

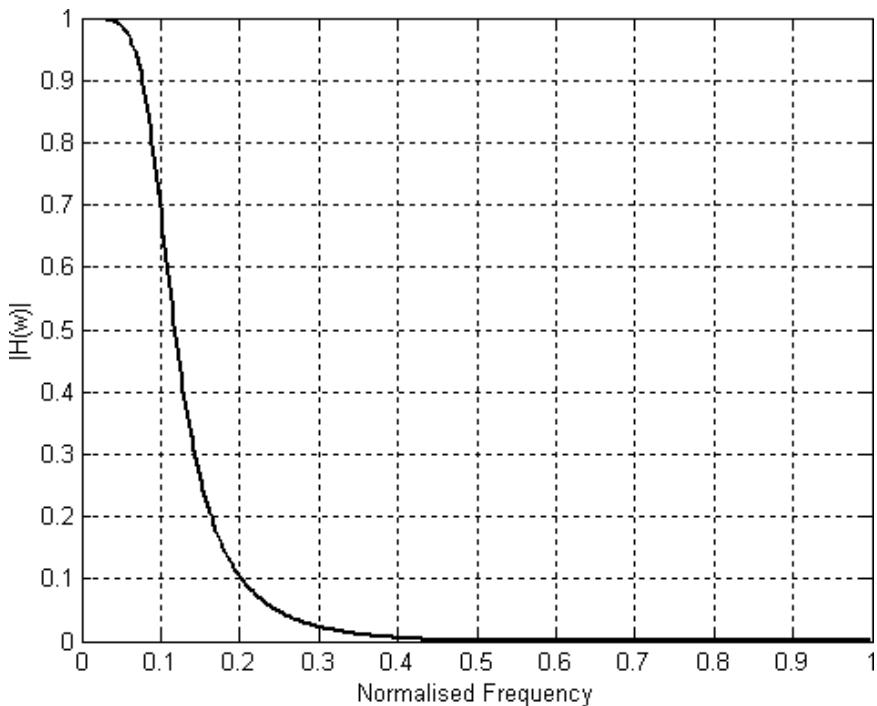
Remember that  $a_0 = 1$ , so

$$\begin{aligned} A &= (1, -2.3741, 1.9294, -0.5321) \\ B &= (0.0029, 0.0087, 0.0087, 0.0029) \end{aligned}$$

The frequency response can then be obtained and plotted, as shown in Figure 5-52, using the following MATLAB code:

```
[h,w] = freqz(B,A,512); % extract the transfer function from DC to fs/2
freq = (0:511)/512;
plot(freq,abs(h));
xlabel('Normalized Frequency')
ylabel('|H(w)|')
```

**FIGURE 5-52** ■ Frequency response of a low-pass Butterworth filter.



### 5.5.5.2 Synthesizing Infinite Impulse Response (IIR) Filters

There is a large body of literature that deals with the synthesis of the common and less common IIR filter algorithms. However, once again MATLAB synthesizes all common filter types with minimum effort. For example, the filter for the previous example was generated with a single command:

```
[B, A]=butter(3, 0.1)
```

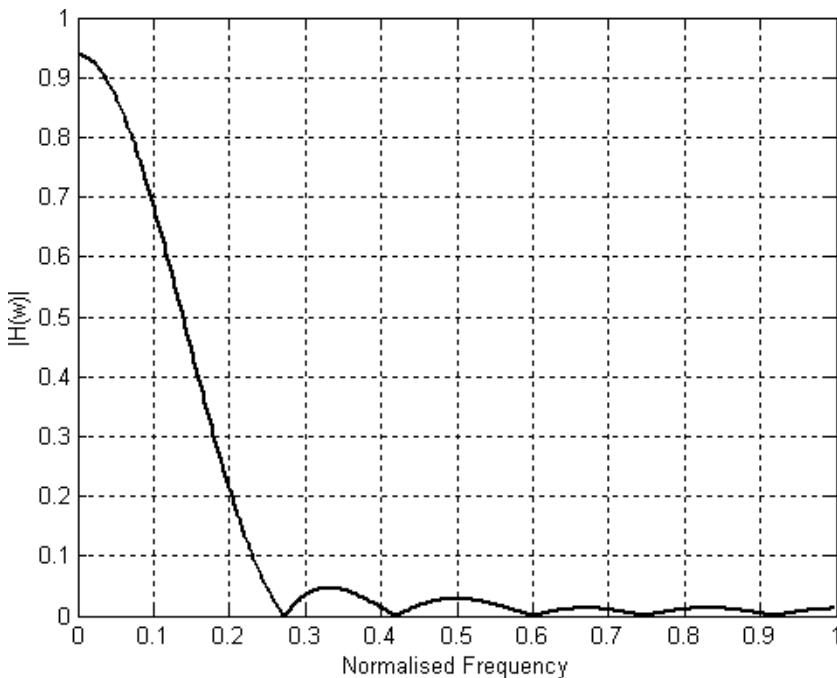
This generates the coefficients of a third-order Butterworth filter with a cutoff frequency of 0.1 (normalized to  $f_s/2$ ).

In addition to a range of Butterworth filters, MATLAB synthesizes Bessel, Chebyshev, and elliptical low-pass, high-pass, and band-pass filters of whatever order is required. However, some care needs to be exercised as numerical issues tend to make high-order filters unstable.

### 5.5.5.3 Synthesizing Finite Impulse Response (FIR) Filters

In finite impulse response filters, only the input values are used to produce an output; they are therefore not recursive.

In this implementation, the shape of the filter is described roughly in terms of the required amplitude response. So, if the frequency response of the Butterworth filter described in the previous section was to be synthesized using a FIR filter, then the following



**FIGURE 5-53 ■**  
Frequency response  
of a tenth-order FIR  
filter.

MATLAB code could be used to produce the frequency response shown in Figure 5-53:

```
f=[0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,1.0]; % define the frequency response
a=[1,0.707,0.1,0.02,0.01,0,0,0,0,0]; % from DC to fs/2
B=firls(10,f,a); % generate the coefficients for 10th order FIR filter
A=[1];
[h,w] = freqz(B,A,512); % extract the transfer function from DC to fs/2
freq = (0:511)/512;
plot(freq,abs(h));
xlabel('Normalized Frequency')
ylabel('|H(w)|')
```

#### 5.5.5.4 Tracking Filters

Consider a prosthetic arm in which only a noisy measurement of the angle of the elbow joint is measured but effective control requires that the angular rate be estimated. This can be achieved in isolation by using a differentiator or using a tracking filter. The latter can be used to produce an estimate of the angular rate (and higher order derivatives) as well as a filtered estimate of the actual angle.

The  $\alpha - \beta$  tracker is a fixed gain formulation of the Kalman filter and is still widely used because it is easy to implement and performs well under most circumstances (Blackman, 1986). The implementation of most tracking filters follows a two-stage process: smoothing then prediction. The equations defining these two stages are shown as follows for the  $\alpha - \beta$  tracker implementation:

Smoothing

$$\hat{R}_n = \hat{R}_{pn} + \alpha(R_n - \hat{R}_{pn}) \quad (5.35)$$

$$\hat{V}_n = \hat{V}_{pn} + \frac{\beta}{T_s}(R_n - \hat{R}_{pn}) \quad (5.36)$$

## Prediction

$$\hat{R}_{p(n+1)} = \hat{R}_n + \hat{V}_n \cdot T_s \quad (5.37)$$

$$\hat{V}_{p(n+1)} = \hat{V}_n \quad (5.38)$$

where

$\hat{R}_n$  = smoothed estimate of position

$\hat{V}_n$  = smoothed estimate of rate

$R_n$  = measured position

$\hat{R}_{p(n+1)}$  = predicted position after  $T_s$  seconds (one sample ahead)

$\hat{V}_{p(n+1)}$  = predicted position rate after  $T_s$  seconds (one sample ahead)

$\hat{R}_{pn}$  = predicted position at the measurement time

$\hat{V}_{pn}$  = predicted rate at the measurement time

$T_s$  = sample time

$\alpha, \beta$  = smoothing constants

The previous equations are only one of the ways that the tracking filter can be described. Other ways include the block diagram representation (Mahafza, 2000), state-space representations, or rewriting the equations as a position transfer function from the difference equations.

$$R_p(z) = R(z)z^{-1} + T_s \dot{R}(z)z^{-1} \quad (5.39)$$

$$R(z) = R_p(z) + \alpha [R_m(z) - R_p(z)] \quad (5.40)$$

$$\dot{R}(z) = \dot{R}(z)z^{-1} + \frac{\beta}{T_s} [R_m(z) - R_p(z)] \quad (5.41)$$

where  $R_p(z)$  is the predicted position,  $R(z)$  is the estimated position,  $\dot{R}(z)$  is the rate estimate, and  $R_m(z)$  is the measured position.

The position transfer function will therefore be

$$H_1(z) = \frac{R(z)}{R_p(z)} \quad (5.42)$$

and it can be obtained by manipulation of equations (5.38) to (5.40).

$$H_1(z) = \frac{\alpha + (\beta - \alpha)z^{-1}}{1 + (\beta + \alpha - 2)z^{-1} + (1 - \alpha)z^{-2}} \quad (5.43)$$

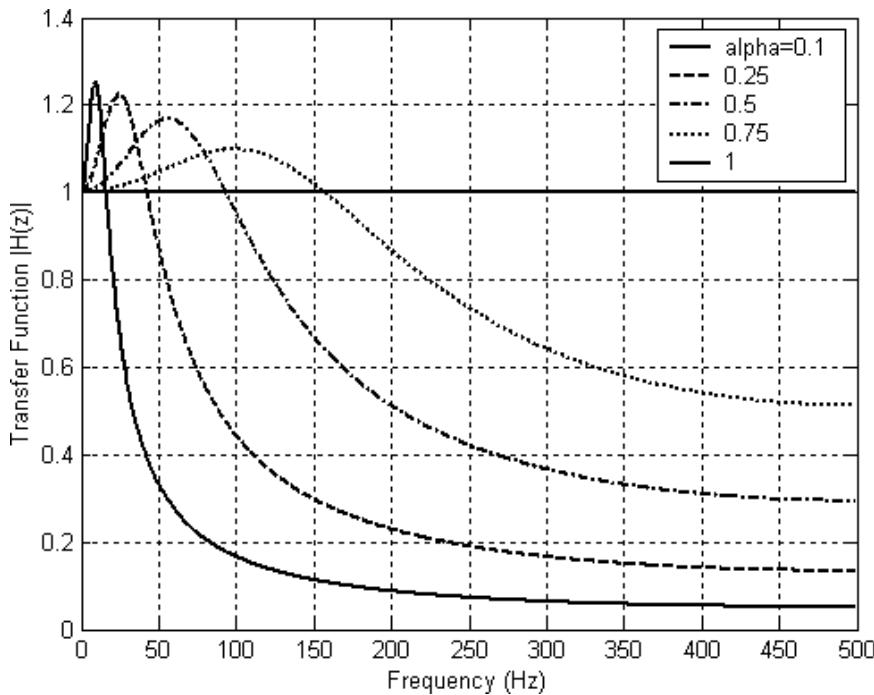
This is a convenient form of the filter to be used in MATLAB where the filter coefficients described earlier are

$$B = [\alpha, (\beta - \alpha)]$$

$$A = [1, (\beta + \alpha - 2), (1 - \alpha)]$$

One possible filter optimization (Benedict and Bordner, 1962) is to minimize the output noise variance at steady state, and the transient response to a changing input is modeled by a ramp function. This results in the following relationship between the two gain coefficients:

$$\beta = \frac{\alpha^2}{2 - \alpha} \quad (5.44)$$



**FIGURE 5-54** ■ Frequency response of an  $\alpha - \beta$  filter at  $f_s = 1$  kHz.

Other criteria can also be used. For example, it has been suggested that the filter should have the fastest possible step response (critical damping), in which case the coefficients are related as

$$\alpha = 2\sqrt{\beta} - \beta \quad (5.45)$$

The actual values of the gains depend on the sample period, the predicted dynamics, and the required loop bandwidth.

Figure 5-54 shows the filter transfer function for the position estimate of the tracking filter using various values of  $\alpha$  and  $\beta$  (Benedict–Bordner relationship). This is obtained using equation (5.42) and the Freqz function in MATLAB. Even though the transfer function does not conform to the classical low-pass filter shape discussed earlier, the half-power bandwidth is still measured at  $|H(z)| = 0.707$ .

```
% plot the alpha beta filter position transfer function
% ab_trans.m
%
% Variables
fs = 1000; % Sample frequency (Hz)
alpha = 0.1 % Position gain
beta = alpha.^2/(2-alpha); % Velocity gain (Benedict Bordner Relationship)

% Calculate the filter coefficients for the position transfer function
B=[alpha,beta-alpha];
A=[1,(beta+alpha-2),(1-alpha)];
[H1,F]=freqz(B,A,1024,fs);
% Plot the magnitude of the transfer function
```

```

plot(F,abs(H1), 'k')
grid
xlabel('Frequency (Hz)')
ylabel('Transfer Function |H(z)|')

```

One of the main disadvantages of the fixed gain  $\alpha-\beta$  filter is that it estimates the position of an accelerating input with a constant lag. However, because the magnitude of this lag can easily be estimated, the filter can use adaptive gains to improve the RMS tracking accuracy under these circumstances. The cost function that is minimized in this case is  $[lag^2 + \sigma_r^2]$ .

For estimates that include accelerations, the  $\alpha-\beta-\gamma$  tracker can be substituted. However, its performance for constant velocity motion is poorer than that for the simple  $\alpha-\beta$  tracker because of the additional noise introduced by the acceleration estimate.

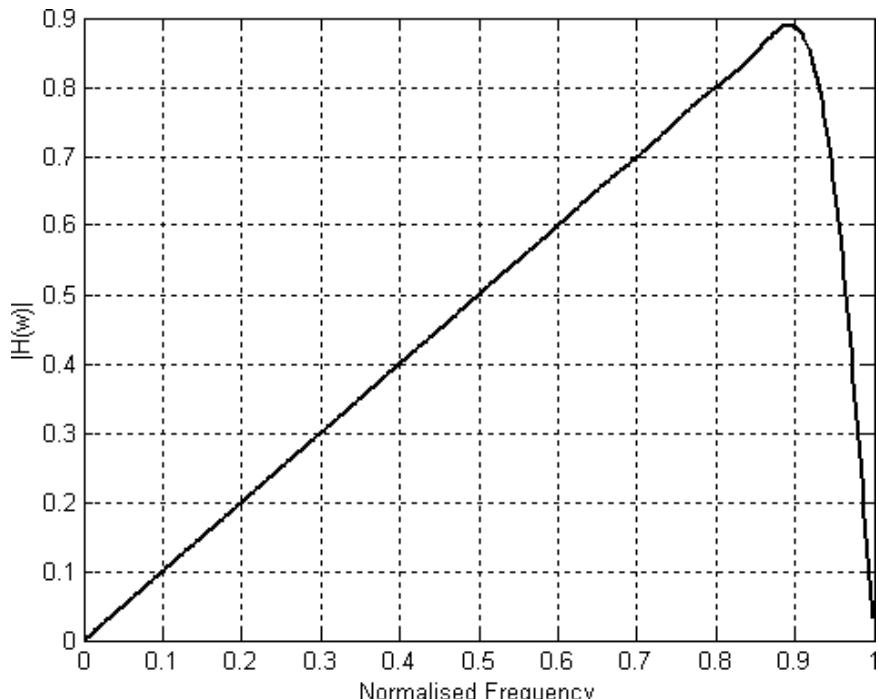
### 5.5.5.5 Differentiator

A simple differentiator can be obtained by taking the difference between successive input samples and dividing by the interval between them. This algorithm produces a high-pass filter and therefore adds noise to the measurement. Some improvement can be obtained by following the differentiator with a higher-order low-pass filter to attenuate high-frequency noise.

An alternative is to use the same MATLAB process that was used to determine the coefficients of the FIR low-pass filter to synthesize a differentiator with good characteristics. The frequency response of the filter generated by the following code is shown in Figure 5-55.

```
b=firls(30,[0,0.9],[0,0.9],'differentiator');
```

**FIGURE 5-55 ■**  
Frequency response  
of a thirtieth-order  
FIR differentiator.



### 5.5.6 Integrator

The common methods of integration use the forward or backward Euler (also known as rectangular) and the trapezoidal algorithms.

The difference equation for the forward Euler integration is defined in terms of the latest input,  $x(n - 1)$ , and the sample interval,  $T$ ,

$$y(n) = y(n - 1) + Tx(n - 1) \quad (5.46)$$

This can be written as a transfer function

$$H(z) = \frac{T}{z - 1} \quad (5.47)$$

For the backward Euler method the difference equation is

$$y(n) = y(n - 1) + Tx(n) \quad (5.48)$$

and the transfer function is

$$H(z) = \frac{Tz}{z - 1} \quad (5.49)$$

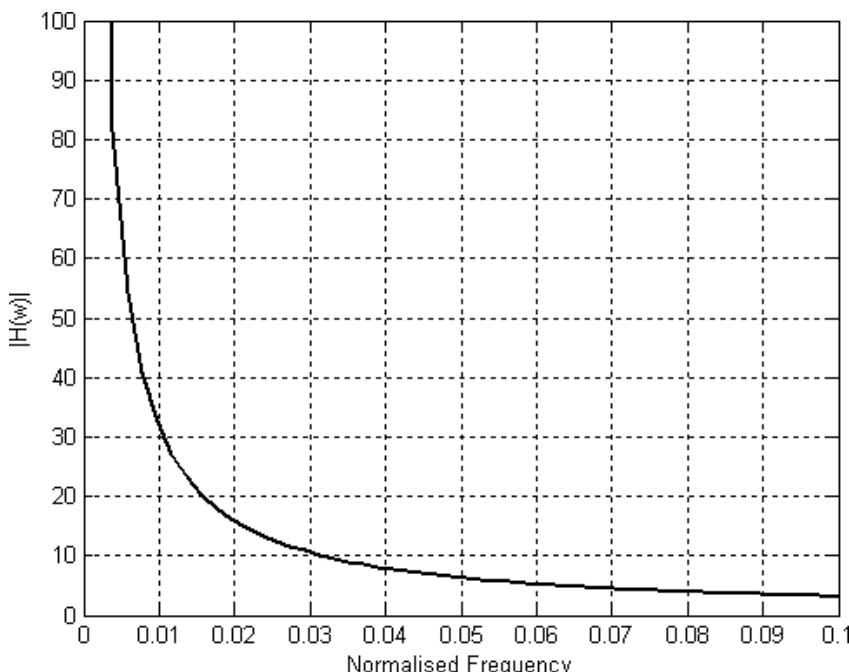
In the trapezoidal case, the difference equation is a little more complicated

$$y(n) = y(n - 1) + \frac{T[x(n) + x(n + 1)]}{2} \quad (5.50)$$

and the transfer function can be written as

$$H(z) = \frac{T(z + 1)}{2(z - 1)} \quad (5.51)$$

The frequency response of the trapezoidal integrator is shown in Figure 5.56.



**FIGURE 5-56 ■**  
Integrator frequency response for  $T = 1$ .

In theory the DC gain will be infinite because the output increases indefinitely for a DC input. However, as shown in Figure 5-56, this gain decreases exponentially with increasing frequency.

### 5.5.5.7 Averager

The process of averaging is one way of generating a simple low-pass filter. This process can be implemented as a block algorithm in which  $N$  input samples are processed to generate a single output

$$y(k) = \frac{1}{N} \sum_{n=(k-1)N}^{kN} x(n) \quad (5.52)$$

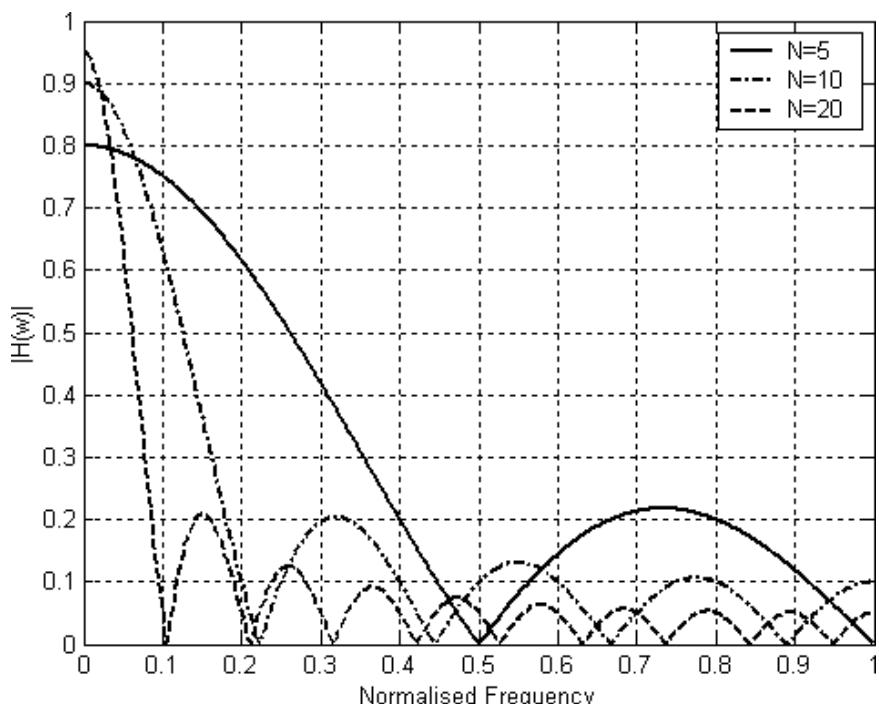
This reduces both the bandwidth and the sample rate of the output. However, if the sample rate is to be maintained, the moving-average algorithm is used. In this case the average is formed by adding the most recent input and subtracting the oldest.

$$y(n) = y(n-1) + \frac{1}{N} [x(n-1) - x(n-N-1)] \quad (5.53)$$

It can be seen that this algorithm is very similar to the integrator discussed in the previous section. The frequency response of the moving averager is shown in Figure 5-57.

If the signal spectral content and that of the noise overlap significantly, there is often no alternative but to use an averaging process. For example, brain potentials generated by external visual or acoustic stimulation, called evoked potentials (EPs), have amplitudes that are much lower than the background EEG activity but contain information in the same bands, so they cannot be separated using conventional frequency-based filters (Bronzino, 2006).

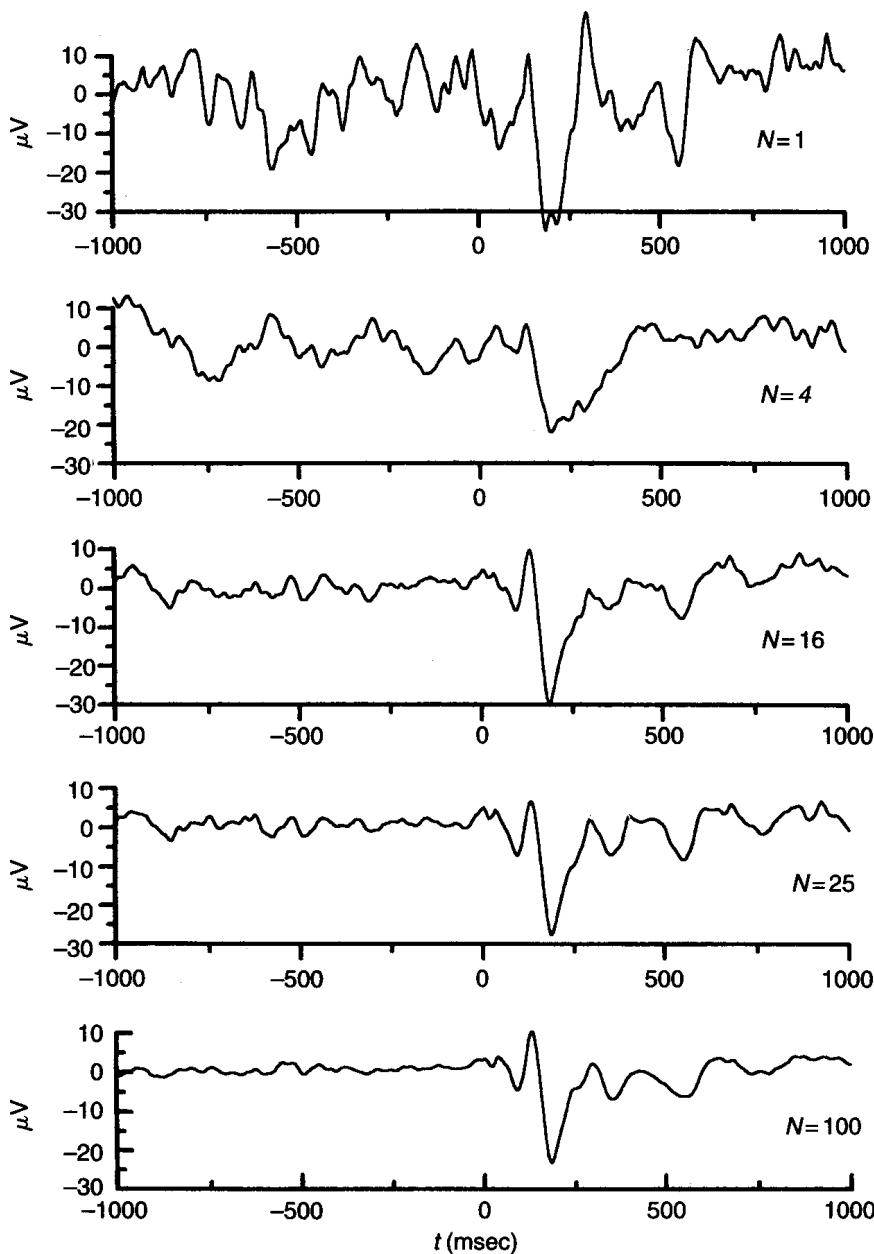
**FIGURE 5-57** ■ Frequency response of a moving averager for different  $N$ .



When the desired signal repeats almost identically at each iteration, a form of block averaging can solve the problem. In this case a sequence of points is sampled, covering the duration and synchronized to the triggered event. This process is repeated a number of times, as illustrated in Figure 5-58, and the sequences are averaged a number of times. As  $N$  increases, the signal-to-noise ratio is improved by a factor  $\sqrt{N}$  (in RMS value) until the EP characteristics emerge from the random ECG background.

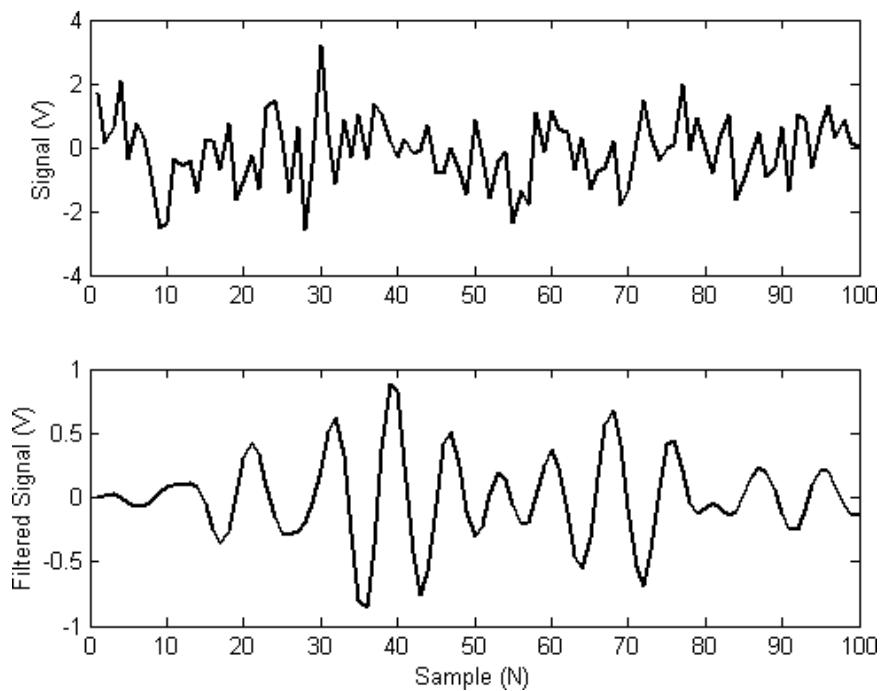
This averaging process can be described by

$$y(n) = \frac{1}{N} \{x(n) + x(n - h) + x(n - 2h) + x(n - 4h) + \dots + x(n - [N - 1]h)\} \quad (5.54)$$



**FIGURE 5-58** ■ Block averaging process used to extract evoked potentials from background EEG activity. [Adapted from (Bronzino 2006).]

**FIGURE 5-59** ■  
Band-pass filter  
response to white  
noise.

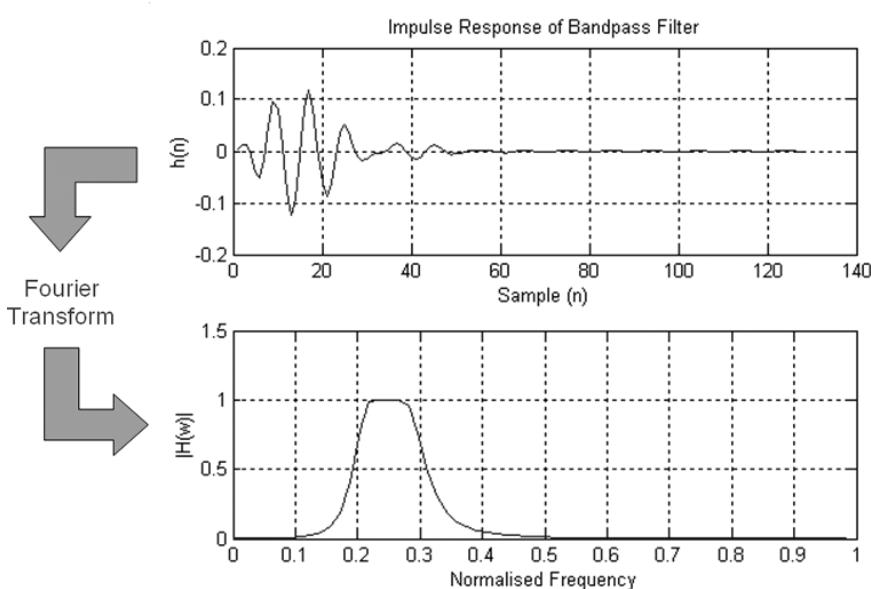


where  $N$  is the number of elements in the averaging process, and  $h$  is the number of points in the sequence.

### 5.5.6 Filter Time-Domain Response

The time-domain response of both analog and digital filters is obtained by exciting the filter with the signal of choice and monitoring the output waveform. Common methods of stimulating filters include white noise, impulses, steps, and ramps. All of these inputs are easily generated in MATLAB, as is the filter. Figure 5-59 shows the response of a Butterworth band-pass filter to white noise, as generated by the following code:

```
% Band-pass filter
% bandpass2.m
fs = 200e+03; % sample frequency (Hz)
ts = 1/fs; % sample interval (s)
fmat = 25.0e+03; % center frequency of filter (Hz)
bmat = 10.0e+03; % bandwidth of filter (Hz)
wl=2*ts*(fmat-bmat/2); % lower band limit
wh=2*ts*(fmat+bmat/2); % upper band limit
wn=[wl,wh];
[b,a]=butter(3,wn); % 6th order Butterworth band-pass filter
sig = randn(1,100); % normally distributed white noise
sig_fil = filter(b,a,sig); % time domain implementation of the filter
subplot(211),plot(sig),ylabel('Signal (V)');
subplot(212),plot(sig_fil),ylabel('Filtered Signal (V)'), xlabel('Sample (N)')
```



**FIGURE 5-60** ■ Relationship between the frequency response and the impulse response for a band-pass filter. [Adapted from (Brooker 2008).]

### 5.5.6.1 Filter Impulse Response

The impulse response is the time-domain output of the filter if it has been excited by a unit impulse (Dirac function) at its input. Because a true impulse contains equal contributions from all frequencies (as shown in Figure 5-15), it is equivalent to injecting white noise into the filter and hence characterizes it over the full frequency band.

Frequency response and impulse response methods are equivalent and are related by the Fourier transform, as shown in Figure 5-60, for the band-pass filter described in the previous section.

In general, because the impulse response dies down quickly it is a convenient and quick method of determining how effective filters are at attenuating white noise. This is achieved in the time domain by inputting an impulse with a known variance into the filter and measuring the variance of the output. The ratio of the two is known as the variance reduction ratio (VRR).

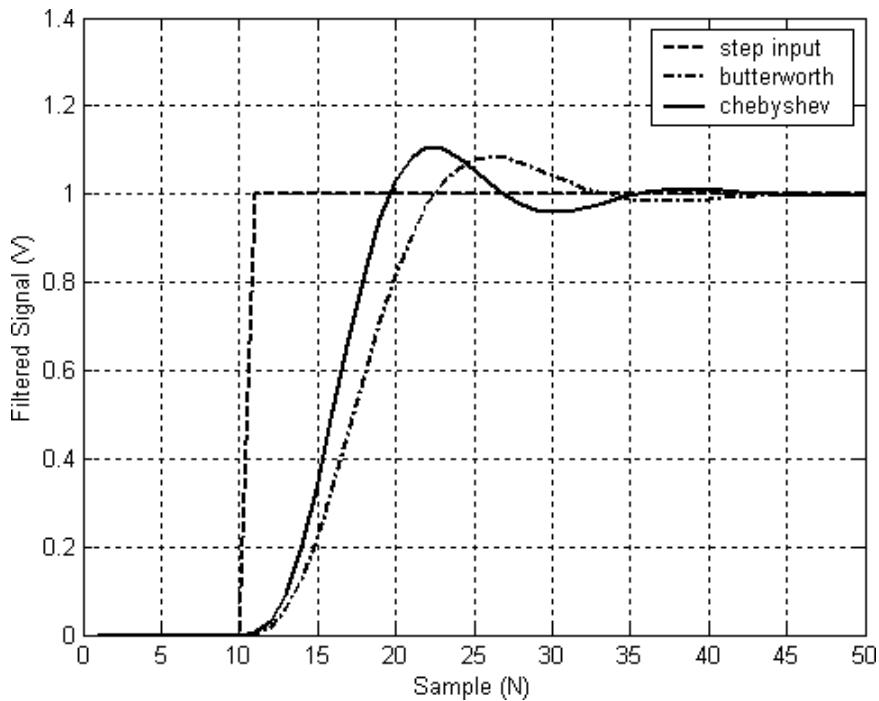
### 5.5.6.2 Filter Step Response

The step response of filters is often used to quantify their dynamic performance in terms of a rise time and overshoot. The rise time is an indication of the bandwidth of the filter and is one way of determining how quickly it will track a changing input, while the overshoot is an indication of the amount of damping. Figure 5-61 shows the responses of a third-order Butterworth filter and the same order Chebyshev filter with a 0.1 dB ripple in the passband.

### 5.5.7 Envelope Detection

Two different methods are commonly used to extract the envelopes from amplitude modulated signals. The conventional method is to rectify the signal followed by a low-pass filter with a cutoff frequency dependent on the frequency content of the modulation. The second method uses the Hilbert transform.

**FIGURE 5-61** ■ Comparison between the step responses of a number of low-pass filters.



In a digital implementation of the conventional method, the rectification process is achieved by taking the absolute value of the signal before it is processed through a low-pass algorithm such as the Butterworth filter discussed earlier in this chapter. The cutoff frequency of the low-pass filter determines the modulation depth of the envelope, with lower cutoff frequencies resulting in smaller modulation depths.

The Hilbert transform is a mathematical function that can represent a time waveform as the product of a slowly varying envelope and a carrier containing fine structure information. In mathematical terms, the filtered waveform,  $x_i(t)$ , can be represented as

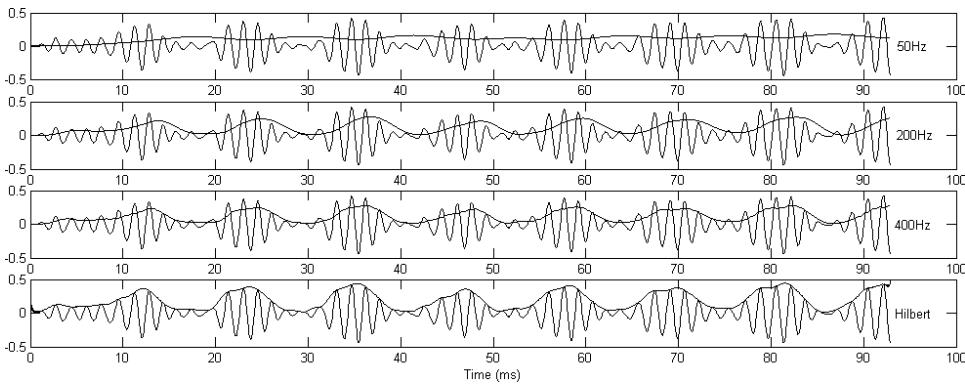
$$x_i(t) = a_i(t) \cos f_i(t) \quad (5.55)$$

where  $a_i(t)$  represents the envelope, and  $\cos f_i(t)$  represents the fine structure waveform. Note that  $f_i(t)$  is the instantaneous phase of the signal, and its derivative produces the instantaneous frequency (the carrier frequency).

Figure 5-62 shows the outputs of a number of different envelope detectors used in a cochlear implant. It can be seen that the Hilbert transform method reproduces the envelope more accurately than the conventional envelope detector even with a filter bandwidth of 400 Hz.

### 5.5.8 Spectral Estimation

The traditional method of frequency analysis is based on the Fourier transform, which is generally evaluated using the discrete Fourier transform (DFT) or the fast Fourier transform (FFT) if the number of points corresponds to  $k = 2^N$ , where  $N$  is an integer.



**FIGURE 5-62** ■ Envelope detection with various low-pass filter cutoff frequencies.

The expression for the PSD as a function of the frequency,  $P(f)$ , can be obtained directly from the time series  $y(n)$  using the periodogram expression

$$P(f) = \frac{1}{T_s} \left| T_s \sum_{k=0}^{N-1} y(k) e^{-j2\pi f k T_s} \right|^2 = \frac{1}{NT_s} |Y(f)|^2 \quad (5.56)$$

where  $T_s$  (sec) is the sample period,  $N$  is the number of samples, and  $Y(f)$  is the discrete Fourier transform of  $y(n)$ .

The PSD can also be obtained from the FFT of the autocorrelation function  $\hat{R}_{yy}(k)$  of the signal, where  $\hat{R}_{yy}(k)$  is estimated as

$$\hat{R}_{yy}(k) = \frac{1}{N} \sum_{i=0}^{N-k-1} y(i) y^*(i+k) \quad (5.57)$$

where  $*$  denotes the complex conjugate. The power spectral density is then

$$P(f) = T_s \sum_{k=-N}^N \hat{R}_{yy}(k) e^{-j2\pi f k T_s} \quad (5.58)$$

Both the autocorrelation function and the Fourier transform are theoretically defined over infinite data sequences, so their truncation can lead to errors, particularly in respect to spectral leakage caused by the implicit rectangular windowing of the data.

### 5.5.8.1 Fourier Transform

All continuous periodic signals can be represented by a fundamental frequency sine wave and a collection of sine or cosine harmonics of that signal.

The Fourier series for any waveform can be expressed as (Carr, 1997; Edminster, 1972)

$$v(t) = \frac{a_0}{2} + \int_{n=1}^{\infty} a_n \cos(n\omega t) + \int_{n=1}^{\infty} b_n \sin(n\omega t), \quad (5.59)$$

where  $a_n$  and  $b_n$  are the amplitudes of the harmonics, which can be zero, and  $n$  is an integer.

The amplitude terms for each spectral component can be calculated by integrating the product of the time-domain signal with a sample phasor of the correct frequency

$$\begin{aligned} a_n &= \frac{2}{T} \int_0^t v(t) \cos(n\omega t) dt \\ b_n &= \frac{2}{T} \int_0^t v(t) \sin(n\omega t) dt \\ c_n &= \sqrt{a_n^2 + b_n^2} \\ \theta_n &= \tan^{-1} \left( \frac{a_n}{b_n} \right) \end{aligned} \quad (5.60)$$

The magnitude,  $c_n$ , of the  $n$ -th harmonic can be calculated, along with its phase,  $\theta_n$ , as shown in (5.60).

If there is a component of the signal at or near the selected frequency, this phase-sensitive integration process will produce a nonzero amplitude. This process examines the signal only at discrete integer frequencies determined by the value of  $n$  and at DC where the term  $a_0/2$  is the average value of  $v(t)$  over a complete cycle.

Care must be used when interpreting the results of this spectral analysis, as the process assumes that individual harmonics are present with constant amplitude and phase for the duration of the measurement. However, in reality harmonics are seldom constant in amplitude or phase, so the coefficients, which represent only the averages, may be significantly in error.

Depending on the nature of the signal and the information that must be reproduced, the number of frequency components that are required will vary. For example, a square wave may require up to 1000 harmonics to reproduce the sharp transitions that define the switching points (Carr, 1997). The harmonic analysis in Figure 5-63 shows that the amplitudes of the coefficients are reduced by progressively smaller amounts with the result that their individual contributions become progressively smaller.

However, as discussed earlier in this chapter, an ECG trace can be reproduced using spectral components up to about 100 Hz.

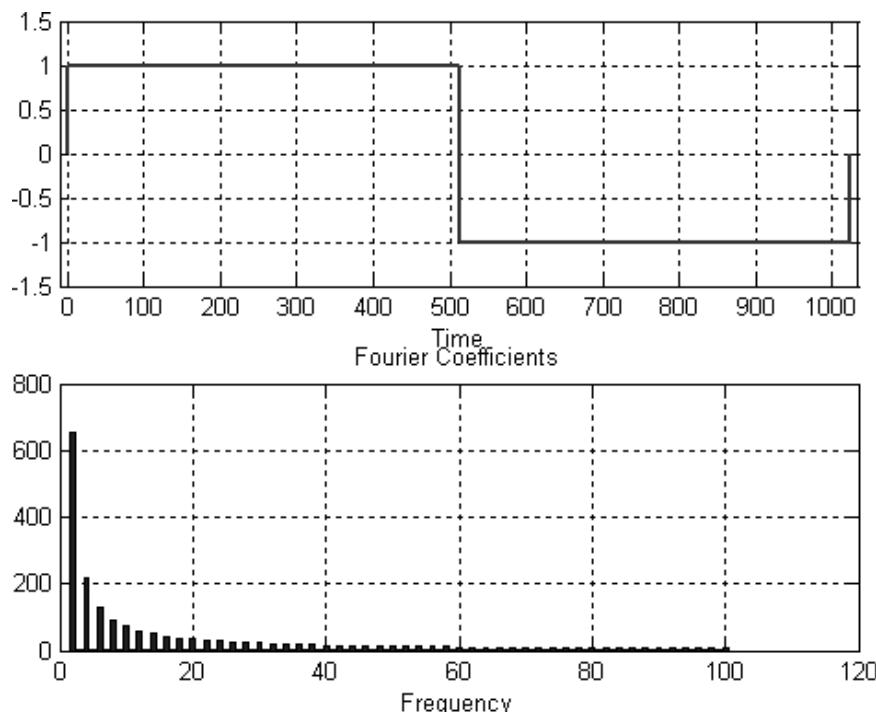
The effect of truncating the series is demonstrated in Figure 5-64, where the waveform is reconstructed by summing the appropriately scaled sinusoidal components. For example, when five components are used, the signal is reconstructed as

$$v(t) = \sum_{n=1}^5 a_n \cos(n\omega t) + b_n \sin(n\omega t) \quad (5.61)$$

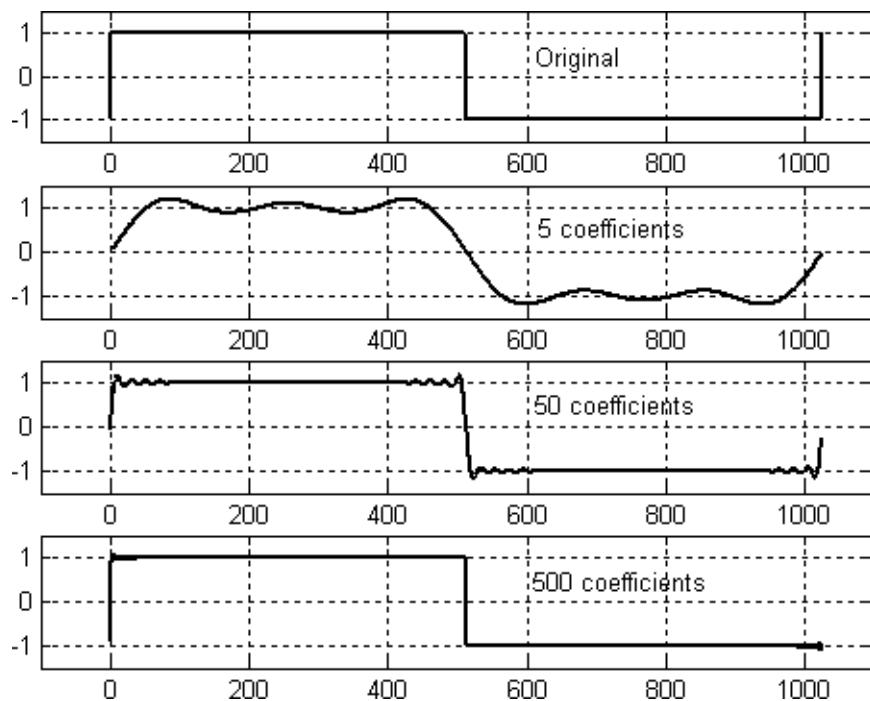
Figure 5-64 shows the effectiveness of the reconstruction for 5, 50, and 500 coefficients. It can be seen that the transitions are still not perfect even after the largest number.

The spectrum of a repeated time-domain sequence such as an ECG produces a comb of frequency components at intervals corresponding to the reciprocal of the repeat period. As shown in Figure 5-65, these fall within the envelope of the spectrum of a single ECG cycle.

To understand this, consider that the repeated time-domain signal can be generated by convolving a single ECG cycle with a sequence of impulses at the correct time intervals,  $\Delta t$ . It can be shown that the spectrum of a sequence of impulses is a regularly spaced

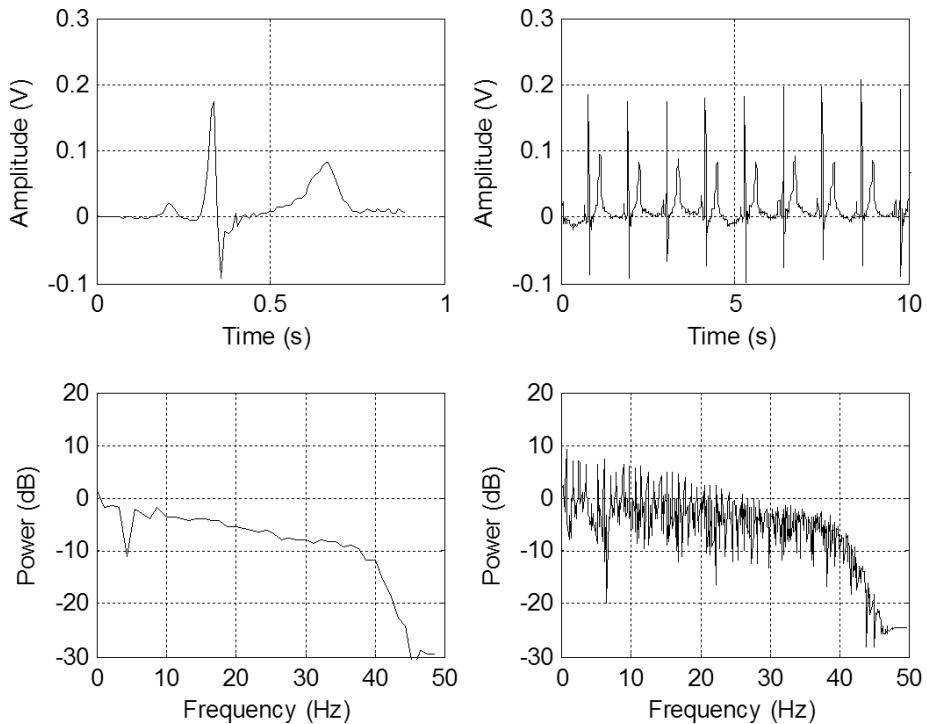


**FIGURE 5-63** ■ Amplitudes of Fourier coefficients to produce a square wave. [Adapted from (Brooker 2008).]



**FIGURE 5-64** ■ Effect on the reconstructed signal of limiting the number of Fourier coefficients.

**FIGURE 5-65 ■**  
Comparison  
between the spectra  
of (a) a single ECG  
cycle and (b) a  
sequence of cycles.



comb of frequency components with spacing  $\Delta f = 1/\Delta t$ . In addition, in the frequency domain the convolution process is replaced by the product of the two spectra; therefore, the composite spectrum will be the product of the spectrum of a single ECG cycle and the spectrum of a sequence of impulses as shown.

## 5.6 | STATISTICAL TECHNIQUES AND MACHINE LEARNING

Living organisms are so complex that it is often difficult, if not impossible, to use any of the conventional signal processing methods discussed previously to obtain useful outputs. In such cases automated processes such as statistical methods, pattern recognition, data mining, or machine learning are often used to extract useful information from raw measurements. Broadly speaking, these techniques have similar goals—modeling for classification or hypothesis testing.

### 5.6.1 Statistical Techniques

Statistical methods have traditionally emphasized models that can be solved analytically, which places some restrictions on their expressive power. However, these methods have been around for a long time and are well understood.

#### 5.6.1.1 Regression Analysis

Regression analysis is probably the oldest and best-known statistical method. It is a technique that provides a mathematical description of the relationship between two or

more measured variables. It is particularly useful when dealing with continuous quantitative data like the amplitude of myoelectric signal levels as a function of muscle force, for example.

In its simplest form, for a given series of paired variables  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  where  $n \geq 2$ , and following a relationship known to be linear

$$y = a + bx \quad (5.62)$$

The best-fitting curve in a least squares sense minimizes the residual,  $\varepsilon$ , to determine the coefficients  $a$  and  $b$

$$\varepsilon = \sum_{i=1}^n [y_i - f(x_i)]^2 = \sum_{i=1}^n [y_i - (a + bx_i)]^2 = \min \quad (5.63)$$

To obtain the least squared error, the partial derivatives of the residual with respect to  $a$  and  $b$ , respectively, must both be zero.

$$\begin{aligned} \frac{\partial \varepsilon}{\partial a} &= 2 \sum_{i=1}^n [y_i - (a + bx_i)] = 0 \\ \frac{\partial \varepsilon}{\partial b} &= 2 \sum_{i=1}^n x_i [y_i - (a + bx_i)] = 0 \end{aligned} \quad (5.64)$$

Expanding and solving for  $a$  and  $b$

$$a = \frac{\sum_{i=1}^n y_i \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2} \quad (5.65)$$

$$b = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2} \quad (5.66)$$

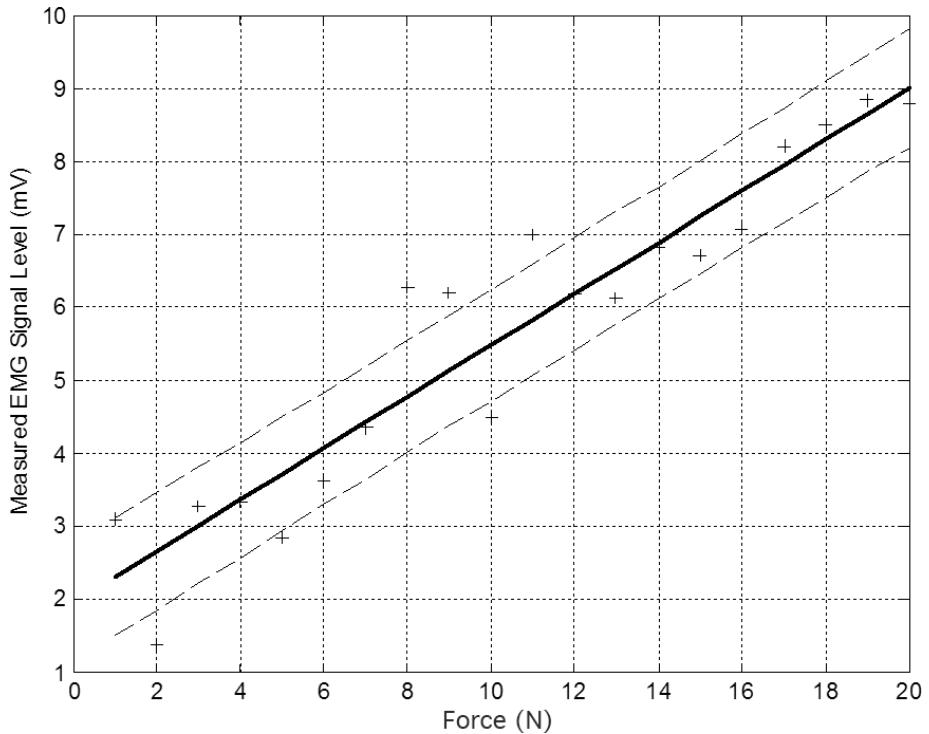
The correlation coefficient,  $r$ , is a measure of the degree of linear association between the two variables and is defined as

$$r = \frac{\sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n}}{\sqrt{\left[ \sum_{i=1}^n x_i^2 - \frac{\left( \sum_{i=1}^n x_i \right)^2}{n} \right] \left[ \sum_{i=1}^n y_i^2 - \frac{\left( \sum_{i=1}^n y_i \right)^2}{n} \right]}} \quad (5.67)$$

It is seldom necessary to use these equations directly, as many scientific calculators and computer software packages are already programmed to perform regression analysis.

Consider the measured EMG signals from the biceps as a function of the applied force, shown as the + markers in Figure 5-66. The coefficients  $a$  and  $b$  are obtained in MATLAB and are used to plot the solid best-fit line  $y = 1.9503 + 0.3531x$ . The dashed

**FIGURE 5-66 ■**  
Measured relationship between applied force and measured EMG signal.



lines above and below the solid line show  $y + /-\delta$ , and they will contain at least 50% of future observations at  $x$ .

```
% Linear fit to measured force and EMG data
[P,S] = polyfit(force,emg,1); % obtain P=(b,a) from the measured data
x = (min(force):1:max(force)); % generate the x vector
[y,delta] = polyval(P,x,S); % generate the y vector using a and b
plot(force,emg,'+',x,y,x,y+delta,x,y-delta); % plot and label the graph
grid
xlabel('Force (N)')
ylabel('Measured EMG Signal Level (mV)')
```

Multiple regression provides an estimate of the coefficients of a function when more than one control variable is involved. For example, in the case where there are two independent variables,  $x$  and  $y$ , and one dependent variable,  $z$ , the linear equation is

$$z = a + bx + cy \quad (5.68)$$

For a given data set  $(x_1, y_1, z_1), (x_2, y_2, z_2), \dots, (x_n, y_n, z_n)$  where  $n \geq 3$ , the best-fitting curve in a least squares sense minimizes the residual,  $\varepsilon$ , to determine the coefficients  $a$ ,  $b$ , and  $c$

$$\varepsilon = \sum_{i=1}^n [z_i - f(x_i, y_i)]^2 = \sum_{i=1}^n [z_i - (a + bx_i + cy_i)]^2 = \min \quad (5.69)$$

Equating the partial derivatives with respect to the coefficients as before allows us to construct a set of simultaneous equations that can be solved to find the coefficients

$$\begin{aligned}\sum_{i=1}^n z_i &= a \sum_{i=1}^n 1 + b \sum_{i=1}^n x_i + c \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i z_i &= a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2 + c \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n y_i z_i &= a \sum_{i=1}^n y_i + b \sum_{i=1}^n x_i y_i + c \sum_{i=1}^n y_i^2\end{aligned}\tag{5.70}$$

When working with categories of outputs rather than numbers, classification algorithms are a better choice than the regression methods so far discussed.

## 5.6.2 Data Mining

Data mining has been defined as the process of extracting patterns from very large data sets by using a combination of statistical and machine learning techniques in conjunction with database management. Typically, the focus of data mining is primarily on data storage, integration, and retrieval, with the analysis functions being of secondary importance.

## 5.6.3 Machine Learning

Machine learning (ML) uses computationally powerful methods to learn very complex nonparametric (or quasi-parametric) models of the data. The models used are often close to human representations, and the process is one of learning or improved performance with “experience.”

ML techniques include the following:

- Artificial intelligence
- Decision/classification trees
- Clustering
- Support vector machines
- Expert systems
- Neural networks
- Genetic algorithms
- Bayesian (belief) networks
- Bayesian statistics and nets

ML can be divided into two broad categories: supervised and unsupervised learning. In supervised learning, the algorithms are presented with a number of examples. This allows the algorithm to learn to predict the correct output corresponding to previously seen inputs as well as previously unseen ones. In unsupervised learning the algorithm is required to generate its own categories of outputs.

### 5.6.3.1 Decision/Classification Trees

Traditional classification methods such as discriminant analysis, cluster analysis, non-parametric statistics, and nonlinear estimation are commonly used where assumptions regarding the statistical nature of the measurements are met. However, in cases where traditional methods fail classification trees can be used.

Classification trees predict membership of objects into classes of a categorical dependent variable from measurements of predictor variables using a hierarchical process. An example given in Breiman, Friedman et al. (1984) that was used to identify heart attack victims most at risk of dying was a simple three-question binary classification tree, shown in Figure 5-67.

Discriminant analysis of the heart attack data would produce a set of coefficients for a linear combination of blood pressure, patient age, and heart rate that best differentiates between low- and high-risk patients. The sum of each of the measurements weighted by the respective discriminant coefficient provides a single score based on the simultaneous evaluation of the three measurements.

For example, if  $BP_{min}$  is the minimum systolic blood pressure measured over a 24-hour period,  $A$  is the patient's age,  $T$  is a binary value (1, 0) representing the presence or absence of tachycardia are the predictor values, and  $b$ ,  $a$ , and  $t$  are the discriminant function coefficients, then a decision equation can be written based on a threshold value,  $Thresh$ , that is used to decide whether the patient is at high risk.

If

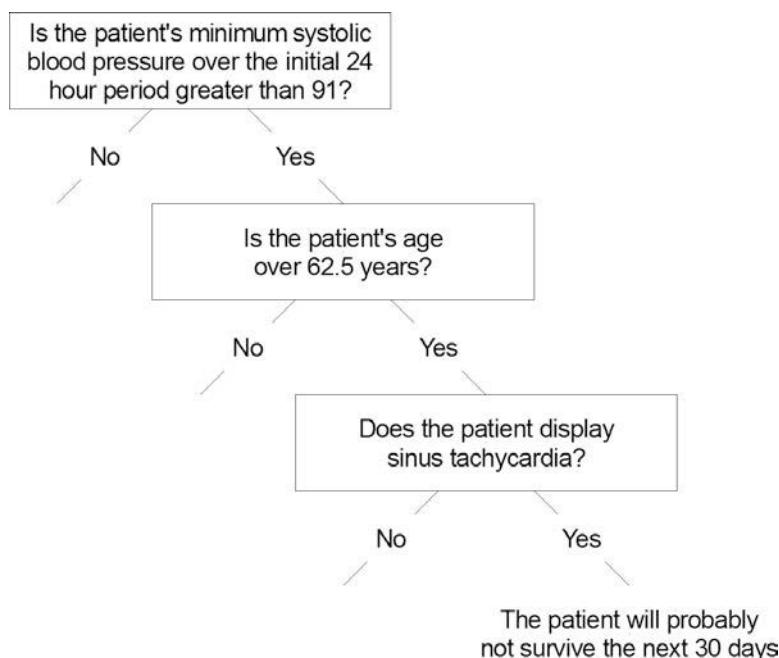
$$bBP_{min} + aA + tT > Thresh$$

Then

The patient is at high risk

End

**FIGURE 5-67** ■ Example of a simple classification tree used to determine the likelihood of a heart attack victim dying in the next 30 days. [Adapted from Breiman, Friedman et al. (1984).]



### 5.6.3.2 Clustering

In the case where the user has no prior knowledge of the relationships between groups, then  $k$ -means clustering is a simple ML algorithm commonly used in biomechatronics and medical imaging.

$k$ -means is an evolutionary unsupervised learning algorithm that clusters observations into  $k$  groups, where  $k$  is provided as an input parameter. Each observation is assigned to a specific cluster based on its proximity to the mean of that cluster. The cluster's mean is then recomputed and the process is repeated as described:

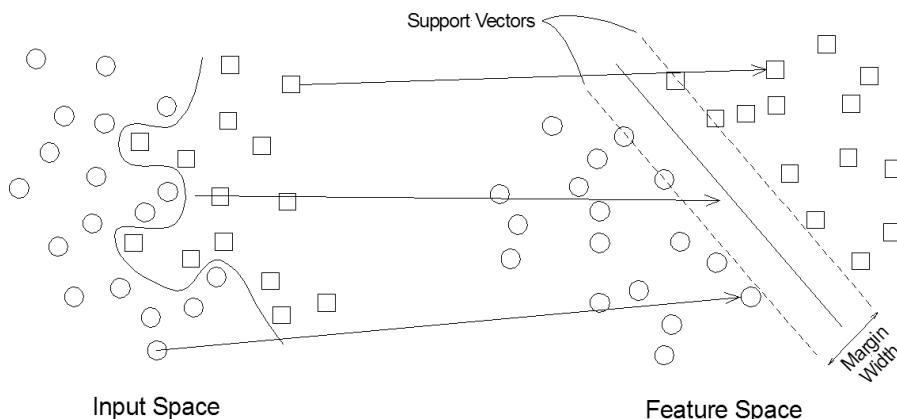
1. The algorithm arbitrarily selects  $k$  points as the initial cluster means.
2. Each point in the data set is assigned to a cluster based on the Euclidean distance between that point and each cluster mean.
3. Each cluster mean is recomputed as the average of the points in that cluster.
4. Steps 2 and 3 repeat until the clusters converge (typically when no change in the mean is seen when steps 2 and 3 are repeated).

A disadvantage of  $k$ -means is that the user specifies the number of clusters as an input to the algorithm. As designed, the algorithm is not capable of determining the appropriate number of clusters and depends on the user to identify this in advance. For this reason, it is a good idea to experiment with different values of  $k$  to identify the value that best suits the data.

### 5.6.3.3 Support Vector Machines

Support vector machines (SVMs) are a set of kernel-based supervised learning methods that perform their classification by constructing an  $N$ -dimensional hyperplane that separates the measured data into two categories. In most cases the raw data cannot be separated by a hyperplane, so it must be transformed by the kernel into another “feature” space that maximizes the separation between the two classes.

In the two-dimensional (2-D) example shown in Figure 5-68, the two different categories of data are represented by circles and squares. In the original data shown in the *input space*, a complex curve is required to separate the categories. However, the appropriate kernel transforms the data in such a manner that a straight line separates the two categories as shown.



**FIGURE 5-68 ■**  
Kernel transformation for a 2-D SVM.

In this example, the support vectors are shown as dashed lines that define the edges of each of the categories. Their orientations are defined to maximize the margin (distance between the two vectors).

## 5.7 | ISOLATION BARRIERS

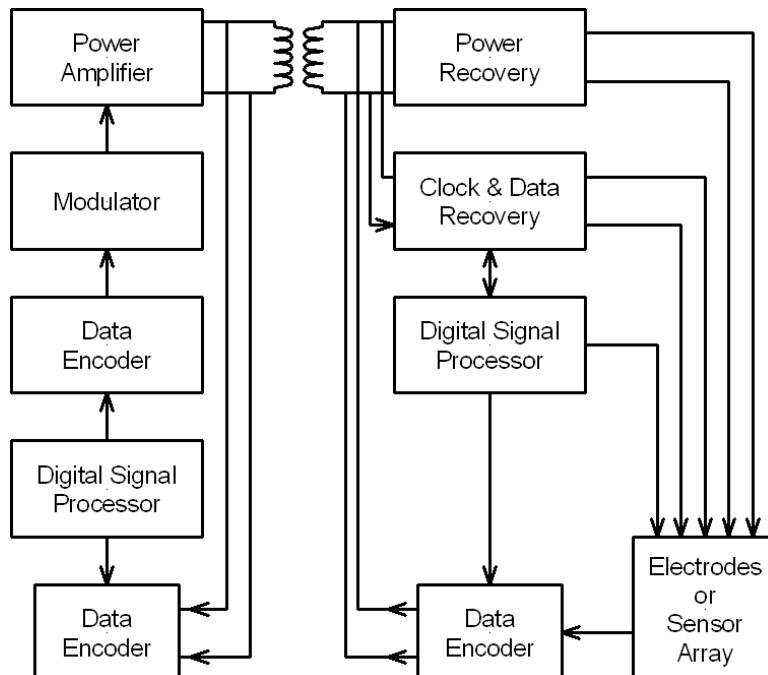
Many biomedical applications require that power be conveyed one way across an isolation barrier and that signals be conveyed in the other direction. These serve two purposes. In the first, isolation is required to ensure that, in the case of electronic failure, high voltages cannot be introduced into the patient through electrodes attached to the body. In the second, electronics are mounted under the skin or within the body cavity, so a method of charging internal batteries and obtaining telemetry is required.

### 5.7.1 Implant Systems

Most implant systems share a common framework, as shown in Figure 5-69, which consists of an external signal processing unit for biological sensory information (e.g., sound, images), a bidirectional telemetry unit, an internal signal processing module, a stimulus generator/driver, and an electrode array to interface to tissue or nerves (Finn and LoPresti, 2003).

It is envisaged that ultimately most prostheses will be totally internal, with sensors, processing, and actuation electrodes all self-contained within a biocompatible module. However, during the development phase or while the technology is immature, there are myriad advantages of including both internal and external modules, including decreased risk of adverse reaction to implanted materials, lower internal heat dissipation and, of course, ease of refinement to both the hardware and software.

**FIGURE 5-69** ■  
Schematic diagram for a generic implantable system  
[Adapted from (Finn and LoPresti 2003).]



Percutaneous connectors (physical wired links between implanted and exterior components) raise the risk of infection due to perpetual breach of the skin through which the wires must pass. In addition, anchor points can be a problem, and movement between the internal and external components can result in unnecessary stresses to the body and to the connection itself.

Transcutaneous (wireless) connections are therefore justified in most cases for bidirectional telemetry and to supply power to the embedded portion of the prosthesis. Three types of telemetry are considered to be appropriate under these circumstances for providing power and communication: optical, magnetic (inductive), and electromagnetic (radio frequency). They are discussed in the following sections (Finn and LoPresti, 2003).

Much work is being done in this field to improve the efficiency of radio frequency identification (RFID) tags, which work on similar principles, albeit in applications where the tag requires very little power other than to generate a response.

### 5.7.1.1 Optical Telemetry

Optical telemetry appears most commonly among research groups developing ocular prostheses as the cornea and lens are usually transparent. This form of telemetry usually uses high-energy light from a laser to power the photoelectronics that form part of the implant. In a prosthesis created at Massachusetts Institute of Technology (MIT) and Harvard University, an array of photodiodes was illuminated by a 30 mW 820 nm laser to produce a current of 300  $\mu$ A at 7 V.

### 5.7.1.2 Inductive Telemetry

Inductive telemetry has been the standard means of wireless connection to implanted devices for many years. It is based on the coupling between a pair of coaxial coplanar coils. The secondary is mounted subcutaneously along with the internal modules, while the primary coil remains exterior. A low-frequency carrier, usually between 1 and 10 MHz, supplies the primary coil, and power is coupled into the secondary coil where it is rectified and filtered to supply power. Additionally, the power carrier can be modulated to transmit data to the implant.

Problems with this method of power transfer include the fact that, because the coils have no cores, very little of the energy couples through to the other coil, and most of it radiates omnidirectionally. This results in a large drain on the batteries and the potential for electromagnetic interference.

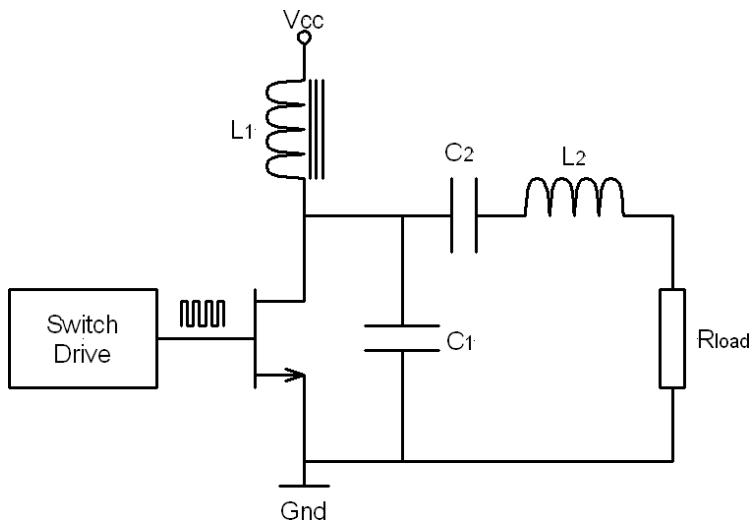
The coupling between coils is dependent on coil loading, the excitation frequency, and number of turns as well as coil alignment and spacing. Typical values for coil coupling vary from 0.01 and 0.1. However, this can be improved if the coils are series or parallel resonant using an appropriate capacitor. If  $L$  (H) is the self-inductance of the coil, then the capacitance,  $C$  (F), should be

$$C = \frac{1}{(2\pi f)^2 L} \quad (5.71)$$

The resonant coil is driven by a class-E amplifier (usually just a field-effect transistor [FET] switch) that shunts the supply to ground, as shown in Figure 5-70. As with conventional switch-mode power supplies, unused energy is not dissipated as heat but is traded back and forth between electric and magnetic fields in the resonant structure.

The low coupling coefficient to the secondary means that shifts in the load impedance have very little effect on the primary circuit. However, coil distortion and the proximity

**FIGURE 5-70** ■ Class-E amplifier driving a resonant coil.



of any metallic objects do have an effect, so feedback using a sensing coil is often used to maintain performance.

Data modulation usually follows standard digital protocols, and this is accomplished using amplitude, frequency, or phase-shift keying, though amplitude shift keying (ASK) is the most common. This can be achieved by altering the supply voltage or by shifting the switching frequency slightly to move off resonance. In this case a 10 Hz shift on a carrier frequency of 760 kHz is sufficient. However, because the resonant circuit is high Q, the data rate is limited.

Secondary coils are also resonant to maximize power transfer efficiency, and then conventional diodes are used to rectify the AC signal for use by the prosthesis.

### 5.7.1.3 Radio frequency Telemetry

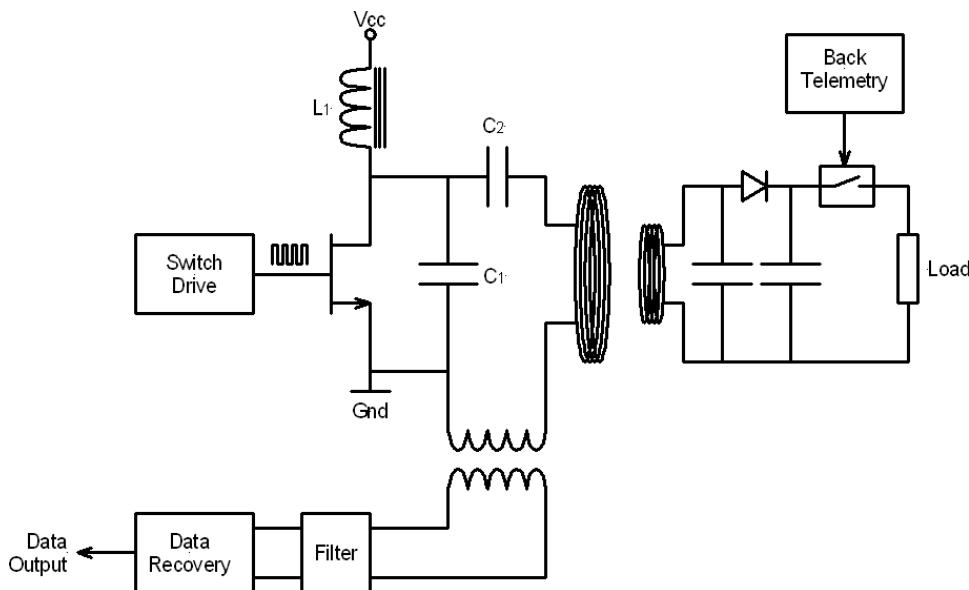
Radio frequency telemetry is similar to inductive coupling except that a traveling electromagnetic (EM) wave is employed rather than magnetic induction and antennas are used rather than coils. Problems with this method include the size of the antennas required and the dielectric constant and conductivity of the tissue, which introduces high losses.

### 5.7.1.4 Back Telemetry

Back telemetry refers to the capability of a prosthesis to relay information from the implanted electronics out to the exterior of the body. Optical means can be used in which a light-emitting diode (LED) is modulated to provide the signal or in magnetic systems a coil-based inductive link. An alternative is to use a passive method with a power transfer link. This is achieved by changing the loaded Q of the inductive link by altering the load impedance, as shown in Figure 5-71.

## 5.7.2 Isolation Amplifiers

Signal isolation enables digital or analog signals to transmit without galvanic connections between the transmitting and receiving side of the barrier. This in turn allows ground or reference levels on the transmitting and the receiving side to differ by thousands of volts and prevents circulating currents between differing ground potentials that might contaminate



**FIGURE 5-71** ■  
Back telemetry implemented by altering the secondary load.

signals. Noise on a signal ground may corrupt the signal; isolation can separate the signal to a clean signal subsystem ground.

In addition, a galvanic connection between reference levels might provide a current path that would present a safety risk to an operator or medical patient. To avoid this, safety regulations in medical care prescribe maximum leakage currents in case a patient should touch ground or mains power during a bioelectric recording (Webster, 1978). Although the regulations vary among countries, sufficient galvanic isolation and an isolation capacitance less than 800 pF are usually demanded (maximum leakage current is  $50 \mu\text{A}_{rms}$  when the patient touches the line voltage of  $220 \text{ V}_{rms}$ , 50 Hz).

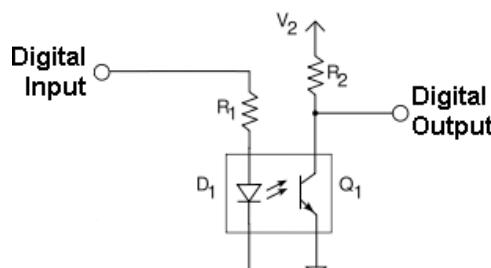
Isolation can be realized using three different methods: transformer isolation, capacitor isolation, and opto-isolation. Because there is no current flow across the barrier, the potential difference across it is determined by the difference between the input and output grounds (common points).

Power transfer is not a problem in instrumentation because magnetic coupling between the primary and secondary coils can be enhanced by using magnetic cores. Conventional DC/DC converters with isolation voltages of many kV are available with efficiencies of 70 to 90%. These are discussed in detail in Chapter 2. In some cases, particularly if a fiber optic link is used for signal isolation batteries, this technique can be used to provide an isolated power source.

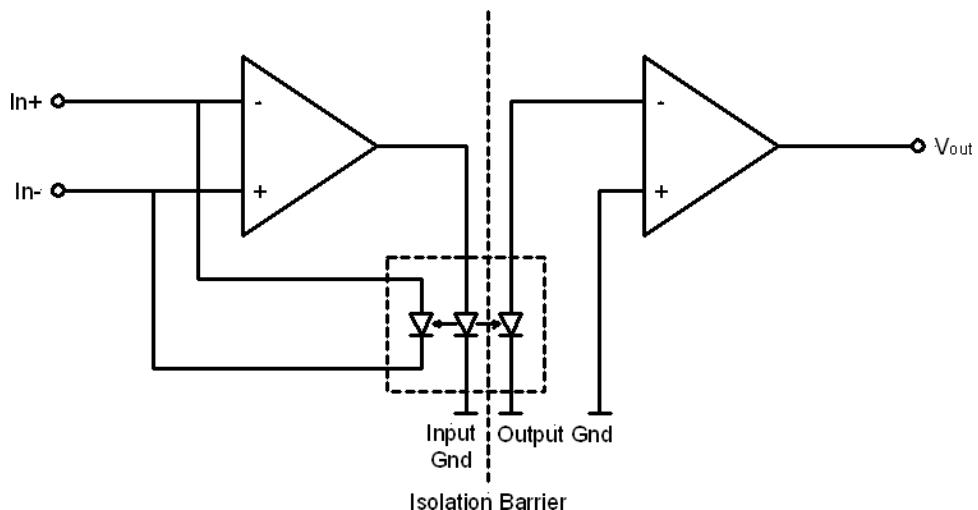
Personal computers (PCs) for display and analysis of medical data pose a major isolation problem. The design of power supplies in PCs permits high-leakage currents that are inadequate for medical environments, and even if the patient is separated from the PC by an isolation barrier care must be taken not to touch them simultaneously because of the danger of electrocution. If PCs are used in this role, then medical-rated devices are the right choice, or, as a last resort, an additional medical-grade isolation transformer should be installed between the PC and the mains supply (Bronzino, 2006).

Galvanic isolation in digital systems is easily achieved using opto-isolators. These devices consist of a LED physically separated from a phototransistor, as shown in Figure 5-72. The digital input on-off modulates the LED output, which in turn changes

**FIGURE 5-72** ■  
Digital opto-isolator schematic.



**FIGURE 5-73** ■  
Opto-isolated analog link.



the current flow through the phototransistor in an identical manner. Note that in this configuration the digital signal is inverted.

Galvanic isolation of analog signals can be achieved in a number of ways. The simplest is to convert the analog signal to a digital one, using pulse-width modulation, voltage-to-frequency conversion, or even an ADC to generate a digital word. This digital information is then transmitted across the barrier using the method shown in Figure 5-72 before conversion back to an analog signal.

For pulse-width modulation, the conversion is achieved using a simple low-pass filter, whereas in the frequency-encoded signal a frequency-to-voltage converter is used. In the case where a digital word is transmitted, a conventional DAC converts the word back to its analog equivalent.

An interesting alternative is a true analog link, where the intensity of the LED encodes for the analog voltage, as shown in Figure 5-73. In this configuration, the isolation consists of a pair of matched photodiodes and an LED. The signal from the photodiode is fed back to the input of the amplifier to counteract the nonlinear characteristics of both the LED and the photodiodes.

## 5.8 | REFERENCES

- Alciatore, D. and M. Histand. (2003). *Introduction to Mechatronics and Measurement Systems*, 2d ed. Boston: McGraw Hill.
- Benedict, T. and G. Bordner. (1962). "Synthesis of an Optimal Set of Radar Track While Scan Smoothing Equations." *IRE Transactions on Automatic Control* AC-7: 27–32.

- Blackman, S. (1986). *Multiple-Target Tracking with Radar Application*. Norwood, MA: Artech House.
- Breiman, L., J. Friedman, et al. (1984). *Classification and Regression Trees*. Boston: Wadsworth.
- Bronzino, J. (Ed.). (2006). *Medical Devices and Systems*. Boca Raton, FL: CRC Press.
- Brooker, G. (2008). *Introduction to Sensors for Ranging and Imaging*. Raleigh, NC: SciTech.
- Carlson, G. (1998). *Signal and Linear System Analysis*, 2d ed. New York: John Wiley & Sons, Inc.
- Carr, J. (1997). *Electronic CircuitGuidebook: Sensors*. Indianapolis: Prompt.
- Cromwell, L., F. Weibell, et al. (1973). *Biomedical Instrumentation and Measurements*. Englewood Cliffs, NJ: Prentice-Hall, Inc.
- Edminster, J. (1972). *Shaum's Outline Series: Theory and Problems of Electric Circuits*. New York: McGraw Hill.
- Finn, W. and P. LoPresti (Eds.). (2003). *Handbook of Neuroprosthetic Methods*. London: CRC Press.
- Kuphaldt, T. (2003). "Lessons in Electric Circuits: Volume III—Semiconductors." Retrieved August 2008 from <http://www.faqs.org/docs/electric/Semi/index.html>
- Mahafza, B. (2000). *Radar Systems Analysis and Design Using MATLAB*. Boca Raton, FL: Chapman & Hall/CRC.
- Stanley, W. (1975). *Digital Signal Processing*. Reston, VA: Reston Publishing Company, Inc.
- Walpole, R. and R. Myers. (1978). *Probability and Statistics for Engineers and Scientists*, 2d ed. New York: Macmillan.
- Webster, J. (Ed.). (1978). *Electrical Safety in Medical Instrumentation: Application and Design*. Boston: Houghton Mifflin Co.
- Young, P. (1990). *Electronic Communication Techniques*, 3d ed. New York: Merrill.



# Hearing Aids and Implants

## Chapter Outline

6.1	Introduction .....	277
6.2	What Is Sound? .....	278
6.3	How Hearing Works .....	281
6.4	Hearing Loss .....	285
6.5	Hearing Aids .....	289
6.6	Bone Conduction Devices .....	300
6.7	Middle Ear Implants .....	302
6.8	Direct Acoustic Cochlear Stimulatory Devices .....	312
6.9	Cochlear Implants .....	314
6.10	Auditory Brainstem Implants .....	328
6.11	References .....	330



## 6.1 | INTRODUCTION

This chapter examines some of the reasons for hearing loss as well as the amazing technological advances made over the past 100 years that have led to the development of sophisticated prostheses permitting the deaf and hard of hearing to hear again. These

range from conventional hearing aids for people who retain some hearing to cochlear and brainstem implants for those with physical damage to the cochlea and auditory nerves, as well as myriad other devices that address damage to different sections of the auditory pathway.

## 6.2 | WHAT IS SOUND?

It is interesting and informative to understand the nature of sound before considering its interaction with the auditory system. Unlike electromagnetic radiation, which can propagate through space, acoustic waves are transmitted through alternate compression and expansion of the air (or another medium) that propagate in the direction of the gradient, as shown schematically in Figure 6-1. These pressure fluctuations are tiny. For example, a very loud sound may involve a change in pressure of 20 Pa, which equates to a localized change in atmospheric pressure<sup>1</sup> of only 0.02%.

### 6.2.1 Characteristic Impedance ( $Z$ ) and Sound Pressure

As sound propagates through a medium, each small volume element of that medium oscillates about its equilibrium position. For pure harmonic motion, the displacement is

$$y = y_m \cos \frac{2\pi}{\lambda} (x - ct) \quad (6.1)$$

It can be shown that the pressure exerted by the sound wave is

$$p = (k\rho_o c^2 y_m) \sin(kx - \omega t) \quad (6.2)$$

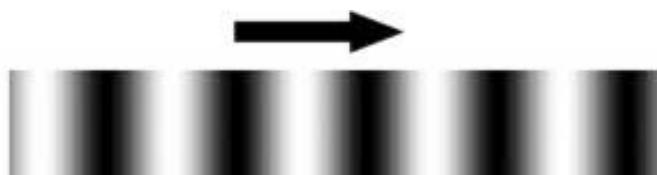
where  $k = 2\pi/\lambda$  is the wave number,  $\rho_o$  (kg/m<sup>3</sup>) is the air density,  $c$  (m/s) is the speed of sound, and  $\omega$  (rad/s) is the angular frequency.

The amplitude of the wave is the acoustic pressure  $p_m = k\rho_o c^2 y_m$ , so a sound wave may be considered a pressure wave with a phase offset of 90° to that of the displacement wave. The pressure at any point changes sinusoidally with time, around a mean value, and the root mean square (RMS) (effective) value of this fluctuating component is known as the acoustic pressure,  $P$ . This can be measured using an appropriately calibrated microphone.

The characteristic impedance of the medium is defined as the ratio of the acoustic pressure and the RMS volume velocity,  $\xi$ , as described in equation (6.3). Note that this is not the wave velocity.

$$Z = \frac{P}{\xi} \quad (6.3)$$

**FIGURE 6-1** ■  
Schematic snapshot  
of sound as a  
compressive wave.



<sup>1</sup>Atmospheric pressure is  $1.01325 \times 10^5$  Pa.

This is a complex quantity, but for propagation in an idealized medium (with no loss)  $Z$  is real and reduces to a proportionality factor that equates intensity to the square of the acoustic pressure.

$$Z = \rho_o c \quad (6.4)$$

For air with  $\rho_o = 1.3 \text{ kg/m}^3$  and  $c = 332 \text{ m/s}$ , this makes  $Z_{\text{air}} \approx 400$  acoustic ohms, whereas in water  $\rho_o = 1000 \text{ kg/m}^3$  and  $c = 1400 \text{ m/s}$  makes  $Z_w = 1.4 \times 10^6$  acoustic ohms. This large difference in the characteristic impedance of the two mediums requires that the ear performs an effective impedance transformation to maximize power transfer as sound travels from the outer to the inner ear.

The proportion of the sound reflected at the interface between air and water is given by

$$\Gamma = \left( \frac{Z_w - Z_{\text{air}}}{Z_w + Z_{\text{air}}} \right)^2 \quad (6.5)$$

The proportion transmitted,  $T$ , is the remainder

$$T = 1 - \Gamma = \frac{4Z_w Z_{\text{air}}}{(Z_w + Z_{\text{air}})^2} = 0.0011 \quad (6.6)$$

Therefore, if the outer and inner ears were just separated by a membrane without the intervening ossicles, only a tiny fraction of the sound would propagate through. The impedance matching mechanism that maximizes acoustic power transfer to the inner ear is discussed later in this chapter.

## 6.2.2 Sound Intensity ( $I$ )

Intensity of a sound is defined as the power transferred per unit area, or the sound pressure level (SPL), and can be expressed in terms of the acoustic impedance in an equation that is similar to ohms law for electrical circuits.

$$I = P\xi = \frac{P^2}{Z} \quad (6.7)$$

where  $I$  ( $\text{W/m}^2$ ) is the sound intensity (SPL),  $P$  ( $\text{N/m}^2$  or Pascal) is the acoustic pressure, and  $\xi$  ( $\text{m}^3/\text{s}$ ) is the volume velocity.

It is common to specify the sound level not as an intensity but as a ratio,  $\beta$ , in dB relative to the minimum audible threshold, the softest sound that can be heard by a human being and equal to  $P_o = 20 \times 10^{-6} \text{ N/m}^2$  ( $2.9 \times 10^{-9} \text{ psi}$ ). In reality, this threshold is a function of frequency as described by the Fletcher–Munson curve, with peak sensitivity occurring between 2 and 4 kHz.

The reference intensity,  $I_o$ , can be determined from  $P_o$  using

$$I_o = \frac{P_o^2}{Z} = \frac{(2 \times 10^{-5})^2}{400} = 10^{-12} \text{ W/m}^2$$

The ratio is therefore

$$\beta = 10 \log_{10} \frac{I}{I_o} = 20 \log_{10} \frac{P}{P_o} \quad (6.8)$$

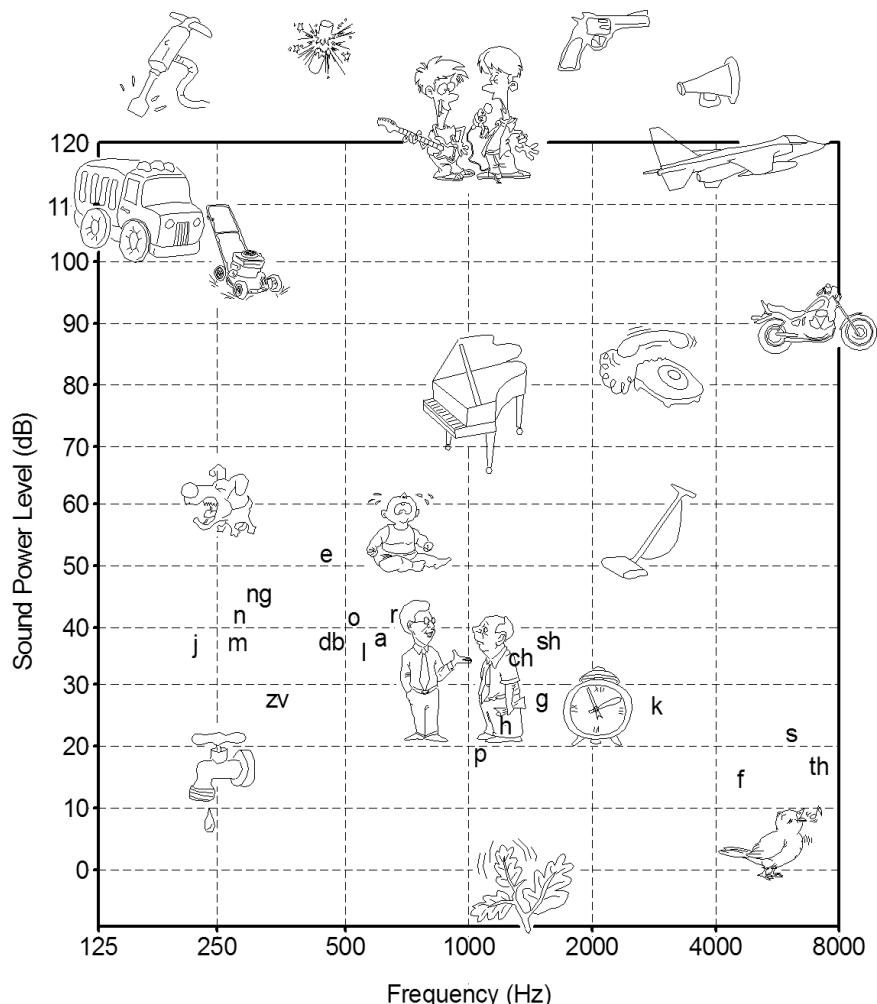
Table 6-1 and Figure 6-2 together give some insight into the relative levels of some common sounds. It shows that the normal range of auditory input extends from 0 up to about 120 dB, with the pain threshold 10 dB higher than this. What is really incredible is that the range of human hearing spans 12 orders of magnitude.

**TABLE 6-1** ■ Sound Intensity Levels

Sound	Level (dB)
Stream flow, rustling leaves	15
Watch ticking, soft whisper	20–30
Quiet street noises	40
Normal conversation	45–60
Normal city or freeway traffic	70
Vacuum cleaner	75
Hair dryer	80
Motorcycle, electric shaver	85
Lawn mower, heavy equipment	90
Garbage truck	100
Screaming baby	115
Racing car, loud thunder, rock band	120–130
Jet airplane's takeoff from 35 m	120
Pain threshold	130
Rocket launch from 45 m	180

**FIGURE 6-2** ■

Illustrated audiogram showing the SPL and frequency.



## 6.3 | HOW HEARING WORKS

### 6.3.1 The Outer Ear

The ear provides one of the five senses of the human body. The outer ear, comprising the pinna and the outer ear canal, is the sound-gathering portion of the anatomy, as seen in Figure 6-3. The pinna performs a number of important functions, including the first of a number of impedance transformations, sound amplification, and direction finding. Our sense of directionality is determined by the differential arrival times at the outer ear in conjunction with the binaural difference in arrival times and loudness at the auditory complex of the brain.

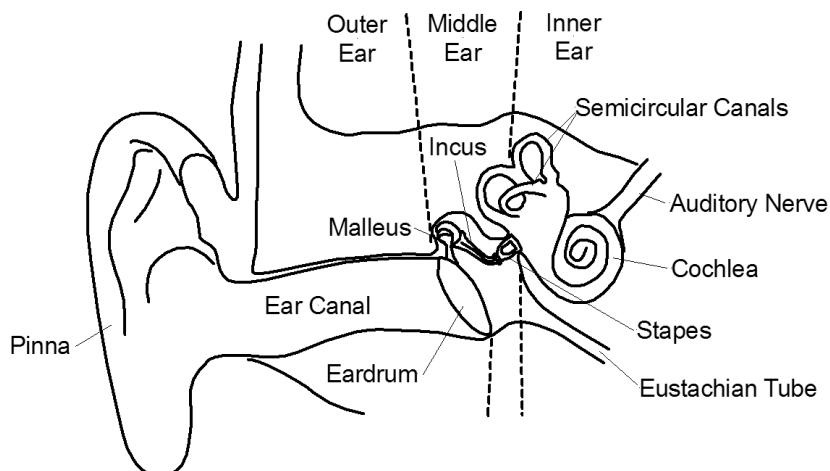
### 6.3.2 The Middle Ear

The outer ear is connected to the middle ear by the tympanic membrane (eardrum) through the ear canal. The middle ear houses the ear ossicles (malleus, incus, and stapes), which couple sound vibrations through from the tympanic membrane via the oval window into the cochlea.

The handle of the malleus is attached to the tympanic membrane, and the tensor tympani muscle, also attached to the malleus, regulates the tension on this membrane. The incus is attached to the malleus and to the third ossicle in the chain, the stapes, which is in turn attached via its footplate to the oval window of the cochlea. The stapedius muscle regulates the range of motion of the stapes. The two muscles (tensor tympani and stapedius) adjust the amplitude sensitivity of the ear and are instrumental in achieving its incredible dynamic range as well as reducing the perceived amplitude of our own voices as we speak.

It was shown earlier in this chapter than only a small fraction of the acoustic power would be transmitted from the air through to a fluid due to the large difference in their relative acoustic impedances ( $T_{dB} = 10\log_{10} 0.0011 = -29.6 \text{ dB}$ ) unless the impedances are matched in some way.

The middle ear converts variations in air pressure into variations in fluid pressure in the inner ear (perilymph in the cochlea). This impedance change is accomplished by the

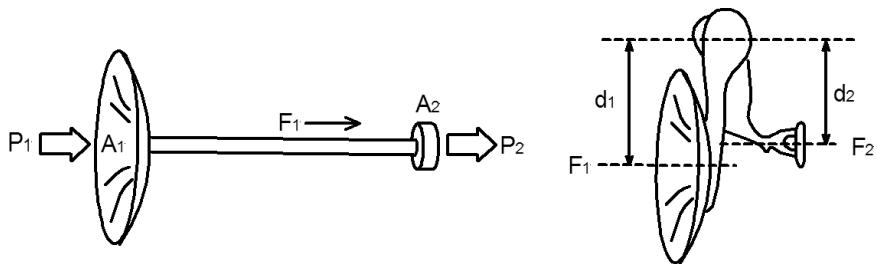


**FIGURE 6-3 ■**  
Schematic diagram showing the various parts of the ear.

**FIGURE 6-4 ■**

Mechanisms that increase the sound pressure from the tympanic membrane to the oval window.

- Decrease in area.
- Lever mechanism.



anatomy and mechanics of the middle ear, with the primary conversion being the relative surface areas of the tympanic membrane and the footplate of the stapes, as shown in Figure 6-4.

In Figure 6-4a, the force remains unchanged, but the cross sectional area decreases from  $A_1$  (the area of the tympanic membrane) to  $A_2$  (the area of the footplate of the stapes); therefore,

$$F_1 = P_1 A_1 = P_2 A_2 \quad (6.9)$$

which increases the pressure at the oval window in the ratio of the two cross sectional areas

$$P_2 = P_1 \frac{A_1}{A_2} \quad (6.10)$$

In Figure 6-4b the lever arm effect is determined by equating the moment about the rotation axis at the head of the malleus

$$F_1 d_1 = F_2 d_2 \quad (6.11)$$

which makes the output force

$$F_2 = F_1 \frac{d_1}{d_2} \quad (6.12)$$

For a typical human being the surface area of the tympanic membrane is  $A_1 = 60 \text{ mm}^2$ , and that of the footplate of the stapes is  $A_2 = 3.2 \text{ mm}^2$ . This provides a pressure gain of

$$G_{area} = 20 \log_{10} \frac{60}{3.2} = 25.5 \text{ dB}$$

The second effect is obtained by a small reduction in displacement of the oval window compared with that of the tympanic membrane because the long arm of the malleus,  $d_1$ , is about 1.3 times longer than the incus,  $d_2$ . This provides an increase in force and ultimately a pressure gain of

$$G_{lever} = 20 \log_{10} 1.3 = 2.3 \text{ dB}$$

Together, these effects provide a pressure gain,  $G = 27.8 \text{ dB}$ , which makes up for much of the mismatch loss.

For normal sound transmission at reasonably low frequencies, the ossicles act as a solid body, and the footplate acts as a piston pushing on the oval window. The transfer function for the system describes the relationship between the acoustic input pressure

at the tympanic membrane and the output pressure applied to the stapes footplate that couples into the vestibular canal. This transfer function is important for prosthetic devices that drive the ossicles or the oval window electrically if the tympanic membrane has been destroyed.

The forces involved are very small but can be calculated because the area of the tympanic membrane is known and the acoustic pressure can be measured. For a sound signal at the minimum audible threshold, the pressure  $P_o = 20 \times 10^{-6} \text{ N/m}^2$  is applied to the tympanic membrane area of  $59.4 \text{ mm}^2$ , making the RMS force

$$F = P_o A = 2 \times 10^{-5} \times 59.4 \times 10^{-6} = 1.2 \text{ nN}$$

For normal speech levels of 60 dB above the minimum audible threshold, the acoustic pressure increases to  $P = 0.025 \text{ N/m}^2$ , and the force increases to  $1.5 \mu\text{N}$ .

### 6.3.3 The Inner Ear

The inner ear includes the cochlea and the semicircular canals. The cochlea is a fluid-filled chamber where fluid movement is converted into nerve action potentials by hair cells. Each hair cell has fine rods of protein, called stereocilia, emerging from the one end. Some 30,000 of these hair cells, arranged in four rows, are attached to the top of the basilar membrane in a matrix of cells called the organ of Corti. They operate as miniature displacement transducers responding to displacements of the basilar membrane relative to the perilymph.

Deflection of the hairs in one direction increases the release of a chemical transmitter at the base of the hair cell, while deflection in the other direction inhibits its release. Variations in the concentration of this chemical transmitter alter the discharge rate of nearby neurons that make up the spiral ganglion. Changes in this neural activity are transmitted to the brain via the auditory nerve.

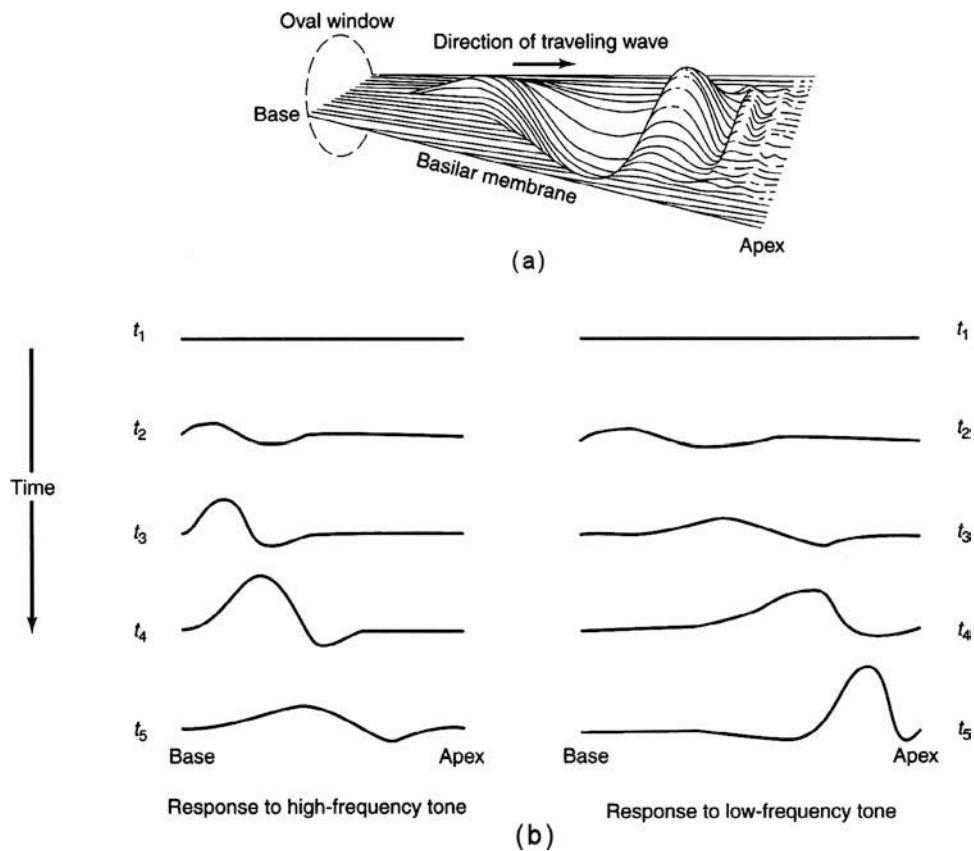
The cochlea consists of three spiral canals, two of which are separated by the basilar membrane, part of the organ of Corti. The basilar membrane, which gets wider and more flaccid as the cochlea gets narrower, conveys the vibratory movement along its length as a traveling wave (much like a snapping rope). This mechanism is illustrated in Figure 6-5. The amplitude of this wave reaches a peak at a location that depends on frequency, as shown in the two examples in Figure 6-4. High-frequency peaks occur toward the base of the basilar membrane, up to 20 kHz (where the membrane is stiffest and narrowest), whereas the low-frequency peaks occur toward the apex, down to 20 Hz. This spatial relationship with frequency sensitivity is called tonotopic organization.

The outer hair cells on the basilar membrane also have a contractile function (actin filaments) and serve as controllable amplifiers for the inner hair cells, which send action potentials to the auditory cortex via the spiral ganglion. Loss of the outer hair cells results in about a 60 dB loss in hearing.

### 6.3.4 Hearing Statistics

The undamaged ear is incredibly sensitive, with the threshold of hearing equating to a displacement of only  $10^{-11} \text{ m}$  of the tympanic membrane. It can also accommodate a dynamic range of 120 dB in sound pressure before damage occurs.

**FIGURE 6-5** ■ The basilar membrane of the cochlea. (a) Uncoiled and flattened showing a travelling wave in 3D. (b) 2D depictions of the resonances for traveling waves of different frequencies.



Some interesting physical characteristics are as follows.

#### Auditory Canal

Cross section	0.3–0.5 cm <sup>2</sup>
Diameter	0.7 cm
Length	2.7 cm
Volume	1 cc

#### Tympanic Membrane

Area	0.5–0.9 cm <sup>2</sup> (roughly circular)
Thickness	≈0.1 mm
Displacement	$10^{-8}$ mm (threshold of hearing at 1 kHz) $10^{-1}$ mm (threshold of discomfort, low-frequency tones)

#### Middle Ear

Total Volume	2 cc
Malleus	23 mg, 8–9 mm long
Incus	27 mg, 5 mm by 7 mm
Stapes	3 to 4 mg, 3.5 mm high, footplate 1.4 × 3.2 mm

**Cochlea**

Length	35 mm (cochlear channels)
Tympanic canal	1 mm high
Round window	2 mm <sup>2</sup>

To give some perspective to the incredible sensitivity of the ear, consider that the radius of a hydrogen atom is  $5.3 \times 10^{-11}$  m (Bohr radius), which is five times larger than the required displacement of the tympanic membrane at the threshold of hearing.

## 6.4 | HEARING LOSS

Hearing loss is defined as a deterioration in hearing ability, which, when it becomes profound, is referred to as deafness. Hearing loss, which becomes much more common among older people, has many causes, some of which are discussed in the following section. Between 30 and 40% of people aged 65 and older are significantly affected. Some children are born deaf or develop hearing loss in their first few years, and as this is detrimental to language and social development it is important to diagnose any hearing issues as early as possible.

### 6.4.1 Causes

The most common cause of hearing loss is the long-term exposure to loud noise. However, even the brief exposure to very loud noise can permanently harm hearing. It is interesting to note that as industrial exposure to loud noise has been reduced thanks to increased awareness and institutional use of personal protective equipment (PPE), hearing loss caused by listening to loud music has become more common.

Degradation in hearing capability may be caused by a mechanical problem in the ear canal or the middle ear that disrupts the conduction of sound. This is known as conductive hearing loss. Blockage in the external ear canal can be caused by a buildup of wax, the insertion of a foreign object, or the growth of a tumor. The most common cause of conductive hearing loss in the middle ear is an accumulation of fluid. This is most often a childhood problem brought about by an infection that blocks the Eustachian tube. This tube drains fluid from the middle ear and helps equalize the pressure across the eardrum.

Hearing loss may also be caused by damage to the eardrum itself or to the bones that conduct sound through the middle ear into the cochlea. Damage to the hair cells of the inner ear, the auditory nerve, or auditory nerve pathways to the brain—known as sensorineural hearing loss—lead to hearing loss and can be caused by drugs, infection, tumors, or injuries to the skull.

Age-related hearing loss is called presbycusis and occurs either as the structures of the ear become less elastic and undergo changes that make them less able to respond to sound waves or to the die-off of sensory hair cells. In many people, this process is hastened by exposure to loud noise over many years. In most people, age-related hearing loss starts at about 20 but doesn't become noticeable for a further 30 years.

As a person ages, hearing loss first affects high-frequency sensitivity and slowly works down the spectrum. Loss of ability to hear high frequencies often makes understanding speech difficult because it reduces the intensity of some of the consonants, particularly C, D, K, P, S, and T, as illustrated in Figure 6-2. Sounds become muffled, and to people with

hearing loss it appears that everyone around them is mumbling. The voices of women and children, which are higher, tend to be the most difficult to understand.

At some stage during their lives, most people are exposed to noise levels that can cause hearing loss. Loud sounds generally damage the sensory hair cells in the inner ear, with the rate of destruction being proportional to the sound intensity and the duration. A single extremely loud noise can lead to immediate hearing loss, and though it is often only temporary, lasting a few hours to a day, it can become permanent if the sound is repeated often enough. Common sources of noise, as listed in Table 6-1, include loud music, particularly if listened to through headphones, power tools, machinery, and many petrol-powered vehicles, particularly chainsaws and lawn mowers. Explosions and gunfire can also damage hearing.

As mentioned earlier, young children often suffer from some degree of conductive hearing loss after an ear infection (otitis media) because the infection often leads to a buildup of fluid in the middle ear. Chronic infections of the middle ear often result in both conductive and sensorineural hearing loss.

Autoimmune disorders such as rheumatoid arthritis, systemic lupus erythematosus, and polyarteritis nodosa may also lead to hearing loss. These generally result in fluctuations in hearing ability that progressively become worse as the immune system attacks cells in the cochlea.

Drugs, particularly intravenous antibiotics such as vancomycin, can destroy sensory hair cells and cause permanent hearing loss if given in high doses. Other drugs implicated include quinine and chemotherapy drugs like cisplatin. Aspirin can also cause reversible hearing loss.

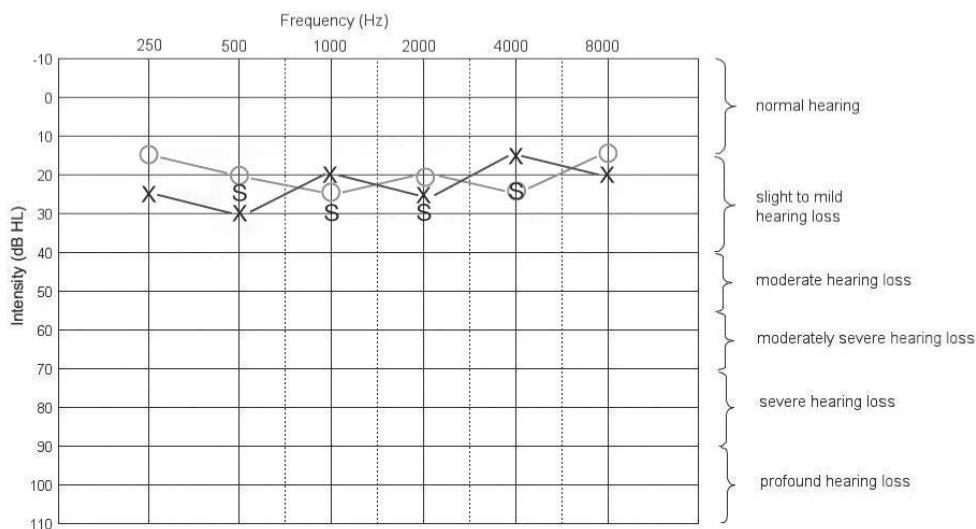
Sudden deafness is the type that occurs within minutes or over a few hours. It may be caused by something as trivial as wax accumulation or by head trauma, sudden changes in pressure (as occurs in aircraft), or internal pressure changes caused by severe straining (as may occur with weight lifting or blowing one's nose).

Hearing loss that is more severe in one ear than the other may be caused by a tumor. These are mostly benign but should be removed as they can lead to tinnitus, difficulty with balance, and facial numbness as well as deafness if left to grow unchecked.

### 6.4.2 Diagnosis

Audiologists perform hearing tests that can determine the degree of hearing loss over a specific frequency range. As a first step, a sequence of sounds is generated with random pitch and amplitude and the patient indicates whether each can be heard. During these tests, which are conducted one ear at a time, white noise is played into the other ear. This test determines the frequency response of each ear and is presented relative to the normal hearing range, as shown in Figure 6-6. In this figure, X shows the response of the left ear, O shows the response of the right, and S shows the response of both ears together.

Other tests include speech threshold audiometry in which the patient identifies two-syllable words at different volumes. The volume above which 50% of the words are correctly identified is considered to be the threshold. Discrimination tests involve presenting pairs of similar one-syllable words. People with conductive hearing loss usually have a normal discrimination score, albeit at a higher volume, whereas people with sensorineural hearing loss often have abnormal discrimination at all volumes.



**FIGURE 6-6 ■**  
Audiogram showing  
mild hearing loss.

The following MATLAB script plays a sequence of sounds of different frequencies in random order. If played through an audio system with a frequency response that extends beyond 20 kHz, it can be used as a rough method to identify hearing loss.

```
%hearing test
%audio_tones.m

fs = 44.1e3; % Sample frequency (Hz)
duration = 5; % Duration of each tone (s)
ts = 1/fs:1/fs:duration;
window = hanning(length(ts))';

% random selection of frequencies between 0 and 22000 (Hz)
freq = [500, 1500, 14000, 2500, 0, 300, 16000, 11000, 16500, 15000, ...
    17500, 20000, 14500, 13000, 700, 15500, 18000, 17200, 14800, 3000, ...
    4400, 0, 19000, 16800, 12000, 17500, 15800, 7000, 22000, 250];

for i = 1:length(freq)
    tone = sin(2*pi*freq(i).*ts);
    tone = tone.*window;
    sound(tone, fs)
    disp(['Playing Tone ' num2str(i)]);
    pause(duration)
    disp(['Complete - Press key for next tone']);
    pause
end
```

More complicated tests include tympanometry, which determines how well sound passes through the eardrum into the middle ear by examining sound reflection from the eardrum at different pressures. The Rinne tuning fork test compares how well the patient

hears sounds conducted through the air compared with how well sounds are heard when conducted through the bones of the skull. If there is hearing loss through both the air and bone, then the problem is probably sensorineural.

Auditory brain stem response is a test that measures nerve impulses in the brain stem resulting from sounds played into the ears. This information helps determine what kind of signals the brain is receiving, and abnormalities can be indicative of sensorineural problems or many types of brain tumor.

Electrocotchleography measures the activity of the cochlea and the auditory nerve by means of an electrode placed on, or through, the eardrum. This test and the previous one can be used to measure hearing in people who cannot or will not respond voluntarily to sound. For example, it is used to determine whether infants or young children are deaf or whether a person is faking or exaggerating hearing loss.

Another interesting test uses sound to stimulate the cochlea. The ear then generates a low-intensity sound that matches the stimulus. These cochlear emissions can be recorded using sophisticated electronics and are used routinely to screen newborns for congenital hearing loss as well as other brain functions.

Imaging tests such as computed tomography (CT) or magnetic resonance imaging (MRI) scans can be used to look for physical damage to the middle or inner ear and the Eustachian tubes.

### 6.4.3 Treatment

Treatment depends on the cause. Blockages to the external ear are easily removed using solvents or minor surgery. Fluid buildup in the middle ear caused by infection may require the insertion of a grommet into the eardrum, a procedure known as a tympanostomy, to reduce the pressure and ensure that permanent damage is not done. Damage to the eardrum or the bones of the middle ear may require reconstructive surgery, the insertion of a middle ear implantable hearing device (MEIHD), or direct acoustic cochlear stimulation (DACS). Brain tumors causing hearing loss may, in some cases, be removed and the hearing preserved.

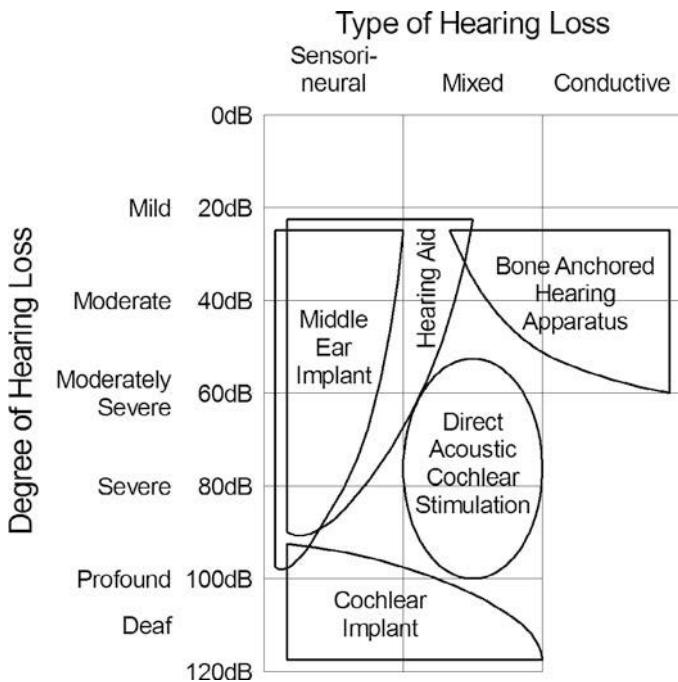
Most other causes of hearing loss have no cure, and in these cases treatment involves compensating for hearing loss using hearing aids or in some cases cochlear or brain stem implants. In moderate to severe hearing loss, the reduction in hair cell numbers is such that the sensitivity to sounds is reduced, but enough still remain for sound amplification or increased ossicular movement to restore adequate hearing.

If the hair cell population is so low that neither acoustic amplification by hearing aids nor mechanical overdrive by MEIHDS or DACS can provide adequate restoration, then direct stimulation of the spiral ganglion cells by passing an electric current across these nerves can be used. This bypasses the hair cell transducers and initiates action potentials to the first relay in the auditory pathway to the brain.

In the case where the eighth cranial nerve has been damaged, usually during surgery to remove tumors, then neural prostheses are available that stimulate the cochlear nucleus in the brain stem.

Figure 6-7 relates the degree of hearing loss and the type to one of a range of prosthetic methods that can be applied to correct for the loss.

These electrical and electromechanical prostheses are of interest to biomechatronics and are discussed in some detail in later sections of this chapter.

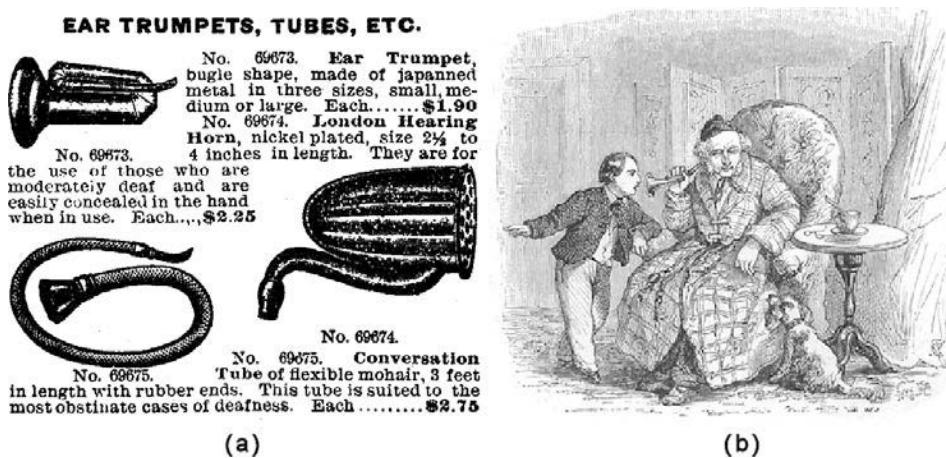


**FIGURE 6-7** ■ Therapies according to the degree and type of hearing loss. [Adapted from (Bernhard, Steiger et al., 2011).]

## 6.5 | HEARING AIDS

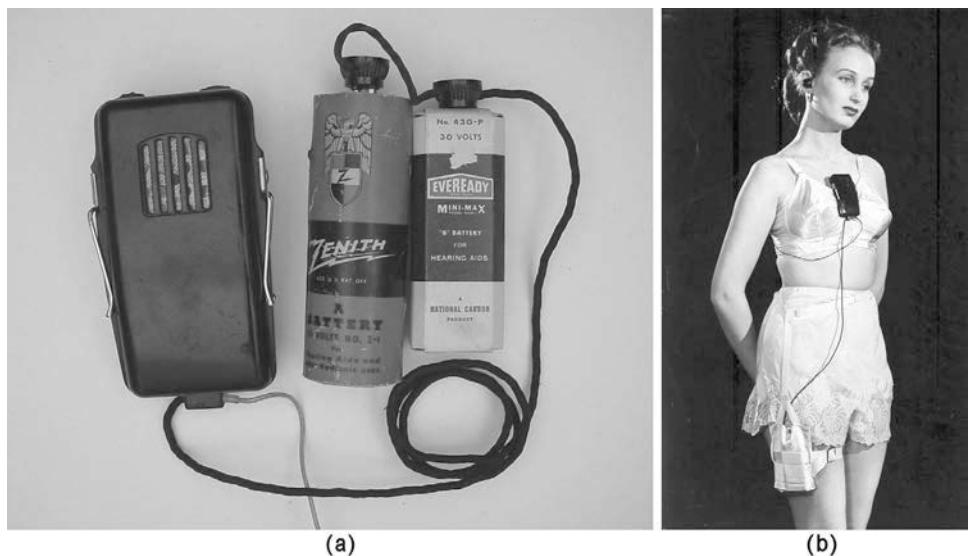
### 6.5.1 History

Until the advent of electronic amplification, hearing aids consisted of horns (ear trumpets) of various sizes and sophistication, as shown in Figure 6-8, which increased the size of the sound collection aperture and therefore provided some amplification. The effective gain of these horns is approximately equal to the ratio of the input to the output cross sectional area.



**FIGURE 6-8** ■ Early hearing aids.  
(a) Advertisements for ear trumpets and tubes. (b) Engraving showing an ear trumpet in use. (Ganot 1872), with permission.

**FIGURE 6-9 ■**  
 Western Electric vacuum tube hearing aid. (a) Hearing aid amplifier with batteries. (b) Model wearing the hearing aid. (Reproduced courtesy of Becker Medical Library, Washington University School of Medicine, with permission.)



The introduction of vacuum tubes in the 1920s was a major technological advance in the evolution of hearing aids. Vacuum tubes provided an efficient method of amplifying electrical signals that allowed for more gain and therefore provided a greater benefit for hard of hearing or deaf users. Early devices were too large for portable operation and were confined to tabletop use.

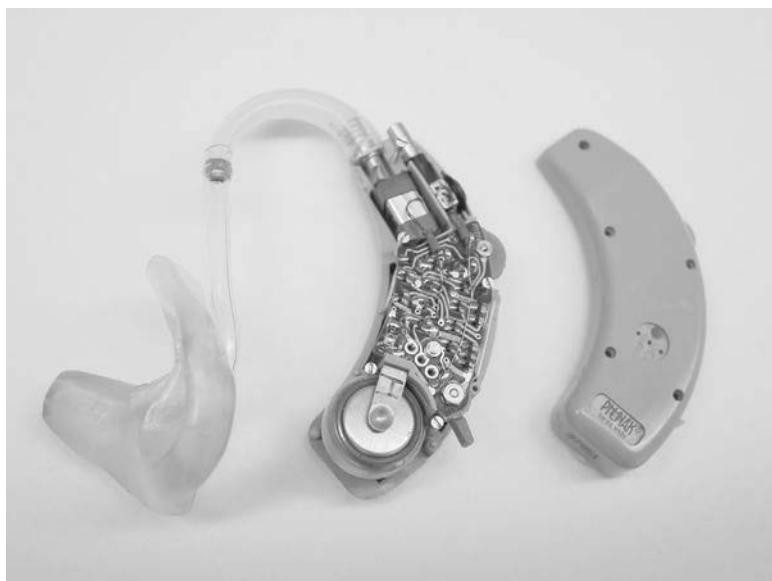
The development of smaller tubes during WWII led to the miniaturization of hearing aids and made them far more portable. Unfortunately, vacuum tubes require one high-current battery to heat the filament and another high-voltage battery to power the circuitry. These batteries together were larger than the amplifier and weighed a kilogram or more. An example of the Western Electric Model 134, manufactured in 1944, is shown in Figure 6-9.

By the late 1940s, miniaturization of vacuum tube and battery technology was able to produce hearing aids that fitted into a top pocket or even, for women, under a ponytail. In 1952, the hearing aid was the first commercial product to employ the transistor—2 years before the first transistor radio was manufactured. Hearing aids could now be hidden in spectacle frames, and within a few years the first behind-the-ear (BTE) models, such as the Zenith Diplomat, were appearing (Figure 6-10).

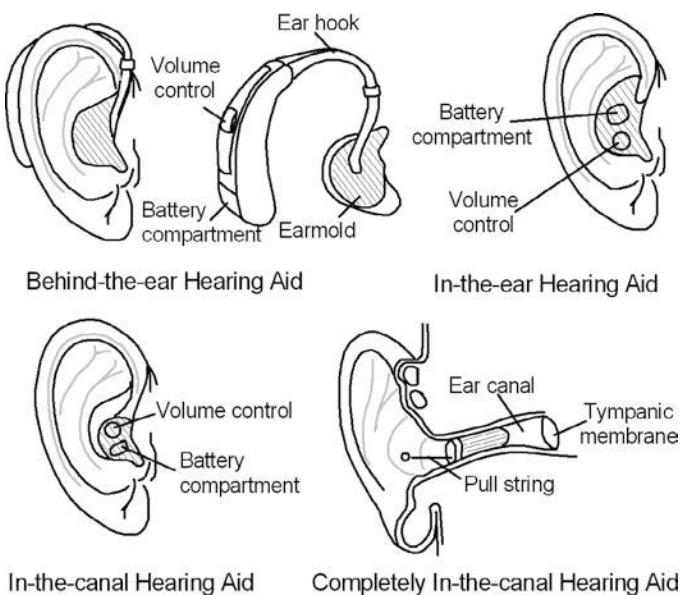
This unit had a four-transistor amplifier, an external receiver, and microphone openings on both sides of the flesh-colored plastic case. It weighed less than 30 g and was described as “tiny, feather-light, tinted and contoured to fit snugly right at the ear.”

In-the-ear (ITE) and in-the-canal (ITC) hearing aids were introduced in the late 1950s and early 1960s, and even smaller completely-in-the-canal (CIC) hearing aids started to appear in the 1980s as integrated circuit and battery technology became more sophisticated. Some of these options are shown in Figure 6-11.

By 1996, hearing devices housing all the hearing aid components completely within a custom-made ear mold represented most new sales (Figure 6-12). These devices offered new features such as directional microphones, digital programming, and adaptive filtering to provide users with an effective hearing aid that was also virtually unnoticeable.



**FIGURE 6-10 ■**  
Early BTE hearing aid. (Reproduced courtesy of Becker Medical Library, Washington University School of Medicine, with permission.)



**FIGURE 6-11 ■**  
Different hearing aid configurations [Adapted from (Fidelis 2009).]



**FIGURE 6-12 ■**  
Custom-molded and standard ITE hearing aids.

In 1987 two manufacturers introduced hearing aids with digital signal processing (DSP). While sophisticated for their time, these hearing aids had little success and were soon abandoned due to their large size and high battery drain.

By the beginning of the new millennium, the technology had improved, and digital hearing aids were produced in a range of popular styles, from BTE to CIC. Despite their higher cost, they were well received by patients. This early success, combined with the promise of highly advanced signal processing, ensured that digital hearing aid technology had come of age. How this technology works is discussed later in this chapter.

Typically, BTE devices are still required for severe hearing loss as they can provide the highest output powers with sound level outputs of up to 130 dB. Smaller ITC and CIC devices are suitable for mild to moderate hearing loss with output sound levels ranging between 100 and 115 dB. At these levels, feedback problems are minimized, and battery life is still adequate.

### 6.5.2 Hearing Aid Operation

All hearing aids have a microphone to convert sound into an electrical signal, a battery-powered amplifier to increase the volume, and a means of transmitting the sound to the person. Most hearing aids transmit the sound through a small speaker placed in the ear canal. If the eardrum is damaged, MEIHDs convey sounds directly to the bones of the middle ear (ossicles) or, in some cases, directly to the bones of the skull.

In addition to amplification, hearing aids must be capable of shaping the output frequency response to suit the hearing loss characteristics of the patient. Most people lose high-frequency response as they age, and therefore the amplifier emphasizes high-frequency signals to restore the overall response characteristic and improve speech recognition.

#### 6.5.2.1 Microphones

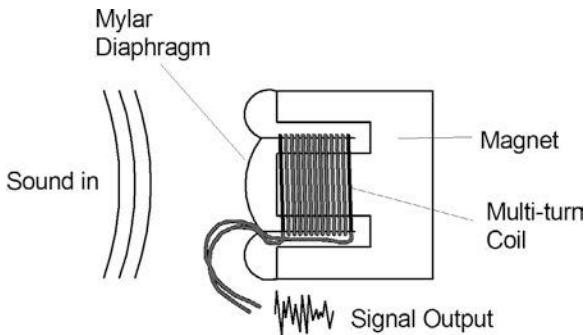
In the earliest vacuum-tube-based hearing aids, carbon microphones were used. These relied on the change in resistance with sound intensity of loosely packed carbon granules mounted behind a diaphragm. With the advent of the transistor amplifier, the smaller dynamic microphone became the sensor of choice. However, as the size of hearing aids has continued to drop, miniature electret microphones or microelectromechanical system (MEMS)-based capacitor microphones have replaced these to a large extent.

#### WORKED EXAMPLE

#### Dynamic Microphone

As shown in Figure 6-13, a dynamic microphone consists of a diaphragm supporting a coil held between the poles of a permanent magnet. Sound waves cause the diaphragm and coil to move, and the motion of the coil in the magnetic field induces an electromotor force (EMF) that can be measured as a voltage at its terminals.

It can be shown that the induced EMF,  $V_{out}$ (V), is a function of the magnetic field,  $\beta$  (Weber/m<sup>2</sup>), the total length of the conductor in the magnetic field,  $l$  (m), and its velocity



**FIGURE 6-13** ■  
Schematic of a dynamic microphone.

orthogonal to the direction of the field  $v$  (m/s):

$$V_{out} = \beta lv \quad (6.13)$$

It was stated earlier in this chapter that the peak acoustic pressure and the displacement of the air are related by  $p_m = k\rho_o c^2 y_m$ , where  $k = 2\pi/\lambda$  is the wave number,  $\rho_o$  (kg/m<sup>3</sup>) is the air density,  $c$  (m/s) is the speed of sound, and  $y_m$  (m) is the maximum displacement of the air particles.

Rewriting the equation to obtain the air displacement,  $y$  (m), as a function of time gives

$$y = \frac{p_m}{k\rho_o c^2} \sin(\omega t) \quad (6.14)$$

The velocity of the air,  $\dot{y}$  (m/s), can be obtained by differentiating (6.14) with respect to time

$$\dot{y} = \frac{\omega p_m}{k\rho_o c^2} \cos(\omega t) \quad (6.15)$$

This reaches a maximum,  $\dot{y}_p$  (m/s), when  $\cos(\omega t) = 1$ .

$$\dot{y}_p = \frac{\omega p_m}{k\rho_o c^2} \quad (6.16)$$

After simplification

$$\dot{y}_p = \frac{p_m}{\rho_o c} = \frac{p_m}{Z} \quad (6.17)$$

It can be seen that the peak air particle velocity is independent of frequency and dependent only on the peak acoustic pressure and the acoustic impedance,  $Z = \rho_o c$ , (acoustic ohms).

At normal conversational levels, the SPL is about 60 dB. This equates to an RMS sound pressure of

$$\begin{aligned} P &= P_o 10^{60/20} \\ &= 20 \times 10^{-6} \times 10^3 \\ &= 0.02 \text{ Pa} \end{aligned}$$

The peak acoustic pressure is  $p_m = \sqrt{2}P = 0.028 \text{ Pa}$ , so for  $Z = 400 \Omega$  the peak velocity of the air particles is

$$\dot{y}_p = 0.028/400 = 0.071 \text{ mm/s}$$

If the inertia of the diaphragm-coil combination can be ignored, the air particle velocity is transferred to the diaphragm and to the coil. This is true at low frequencies, but as the frequency increases beyond about 10 kHz inertial effects become significant and the sensitivity of the microphone is reduced.

In a typical dynamic microphone,  $N = 450$  turns of AWG 50 enameled copper (diameter,  $\phi = 0.025$  mm) wire are wound onto a former with a diameter  $d = 10$  mm. The total length of wire in the coil is therefore

$$\begin{aligned} l &= N\pi d \\ &= 450 \times \pi \times 10 \times 10^{-3} \\ &= 14 \text{ m} \end{aligned}$$

The total resistance of the coil is the product of the resistivity of copper,  $\rho_{cu} = 1.72 \times 10^{-8} \Omega\text{m}$ , and the length of the wire divided by its cross sectional area,

$$\begin{aligned} R &= \frac{\rho_{cu}l}{A} = \frac{4\rho_{cu}l}{\pi\phi^2} \\ &= \frac{4 \times 1.72 \times 10^{-8} \times 14}{\pi \times (0.025 \times 10^{-3})^2} \\ &= 490 \Omega \end{aligned}$$

which is close to the nominal 500  $\Omega$  impedance that is normal for dynamic microphones.

A permanent magnet with  $\beta = 0.2$  Weber/m<sup>2</sup> is typical for a dynamic microphone, so the peak voltage generated is

$$\begin{aligned} V_{out} &= \beta lv \\ &= 0.1 \times 14 \times 0.071 \times 10^{-3} \\ &= 100 \mu\text{V} \end{aligned}$$


---

Electret microphones rely on a diaphragm comprising a thin metallized Teflon film that has been manufactured to maintain its charge permanently. This charged diaphragm forms one plate of a capacitor. As the capacitance,  $C$  (Farad), is a function of the permittivity of free space,  $\epsilon = 8.854 \times 10^{-12}$  F/m, area of the plate,  $A$  (m<sup>2</sup>), and inversely proportional to the distance,  $d$  (m), between the two plates

$$C = \frac{\epsilon A}{d} \quad (6.18)$$

it will change slightly as the diaphragm vibrates in response to the incident sound waves.

However, because the charge,  $Q$  (Coulomb), stored by the electret is constant, changes in capacitance are reflected as changes in the voltage measured across the two plates

$$V = \frac{Q}{C} \quad (6.19)$$

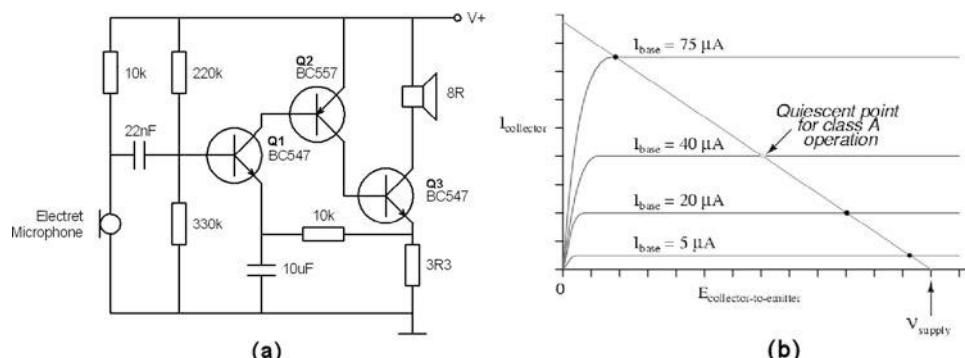
Electret microphones often incorporate a field-effect transistor (FET) into the package to provide some gain and decrease the output impedance to a suitable level.

### 6.5.2.2 Analog Amplifiers

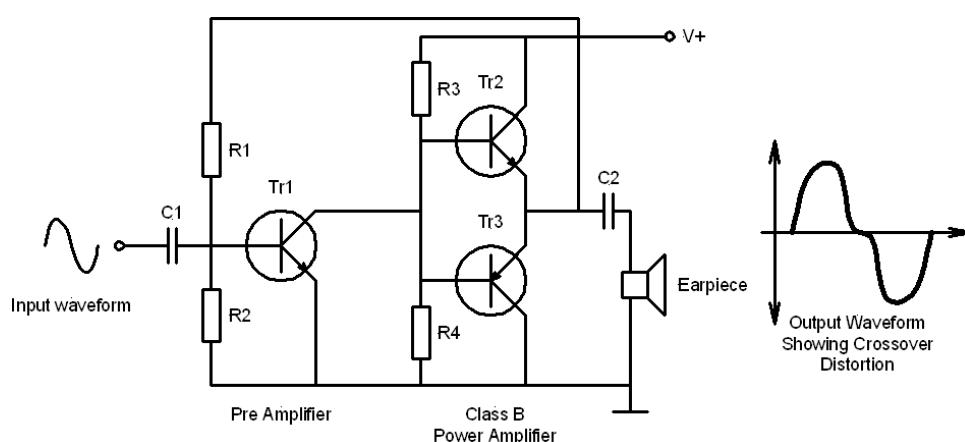
Depending on battery life, amplifier gain, required output power, and distortion, different designs for hearing aids are available. There are five major classifications, referred to as classes: A, B, D, sliding class-A (also known as class-H), and digital amplifiers.

Class-A amplifiers were the original amplifier types developed for hearing aids and are therefore the best understood and lowest-cost option. As can be seen from the simplified schematic shown in Figure 6-14, each of the transistors in the amplifier is biased into its linear region. This results in low harmonic distortion at low input powers, but high quiescent current and therefore a short battery life. Output powers are limited, with the result that harmonic distortion increases if the volume is set too high.

As shown in Figure 6-15, a class-B amplifier generally consists of a class-A preamplifier followed by a class-B driver stage. At the output of the preamplifier, the signal is split into two with the positive half-cycle driving one transistor and the negative half-cycle driving the other transistor. This allows these transistors to remain unbiased until they are needed. Class-B amplifiers are called push-pull circuits because of the way the transistors are driven alternately. However, because the transistor does not begin to turn on until the drive voltage has reached 0.7 V for the positive-going half-cycle and to  $-0.7$  V for the negative-going half-cycle, a dead band occurs. This causes crossover distortion.

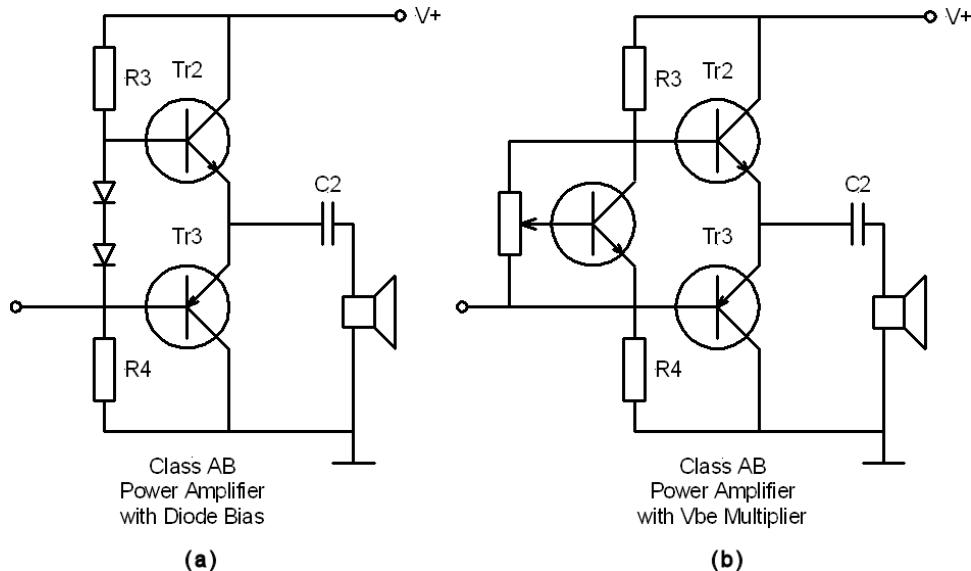


**FIGURE 6-14** ■  
Class-A amplifier. (a)  
Circuit diagram (b)  
Operating point for  
Transistor Q1.



**FIGURE 6-15** ■  
Class-B amplifier.  
(a) Schematic.  
(b) Crossover  
distortion.

**FIGURE 6-16** ■ Class-AB amplifier options. (a) Diode bias. (b) V<sub>be</sub> multiplier adjustable bias.



However, this distortion can be minimized by applying the correct bias and using sufficient negative feedback. The simplest bias consists of a potentiometer, but this cannot compensate for variations in the output transistors' characteristics with temperature; thus, in general two diodes in series and having the correct temperature coefficients are used, or alternatively a V<sub>be</sub> multiplier circuit can perform the same function. These options are referred to as class-AB and are shown in Figure 6-16.

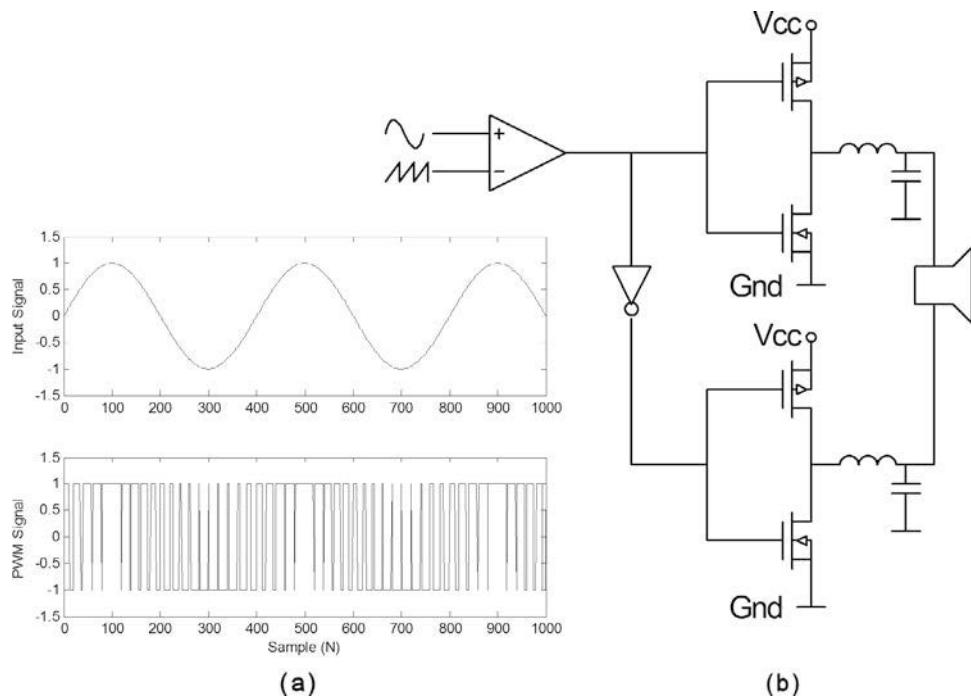
Because battery drain is proportional to output power, battery life is dependent on the output amplitude and the length of time that there is an input. For example, battery life will be much longer in a quiet home environment than at a noisy party.

The high output power capability makes this class of amplifier ideal for severe to profoundly hearing impaired patients.

Class-D amplifiers operate in a different manner to the previously discussed types (Dapkus, 2000). In one common configuration, shown in Figure 6-17, they start with an oscillator running at 220 kHz, divided by two to produce a square wave running at 110 kHz. This is then converted to a symmetrical triangular wave by an integrator with some feedback to ensure that the outputs do not drift. The signal from the hearing aid microphone passes through a class-A preamplifier so that the maximum amplitude is comparable to that of the triangular wave. The triangular wave and the audio signal are then applied to different inputs of a comparator to generate a pulse-width modulated (PWM) output with high-amplitude signals producing wide pulses and low-amplitude signals producing narrow ones. The output is therefore a digital signal whose average value reproduces the analog input. Because of the high sample rate, way beyond the response of any loudspeaker, only the average signal is reproduced with minimal additional filtering.

Because the switching transistors are either on or off, they dissipate very little power as heat and are therefore very efficient, with the result that battery life is extended. In addition, because they are switching the rail voltage into the loudspeaker, the peak output power can be higher than that of class-B amplifiers.

Notwithstanding this unusual method of amplification, power consumption is low and output distortion very low even at high output powers, making class-D hearing aids very popular with patients.



**FIGURE 6-17** ■ Class-D amplifier. (a) Input and output waveforms. (b) Circuit diagram. (Adapted from [Dapkus 2000].)

The sliding class-A or class-H amplifiers offer good frequency response, low distortion, and long battery life. This performance is achieved by adjusting the supply voltage to suit the output signal amplitude. Power drain then depends on the input signal level.

Because the hearing profile of each patient is different, hearing aids must be individually tailored. In the past, an audiologist would determine the gain and frequency response characteristics required, and then a laboratory would build the hearing aid to those specifications. Most modern analog devices are remotely programmable with a number of settings available to the patient depending on the listening environment. These hearing aids are sometimes referred to as digitally programmable but should not be confused with true digital systems.

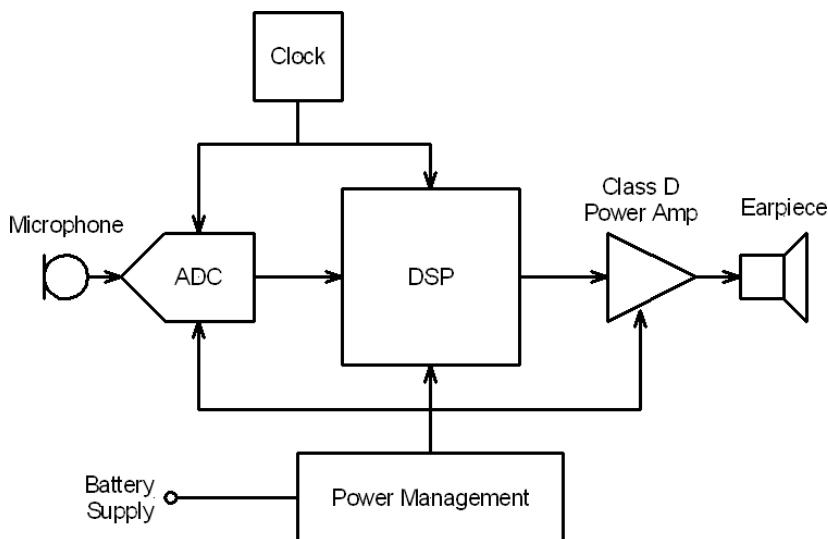
One of the main drawbacks with analog hearing aids is the limited degree by which the frequency response can be tailored to suit individual requirements. The response of each frequency band is determined using resistors and capacitors, and as the latter are large only a few bands are accommodated within the limited space in the device.

### 6.5.2.3 Digital Amplifiers

As shown in Figure 6-18, a digital hearing aid amplifier contains a DSP to perform all of the frequency response shaping and compression required by the hearing aid. The input signal is first amplified and then digitized at a sufficiently high rate to avoid temporal aliasing. This digital signal is clocked through the DSP where numerical algorithms operate on each digital word sequentially to implement the required characteristics. The output controls a class-D amplifier that drives the loudspeaker.

Digital hearing aids used to consist of three integrated circuit chips stacked one on top of another. These chips were nonvolatile electrically erasable programmable read-only memory (EEPROM), a DSP device, and an analog one. However, modern very-large-scale

**FIGURE 6-18 ■**  
Digital signal processor based hearing aid amplifier schematic.



integration (VLSI) processes have been able to combine all of the functions onto a single chip.

Battery voltage ranges from 1.35 V down to 0.9 V, so hearing aid electronics is designed to operate from 0.9 V. Battery management electronics is often included to provide a low-battery warning and graceful shutdown when the battery voltage drops too low.

In a typical hearing aid the microphone is followed by a preamplifier with some analog compression and an antialias filter followed by a sigma-delta analog-to-digital converter (ADC). The ADC has a cutoff frequency of 20 kHz with 16 bits of resolution (14-bit linear).

The digital portion of the hearing aid includes the DSP, logic support functions, a programming interface, and the output stage. The output is typically all digital using a PWM-driven class-D amplifier.

Digital systems have low distortion and excellent signal-to-noise ratios and, because they offer improved processing capability, can be better matched to the requirements of the individual.

One of the first innovations that became practical with the advent of digital hearing aids was an algorithm that identified noise in specific bands and used adaptive processing algorithms to reduce the noise in that band (Graupe and Causey, 1975). It was incorporated in several hearing aids and is known as the zeta noise blocker. Since then many other algorithms have been developed to enhance the capability of these hearing aids, as discussed.

**Gain processing.** One of the primary benefits associated with flexible gain-processing schemes is the potential for increased audibility of sounds of interest with minimal distortion and a reduction in the probability of the output amplitude being too high. While this is more generally a benefit of compression, the greatly increased flexibility and control of compression processing provided by DSP—such as input signal-specific band dependence, greater numbers of channels, and kneepoints with lower compression thresholds—can lead to improved audibility with less expert tuning. Expansion, the opposite of compression,

has also been introduced in digital hearing aids. This processing can lead to greater listener satisfaction by reducing the intensity of low-level environmental sounds and microphone noise that may have been annoying to the user.

**Digital feedback reduction (DFR).** The most advanced feedback reduction schemes monitor for feedback while the listener is wearing the hearing aid. Moderate feedback is then reduced or eliminated through the use of an out-of-phase cancellation system or notch filtering. DFR can substantially benefit users who experience occasional feedback, such as that associated with jaw movement and proximity to objects.

**Digital noise reduction (DNR).** This processing is intended to reduce gain, either in the low frequencies or in specific bands, when steady-state signals (noise) are detected. Although research findings supporting the efficacy of DNR systems are mixed, they do indicate that the DNR can work to reduce annoyance and possibly improve speech recognition in the presence of nonfluctuating noise. DNR is sometimes advocated as complementary processing to directional microphones. While directional microphones can reduce the levels of background noise regardless of its temporal content, they are limited to reducing noise from behind or to the sides of the user.

**Digital speech enhancement (DSE).** These systems act to increase the relative intensity of some segments of speech. Current DSE processing identifies and enhances speech based either on temporal or, more recently, spectral content. DSE in hearing aids is still relatively new, and whether it is particularly effective is unknown.

**Directional microphones and DSP.** The ability of directional hearing aids to improve the effective signal-to-noise ratio provided to the listener is now well established. In some cases, however, combining DSP with directional microphones can act to further enhance this benefit. In some hearing aids, DSP is used to calibrate microphones, to control the shape of the directional pattern, to automatically switch between directional and omnidirectional modes, and through expansion to reduce additional circuit noise generated by directional microphones.

**Frequency shifting.** A new innovation has been to shift some of the high-frequency components of the audio signal into a lower-frequency band where the patient is still able to hear. One of the Phonak hearing aids uses a nonlinear frequency compression technique to downshift the 3 kHz to 6 kHz band down to the band spanning 1.5 kHz to 2 kHz. Though this process takes some time to get used to, it offers significant advantages in regard to speech comprehension.

**Digital hearing aids as signal generators.** Since digital hearing aids have a DSP at their heart, they are able to generate, as well as to process, sound. Current digital hearing aids use this capability to perform loudness growth and threshold testing to obtain fitting information specific to an individual patient in combination with a specific hearing aid. Sound levels also can be verified through the hearing aid once it is fitted. This technology has the potential both to increase accuracy of hearing aid fittings and to streamline the fitting process by reducing the need for external equipment (Ricketts, 2008).

Other capabilities include improved frequency shaping with fine control of up to 16 overlapping channels, binaural processing, pinna and ear canal filtering, as well as

reverberation reduction. Digital hearing aids can also provide for direct digital input from a digital telephone, TV, or MP3 player.

#### 6.5.2.4 Signal Compression

The healthy ear responds over a dynamic range of 120 dB, way in excess of any electronic device. Signal compression is therefore a common feature of most modern hearing aids. It reduces the required dynamic range and can therefore unify conversational levels of speech and improve intelligibility by reducing peak clipping. In addition, it protects the ears of patients such as infants who cannot use the hearing aid volume control.

Compression is achieved not by limiting, which can be wasteful of power, but by feeding back an out-of-phase portion of the signal to reduce the output amplitude. The following terms are commonly used to describe some of these functions:

- Kneepoint: The level of the incoming signal above which compression occurs
- Compression ratio: The ratio by which the sound level is reduced
- Attack time: The time delay between the onset of a signal loud enough to trigger compression and the reduction of gain
- Release time: The amount of time necessary for the gain to return to the precompression level after the input signal amplitude is reduced to below the kneepoint

#### 6.5.2.5 Power Consumption

Power consumption in the current generation of analog and digital hearing aids is about the same, with analog devices drawing between 0.7 mA and 1.0 mA and digital units drawing between 0.5 mA and 0.7 mA. A single 1.35 V zinc-air battery with a 30 to 65 mAh capacity and a 50  $\mu$ A self-discharge rate can power these devices for a couple of days before needing replacement.

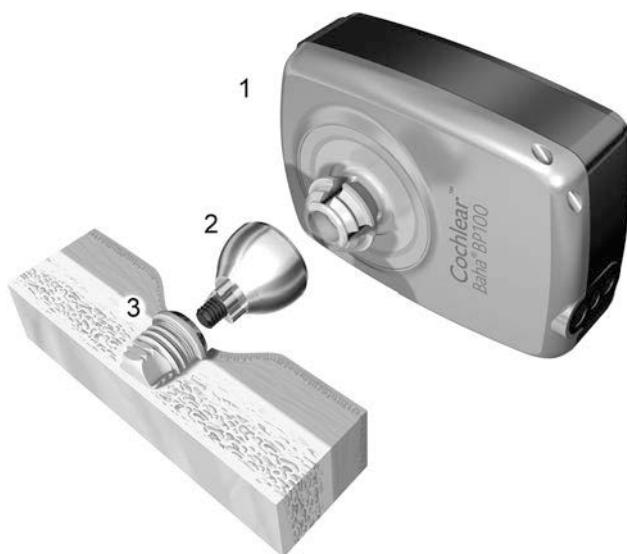
This is hardly a fair comparison of overall capability because digital hearing aids are far more sophisticated and offer more functions than analog devices do.

## 6.6 | BONE CONDUCTION DEVICES

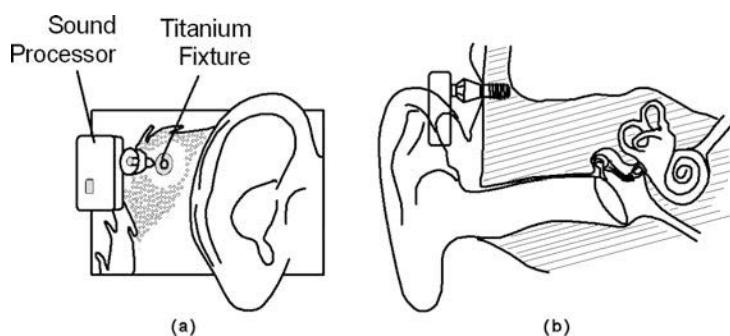
---

The bone anchored hearing apparatus (BAHA), shown in Figure 6-19, has been available in Europe since 1977 and is approved by the U.S. Food and Drug Administration. It is a 3 to 4 mm percutaneous titanium implant with an attached abutment that is surgically implanted in the skull behind the ear. An external digital sound processor containing directional microphones, an amplifier, and an electromagnetic actuator is snap coupled to the abutment.

The attachment is based on the physiological phenomenon of osseointegration (or osteofixation) where the titanium interface becomes attached to the living bone tissue through the cell matrix substance. This interesting and extremely useful phenomenon was discovered by Professor Per Ingvar Bränemark in the early 1950s but was not exploited for more than a decade, at which time it was used for dental prostheses. In 1977 it was suggested by Bränemark and his associates in the ear, nose, and throat (ENT) department at Sahlgrenska University Hospital that bone-anchored and skin-penetrating implants could probably improve the situation for patients with a significant conduction hearing loss, so the BAHA system was born (Håkansson, 2009). This process of osseointegration takes about 3 months to complete, after which the sound processor can be attached.



**FIGURE 6-19** ■ The Baha BP100 osseointegrated hearing device. (Courtesy of Cochlear Ltd.)



**FIGURE 6-20** ■ Installation of a BAHA osseointegrated hearing device. (a) External view. (b) Cross section through the ear showing attachment method.

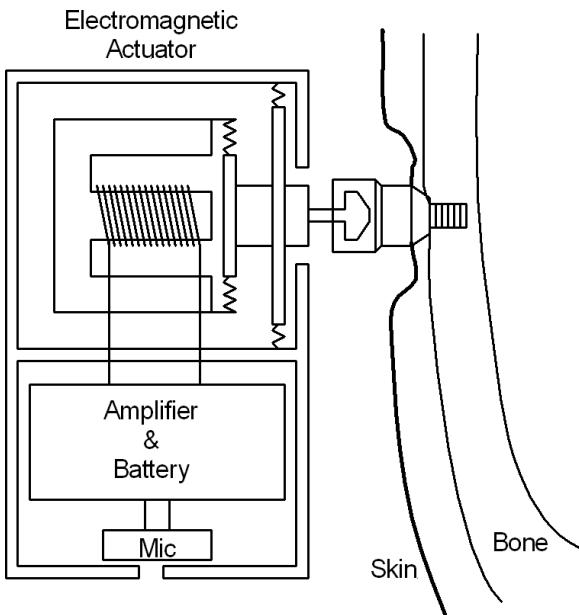
The principle of bone conduction amplification is that the external sound processing device converts auditory signals into mechanical vibratory signals and conveys them through the abutment and the BAHA implant into the mastoid bone, as shown in Figure 6-20. There is no airborne sound transmission, but the vibrations in the skull are conveyed directly to the inner ear where the hair cells are deflected by the normal physiological mechanism (Finn and LoPresti, 2003).

This process bypasses both the outer and middle ear and is therefore ideal in cases of conductive hearing loss. In addition, because of the efficiency of the device and the available gain it is also useful in cases where some sensorineural hearing loss exists.

In cases of single-sided deafness (SSD), patients find that localization of the sound source is impossible and discrimination is impaired. In these cases the BAHA device is worn on the deaf side, and the sound seamlessly transfers to the other ear via bone conduction, thus eliminating head-shadow problems (Flynn, 2007).

Figure 6-21 shows a simplified schematic diagram of the process. Sound picked up by the microphone is amplified and processed to restore the appropriate spectral response for the patient. This electrical signal then drives an electromagnetic actuator similar to a voice coil that introduces vibrations into the mastoid bone through the titanium implant. It is interesting to note that because the BAHA device is suspended from the titanium abutment

**FIGURE 6-21** ■  
BAHA operational principles.



the system relies on the inertia of the unit to facilitate the transfer of sound vibration to the skull. This sound transfer process is extremely efficient and distortion free.

The external processor is available in three sizes and is chosen depending on the bone thresholds of the ear to be aided. At the time of writing, a body-worn external processor allows for the implantation for pure-tone-average bone thresholds as low as 65 dB HL. The two small processors (shown in Figure 6-21) are for pure-tone-average bone thresholds better than 55 dB HL.

One of the main advantages of the BAHA device is that, unlike the other invasive techniques, its efficacy can be determined prior to surgery using a vibratory device held against the mastoid bone with a spring clamp.

At present, Cochlear Ltd. (up to 2005 Entific Medical Systems, a Swedish company) is the only manufacturer of BAHAs.

## 6.7 | MIDDLE EAR IMPLANTS

The history of middle ear implants started in 1935 when Wilska stuck iron filings onto the tympanic membrane of a volunteer. The generation of a fluctuating magnetic field from an earphone without a diaphragm caused the eardrum to vibrate, and this was perceived as sound in the normal fashion. By the late 1950s Rutschmann was successfully stimulating the ossicles by gluing a small magnet (10 mg) onto the umbo of the malleus (Traynor and Fredrickson, 2008).

It was not until the 1970s that active devices were actually placed in the middle ear, and this research has led to the development of what are now known as MEIHDS. Extensive trials are in progress worldwide to determine their efficacy in treating sensorineural and conductive hearing loss.

The rationale for developing these devices is multifaceted. They can improve fidelity by directly stimulating the ossicles and comfort by allowing the ear canal to remain open.

In addition, most implantable middle ear devices almost completely eliminate feedback, one of the most annoying adverse effects of conventional aids. Finally, some devices may allow patients to continue receiving amplification while swimming or bathing.

Implantable middle ear devices are generally available in two types: piezoelectric and electromagnetic. Electromagnetic devices are characterized by high efficiency and variable placement of the implant magnets, while piezoelectric MEIHDs are characteristically small and simple from an electronic perspective but are less efficient than their electromagnetic counterparts (Shohet, 2008).

When attached to the stapes, a driving force of between 0.16 and 16  $\mu\text{N}$  is required to produce a displacement of between 0.1 and 10 nm in the oval window. If the system is modeled as a spring–mass–damper as described in Chapter 4, the relationship between force,  $F$  (N), and the steady-state displacement,  $x$  (m), is

$$F = kx \quad (6.20)$$

For the numbers quoted, the spring constant  $k = 1.6 \text{ kN/m}$ .

### 6.7.1 Piezoelectric Devices

As discussed in Chapter 2, piezoelectric devices operate by applying an oscillating voltage to a piezoceramic crystal that changes its length and thereby produces a vibratory signal. The major disadvantage is that output power is directly proportional to the size of the crystal and there is very little space in the middle ear. Two configurations of actuators are available: the monomorph and the bimorph. The monomorph uses expansion and contraction to provide displacement directly, and the bimorph uses two sheets of piezoelectric material bonded together with opposite polarities to cause the structure to bend.

#### WORKED EXAMPLE

Consider a monomorph lead zirconate titanate (PZT) crystal 4 mm in diameter with a thickness  $t = 0.25 \text{ mm}$  driven by 0.5 V. What is the maximum force available, and what is the displacement?

First calculate the area,  $A$  ( $\text{m}^2$ ), or the piezoelectric element

$$A = \frac{\pi d^2}{4} = \frac{\pi \times (4 \times 10^{-3})^2}{4} = 12.6 \times 10^{-6} \text{ m}^2$$

Now calculate the force per unit volt from equation (3.50). Note that because the piezoelectric material operates as a capacitor, it takes on the same form, with the addition of a piezoelectric constant  $d_{33} = 110 \times 10^{-12} \text{ C/N}$ . In this equation  $\epsilon$  is the dielectric constant of the material (1700 for PZT) and  $\epsilon_0 = 8.8542 \times 10^{-12} \text{ C}^2/\text{Nm}^2$  is the permittivity of free space.

$$\begin{aligned} F_x &= \frac{\epsilon \epsilon_0 A}{d_{33} t} \\ &= \frac{8.85 \times 10^{-12} \times 1700 \times 12.6 \times 10^{-6}}{110 \times 10^{-12} \times 250 \times 10^{-6}} = 6.9 \text{ N/V} \end{aligned}$$

So, for an applied voltage of 0.5 V, the force is 3.4 N.

The strain can be determined from the force, the cross sectional area and Young's modulus,  $E$ ,

$$\begin{aligned}\frac{\Delta t}{t} &= \frac{F}{AE} \\ &= \frac{3.4}{12.6 \times 10^{-6} \times 83 \times 10^9} = 3.25 \times 10^{-6}\end{aligned}$$

For a slab thickness of 0.25 mm, the displacement is equal to 0.8 nm.

It can be seen that this is within the range of displacements required to drive the stapes (0.1 to 10 nm), albeit on the low side. To increase the displacement to 10 nm would require a stack of elements  $10 \times 0.25/0.8 = 3.125$  mm thick.

---

Studies of early designs indicate that such an approach benefits people with only up to moderate hearing loss (about 60 dB HL). These are the same individuals who benefit from the small CIC type of hearing aid. Piezoelectric transducers have the advantage of being inert in a magnetic field and therefore compatible with MRI. However, this advantage is negated if transcutaneous magnetic coupling of signals and power is used.

Direct-drive MEIHDS mostly have a maximum power output of between 90 and 110 dB SPL at 4 kHz and a maximum functional gain of 50 dB at 3 kHz. Functional gain is defined as the improvement over the hearing performance in the preexisting condition.

### 6.7.1.1 Rion Device, E-Type

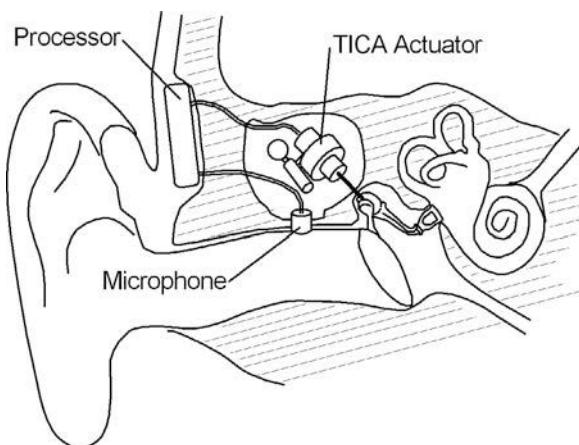
One of the earliest piezoelectric devices, the Rion Device E-type (RDE), originally developed by Yanagihara, has been used for both conductive and sensorineural losses. It is a partially implantable hearing aid (PIHA) composed of an external, ear-level microphone and amplifier, an internal electromagnetic coil to couple signals and power into the device, and a piezoelectric vibrator element. The vibrator element is anchored to the squamous portion of the temporal bone with a titanium screw. It is coupled to the stapes through a hydroxyapatite tube, which is interposed between the tip of the vibrator and the head of the stapes (Shohet, 2008).

In 39 patients in Japan, an initial functional gain of 36 dB at 3 months after surgery was measured. This eventually decreased to 21 dB in the long-term. The reason for diminished performance is thought to be due to a decrease in the sensitivity of the ossicular vibrator caused by aging and tissue reaction around the vibrator element.

### 6.7.1.2 Totally Integrated Cochlear Amplifier

The totally integrated cochlear amplifier (TICA), developed by Implex American Hearing Systems (now owned by Cochlear Corporation), is totally implantable. The microphone is implanted subcutaneously in the external ear adjacent to the tympanic membrane. A conventional digitally programmable processor located subcutaneously on the mastoid bone processes the signal. The piezoelectric transducer is coupled to the body of the incus using a titanium rod and so drives the ossicular chain, as can be seen in Figure 6-22.

A rechargeable lithium battery allows 60 hours of continuous operation and requires 90 minutes to recharge. It has a predicted lifetime of 5 years *in vivo* before it requires replacement.

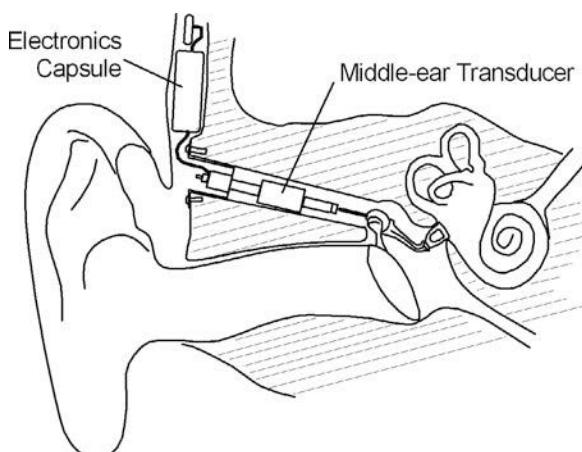


**FIGURE 6-22 ■**  
TICA. [Adapted from  
(Shohet 2008).]

### 6.7.1.3 Middle Ear Transducer

The middle ear transducer (MET; Otologics LLC, Boulder, CO) is based on research by John Fredrickson on Rhesus monkeys in 1995. It was first introduced as a semi-implantable device but has recently been made totally implantable. The original device consists of an implanted piezoelectric transducer coupled to a laser-drilled hole in the body of the incus. This aluminium oxide vibrating probe attaches to the incus by a fibrous union (scar tissue). The transducer translates the electrical signals into a mechanical motion that directly stimulates the ossicles and enables the wearer to perceive sound. The transducer is coupled with an externally worn audio “button” processor containing the microphone, battery, and signal processor.

The fully implantable MET device, shown in Figure 6-23, consists of four primary components: the implant, the programming system, the charger, and the remote control. The implant component consists of the electronics capsule and the MET. The electronics capsule contains the microphone, battery, magnet, digital signal processor, and connector. A sensitive microphone located under the skin picks up sounds, which are amplified and filtered according to the patient’s needs, converted into an electrical signal, and sent down the lead and into the transducer.



**FIGURE 6-23 ■**The  
MET fully  
implantable  
ossicular stimulator.  
[Adapted from  
(Shohet 2008).]

The programming system coil is placed over the implant site and held in place magnetically. The coil couples with the implant by means of a radio frequency signal used to program the device in the same manner as a traditional digital hearing aid. The programming system also allows for extensive testing and diagnostics of the stimulator.

The charger system consists of the base station, charging coil, and charger body. To charge the implant, the wearer removes the charger body from the base station and places the coil on the skin over the implant site. The charger body contains a clip that allows the charger to be attached to the belt of the wearer during charging. Typically, charging time will be about 1 hour if performed daily. While recharging the implant, the wearer can perform normal daily activities, turn the implant on and off, and adjust the volume.

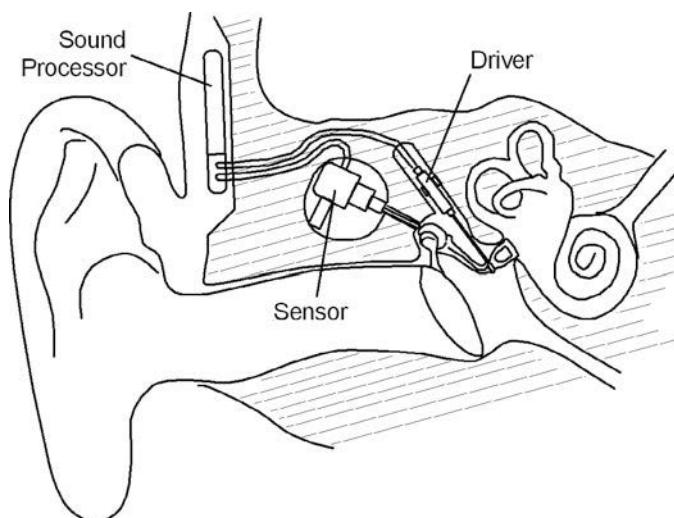
A remote is used to control the stimulator when the device is not being charged. It allows the wearer to turn the implant on and off and to adjust the volume. To use this facility, the wearer holds the remote against the skin over the implant.

The U.S. Phase I trial results yielded a 15–20 dB functional gain across audiometric frequencies in 20 patients. The pure-tone averages and monaural word recognition scores were better with the hearing aid in the same ear preoperatively, whereas the patients generally perceived more benefit in the postoperatively implant-aided conditions (Shohet, 2008; Traynor and Fredrickson, 2008).

#### 6.7.1.4 Envoy System

Another totally implantable piezoelectric device is the Esteem by Envoy Medical (originally St. Croix Medical), shown in Figure 6-24. This device uses the eardrum as the microphone, taking advantage of the natural acoustics of the ear canal without obstruction, interference, or any external devices. Therefore, the input signals are identical to those received by a person with normal hearing. This mechanical signal is converted to an electrical signal by a piezoelectric transducer (the sensor) at the head of the malleus or incus. The electrical signal is amplified and filtered by a programmable digital processor before being converted back to a vibratory signal using another piezoelectric transducer (the driver) attached to the head of the stapes. The incus lenticular process is removed to prevent feedback to the sensor, and this is one of the main disadvantages of the system.

**FIGURE 6-24** ■  
Envoy medical system. [Adapted from (Shohet 2008).]



The piezoelectric transducer can provide an output close to 110 dB SPL (Kroll, Grant et al., 2002; Shohet, 2008).

An audiologist programs the implant using a device called a *commander*. After the device is programmed, patients are given a personal programmer that allows them to turn the device on or off, to adjust the volume, and to remotely modify background noise filters.

The advantages of such a device are notable. Without any appliance in the external auditory canal, the occlusion effect is eliminated. Uncoupling of the sensor and driver eliminates most feedback.

The Envoy device faces some hurdles as development progresses. Its battery life is an issue with an estimated life of only 3–5 years depending on use. However, it can be replaced under local anesthetic. More importantly, removal of a portion of the incus permanently alters the ossicular mechanism and prohibits full recovery of hearing to preimplantation baseline levels if the device fails or is switched off. Modern ossicular reconstruction techniques are not perfect and can restore hearing to only within 10 dB. Finally, functional gain decreased at frequencies above 3000 Hz in the phase I study of the Envoy device. This appears to have been substantially improved in the phase II studies.

## 6.7.2 Electromagnetic Hearing Devices

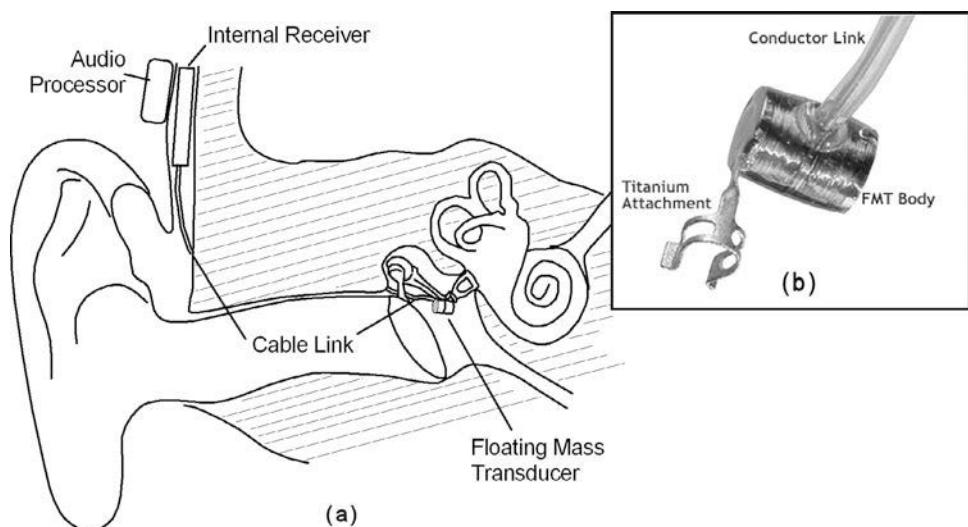
Electromagnetic hearing devices function by passing an electric current through a coil to generate a magnetic flux that attracts or repels an adjacent magnet, as discussed in Chapter 2. In a typical application the small, 50 mg magnet is attached to one of the vibratory structures of the middle ear (e.g., tympanic membrane, ossicles) and induces the structure to move in proportion to the strength of the applied magnetic field.. To date, all of the research programs using this method use devices that are only partially implantable and that still require an external hearing-aid shell to house the electromagnet. In some cases, the external coil is sufficiently small to be housed in a CIC type of hearing-aid shell, thus putting the electromagnet as close to the permanent magnet as possible. The major disadvantage of this setup is that as power transmission is determined by the alignment of the two elements so a slight shift of coil position in the outer ear can result in unpredictable or inadequate power coupling. In addition, the anatomy of the middle ear space restricts the size of the magnet and the coil so the performance of these devices is limited.

### 6.7.2.1 Vibrant Soundbridge Device

One example of an electromagnetic device is the Vibrant Soundbridge, which is shown in Figure 6-25. Originally developed by Symphonix Devices, Inc., the Soundbridge was the first implantable middle-ear hearing device approved by the U.S. Food and Drug Administration (FDA) to treat sensorineural hearing loss. It was marketed and implanted in the United States for a few years until the technology was purchased by Med-El of Austria. It is now marketed by Vibrant Med-El and available for implantation in Europe. Over 1400 such devices have been implanted worldwide (Traynor and Fredrickson, 2008).

The Soundbridge device is a semi-implantable device consisting of an external sound processor and amplifier, an audio processor, and an internal vibrating ossicular prosthesis (VORP). Sound passes into a microphone within the postauricular audio processor and is transmitted through the skin to an implanted receiver on the VORP using amplitude modulation. The VORP, which is implanted just behind the ear, conducts the sound to a magnet surrounded by a coil called the floating mass transducer (FMT). The transducer is

**FIGURE 6-25 ■**  
Vibrant Soundbridge device. (a) Cross section showing installation. (b) Photo of FMT. [Adapted from (Shohet 2008).]



attached to the long process of the incus, and the magnet is aligned with the long axis of the stapes, which causes it to vibrate along that axis.

One of the disadvantages of this device is that the confined space of the middle ear restricts the dimensions and crucial mass of the transducer, limiting the vibratory power output.

The phase III FDA trial was completed in 2000, and results from the 53 patients submitted to the FDA were published in 2002. The device was safe, with no notable change in preoperative and postoperative bone thresholds. Additionally, functional gain and word-recognition scores were improved with the Vibrant device compared with the patients' conventional hearing aids. Self-assessment indicated that 94% of the patients believed that the overall sound quality of the device was better than that of their conventional hearing aid.

As of 2007, the use of the Soundbridge has been expanded. It was successfully implanted on the round window membrane in patients with aural atresia and mixed hearing losses. It has also been applied to the incus in a more traditional configuration in patients with otosclerosis.

The audio processor has evolved from the original, analog Vibrant P unit to a digital, three-channel Vibrant D unit to the current digital, eight-channel Vibrant Signia unit. The Signia device modestly increases functional gain and speech-in-noise understanding results compared with the Vibrant D device (Shohet, 2008).

### WORKED EXAMPLE

A VORP that is 2 mm long and 1.5 mm in diameter contains an outer coil that uses the finest commercially available insulated copper wire, AWG-60 (American Wire Gauge), with a bare copper diameter of 0.00786 mm and a final diameter of 0.01mm with insulation. The cross sectional area of the copper is  $A = 4.85 \times 10^{-11} \text{ m}^2$ . A total of  $N = 200$  turns can be accommodated along the 2 mm length of the VORP, with a total wire length,  $l$  (m), of

$$l = N\pi d = 200 \times \pi \times 1.5 \times 10^{-3} = 0.942 \text{ m}$$

The resistivity of copper  $\rho = 1.68 \times 10^{-8}$  ohm.m, making the resistance of the coil

$$R = \frac{\rho l}{A} = \frac{1.68 \times 10^{-8} \times 0.942}{4.85 \times 10^{-11}} = 326 \Omega$$

An applied voltage of up to 0.5 V is applied to the coil to produce a maximum current of

$$I = \frac{V}{R} = \frac{0.5}{326} = 1.5 \text{ mA}$$

The Neodymium-30 rare-earth magnet, which constitutes the floating mass, has a diameter of 1.3 mm and a length of 1.6 mm. It is supported in the center of the coil by a pair of fine rubber springs. It has a strength of  $B = 1.1 \text{ Wb/m}^2$ .

Assuming that the magnetic field surrounding the coil is equal to  $B$ , the force on the magnet can be determined from the total length of the conductor and the current flowing through it

$$F = BIl = 1.1 \times 1.5 \times 10^{-3} \times 0.942 = 1.55 \text{ mN}$$

Unfortunately, this force cannot be directly coupled to the head of the incus as it is balanced by an opposing force on the coil. However, if the force results in an acceleration of the magnet relative to the coil, and hence to the incus, Newton's third law states that this action must be balanced by an equal and opposite reaction and the ossicles will accelerate in the opposite direction.

The density of Neodymium is 7.4 g/cm<sup>3</sup>, and the volume of the magnet  $V = 0.0021 \text{ cm}^3$ , making the mass of the magnet 15.7 mg.

The acceleration of the magnet is then

$$a_{mag} = \frac{F}{m} = \frac{1.55 \times 10^{-3}}{15.7 \times 10^{-3}} = 0.1 \text{ m/s}^2$$

The acceleration of the ossicles with a combined mass of 53 mg will be

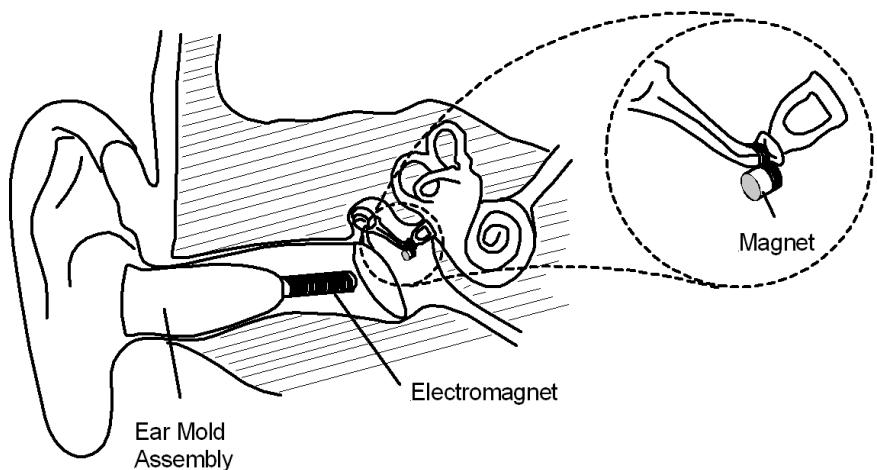
$$a_{os} = \frac{m_{mag}}{m_{os}} a_{mag} = \frac{15.7}{53} \times 0.1 = 0.03 \text{ m/s}^2$$

Obviously, the applied voltage will be amplitude modulated to reproduce the received acoustic signal so it will result in the magnet accelerating back and forth within the coil, and conveying this vibration to the ossicles where it is transferred into the cochlea and perceived as sound.

### 6.7.2.2 Soundtec Direct System

The Soundtec implant, shown schematically in Figure 6-26, was introduced to the U.S. market in 2001 after initial trials were successful (Hough, Dyer et al., 2001). It was voluntarily withdrawn in 2004. This semi-implantable device converts sound energy to electromagnetic energy to directly stimulate the ossicles. The only internal item is a surgically implanted neodymium-iron-boron (NdFeB) magnet in a titanium canister that is attached to the ossicular chain by positioning a collar around the neck of the stapes. An ear-mold coil assembly consisting of an acrylic skeleton mold with an embedded electromagnetic coil stimulates the magnet. The coil assembly is inserted deeply into the ear canal, ideally approximately 2 mm away from the tympanic membrane. It is attached to an analog or digital sound processor that is fitted either in the canal or behind the ear, similar to an ordinary hearing aid.

**FIGURE 6-26 ■**  
Soundtec direct system. [Adapted from (Shohet 2008).]



Such a design offers several advantages. Because it works by electromagnetic energy through the ear canal, the Soundtec device does not require an acoustic seal, which often leads to the occlusion effect or alters the resonance qualities of the ear canal. Also, functional gain can be improved without necessarily precipitating feedback, a common problem with traditional aids that occurs when sound pressure escapes the ear canal and cycles back through the microphone.

A relative disadvantage is that the procedure requires separation and then reconstitution of the incudostapedial joint, a process that may be responsible for as much as a 4 dB increase in air-conduction thresholds after surgery. Studies have also demonstrated a loss of average bone conduction of pure tones after implantation. This effect may be a result of movement of the mobile stapes into the vestibule during disarticulation of the incudostapedial joint, which may cause some sensorineural hearing loss.

Early reports indicated that, though the Soundtec did not provide a significant difference from optimal traditional amplification, it provided statistically significant high-frequency functional gain. Additionally, reports from patients indicated a cleaner, more natural sound than that achieved with traditional amplification.

The phase II FDA clinical trial included 103 patients with moderate to moderately severe sensorineural hearing loss who had previously worn hearing aids for at least 45 days. An average 7.9 dB increase in functional gain in the speech frequencies and a 9.6 dB increase in the high frequencies were reported. In addition, an increase of 5.3% in speech discrimination and a number of subjective improvements were also reported. Other surveys have failed to find a significant difference between the perceived performance of the Soundtec and conventional hearing aids.

A long-term follow-up retrospective review of 64 patients who received the Soundtec device revealed a significant average functional gain of 26 dB. About 55% of patients complained of hearing the magnet move when the processor was not being worn. This effect was diminished but not completely eliminated by further stabilizing the implant by placing adipose tissue between the implant collar and the neck of the stapes.

To date, around 600 Soundtec devices have been implanted, mostly in the United States. The device was voluntarily withdrawn from the market in 2004 when the company identified ways to improve it and to eliminate the distortion some patients experienced. The distortion occurred in as many as 7% of patients and was most irritating when the external

processor was not in use. It is thought to occur from movement of the magnet around its single point of fixation with the ossicular chain. The new magnet has an additional point of fixation onto the ossicular chain to further stabilize it. The company is retrofitting a digital processor and making variable-sized interchangeable coils to allow for tailoring of the coil length to the external auditory canal.

Additional studies of the Soundtec in a magnetic field indicate that it should be mechanically stable and nondestructive during 0.3 T open MRI with a modified MRI protocol (Shohet, 2008).

### 6.7.2.3 Other Semi-Implantable Devices

Another electromagnetic device is the semi-implantable middle ear electromagnetic hearing device (SIMEHD). This device consists of an external unit, which is similar in appearance to a behind-the-ear hearing aid, and an internal device, which is anchored to the mastoid. The internal device contains a driving coil and a magnet cemented to the incus.

In ongoing investigations, a magnet attached to the tympanic membrane is being used as a sound detector. As the magnet vibrates, it induces a small voltage in the drive coil, which is amplified and can be used as an input to other hearing aid devices such as cochlear implants.

Other researchers are investigating the attachment of a small magnet on the membrane covering the round window. This membrane moves in the opposite direction to the oval window membrane under the footplate of the stapes, so it will induce the same pressure waves into the cochlea.

A problem with these devices is that of permanently attaching a magnet to the constantly growing epithelium of the tympanic and round window membranes.

An experimental device known as an ear lens consists of a disk-shaped magnet held to the outside of the tympanic membrane using surface tension of a drop of oil and a silicone rubber plug (the lens). An electromagnet is placed externally at ear level or around the neck as a collar. This was not a success because of the large amount of power required to drive the electromagnet so far from the magnet.

### 6.7.3 Issues with Implantable Middle Ear Devices

A few major biomechanical issues must be addressed in developing implantable middle ear devices. Ideally, the device should not affect normal function of the middle ear in any way. It should not alter air-conduction thresholds, and if the implant is unsuccessful the added mass of the unit attached to the vibratory structure of the middle ear should not affect that structure's ability to vibrate.

Another important issue is the anchoring of the device to the ossicular chain. Even a small amount of slack at the interface between the prosthesis and bone tends to reduce the transmitted acoustic signal enough to render the device ineffective. Long-term stability of the fixation must be considered as well, particularly in respect of the mechanical forces that act at the interface, as these can affect the life expectancy of the device.

Finally, the direction of the transducer's action must be coincident with the axis of normal sound transmission through the ossicular chain. This requires that the device be attached to the tympanic membrane, the long process of the incus, or the head of the stapes.

Mild to moderate hearing loss is usually adequately improved with conventional audio type hearing aids, but the amplifier gain requirements for severe cases are limited by

feedback. In contrast, given adequate output production, an implantable aid can accommodate a severe hearing loss and still avoid feedback.

Some challenges remain in developing the ideal implantable hearing device. Limitations in battery capacity and charging times necessitate good transducer energy efficiency. Restrictions due to the small size of the middle ear constrain the available gain and limit performance where severe hearing loss is concerned. Costs associated with the development of these devices, as well as surgical implantation, make these devices more expensive than conventional hearing aids, and this may limit their widespread acceptance.

Some devices are currently being studied in FDA trials, but only the Vibrant Soundbridge and Soundtec direct systems have been approved so far. As these devices become more mature they will come to represent a new era in hearing augmentation akin to the largely successful cochlear implants. Although several hurdles still remain, the potential advantages and increasing number of people who may benefit from such devices continue to support their development (Shohet, 2008).

## **6.8 | DIRECT ACOUSTIC COCHLEAR STIMULATORY DEVICES**

---

As is shown in Figure 6-7 conventional hearing aids and MEIHDS can be used to treat moderate to severe sensorineural hearing loss, and BAHA devices cover conductive hearing loss. However, mixed hearing loss is best treated using Direct Acoustic Cochlear Stimulation (DACS) because the gain and power requirements for conventional devices are both difficult to achieve and can result in instabilities and feedback problems. By bypassing the region of conductive hearing loss, required gain and output power of the DACS device is reduced.

The principles involved are similar to those used by MEIHDS, but instead of actuating the intact ossicular chain a stapedectomy is performed, and the stapes is replaced by a prosthesis that is driven by a fixed actuator.

### **6.8.1 Actuator Design**

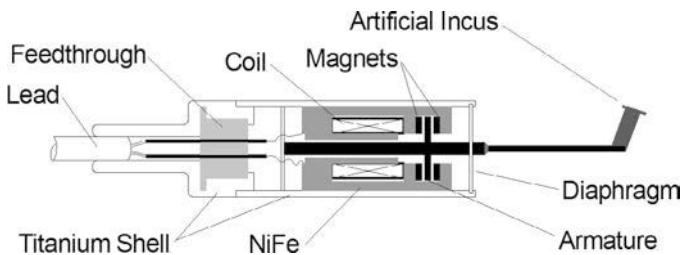
According to Bernhard, Steiger et al. (2011), the actuator should be capable of generating a signal equivalent to a 125 dB SPL over the whole specified frequency range from 100 Hz to 10 kHz while using a maximum power of only 1 mW. In addition, to fit best within the middle ear it should be cylindrical with a diameter of less than 3.6 mm and a maximum length of 14 mm.

To make the interface as efficient as possible, the actuation geometry mimics the operation of the incus, and the frequency response is designed to be as similar as possible to that of the middle ear. This eliminates the requirement for corrective preemphasis, which wastes energy.

The middle ear transfer function, from the tympanic membrane through to the oval window, corresponds to a second-order low-pass filter with a cutoff frequency of about 1 kHz. From a mechanical perspective this corresponds to the oscillation characteristics of a suspended mass with a resonant frequency,  $f_{res}$ , of 1 kHz as described by

$$f_{res} = \frac{1}{2\pi} \sqrt{\frac{k}{m}} \quad (6.21)$$

where  $k$  (N/m) is the spring constant, and  $m$  (kg) is the suspended mass.



**FIGURE 6-27** ■ Schematic diagram of the electromagnetic actuator showing the drive coils and balanced armature mechanism. [Adapted from (Bernhard, Steiger et al., 2011).]

To minimize the force required to drive this system, and hence to keep energy expenditure to a minimum, the mass is made as small as possible. In a human adult, the mass of the three ossicles together is about 53 mg, which requires a spring constant of about 2.1 kN/m to obtain the correct resonant frequency.

According to Bernhard, Steiger et al., (2011), it proved to be impractical to manufacture a diaphragm with a sufficiently low spring constant and still remain robust enough to survive in the body. To compensate, a disk connected to the actuator arm is placed in the field between two strong magnets as shown in Figure 6-27. When the armature is equidistant from the magnets it is in an unstable equilibrium. However, as soon as it moves it becomes more attracted to the nearer magnet, and this generates a negative spring constant to reduce the diaphragm spring constant of 6.9 kN/m to the correct value. This construction is known as a balanced armature mechanism.

A set of coils mounted within a NiFe magnetic circuit convert the electrical input into the equivalent motion of the actuator rod, which is attached to the artificial incus. This, in turn, conveys these vibrations into the prosthetic stapes and from there through the oval window into the cochlea.

Depending on the study (Chien, Rosowski et al., 2009), the normalized RMS velocity of the stapes at 1 kHz can lie between 30 and 100  $\mu\text{m/s/Pa}$ . Taking the lower, more recent value as more accurate, this equates to a velocity of 1.1 mm/s for an SPL of 125 dB (35.5 Pa).

One important consideration in regard to the operation of the actuator is the size of the prosthetic stapes head. In this case the head diameter was only 0.6 mm, which equates to a cross sectional area one tenth that of a natural stapes. This requires that the RMS actuator velocity should be 10 times higher than that of the natural system. The required RMS velocity for the prosthesis will therefore be about 11 mm/s at an SPL of 125 dB.

### WORKED EXAMPLE

For the case of a sinusoidal excitation at 1 kHz, with an RMS velocity of 11 mm/s, determine the peak displacement.

The displacement,  $x$  (mm), can be written as

$$x = a \cos 2\pi ft$$

The velocity is obtained by differentiation

$$\dot{x} = 2\pi fa \sin 2\pi ft$$

The RMS velocity is therefore  $\dot{x}_{rms} = \frac{2\pi fa}{\sqrt{2}}$ , which makes the peak displacement

$$\begin{aligned} a &= \frac{\sqrt{2}\dot{x}_{rms}}{2\pi f} \\ &= \frac{\sqrt{2} \times 11}{2\pi \times 1000} = 2.5 \mu\text{m} \end{aligned}$$

This would appear to be much larger than the typical displacements (0.1 to 10 nm) calculated earlier in this chapter. This is because a lower SPL and a conventional stapes were used. For an SPL of 60 dB and a stapes footplate area of  $3.2 \text{ mm}^2$ , the stapes velocity would be  $1.1/1780 = 0.6 \mu\text{m/s}$ , making  $a = 140 \text{ nm}$ .

---

## 6.9 | COCHLEAR IMPLANTS

### 6.9.1 Historical Background

Interest in electrical stimulation of hearing began with experiments conducted by Count Volta in the eighteenth century. In 1800 he described how he completed a circuit through metal rods stuck into his ears and received a “jolt to the head” followed by “a kind of crackling, jerking or bubbling as if some dough or thick stuff was boiling.” Not surprisingly, he found the experience rather uncomfortable and did not repeat the experiment very often.

Fifty years later, Duchenne of Boulogne performed a similar experiment using alternating current and heard what he described as the sound of an insect trapped between a glass pane and a curtain.

The first direct stimulation of the auditory nerve in a human being was performed during an operation by Lundberg in 1950, and the patient became aware of noise. Some years later in 1957 Djurono and Eyries implanted an electrode connected to an induction coil in the head of a deaf person. One end of the electrode was placed on the stump of the auditory nerve or adjacent brain stem and the other within the temporalis muscle (Wilson and Dorman, 2008). They were able to induce a voltage in the electrode from an external antenna, and the patient heard sounds like the chirping of a grasshopper or cricket. He was also able to recognize simple sounds like *mama*, *papa*, and *allo*. The patient used the device for a number of months before it failed, and he was able to sense the presence of environmental sounds but could not discriminate between them or understand speech.

In 1961, American surgeons John Doyle and William House and electronic engineer James Doyle inserted a single gold electrode a short distance into the cochlea of their first deaf patient. They also undertook an important trial with multiple electrodes but concentrated on the single-electrode option as an effective aid to lip reading, and by 1975 had undertaken many more implants. In 1964, Dr. F. Blair Simmons and his team implanted six electrodes into the cochlea. They showed that electrical stimulation of the auditory nerve using a bipolar stimulating electrode could produce a sensation of sound and some discrimination of pitch for pulse repetition frequencies below 1000 Hz (Møller, 2006).

Australian Graeme Clark was inspired by Simmons’s work with multiple electrodes, and in 1967 he started a Ph.D. at Sydney University and reviewed the available work to investigate whether a single or multiple-channel cochlear implant would be possible for

the management of a profound hearing loss (Clark, 1969). For the following 10 years he proceeded with animal experiments. It was only in 1978 that he performed his first implant on a human being, Rod Saunders, who had been deafened by a blow to the back of his head in a car accident 18 months before. Saunders continued to help Clark's team for more than 20 years, and when he died in 2007 his last wish was that his ear bones and brain be used for further research (Carman, 2008).

Meanwhile, in 1977 the U.S. National Institutes of Health (NIH) commissioned a study to determine whether further development of cochlear implants would be wise. One of the conclusions of this Bilger Report (Bilger, Black et al., 1977) was that although the subjects could not understand speech through their prostheses, they did score significantly higher on tests of lip reading and recognition of environmental sounds. In a consensus statement 11 years later, in 1988, the NIH suggested that multichannel implants were more likely to be effective than single-channel implants (Wilson and Dorman, 2008). This is because the cochlea exhibits a tonotopic organization in which the outer sections are sensitive to low-frequency sounds with the frequency sensitivity decreasing monotonically toward the center. At this time only about 3000 patients had received cochlear implants.

All modern cochlear implants separate the sound spectrum using band-pass filters and use these individual outputs to stimulate different regions of the cochlea. However, new and highly effective processing strategies were developed in the late 1980s and early 1990s principally through the national prosthesis program in the United States. These include continuous interleaved sampling (CIS), *m-of-n*, and spectral peak (SPEAK) strategies. Large gains in speech reception performance were achieved, and CIS and *m-of-n* remain in widespread use today (Wilson and Dorman, 2008).

In 1995, by which time about 10,000 patients had received implants, a second NIH consensus development conference was held. Their primary conclusion was that a majority of those individuals with the latest speech processors for their implants scored above 80% correct on high-context sentences, even without visual cues.

By the middle of 1996, the cumulative number of implants exceeded 110,000, as can be seen in Figure 6-28. This is orders of magnitude higher than the numbers for all other neural prostheses, including those for restoration of motor and other sensory functions.

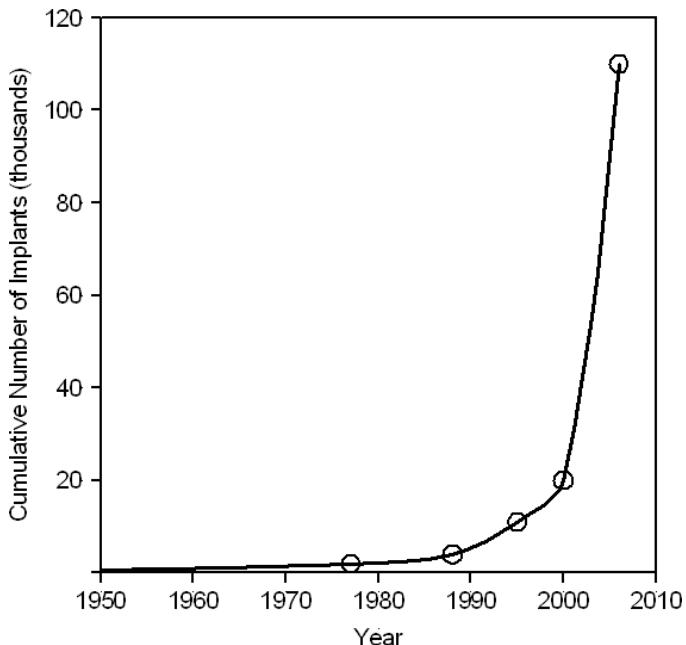
Contemporary manufacturers include AllHear Inc. (Aurora, Oregon), Clarion® (Advanced Bionics, Inc., Sylmar, California), Nucleus® (Cochlear Corporation, Sydney, Australia), Digisonic® (MXM Co., Vallauris, France), Interaid (Symbion, Inc., Provo, Utah), Laura Flex (Antwerp Bionic Systems, Belgium), and COMBI™-40+ (MED-EL Corp., Innsbruck, Austria). In addition, a number of research institutes and universities are also pursuing improved electrode arrays and better signal-processing algorithms.

## 6.9.2 How Cochlear Implants Work

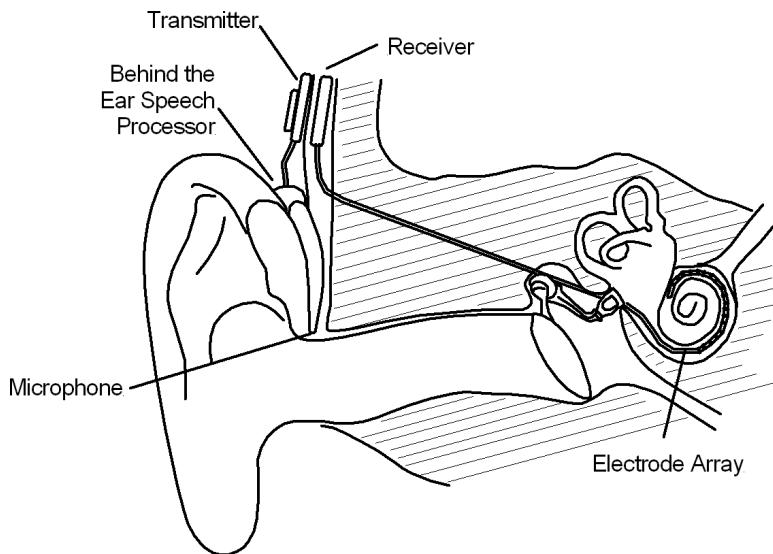
As shown in Figure 6-29, a modern cochlear implant consists of a microphone, external sound processor and power supply, transmitting circuitry, the receiver-stimulator package, and an electrode array.

The microphone picks up sounds, and the signal processor filters and selects the information and converts it into electrical signals that are transmitted to the intracochlear electrode array. Both the encoded signal and the power are transmitted transcutaneously, using radio frequency antennas, to a demodulator that assigns speech information to the electrode array. Single channel systems use only one electrode, while multichannel systems employ up to 31 channels.

**FIGURE 6-28 ■**  
Cumulative number  
of cochlear implants  
over time.

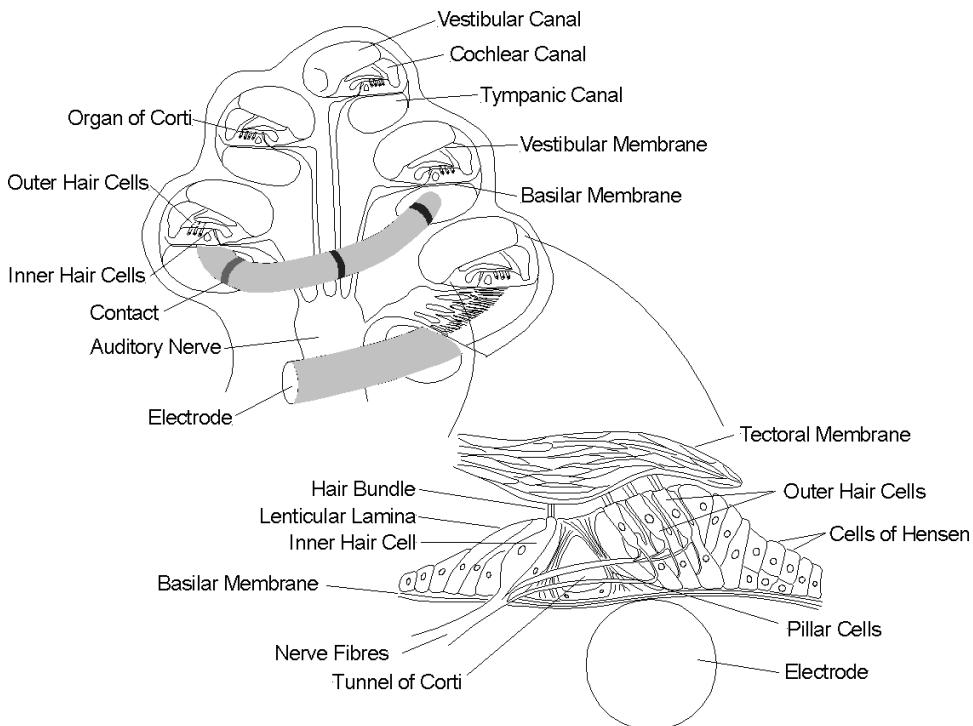


**FIGURE 6-29 ■**  
Cochlear  
implant—overview.

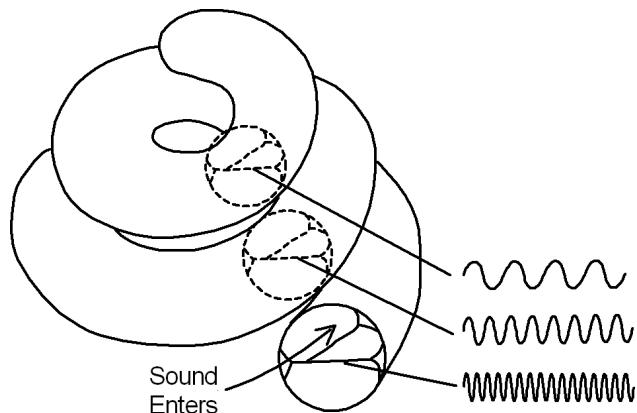


Electrodes are introduced into the tympanic canal, one of the three fluid-filled chambers along the length of the cochlea. A cutaway drawing shown in Figure 6-30 shows the partial insertion of an array of electrodes into the cochlea. The array is generally inserted through a drilled opening made by the surgeon into the bony shell of the cochlea overlying the tympanic canal and close to the base of the cochlea. Alternatively, the electrode array can be introduced through the round window. The depth of insertion is limited by the decreased diameter of the tympanic canal, the curvature of the cochlear spiral, and irregularities in the structure that snag the electrode tip.

The total length of the human cochlea is about 35 mm, and no array has ever been inserted more than 30 mm, with typical insertions much less than that (18 to 26 mm).



**FIGURE 6-30** ■ 3-D view of a cochlear implant—details of electrode insertion. [Adapted from (Loeb, 1985).]



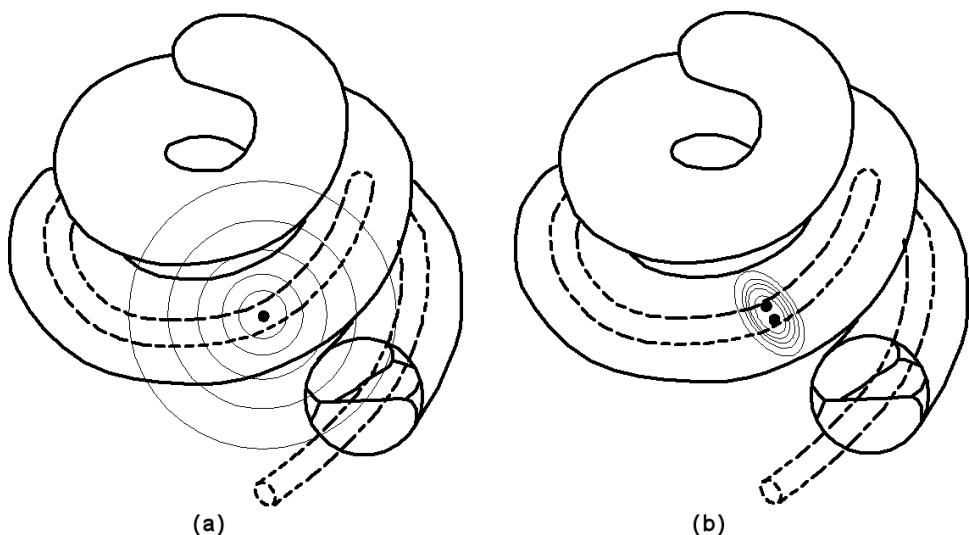
**FIGURE 6-31** ■ Tonotopic response of the cochlea. [Adapted from (Loeb, 1985).]

Because neurons at different positions along the cochlea respond to different frequencies of acoustic stimulation during normal hearing, as shown graphically in Figure 6-31, nerve stimulation can be affected by using different electrodes along the length of the array, separated to excite different frequencies. Electrodes toward the tip of the array will stimulate lower-frequency responses higher frequencies at the base.

Most cochlear implants to date use arrays of single electrodes, each of which is referenced to a remote electrode outside the cochlea, but as can be seen in Figure 6-32 the electric field induced using this method operates over a large volume. In contrast, the field produced by a bipolar electrode is far more constrained. In theory the latter should be capable of stimulating a smaller number of nerve cells that are sensitive to a narrower band of frequencies.

**FIGURE 6-32 ■**

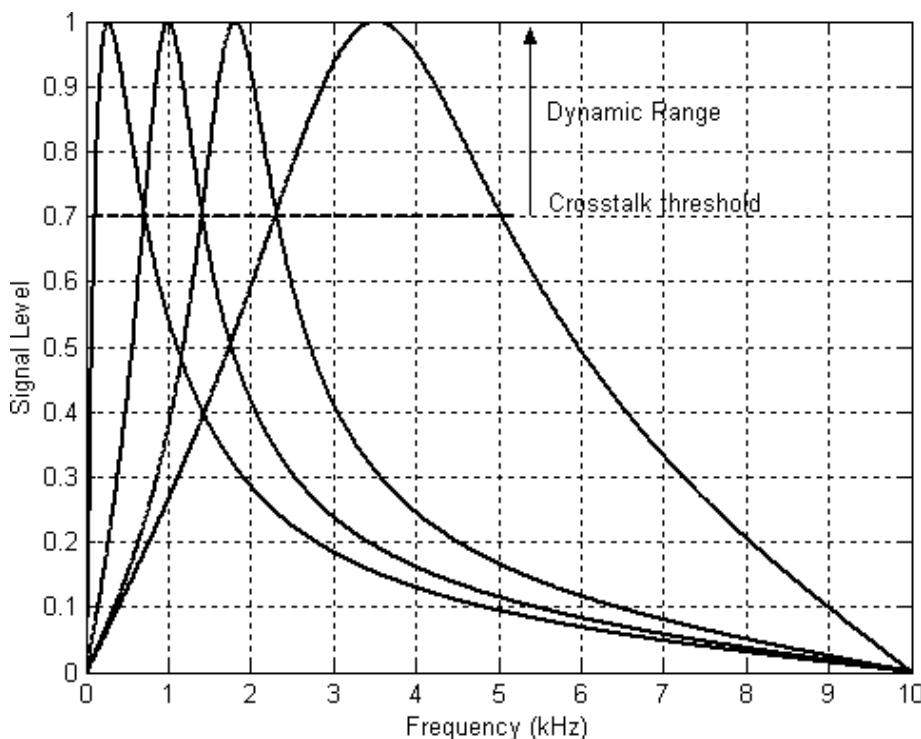
Electric field around a cochlear electrode  
 (a) Monopolar configuration.  
 (b) Bipolar configuration.



An important goal of electrode design is to maximize the number of largely nonoverlapping populations of neurons that can be stimulated by the electrode array. However, evidence suggests that even monopolar arrays with more than 20 electrodes stimulate only between four and eight independent sites.

These overlaps are unavoidable for electrode arrays in the tympanic canal as the electrodes are sitting in the highly conductive perilymph and are relatively far from their target neurons. If the array could be placed closer to the inner wall of the tympanic canal, it would be a little closer to the target neural tissue of the spiral ganglion. This can be achieved to some extent if the electrode array is manufactured with a built-in spiral. However, even if positioned perfectly the shortest distance between an electrode and a neuron is about 1 mm. Assuming the inverse square law relationship between the distance from an electrode and the current density holds, a stimulus of four times the threshold for neurons in one position would start to trigger neurons 2 mm away from the target neuron in both directions. For speech perception the critical frequency range is from 500 Hz to 3 kHz, and this represents a distance of less than 14 mm along the basilar membrane. It is therefore obvious that a maximum of between three and four independent stimulation sites can be accommodated using monopolar electrodes if a reasonable dynamic range is to be achieved. This trade-off between the number of independent sites and the dynamic range is illustrated in Figure 6-33.

With bipolar stimulation the current density gradient is very steep, and it is mostly constrained between the two electrodes, as shown in Figure 6-32, allowing the number of independent sites to be increased significantly. Because each neuron must be stimulated by inducing a current along its length, the electrodes must be oriented at right angles to the long axis of the tympanic canal. The actual number of independent sites that can be stimulated depends on the condition of the remaining auditory nerve fibers as these sometimes die back from the organ of Corti. However, it has been found by experiment that radial bipolar pairs can be positioned about 2 mm apart without significant interaction. This spacing makes it possible to achieve eight independent channels to span the speech frequency sites lying between 10 and 24 mm along the cochlea (Loeb, 1985).



**FIGURE 6-33** ■ Relationship between independent stimulation sites and the signal dynamic range of a monopolar cochlear implant.

If further improvement in the number of independent sites is required to improve speech fidelity, it may require a fundamentally new type of electrode or a different placement method (Wilson and Dorman, 2008).

A cochlear implant relies on the fact that many of the auditory nerve fibers remain intact in patients with sensorineural deafness. These neurons can be made to fire using electrical stimulation of the appropriate strength, duration, and orientation. These induced responses arrive at the brain looking just like the impulses generated by acoustic signals triggered by hair cells.

The perceived intensity of the sound is determined by the number of neurons activated and their firing rate, and these are both dependent on the amplitude of the stimulus current. The pitch is related to the place on the basilar membrane from which those nerve fibers originally derived their acoustic input, in agreement with the place–pitch theory. In principle one should be able to recreate the normal neural response to stimuli if enough independent spectral channels could be excited, and this would result in the patient “hearing” the sounds (Loeb, 1985).

New research conducted by Claus-Peter Richter at Northwestern University in Chicago confirmed, using deaf guinea pigs, that illumination of the neurons of the inner ear using infrared light stimulated activity in the interior colliculus (a neural relay point between the inner ear and the auditory cortex in the brain). Spatial frequency maps of this region were as sharp as those produced by real sounds, in contrast to the blurred maps produced using electrical stimulation (Nowak, 2008).

At present, the mechanism that makes neurons sensitive to infrared light is unknown, though it is speculated that it might be due to a rise in temperature. However, that notwithstanding, if this technique proves to be safe then it opens the way to cochlear implants with hundreds or even thousands of individual points of excitation.

### 6.9.3 Installation of the Electrode

There are a number of technical problems to physically introducing the electrode array into the tympanic canal. The first 24 mm of the basilar membrane are coiled into 1.5 of the 2.5 turns of the cochlear spiral, and because surgical access is through the long straight external ear canal the array must be straight on insertion. Unfortunately, the half-circular cross section of the tympanic canal causes a flexible object pressed against the side wall to be deflected upward into the extremely fragile basilar membrane. If this is damaged, fluid, high in potassium, leaks into the tympanic canal and damages the nerve fibers. Fiber options have included extremely flaccid types and those with a spiral memory and hug the inner curve of the tympanic canal.

A further problem is that the field current must be conducted away from the electrodes using ions. This can result in electrochemical reactions and erosion of even the most biocompatible metals or the production of a number of gases by electrolysis. Analysis conducted in the 1980s identified the strict conditions that would allow electrodes, particularly platinum and iridium, to be safely used to induce ionic currents in this fluid.

#### 6.9.3.1 Safe Current Density

While most forms of charge delivery result in damage to the surrounding tissue, nondamaging stimulation can be achieved using short-duration biphasic current pulses delivered by platinum electrodes. Typical durations are  $<300 \mu\text{s}$  per phase (half-cycle) with charge densities of  $<60 \mu\text{C}/\text{cm}^2$  per phase. Contemporary cochlear implants operating with a monopolar electrode produce charge densities an order of magnitude below this level.

If circular electrodes are used the maximum charge density,  $q^o$  ( $\text{C}/\text{cm}^2$ ), will be at the surface of the electrode and is given by the total charge,  $Q$  (C), divided by the geometric surface area of the electrode

$$q^o = \frac{4Q}{\pi a^2} \quad (6.22)$$

where  $a$  (cm) is the diameter of the electrode.

A biphasic pulse is used to ensure that the electrochemical reactions that take place are reversible and localized to the electrode–tissue interface. In practice, however, it is not possible to produce a perfectly balanced stimulus. Protection against this residual imbalance can be achieved by shorting the electrodes between current pulses or AC coupling the electrodes.

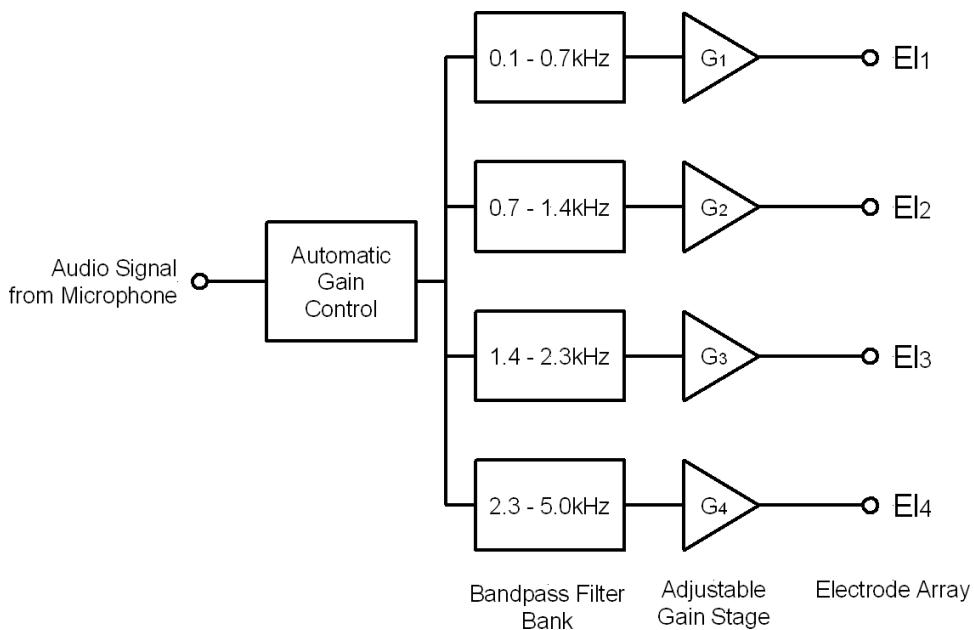
One of the problems with the use of bipolar electrodes is that, because the charge field gradient decreases so quickly, if they are not close enough to the auditory neurons, high charge densities have to be used—with the potential for local tissue damage.

### 6.9.4 Signal Processing and Cochlear Stimulation

The advent of fast low-power microprocessors has made it possible to perform sophisticated signal processing on the audio signal before conversion to an electrical stimulation. Refinements in the algorithms used that have occurred over the past 20 years have contributed considerably to the success of cochlear implants.

#### 6.9.4.1 Band-Pass Filter Bank

The processors of the first implants converted the audio into a high-frequency signal that was applied directly to a single electrode. Contemporary devices all use electrode arrays



**FIGURE 6-34** ■ Basic four-channel cochlear compressed analog processor. [Adapted from (Møller, 2006).]

that stimulate neuron response along a short section of the basilar membrane, with different electrodes being activated by different parts of the sound spectrum.

Because the dynamic range available to electrical stimulation is much smaller than that of normal activation, cochlear implant processors must compress the range of sound intensities to minimize overlap. This is achieved using automatic gain control (AGC) before the signal is passed through a bank of band-pass filters, as shown in Figure 6-34. This is the simplest form of processing and is known as the compressed analog (CA) principle. It simultaneously presents both spectral and temporal information to the cochlea, and both become encoded in the discharge pattern of the stimulated nerve fibers (Møller, 2006).

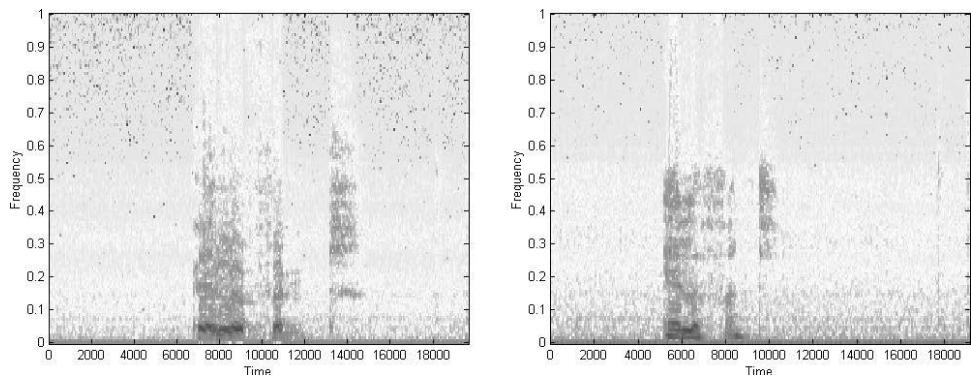
When a small subset of neurons along the basilar membrane is activated with a multichannel electrode array, the noisy sensation takes on a definite pitch, and when several physically separated electrodes are activated sequentially the patient has no difficulty in ordering them into a musical scale. However, patients have likened what they hear to the quacking of ducks or the banging of garbage cans and not at all like the pure tones that are perceived in the case of a sinusoidal tone applied acoustically to a good ear (Loeb, 1985).

A normal auditory nervous system has an extraordinary capacity for extracting underlying information from noisy signals and generalizing across distantly related spectral patterns. That is why it is possible to understand speech by a bass or a soprano from a whisper to a shout irrespective of accent or quality. Examination of the spectrogram of the same word spoken by different people, an example of which is shown in Figure 6-35, shows so much variability that it is difficult for even experienced analysts to identify, so it is not surprising that the crude outputs provided by CA processors were not very effective.

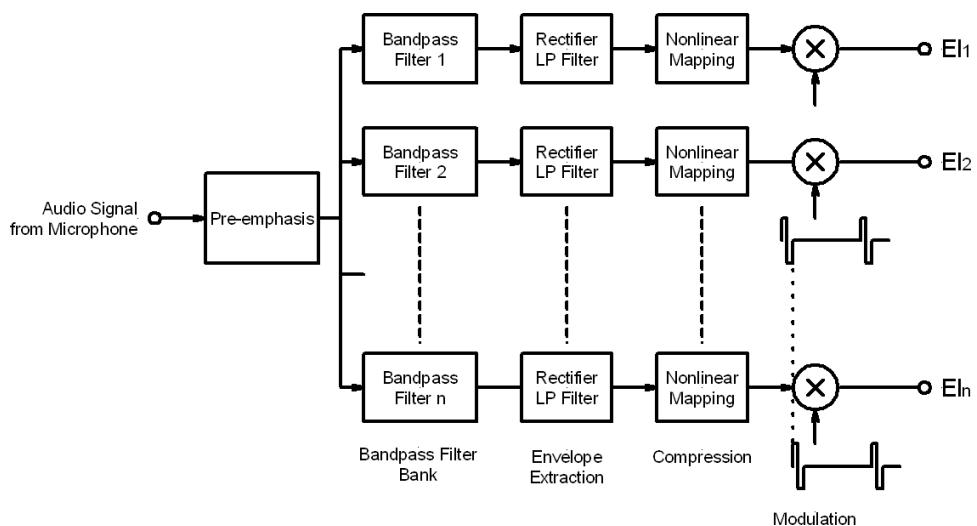
#### 6.9.4.2 Continuous Interleaved Sampling

The first improvement to the CA process, proposed by the Research Triangle Institute, involved applying short impulses to the electrodes rather than the analog signals from each band-pass filter. These electrodes were excited sequentially with small time intervals between them in a process called continuous interleaved sampling (CIS). The output of

**FIGURE 6-35 ■**  
Spectrograms of the word *elephant* spoken by two different people.



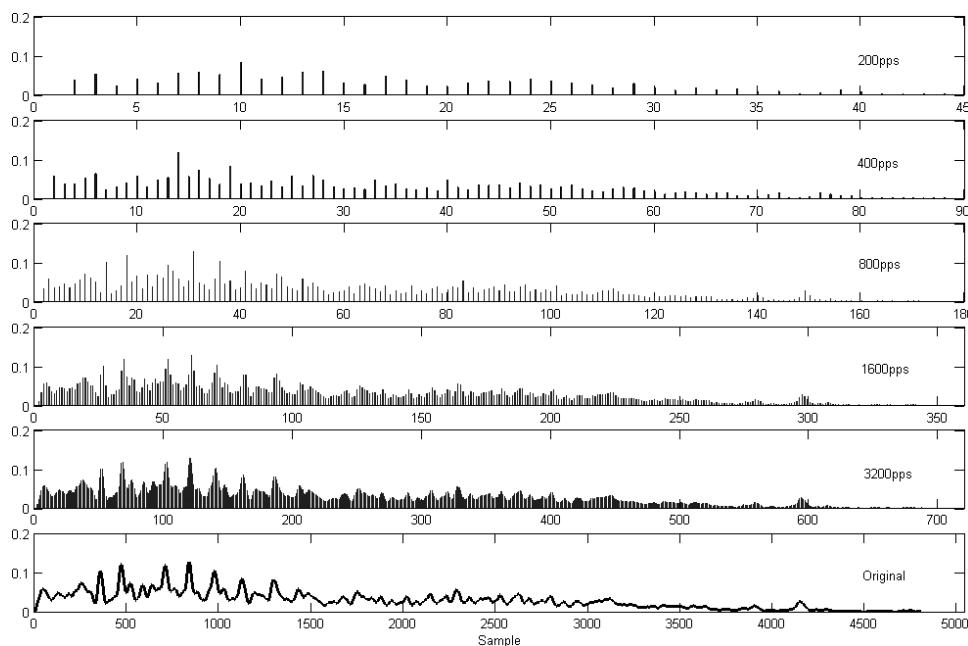
**FIGURE 6-36 ■**  
Schematic diagram showing the CIS strategy for electrode stimulation.



the band-pass filters controlled the amplitude of the pulses applied to produce a channel vocoder (voice encoder).

As shown in Figure 6-36, the CIS strategy, as implemented, starts with a preemphasis filter to attenuate strong components in speech below 1.2 kHz. This is followed by the standard band-pass filter bank after which the envelope from each filter output is obtained by rectification and low-pass filtering. These envelope signals must then be compressed by a nonlinear function (typical logarithmic) to map the wide dynamic range (typically > 90 dB) into the narrow dynamic range for electrically induced hearing, which is typically about 10 dB or a little more. These envelope outputs amplitude modulate the pulse trains applied to the individual electrodes. The biphasic pulse trains for the individual channels are interleaved in time so that the pulses across all of the channels are nonsimultaneous. This eliminates a principal component of electrode interaction of each area on the basilar membrane that would otherwise receive a signal proportional to the weighted vector sum of all of the outputs. The corner frequency of each of the low-pass filters is set to 200 Hz or higher so that the fundamental frequencies of the speech sounds are represented by the modulation waveform.

This algorithm can be implemented in a digital signal processor using the fast Fourier transform (FFT), and the envelope extraction can use the Hilbert transform. Alternatively, it can use a more conventional band-pass filter and envelope detector followed by a low-pass filter.



**FIGURE 6-37 ■**  
Effect of stimulation pulse rates on fine temporal modulation of the syllable /ti/ for a single band-pass channel centered at 3.2 kHz.

**CIS design parameters:** The CIS strategy can be configured in a number of ways to maximize the patient's ability to understand speech. These include filter spacing and envelope cutoff frequencies, the shape of the compression function and stimulation rate (number of pulses delivered by each channel every second).

**Stimulation Rate:** The rate at which current pulses are delivered to each electrode can be as low as 250 pps or as high as 5,000 pps in some devices. It is reasonable to expect that higher pulse rates would result in superior performance of the cochlear implant because high pulse rate stimulation better represents fine temporal modulations present in the signal. This effect is shown in Figure 6-37. However, test results are inconsistent, with some finding a small improvement and some finding no improvement at all at the higher rate. It is possible that these ambiguous results are due to differences in the experimental procedures used or to the actual hardware and algorithms used to implement the CIS strategy.

For example, the Nucleus device uses an FFT to generate each of the envelopes, and because its frame rate is limited frames are repeated if the pulse rate is very high. In that case, no new information is provided with the increase. Results using the Med-El/CIS-Link tests showed that pulse rates above 2100 pps resulted in improved speech and consonant identification compared to the 800 pps base rate (Møller, 2006).

Current commercial implant processors operate at stimulation rates between 800 pps and 2,500 pps with research being undertaken at 5,000 pps. This higher rate aims to restore some randomness to the firing rate of neurons, which is the more natural state.

**Compression function:** Compression of the envelope is essential as it transforms the large dynamic range of the acoustic signal into the small dynamic range required by electrical stimulation. The dynamic range is defined as the range in amplitudes between threshold (barely audible) and uncomfortable loudness (extremely loud). In conversational speech

the amplitude can vary by as much as 50 dB, whereas implant users may have a range of only 5 dB in some cases.

The logarithmic function is commonly used as it matches the amplitude of an electrical signal and the perceived loudness. As a rule of thumb, the loudness of an electrical stimulation current in  $\mu\text{A}$  is analogous to the loudness of an acoustic stimulus in dB.

For log compression a commonly used function is

$$y = A \log(1 + Cx) + B \quad (6.23)$$

As an alternative, power law functions can be used

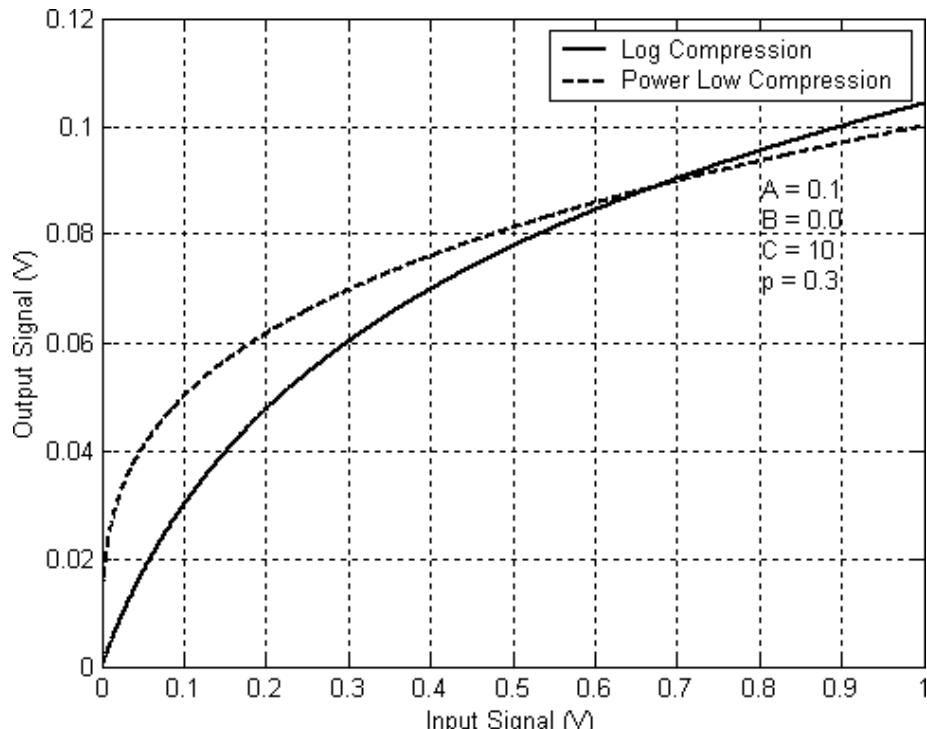
$$y = Ax^p + B \quad (6.24)$$

where ( $p < 1$ ) are becoming more prevalent as the steepness of the compression function can easily be controlled by varying the value of the exponent,  $p$ . The constants  $A$  and  $B$  are chosen so that the input acoustic dynamic range is mapped to the output electrical dynamic range. Examples of these compression curves are shown in Figure 6-38. Experimental studies have shown that the shape of the compression function from  $p = -0.1$  (too compressive) to  $p = 0.5$  (almost linear) has very little effect on performance, with the nearly linear function performing slightly worse (Møller, 2006).

In this example, it can be seen that the input signal has a dynamic range of unity, while the output signal has been compressed down to about 1/10 of that. This equates to a compression ratio of only 10 dB, much smaller than what is used in practice.

**Envelope detection:** Two different methods are commonly used to extract the envelopes from the filtered waveforms. The conventional method is to rectify the signal (full- or

**FIGURE 6-38 ■**  
Comparison  
between log and  
power-law  
compression  
functions.



half-wave) followed by a low-pass filter with a cutoff frequency of between 200 and 400 Hz. The second method, used by the Med-El device, employs the Hilbert transform. No clear advantage has been demonstrated for either method.

In the first method, the low-pass filter also operates as an antialiasing filter, which is required if the signal is down-sampled prior to further processing. Stimulation rate must be at least twice the highest envelope frequency to avoid temporal aliasing, but psychophysics studies have suggested that it should be at least four times higher.

The cutoff frequency of the low-pass filter controls the modulation depth of the envelopes, with lower cutoff frequencies resulting in smaller modulation depths. Studies have found no correlation between cutoff frequency and word, consonant, or melody recognition.

The Hilbert transform is a mathematical function that can represent a time waveform as the product of a slowly varying envelope and a carrier containing fine-structure information. In mathematical terms, the filtered waveform,  $x_i(t)$ , can be represented as

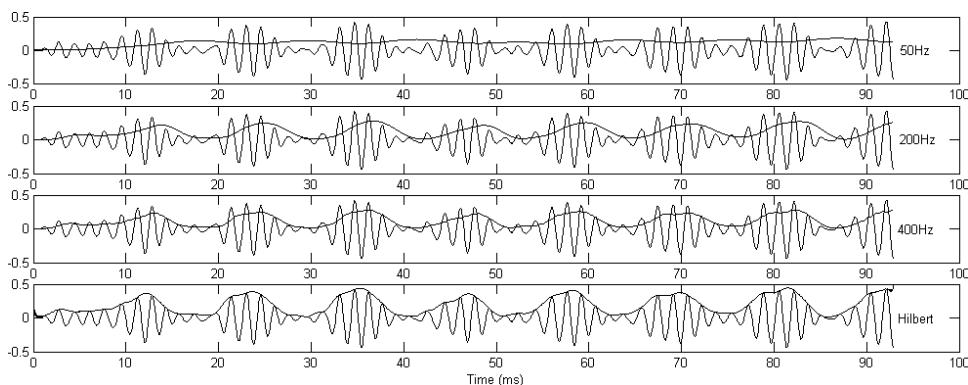
$$x_i(t) = a_i(t) \cos f_i(t) \quad (6.25)$$

where  $a_i(t)$  represents the envelope, and  $\cos f_i(t)$  represents the fine structure waveform. Note that  $f_i(t)$  is the instantaneous phase of the signal, and its derivative produces the instantaneous frequency (the carrier frequency).

Of the two methods of envelope extraction, the Hilbert transform is the more accurate, but it is not significantly better than the rectifier-filter option if the cutoff frequency is high enough. A comparison of these two techniques is shown in Figure 6-39 for different low-pass filter bandwidths. The audio signal is first passed through a band-pass filter with a center frequency of 600 Hz before envelope detection.

Current cochlear implant processors use only the envelope and discard the fine structure. However, simulation studies with normal-hearing listeners have shown that there is useful information in the fine structure. It is not yet clear how this could be used in cochlear implants to improve intelligibility (Møller, 2006).

**Filter spacing:** Over a given frequency range (0–8 kHz), there are a number of ways of allocating filters. Some devices use log spacing, while others use linear spacing at low frequencies (<1.3 kHz) and log spacing thereafter. Research studies have shown that there are some advantages in placing more filters in the F1/F2 region for better representation of the first two formants. The F1 range for most vowel sounds is in the 0–1 kHz band, whereas the F2 band lies in the 1–3 kHz range (Møller, 2006).



**FIGURE 6-39** ■ Comparison between envelope extraction based on full-wave rectification and low-pass filtering and the Hilbert transform for the vowel /a/ output by a single band-pass channel.

### 6.9.5 Spectral Maxima Strategies

Some other strategies have also produced good results. These are different from the CIS method as the channel outputs are scanned and the  $n$  channels with the largest envelope signals are delivered to a subset of the  $m$  available electrodes. They are known as  $m$ -of- $n$ , advanced combination encoder (ACE), or SPEAK strategies. This peak-picking process is designed to reduce the density of stimulation while still representing the important aspects of the acoustic signal. In addition, it is believed that this strategy reduces background noise while maintaining speech levels and therefore improves the overall signal-to-noise ratio of the perceived signal.

In the Nucleus-24 device, the  $m$ -of- $n$  strategy selects 10 to 12 maximum amplitudes of a total of 20 channels. In the ACE (and SPEAK) strategies, a threshold determines how many channels are stimulated. This is typically between 5 and 10 depending on the spectral content of the input signal. Tests have shown that even selecting as few as three channels still achieves a 90% correct level of speech understanding for all stimulus material including sentences, vowels, and consonants. In contrast, the classical CIS process required eight channels to achieve the same level of understanding for consonants and four channels for normal speech (Møller, 2006).

### 6.9.6 Strategies to Enhance Vocal Pitch

The complexity of the sensations evoked by even the most localized stimulation of the basilar membrane has surprised and intrigued many investigators. The place–pitch theory in its simplest form predicts that local activation of an area should elicit a fairly pure tone corresponding to the local resonant frequency of the basilar membrane. Changes in frequency of the stimulus should result only in changes in perceived amplitude as they affect the average rate of firing of the local neurons. However, in practice spectrally complex sensations such as buzzes and clangs arise, and researchers are interested in using this information to improve understanding.

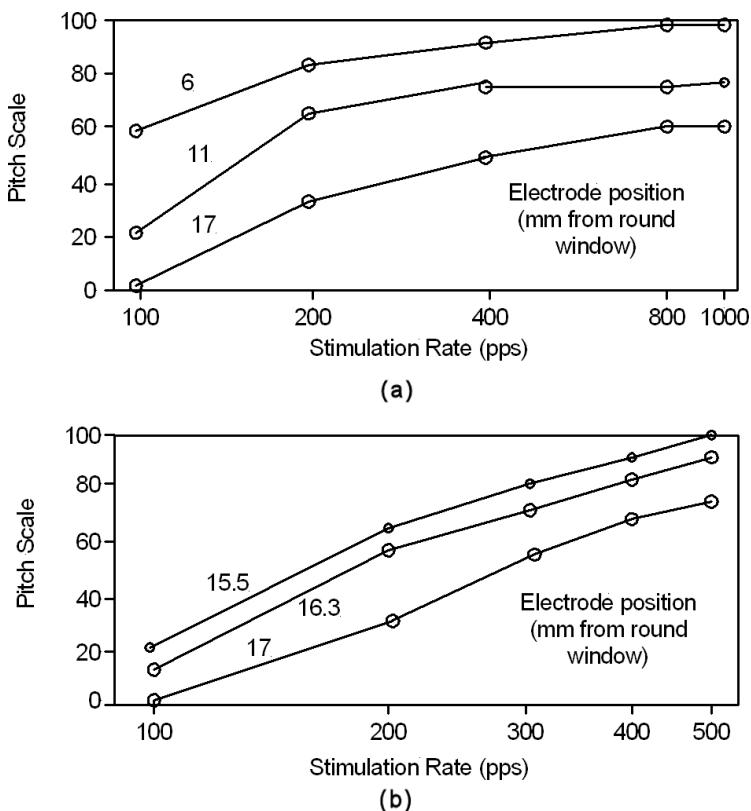
The strategies described in the previous section are designed to convey speech information but fail to convey vocal pitch (F0) information effectively. This means that speakers of tonal languages such as Cantonese and Mandarin find the processing inadequate.

Pitch information can be conveyed by both temporal and spectral (place) cues. Temporal cues are present in the envelope modulations of the filtered waveforms, whereas pitch can be elicited by varying the stimulation rate of the pulse train output on a single electrode. However, once the rate exceeds about 300 Hz patients are unable to use this information directly as it has exceeded the maximum firing rate of the stimulated neurons.

As part of a project to improve music appreciation among cochlear implant users, research was undertaken to determine the relationship between rate and place for pitch perception (Fearn, Carter et al., 1999). Their results, reproduced in Figure 6-40, showed that pitch was strongly dependent on both rate and place at low frequencies but that at rates above a few hundred pps stimulation rate had very little effect and pitch perception was determined solely by the distance of the electrode from the round window.

Little progress has been made in improving music perception since then, so researchers at the Bionic Ear Institute in Melbourne have started working with musicians to compose pieces specifically for cochlear implant users (Hagan, 2010).

It has long been known that auditory nerve activity transmitted to the brain contains detailed information about the exact phase of the motion of the basilar membrane. For



**FIGURE 6-40** ■ Pitch estimate as a function of stimulation rate with the position of the excitation electrode from the round window as a parameter. (a) Wide range of positions 6–17 mm. (b) Narrow range of positions 15.5–17 mm. [Adapted from (Fearn, Carter et al., 1999).]

frequencies below 5 kHz, the exact timing of each nerve impulse is locked to the mechanical phase of the motion sensed by a particular hair cell. Even though the nerve fiber must pause for a few milliseconds between impulses, spectral analysis of the signal reveals the frequency of the stimulation that activated it whether the frequency corresponds to the resonant frequency of that position on the basilar membrane. Specialized neurons processing the output from the auditory nerve can preserve this high-frequency temporal information despite the 300 Hz limitation of the firing rate.

To date no processors are capable of exploiting this phase-locked information, and it is speculated that this failure to correctly reproduce such temporal patterns may be the cause of the complex noisy sensations heard by some patients (Loeb, 1985).

A simpler processing approach now being investigated to enhance spectral (place) cues is based on the *virtual* channel concept. It involves properly manipulating the weighting and phasing (steering) signals delivered to adjacent channels to elicit a pitch intermediate to those obtained by each of the channels individually. A second approach involves modifications to the band-pass filter shape or spacing. To this end, filters have been designed with a triangular transfer function and considerable overlap.

Two different strategies have been developed to enhance temporal cues. One explicitly codes F0 information in the envelope, and the other increases the modulation depth of the filtered waveforms to make F0 cues perceptually stronger. Because these processes rely on the extraction of the F0 component from speech, they can be difficult to achieve (Møller, 2006).

## 6.10 AUDITORY BRAINSTEM IMPLANTS

Patients with neurofibromatosis develop tumors that, when removed, result in deafness and also damage to the auditory nerve. In these cases, the only way to restore some hearing is to use an auditory brain stem implant (ABI). This was first attempted in 1979 when an array was placed over patients' ventral and dorsal cochlear nuclei by William House and William Hitselberger.

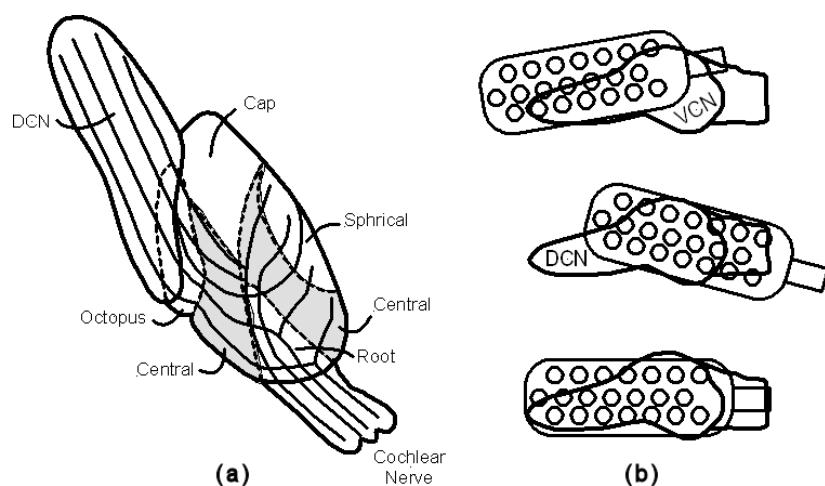
Modern ABIs are similar in design and function to multichannel cochlear implants except for the stimulating arrays. This is possible because the neurons in the cochlear nucleus maintain a frequency sensitivity that maps to their spatial distribution. However, array placement of the electrode array is more challenging than it is with cochlear implants.

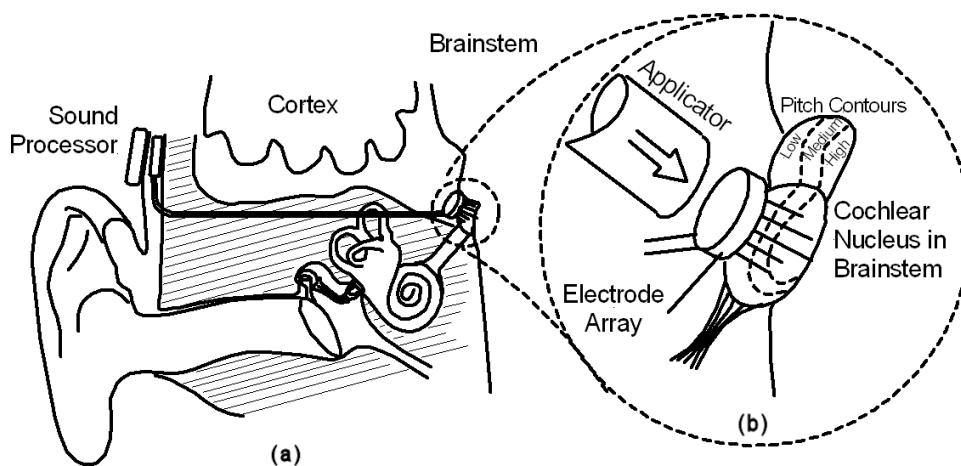
First, the procedure is far more risky than placing implants in the visual cortex to try to restore some vision. Damage to the visual cortex affects only sight, but at the brain stem level every neuron destroyed could damage some other important function. It has taken 15 years of background research, experiments with animals, and surgery on cadavers to convince surgeons that the procedure could be undertaken reasonably safely.

The cochlear nucleus complex is part of the floor of the lateral process of the fourth ventricle, and it is partially obscured by the cerebellar peduncles, making access difficult. The ventral cochlear nucleus (VCN) consists of a rostral area of spherical cells, a caudal region containing octopus cells, and a central region containing a heterogenous mix of globular, multipolar, and small cells. The dorsal cochlear nucleus (DCN) lies within the lateral recess of the fourth ventricle, whereas the VCN extends to the foramen of Luschka. Once the cochlear nucleus is exposed, the electrode array position must be optimized as shown in Figure 6-41 by making use of electrically evoked auditory brainstem responses obtained by stimulating the nucleus and recording from electrodes placed on the scalp. In addition to facial nerve monitoring, the lower cranial nerves are also monitored to detect any nonauditory sensations. Penetrating ABIs can be positioned in a similar manner.

To date, more than 500 patients have had ABIs implanted, but the results have not been as positive as those for cochlear implants. ABIs enable the person to hear, but usually not well enough to understand speech because the implant cannot separately stimulate different groups of nerves corresponding to distinct frequency ranges.

**FIGURE 6-41 ■**  
Auditory brainstem implant. (a) Sketch of the cochlear nucleus. (b) Some surface electrode alignment issues.



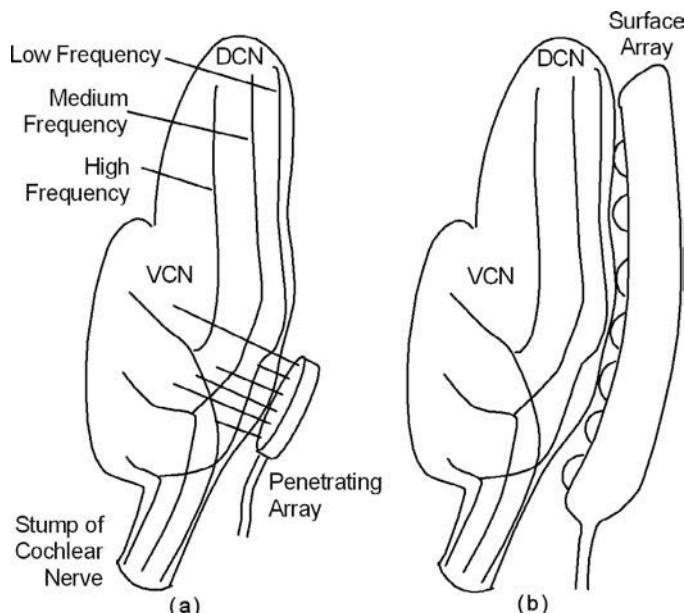


**FIGURE 6-42** ■  
Auditory brainstem implant. (a) Cross section through the ear showing implant position. (b) Insertion of penetrating electrode array. [Adapted from (Graham-Rowe 2004).]

An implant developed by Bob Shannon of the House Ear Institute is expected to perform better as it consists of eight electrodes of different lengths that are inserted into the brain stem, as shown in Figure 6-42, and will be able to stimulate several bundles of nerves individually. These electrodes should be able to excite different regions of the cochlear nucleus that respond to and process signals of different frequencies.

### 6.10.1 Electrodes

Surface ABI electrode arrays consist of a number of closely spaced disk or button-shaped elements embedded in a silicon carrier fixed to a fabric mesh for strength. Penetrating ABIs typically consist of needle microelectrodes, which are carefully designed because if they are too sharp they cut cells and if they are too blunt they crush them (Graham-Rowe, 2004). Examples of these two electrode types are shown in Figure 6-43.



**FIGURE 6-43** ■  
Examples auditory brainstem electrode arrays.  
(a) Penetrating type.  
(b) Surface type.

## 6.10.2 Stimulus Mapping

Unlike cochlear implants, the relationship between electrode position and perceived pitch is not straightforward in the cochlear nucleus, so 1 to 2 months after insertion of the electrode array, a process of mapping must occur to eliminate electrodes that stimulate nonauditory responses and to identify the pitch perceived by each electrode.

During this process, which occurs repeatedly over several months, different electrodes are activated, and patients are asked what kind of sounds they are hearing. This is used to calibrate the machine to provide usable information to patients and also because many other things in this area of the brain might be stimulated with unfortunate results.

## 6.11 REFERENCES

---

- Bernhard, H., C. Steiger and Y. Perriard. (2011). "Design of a Semi-Implantable Hearing Device for Direct Acoustic Cochlear Stimulation." *IEEE Trans. on Biomedical Engineering* 58(2): 420–428.
- Bilger, R., F. Black, N. Hopkinson, E. Myers, J. Payne, N. Stenson, A. Vega and R. Wolf (1977). "Evaluation of subjects presently fitted with implanted auditory prostheses." *Annals of Otology, Rhinology and Laryngology* 86(3 part 2): 1–176.
- Carman, G. (2008). "Eureka Moment from First One to Hear with Bionic Ear: Rod Saunders 1931–2007." *Sydney Morning Herald*, p. 28.
- Chien, W., J. Rosowski, M. Ravicz, S. Rauch, S. J and S. Merchant. (2009). "Measurements of Stapes Velocity in Live Human Ears." *Hearing Research* 249(1–2): 54–61.
- Clark, G. (1969). *Middle Ear and Neural Mechanisms in Hearing and in the Management of Deafness*. PhD dissertation, Sydney, University of Sydney.
- Dapkus, D. (2000). Class-D Audio Power Amplifiers: An Overview. In *International Conference on Consumer Electronics, 2000. ICCE. 2000 Digest of Technical Papers*. Dallas, Texas Instruments, pp. 400–401.
- Fearn, R., P. Carter and J. Wolfe. (1999). "The Dependence of Pitch Perception on the Rate and Place of Stimulation of the Cochlea: A Study Using Cochlear Implants." *Acoustics Australia* 27: 41–43.
- Fidelis. (2009). "Audiology: Hearing Aids." Retrieved August 2009 from <http://www.fidelisdiagnostics.com/audiology/other/hearingaids.html>
- Finn, W. and P. LoPresti (Eds.). (2003). *Handbook of Neuroprosthetic Methods*. London: CRC Prsss.
- Flynn, M. (2007). "Overcoming Conductive/Mixed Hearing Losses and Single-Sided deafness with Bone-Anchored Hearing Devices." Retrieved July 2008 from <http://www.hearingreview.com>
- Ganot, A. (1872). *Natural Philosophy for General Readers and Young Persons*. New York: D. Appleton.
- Graham-Rowe, D. (2004). "First Brainstem Implants Aim to Tackle Deafness." Retrieved July 2008 from <http://environment.newscientist.com/article/dn4540-first-brainstem-implants-aim-to-tackle-deafness.html>
- Graupe, D. and G. Causey. (1975). *Method of and Means for Adaptively Filtering Near-Stationary Noise from Speech*. U.S. Patent 4025721.
- Hagan, K. (2010). "Implant Research Puts New Spin on an Ear for Music." *Sydney Morning Herald*.
- Håkansson, B. (2009). "Technical Development of the BAHA—An Historical Review." Retrieved August 2009 from [http://www.baha-users-support.com/birth\\_of\\_baha.php](http://www.baha-users-support.com/birth_of_baha.php)
- Hough, J., R. Dyer and J. Wolfe. (2001). "Early Clinical Results: SOUNDTEC Implantable Hearing Device Phase II Study." *Laryngoscope* (111): 1–8.
- Kroll, K., I. Grant and E. Javel. (2002). "The Envoy Totally Implantable Hearing System." *Trends in Amplification* 11(4): 73–80.

- Loeb, G. (1985). The Functional Replacement of the Ear. *Scientific American* 252: 86–92.
- Møller, A. (Ed.). (2006). *Cochlear and Brainstem Implants*. Basel, Switzerland: Karger.
- Nowak, R. (2008). “Light Opens Up a World of Sound for Bionic Ears.” *New Scientist*, November 22, p. 10.
- Ricketts, T. (2008). “Digital Hearing Aids: ‘State-of-the-Art.’” Retrieved July 2008 from [http://www.asha.org/public/hearing/treatment/digital\\_aid.htm](http://www.asha.org/public/hearing/treatment/digital_aid.htm)
- Shohet, J. (2008). “Implantable Hearing Devices.” Retrieved July 2008 from <http://www.emedicine.com/ent/TOPIC479.HTM>
- Traynor, R. and J. Fredrickson. (2008). “The Future Is Here: The Otologics Fully Implantable Hearing System.” Retrieved August 2009 from <http://www.middleearimplants.com/>
- Washington University. (2005). “Deafness in Disguise.” Retrieved July 2008. from <http://beckerexhibits.wustl.edu/did/>.
- Wilson, B. and M. Dorman. (2008). “Interfacing Sensors with the Nervous System: Lessons from the Development and Success of the Cochlear Implant.” *IEEE Sensors Journal* 8(1): 131–147.



# Sensory Substitution and Visual Prostheses

## Chapter Outline

7.1	Introduction.....	334
7.2	Anatomy and Physiology of the Visual Pathway .....	335
7.3	Main Causes of Blindness.....	339
7.4	Optical Prosthetics—Glasses, Thermal Imagers, Night Vision.....	339
7.5	Sonar-Based Systems.....	341
7.6	Laser-Based Systems.....	350
7.7	Sensory Substitution .....	350
7.8	GPS-Based Systems .....	370
7.9	Visual Neuroprostheses .....	371
7.10	The Future.....	391
7.11	References .....	392



## 7.1 | INTRODUCTION

According to a 2005 World Health Organization (WHO) report, about 37 million people worldwide are totally blind, and a further 128 million are visually impaired. The statistics for the developed world are somewhat improved because of better healthcare, but nonetheless there are still approximately 10 million people in the United States who are visually impaired, of which more than 1 million are legally blind. The statistics in Australia are similar, with about 480,000 visually impaired and 50,000 blind people. It is obvious that if an effective visual prosthetic device could be developed it would improve the quality of life of many people.

To date, the most successful and widely used visual prosthesis for the blind is the *white cane*. It can be used to detect obstacles on the ground, uneven surfaces, holes, steps, and puddles. Even though the cane is inexpensive and is so light and small that it can be folded and tucked away in a pocket, only a small percentage of blind people actually use one—about 110,000 in the United States. Even fewer use other aids, with only 7000 using guide dogs at present and only about 1500 new users graduating from dog guide schools every year.

High-tech devices have been on the market for many years but appear to lack utility and consequently are not as widely used.

For the partially sighted, visual prosthetics include lenses—eyeglasses and contact lenses that correct for focus and astigmatism. These have been around for more than 700 years; Salvino D'Armate is credited with inventing the first wearable eyeglasses in 1284 in Italy (Figure 7-1). However, the oldest known lens was found in the ruins of ancient Nineveh and was made of polished rock crystal about 40 mm in diameter, so it is quite likely that primitive forms of prosthetic lenses were developed prior to D'Armate's invention (Bellis, 2008).

More advanced visual prosthetics can be divided into three major groups. First, there are the devices that use either ultrasound or a camera to sample the environment ahead of an individual and render the results into either a series of sounds or a tactile display. This process is known as sensory substitution because the sense of hearing or touch is substituted for that of sight. From this substitution users are supposed to be able to discern the shape and proximity of objects in their path. The second major form is retinal enhancement. These sensors supplement functions of the retina by stimulating it with electrical signals, which are then converted to nerve impulses and transferred through the optic nerve to the brain. The third major category of visual prosthetic is a digital camera that samples an

**FIGURE 7-1 ■**

Reproduction of the earliest known pair of wearable eyeglasses invented by Salvino D'Armate in 1284.



image and stimulates the brain directly with electrical signals, either by penetrating into or placing electrodes on the surface of the optic nerve or the visual cortex.

Sensory substitution prostheses based on ultrasound technology have been available for many years because that technology was really the only type that was portable and reasonably low cost. In the last decade, however, the advent of low-cost cameras based on charge coupled devices (CCD), and more recently complementary metal oxide semiconductor (CMOS) arrays has spurred the development of camera-based systems.

Scientists knew as early as 1918 that touching electrodes to the visual cortex of conscious patients produced spots of light (phosphenes), and by the 1940s researchers had established the concept of artificial electrical stimulation of the visual cortex. In the late 1960s, British scientist Giles Brindley experimented with a system that placed electrodes on the brain's surface. When specific areas of the brain were stimulated using relatively high currents, the blind volunteers all reported "seeing" phosphenes that corresponded approximately to where they would have appeared in space.

Encouraged by this work, the National Institutes of Health (NIH) supported research to develop and deploy an interface based on ultrafine wire (25 to 50  $\mu\text{m}$  in diameter) densely populated with electrode sites that could be implanted deep into the visual cortex. This innovation decreased the current density required, compared with Brindley's original design, and was therefore less damaging to the surrounding tissue. New electrode technology was developed that could be safely implanted in animals to electrically stimulate, and passively record, electrical activity in the brain. These efforts produced significant advances in neurophysiology, with the publication of hundreds of papers in which researchers attempted to develop electronic interfaces to the brain.

By 1995, NIH researchers decided they were ready to proceed to human testing of an intracortical visual prosthesis. A total of 38 electrodes, connected to fine wires penetrating the scalp, were implanted in the brain of a 42-year-old woman. Although blinded by glaucoma 22 years earlier, she was still able to sense phosphenes using the electrical stimulation and eventually became adept at perceiving those dots under a variety of stimulation patterns.

The state-of-the-art continues to advance, with more sophisticated retinal and brain implants along with their associated electrode arrays and image processing networks in production. However, it will be many years, if ever, before any system can approach the resolution and dynamic range of human vision.

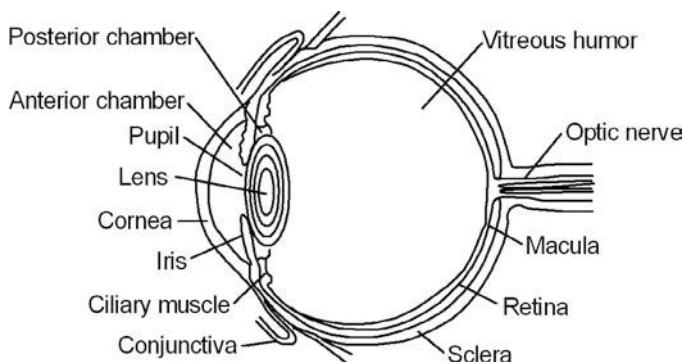
## 7.2 | ANATOMY AND PHYSIOLOGY OF THE VISUAL PATHWAY

---

Light enters the eye through the cornea, a transparent dome on the front surface. The cornea serves as a protective covering for the front of the eye and, because it is curved, also helps focus light on the retina. After passing through the cornea, some of the available light travels through the pupil, a hole through the iris. The iris is the eye's circular, colored area that controls the amount of light that enters by dilating and constricting the pupil like the aperture of a camera lens. This adjustment is controlled by the action of the pupillary sphincter muscle and the dilator muscle.

Behind the iris sits the lens, which by changing its shape focuses light onto the retina. Through the action of small ciliary muscles that surround it, the lens becomes thicker to

**FIGURE 7-2 ■**  
Cross section through the eye showing its internal structure.



focus on nearby objects and thinner to focus on more distant objects. Objects that are more distant than about 7 m are considered to be at infinity, and no focal accommodation takes place.

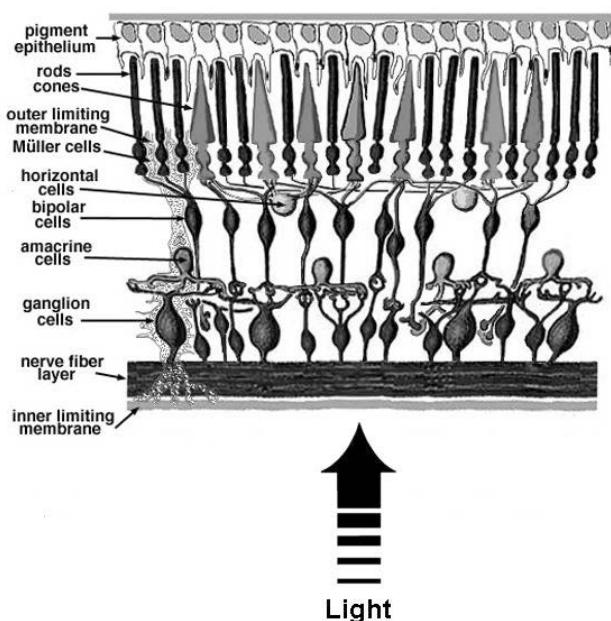
As illustrated in Figure 7-2, the eyeball is divided into two sections, each of which is filled with fluid. The anterior segment extends from the inside of the cornea to the front surface of the lens and is filled with a fluid called the aqueous humor that nourishes the internal structures. The posterior segment extends from the back surface of the lens to the retina and contains a jelly-like fluid called the vitreous humor. The pressure generated by these fluids fills out the eyeball and helps maintain its shape.

The anterior segment is divided into two chambers. The anterior chamber extends from the cornea to the iris, and the posterior chamber extends from the iris to the lens. Normally the aqueous humor, which is produced in the posterior chamber, flows through the pupil into the anterior chamber before draining out of the eyeball through outflow channels located where the iris meets the cornea.

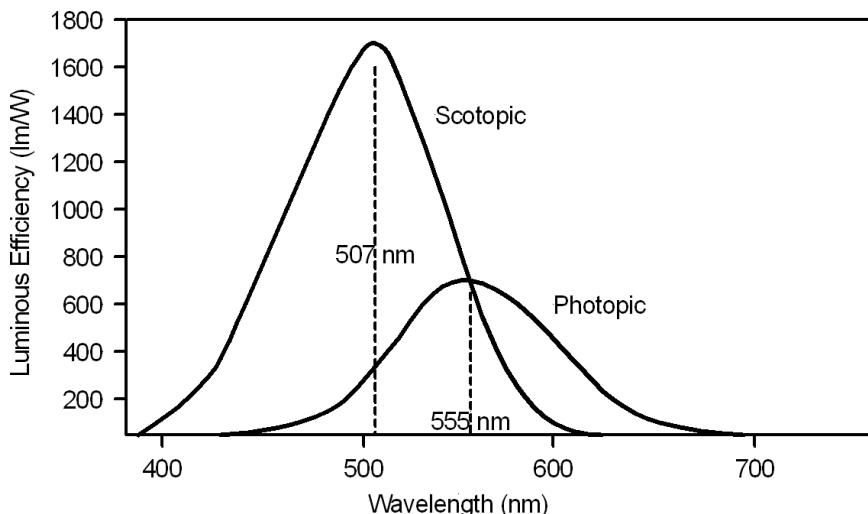
At the back of the eye lies the retina, which contains photoreceptors and the blood vessels that nourish them as well as a dense array of nerve fibers and specialist cells that preprocess the visual information. The most sensitive part of the retina is a small area called the macula, which is about 5 mm in diameter and subtends an external angle of about 20°. At its center is the fovea, a region about 1.5 mm in diameter, which contains the highest density of tightly packed photoreceptors. In this region, there are typically more than 150,000 photoreceptors per mm<sup>2</sup>, which together subtend an external angle of about 6° of central vision (Finn and LoPresti, 2003).

The retina has a laminar neural organization that consists of three layers containing neuronal cells separated by two layers that contain synaptic connections. An inner and outer limiting membrane and an outer epithelial layer surround the neural components, as shown in the simplified diagram in Figure 7-3.

The outermost layer, called the retinal pigmented epithelium (RPE), supports and provides essential nutrients to the photoreceptors in the layer below. This nuclear layer contains the photoreceptor neurons that convert the optical signal into electrochemical energy. There are two main types of photoreceptors: cones and rods. Cones are responsible for the detailed central vision and color (photopic) vision and are clustered mainly in the macula. There are about 5 million of these in the human retina. The rods are responsible for night (scotopic) vision and peripheral vision and are not present at all in the fovea. They are more numerous than the cones (about 100 million) and are much more sensitive to light, as shown in Figure 7-4, but only in the blue-green spectral region and therefore do not register color. Rods are found mainly in the peripheral areas of the retina and do



**FIGURE 7-3** ■ Schematic diagram of the organization of the retina  
[Adapted from (Kolb, Fernandez et al., 2009) with permission]



**FIGURE 7-4** ■ Sensitivity of scotopic vision (rods) and photopic vision (cones) to light intensity

not contribute to detailed central vision as the cones do. Both rod and cone photoreceptors use graded voltage potentials to signal a change in light intensity. This is a change in the membrane potential rather than an action potential.

Photoreceptors form synapses with two classes of neurons (bipolar and horizontal) in the layers below, as shown in Figure 7-3. These also use graded voltage potentials, while a third class of neurons in that layer, amacrine cells, generate action potentials. Both the bipolar and amacrine cells form synapses with retinal ganglion cells (RGCs), the axons of which form the nerve fiber layer that crosses the inner surface of the retina and runs toward the optic nerve. These fibers are unmyelinated until they reach the optic disk.

The final layer of the retinal structure is the inner limiting membrane that separates this nerve fiber layer from the vitreous humor.

Visual information is processed from the outer to the inner layers in a column-wise process. That is, the activity of a ganglion cell is modulated by the photoreceptor and bipolar neurons along a column from the outer to the inner retina. This ensures that the visual map structure is preserved through all of the layers. In addition to this, horizontal and amacrine cells integrate information across neighboring regions of visual space.

The fovea has the highest density of ganglion cells to photoreceptors—a ratio of two to one—hence, this region has the highest visual acuity. The number of ganglion cells per photoreceptor decreases with increasing radius from the fovea, and by the edge of the macular region a large number of photoreceptors drives a single ganglion cell. At the fovea, the receptive field size is of the order of a few minutes of arc, while even at the edge of the macula the receptive field size can be as large as 5°.

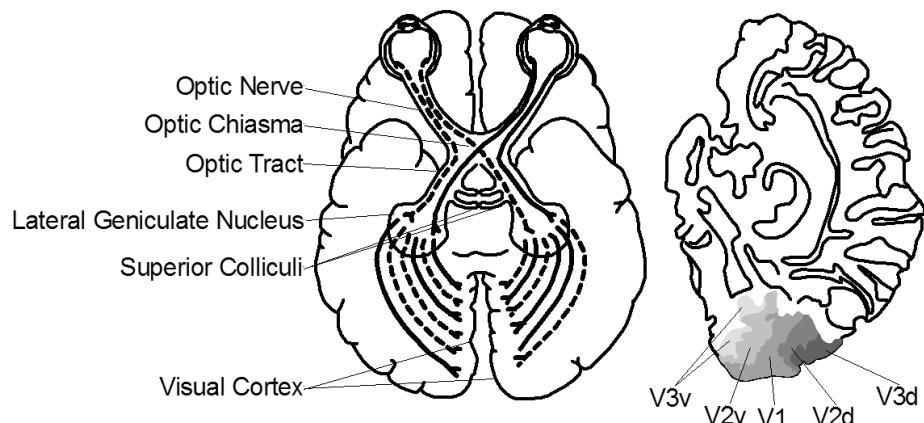
Most RGCs are optimally stimulated by a circular patch of light of one intensity surrounded by an annulus of a different intensity. This takes the form of a patch of one color surrounded by an annulus of another color for 80% of cells, or a light center and a dark annulus or a dark center and a light annulus in about 8% of the cells. The remaining 12% of retinal ganglion cells have an as yet uncharacterized function (Finn and LoPresti, 2003).

As shown in Figure 7-5, nerve signals travel from each eye along the corresponding optic nerve to the back of the brain where vision is sensed and interpreted. The two optic nerves meet at the optic chiasma, which is an area behind the eyes immediately in front of the pituitary gland and just below the cerebrum. There, the optic nerve from each eye divides to form the optic tract with half of the nerve fibers from each side crossing to the other side and continuing to the subcortical targets at the back of the brain. Thus, the right side of this subcortical region receives information through both optic nerves for the left field of vision, and the left side of the brain receives information through both optic nerves for the right field of vision. The center of these fields of vision overlaps.

Indications suggest that the fibers of the optic nerve are visuotopically organized with the upper retina (lower visual field) represented along the dorsal side of the nerve, the central retina along the lateral side, and nasal visual field along the medial side. However, this seems to vary somewhat along the length of the nerve (Finn and LoPresti, 2003).

The axons of the retinal ganglion cells target three subcortical structures: (1) the superior colliculi; (2) the pretectum; and (3) the lateral geniculate nucleus (LGN). The vast majority (>90%) target the LGN. The superior colliculi and the pretectum are located on the roof of the midbrain and are associated with saccadic eye movements and pupillary

**FIGURE 7-5 ■**  
Nerve fiber path to the neural cortex showing potential situations for visual prosthetic implants.  
[Adapted from (Foley and Martin, (2006).]



reflexes, respectively. Neither of these structures is associated with vision as such, so neural prostheses generally target the LGN or the visual cortex.

The LGN is a small structure about  $7 \times 7 \times 2$  mm located on the ventral side of the thalamus and is often considered a relay point receiving input from the RGC and passing the same on to the cortex. However, this is a simplification as the LGN consists of six distinct processing layers that are visuotopically organized so that the image (primarily from the fovea) propagates through the structure undistorted.

Most of the axons from the LGN project to the gyri lining the calcarine fissure shown in Figure 7-5. This 2500 to 3500  $\text{mm}^2$  region is known as the primary visual cortex or visual area (V1), indicating that it is the first region of visual processing at the cortical level. Five cortical areas have a similar laminar structure consisting of six layers arranged in planes extending from the pia mater through to the white matter below. Their total thickness is about 2 mm in human beings. Visual processing is performed in both a columnar and a horizontal manner. In these layers of area V1, the circular sensitivity of the RGCs is elongated into a long ellipse that leads to a preference for bars of a particular orientation.

The other areas—V2, V3, V4, and V5—indicate a somewhat hierarchical structure of visual processing that is not completely accurate as significant feedback occurs between the levels and areas. After area V1, the organization of the visual pathway becomes more complex and optimal visual stimuli become less clear. From a biomechatronic perspective, these regions of higher-level feature extraction are not well enough understood to be used as sites for visual prostheses.

## 7.3 | MAIN CAUSES OF BLINDNESS

The main causes of untreatable blindness are age-related macular degeneration (AMD), retinitis pigmentosa (RP), accidents, and cancers. Other causes include glaucoma, cataracts, and diabetic retinopathy, but these can be treated or sometimes even prevented if treatment starts at an early stage.

AMD is characterized by a progressive loss of photoreceptor cells in the macula and then of central vision. For unknown reasons the pigmented epithelium of the retina degenerates, leading to a subsequent degeneration of the photoreceptor layer and fluid leakage into the neural portions of the retina.

RP is an inherited disease characterized by loss of peripheral and night vision. Once again, the cause of the problem is unknown, but it results in the degeneration of rod photoreceptors. In extreme cases, sufferers are left with only cone cells and therefore have tunnel vision.

Despite the degradation of the outer layers of photoreceptors in both of these diseases, the inner cell layers are often partially preserved, which gives retinal implants the potential of restoring some vision.

## 7.4 | OPTICAL PROSTHETICS—GLASSES, THERMAL IMAGERS, NIGHT VISION

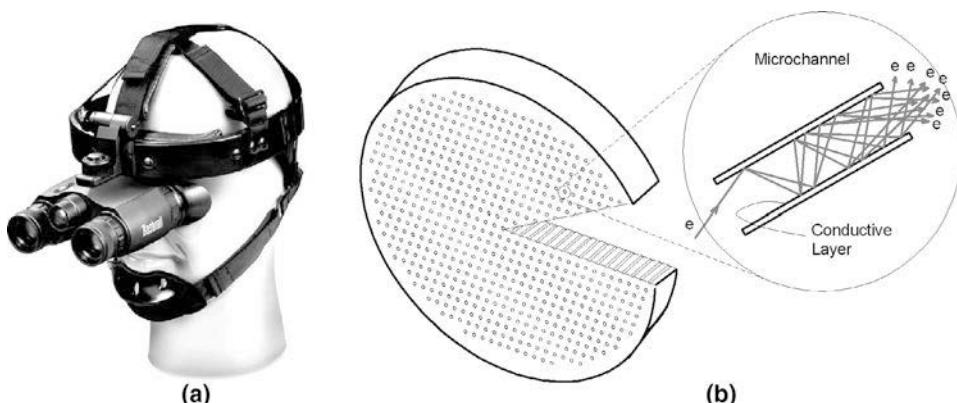
As discussed at the start of this chapter, eyeglasses are the most common optical prosthesis, followed closely by contact lenses. A third extremely common prosthesis is the plastic lens that replaces the cloudy one in cataract sufferers. The lenses in most eyeglasses are

**FIGURE 7-6 ■**

Night vision

(a) Photograph of night vision binoculars.

(b) Schematic diagram of a microchannel plate image intensifier element. (Brooker 2008.)



designed to compensate for near- or farsightedness, in which the eye is unable to focus a sharp image on the retina. In some cases, they also correct for astigmatism, in which distortions of the eyeball or lens structure have resulted in two different focal axes.

One of the goals of visual prosthetic research is to provide a capability that extends the frequency range and acuity of our vision in an unobtrusive package. Lenses and mirrors can be configured to provide magnification and to capture orders of magnitude more light than our own eyes do, but these telescopes are hardly unobtrusive. An alternative is to electronically amplify the numbers of photons striking the eye. Such devices are called image intensifiers or night-vision cameras.

Image intensifiers are sensors that amplify ambient ultraviolet, visible, or near-infrared (NIR) radiation from a scene and then redisplay it in the visible spectral range. The scene is imaged onto a photocathode that has an energy-band structure such that, on absorption of light, electrons are emitted from the surface. The process of amplification involves either adding kinetic energy to each electron using a potential difference or multiplying the number of electrons using the avalanche effect so that many more photons are emitted when the electrons strike a phosphor screen than were originally generated. In most modern devices, the structure that performs the amplification is the microchannel plate (see Figure 7-6). It is a disk a few millimeters thick containing millions of holes lined with a conductive medium that facilitates avalanche multiplication of the number of electrons passing through.

Thermal imagers use a different principle to extend the frequency range of human vision right down into the thermal infrared region. Instead of using a sensitive photocathode, they mostly use cryogenically cooled arrays of light-sensitive elements made from mercury cadmium telluride that convert thermal photons directly to an electric current to produce TV quality images of long-wave infrared radiation (Brooker, 2008). It is in this region that warm objects, including the human body, radiate. Imaging sensors thus can see and use subtle temperature variations in human beings for diagnosing many diseases, including some cancers and even severe acute respiratory syndrome (SARS; Bronzino, 2006).

Ultimately, it may be possible to incorporate this extended frequency range and sensitivity, as well as other features, into implantable prostheses to produce a true bionic visual capability. At present, however, these devices rely on external optics and electronics and can be used to enhance, but not repair, human visual acuity.

## 7.5 SONAR-BASED SYSTEMS

Sonar systems operate by transmitting a brief pulse of ultrasound from a transducer and then listening for an echo. As the speed of sound in air is reasonably constant (about 340 m/s), the round-trip time gives the range to the reflecting surface, and because the sound is radiated and detected over a narrow angular field, the beam pattern of the transducer, this also gives a good measure of direction to the reflecting surface.

For short-range applications, most ultrasonic transmitters are either made from thin wafers of piezoelectric (PZT) material or use electrostatic transducers. A typical piezoelectric transducer similar to the unit shown in Figure 7-7, such as the SensComp 40LT16, has the following characteristics (SensComp, 2004a):

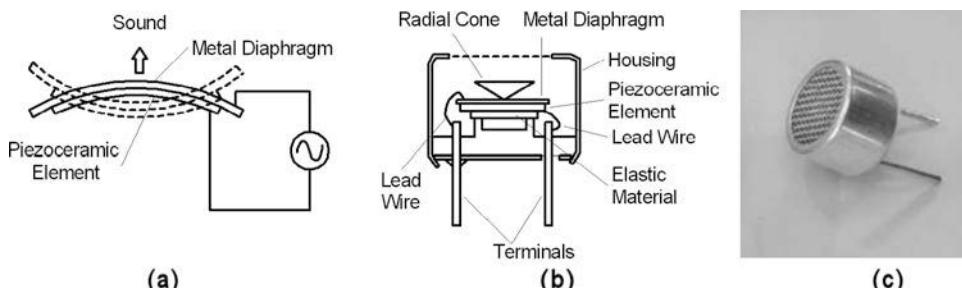
- Center frequency:  $40+/-1$  kHz
- Bandwidth: 2 kHz
- Transmit sound pressure level (SPL): 120 dB min at 40 kHz; 0 dB relative to 0.0002  $\mu$ bar per 10 V root mean square (RMS) at 30 cm
- Maximum drive voltage: 20 V RMS
- Beam angle (6 dB):  $55^\circ$

The matched receiver for this transducer is the 40LR16, which has similar specifications:

- Center frequency:  $40+/-1$  kHz
- Bandwidth: 2.5 kHz
- Sensitivity:  $-65$  dB at 40 kHz relative to 0 dB = 1 V/ $\mu$ bar

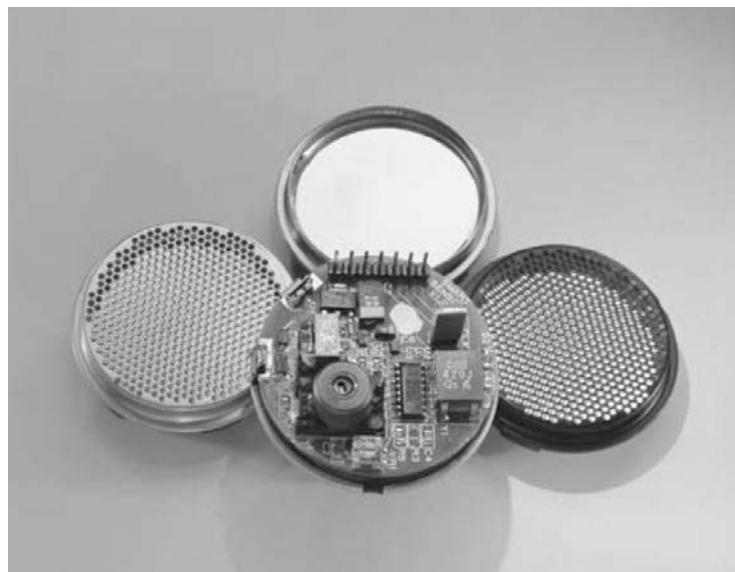
Electrostatic transducers are typically larger than the PZT variety, as can be seen in Figure 7-8, and thus have narrower beamwidths. Because they are not resonant, they also exhibit a much wider bandwidth, as seen in the abridged specifications for the SensComp 600 series reproduced as follows (SensComp, 2004b):

- Center frequency: 50 kHz
- Bandwidth:  $\approx 40$  kHz
- Transmit SPL: 110 dB min at 50 kHz; 0 dB relative to 20  $\mu$ Pa at 1 m (300 VACpp 150 VDC bias)
- Beam angle (6 dB):  $15^\circ$



**FIGURE 7-7** ■ Principles and structure of a PZT ultrasound transducer. (a) Operational principles of the bimorph element. (b) Transducer diagram showing impedance matching cone and electrical connections. (c) Photograph of transducer. [Adapted from (Brooker, 2008).]

**FIGURE 7-8 ■**  
 Photograph of the  
 SensComp 600  
 series transducer.  
 (Courtesy of  
 SensComp  
<http://www.senscomp.com/>.)



In this case the same transducer operates as a receiver with the following specifications:

- Sensitivity:  $-42 \text{ dB}$  at  $50 \text{ kHz}$  relative to  $0 \text{ dB} = 1 \text{ V}/\mu\text{bar}$  (for a  $150 \text{ V}$  direct current [DC] bias)

Unfortunately, the wider bandwidth results in a much poorer receive sensitivity than that of the narrow band resonant transducer variety.

As an acoustic signal propagates through the atmosphere, it attenuates. The main reason for this is the beam divergence, which reduces the signal power density with the square of the distance traveled. The second mechanism is atmospheric absorption, which converts some of the acoustic energy into heat. A reasonably simple model for the atmospheric attenuation,  $\alpha$  (dB/m), which includes the effects of both temperature and humidity, has been derived from measured data and is accurate for relative humidity greater than 30% (Knudsen and Harris, 1950).

$$\alpha_c = \left( \frac{f}{1000} \right)^{3/2} \frac{0.283}{20 + \phi_t} \quad (7.1)$$

where

$$f = \text{frequency (Hz)}$$

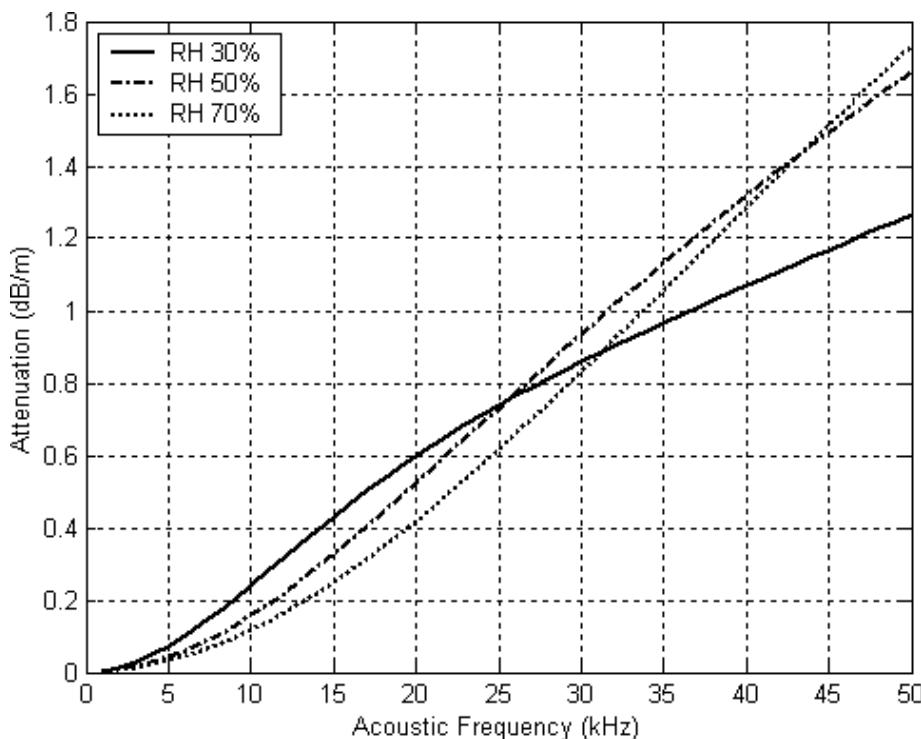
$$\phi_t = \phi_{20}(1 + 0.067\Delta t)$$

$$\phi_{20} = \text{relative humidity \% at } 20^\circ\text{C}$$

$$\Delta t = \text{the temperature difference from } 20^\circ\text{C}$$

Consider the attenuation,  $\alpha_c$  (dB/m), determined using this model at a frequency of  $30 \text{ kHz}$  for a temperature of  $20^\circ\text{C}$  and a relative humidity of 30%.

$$\alpha_c = \left( \frac{30 \times 10^3}{1000} \right)^{3/2} \frac{0.283}{20 + 30} = 0.93 \text{ dB/m}$$



**FIGURE 7-9** ■ Sound attenuation as a function of frequency and relative humidity in air at 20 °C.

A far more comprehensive model (Burnside, 2004) has been used to generate the results shown in Figure 7-9. These results give an attenuation of 0.88 dB for a frequency of 30 kHz at 20 °C and a relative humidity of 30%. This result is in reasonably good agreement with the results from (7.1).

The maximum range that can be achieved by ultrasound sonar systems is determined by a combination of the beam divergence, the atmospheric attenuation, and the target size. If the target is a fixed size, like a street sign, then the received power decreases with  $R^4$ , made up from an  $R^2$  term on the way to the target and a second  $R^2$  term on the way back. For a flat target like a wall, the received power decreases with the square of the range,  $R^2$ , because the reflecting area increases with  $R^2$ , which cancels one of the terms of the previous case. These losses in conjunction with an atmospheric attenuation of about 1 dB/m limit the maximum range to about 5 m. However, that is more than sufficient as an extension to the white cane.

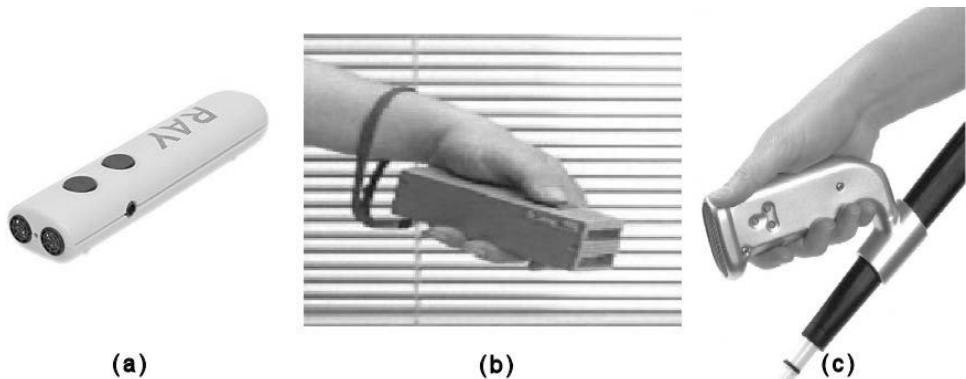
The beamwidths of most ultrasound transducers are diffraction limited, with the result that the larger the transducer diameter the narrower the beam. As a rule of thumb, this relationship can be calculated as

$$\theta_{3dB} = \frac{25000}{df} \quad (7.2)$$

where  $\theta_{3dB}$  (deg) is the half-power beamwidth of the transducer,  $d$  (mm) is the diameter of the transducer, and  $f$  (kHz) is the frequency.

The sizes of the transducers used, and hence their beamwidths, are determined by the application. For example, a unit mounted on the frame of a pair of eyeglasses would have to be much smaller diameter than one that is handheld or attached to a belt, and as a result its beamwidth would be much larger.

**FIGURE 7-10 ■**  
 Photographs of handheld ultrasonic prostheses. (a) Ray electronic mobility aid. (b) Nottingham obstacle detector. (c) K-Sonar.



### 7.5.1 Some Existing Systems

In the past 3 decades a large number of sonar-based visual prostheses have been introduced, with the aim of improving mobility in terms of safety and speed. Some of the more successful devices are discussed in the following sections.

#### 7.5.1.1 Pathsounder

One of the earliest ultrasonic travel aids, the pathsounder consists of two ultrasonic transducers mounted on a board that the user wears around the neck, at chest height. This unit provides only three discrete levels of feedback (series of clicks), coarsely indicating distances to an object.

#### 7.5.1.2 Mowat Sensor and Derivatives

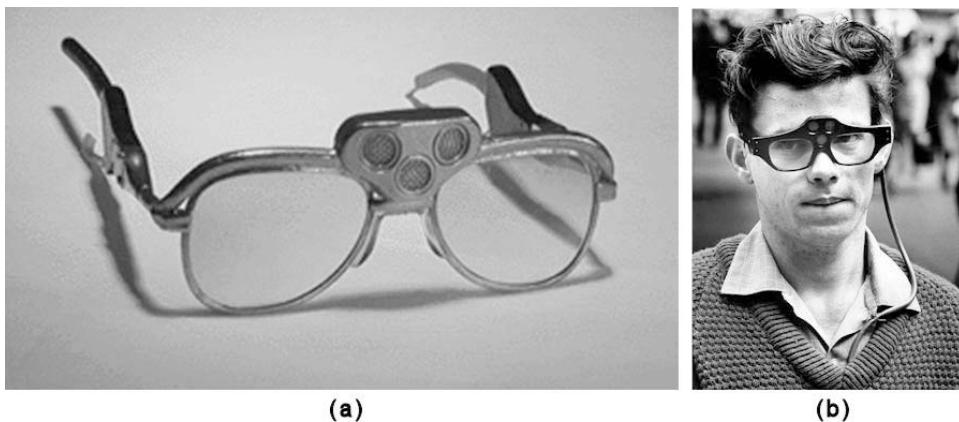
Shown in Figure 7-10 are three of a number of commercially available handheld ultrasonic-based devices that inform the user of the distance to detected objects by means of tactile vibrations or sound. The frequency of the vibration is typically inversely proportional to the distance between the sensor and the object.

The Nottingham obstacle detector (NOD) is an early handheld sonar device that provides an auditory feedback indicating eight discrete levels of distance by different musical tones. The NOD has been commercially available since 1980 (Bissitt and Heyes, 1980).

Other sensors in this genre include the K-Sonar, shown attached to a cane in Figure 7-10. This newer prosthesis uses a wideband signal (one octave), probably based on frequency-modulated continuous wave (FMCW) techniques, to improve resolution and sensitivity. The multiple echoes are converted into “tone-complex” sounds, which the user can learn to interpret for improved obstacle classification (K-Sonar, 2008).

#### 7.5.1.3 Sonic Pathfinder

This is a head-mounted pulse-echo sonar system comprising three receivers and two transmitters, controlled by a microprocessor. Developed in 1984 by Tony Heyes as an improvement to the NOD, it does not give information about surface texture, and normally the auditory display indicates only the nearest object, which is why it should be classified as an obstacle detector rather than as an imaging device. Heyes's approach is rather different from Kay's in that the Sonic Pathfinder deliberately supplies only the most



**FIGURE 7-11** ■ The original binaural sonic aid.  
 (a) Photograph of device mounted in a glasses frame.  
 (b) Device in use.  
 (Courtesy of Zabonne Ltd  
<http://www.zabonne.co.nz/.>)

relevant information needed by the user, whereas Kay strives for more information-rich sonar-based displays.

Heyes argues, for instance, that if some object moves away from users they do not need to know this in regard to safe and efficient travel. In addition, this minimalist approach results in less confusion by minimizing interference with hearing normal environmental sounds (e.g., traffic). The Sonic Pathfinder uses brief tones on a musical scale to denote distance, with pitch descending when approaching an object. Perception of objects on the left and on the right is further supported by tones for the corresponding ear. Two off-center sonar beams are used to detect objects on the left or right to ensure that users are aware of obstacles in their periphery.

#### 7.5.1.4 Binaural Sonic Aid (Sonicguide)

As can be seen in Figure 7-11, this device comes in the form of a pair of spectacle frames, with one ultrasonic wide-beam transmitter mounted between the spectacle lenses and one receiver on each side of the transmitter (Kay, 1974). FMCW signals from the receivers are demodulated and presented separately to the left and right ear. The resulting *interaural amplitude difference* allows the user to determine the direction of an incident echo and thus of an obstacle.

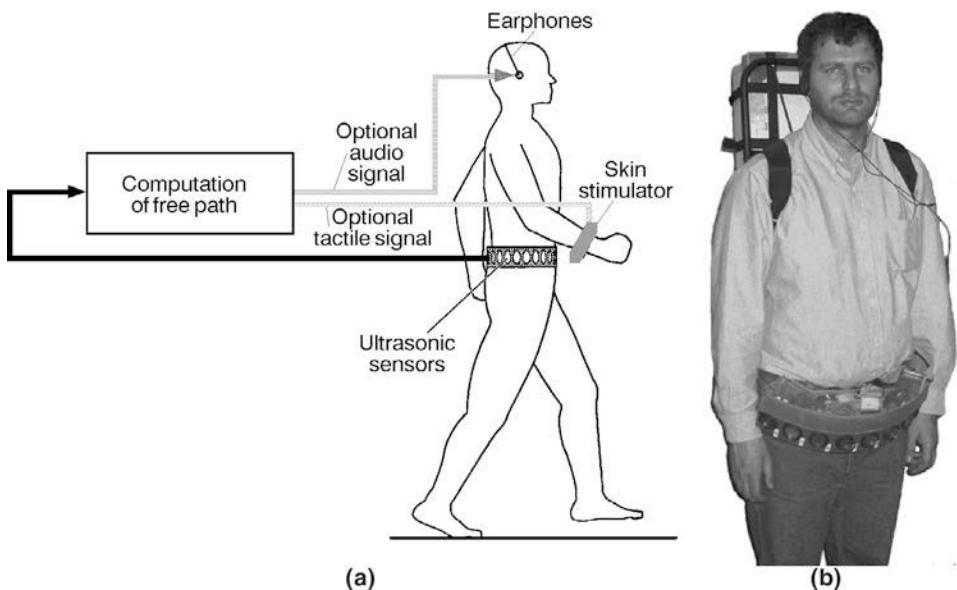
In addition to providing wide field-of-view returns, the newer Trisensor is fitted with a narrow-beam transmitter to provide information about the size and location of objects directly ahead. In both the Sonicguide and the Trisensor, distance to an object is encoded in the frequency of the demodulated low-frequency tone, whereas surface texture is returned by the timbre of the audio signal.

Psychophysical studies conducted at short range (within arm's reach) showed that users were able to judge the direction to a small cylinder to an accuracy of about  $5^\circ$  compared with  $1^\circ$  for a source of natural sound. In contrast, range estimation was improved, with subjects able to judge the distance to an object to an accuracy of 50 mm, which was significantly better than their ability to judge the distance to a natural sound. At longer ranges (up to 5 m), users were able to judge direction to an accuracy of  $6^\circ$  and distance to within 300 mm (Warren and Strelow, 1984).

#### 7.5.1.5 NavBelt

This device was developed as part of a Ph.D. thesis in 1989 and provided imaging and guidance operational modes. In the imaging mode, the device output a crude  $120^\circ$  wide

**FIGURE 7-12 ■**  
**Navbelt.** (a) Graphic showing the operational principle.  
(b) Photograph of the prototype in operation.  
(Borenstein and Ulrich 1997), with permission.



view of obstacles ahead of the user. This was translated into a series of directional (stereo) audio cues that allowed the operator to determine which directions were blocked and which were clear. In the guidance mode, illustrated in Figure 7-12, the NavBelt knows users' positions and their ultimate destination from which its computer would calculate the free path and recommend the direction of travel (Borenstein and Ulrich, 1997).

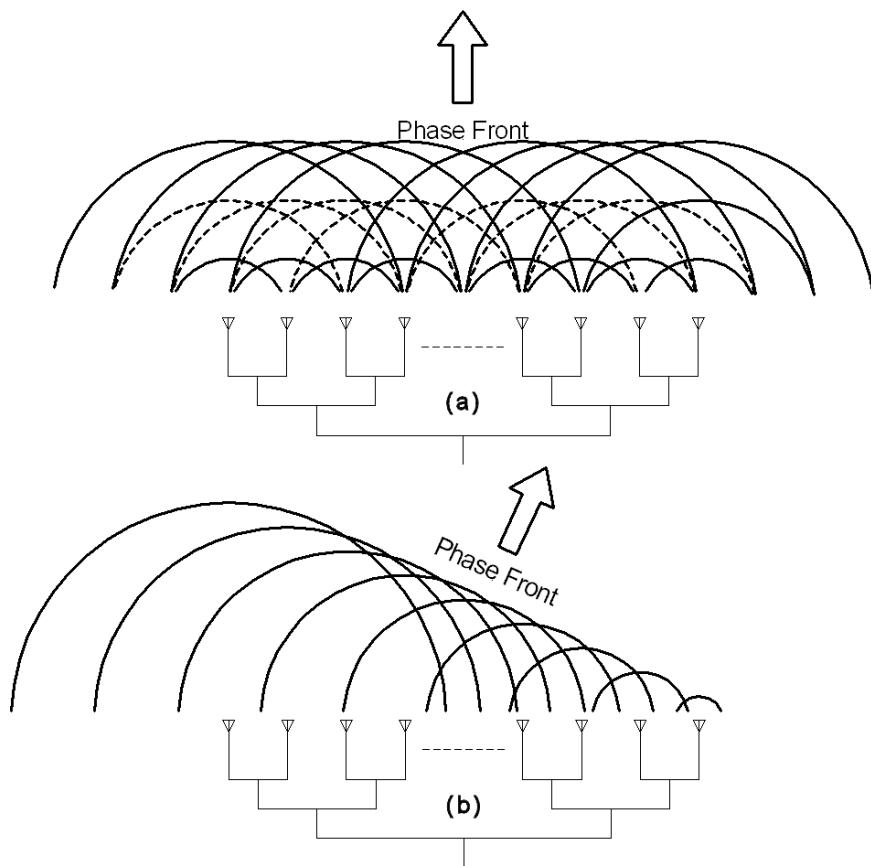
The sonar array in the NavBelt used Polaroid (now SensComp) transducers, each with a beamwidth of  $15^\circ$ , requiring that eight be used to cover the  $120^\circ$  for the situational awareness mode. Each of these sensors outputs the range to a single target within its beam, if there is one within range, to form a crude map of the obstructions around the user.

It is possible to process the phase information returned by individual elements of the array in much more sophisticated ways to improve the angular resolution of the sensor and to direct the beam.

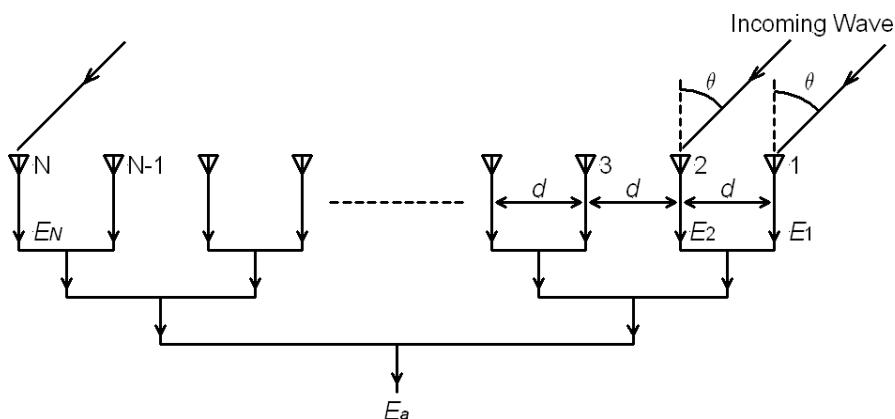
As an introduction to the idea of beam steering, consider the array of ultrasound elements shown in Figure 7-13. In this case, each of the elements is radiating the same frequency signal, with the same phase in (a) but with a different phase—one that increases linearly across the array—in (b). If a line of constant phase is drawn for each of the elements, a linear phase front is produced parallel to the array in the first case and at an angle to the array in the second. This phase front defines the direction in which the beam is synthesized, or the direction from which a signal will be received.

In a conventional linear array, the power received by each element is the sum of the received powers scattered by target,  $P$ , from all the transmit elements. The voltage outputs of all  $N$  elements are summed without delays or phase shifts to give  $E_a$  (V), as shown in Figure 7-14. Since each element observes the same phase,  $\phi_k$  (rad), the output after summing the signals from all  $N$  elements directly with no phasing is given by

$$E_a = \sum_{k=1}^N \sin(\omega t + \phi_k) \quad (7.3)$$



**FIGURE 7-13** ■ Radiation from a phased array. (a) In phase stimulation produces a linear phase front parallel to the array. (b) Linearly increasing phase shifts combine to form a phase front at an angle. (Brooker 2008.)

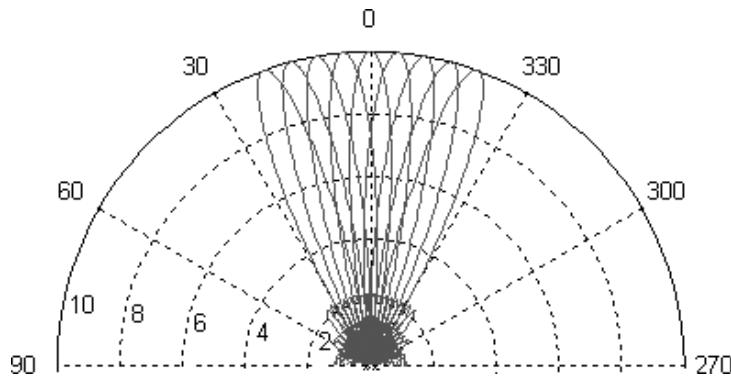


**FIGURE 7-14** ■ The array beam-forming process is achieved by summing the signals from all the array elements in phase. (Brooker 2008.)

It can be shown that the phase shift,  $\psi$ , between elements is constant and proportional to the spacing between the elements,  $d$ , and the sine of the angle of arrival,  $\theta$

$$\psi = \frac{2\pi d}{\lambda} \sin \theta \quad (7.4)$$

**FIGURE 7-15 ■**  
Beam scanning of a 10-element linear array with a 4 mm spacing between elements.



If the relative phase of each of the elements is referred to the center of the array, then the equation for the sum can be written as

$$E_a = \sum_{k=1}^N \sin \left[ \omega t + k\psi - \frac{N+1}{2}\psi \right] = \sin(\omega \cdot t) \frac{\sin(N\psi/2)}{\sin(\psi/2)} \quad (7.5)$$

Substituting for  $\psi$  into equation (7.5) results in the classical equation that defines the voltage output of the one-dimensional (1-D) linear phased array

$$E_a = \sin(\omega t) \cdot \frac{\sin \left\{ \frac{N\pi d}{\lambda} \sin \theta \right\}}{\sin \left\{ \frac{\pi d}{\lambda} \sin \theta \right\}} \quad (7.6)$$

By introducing linearly increasing phase shifts across the array, it is possible to steer the main lobe of the beam over a reasonably wide sector while still maintaining a reasonably narrow beam and low sidelobes, as shown in Figure 7-15.

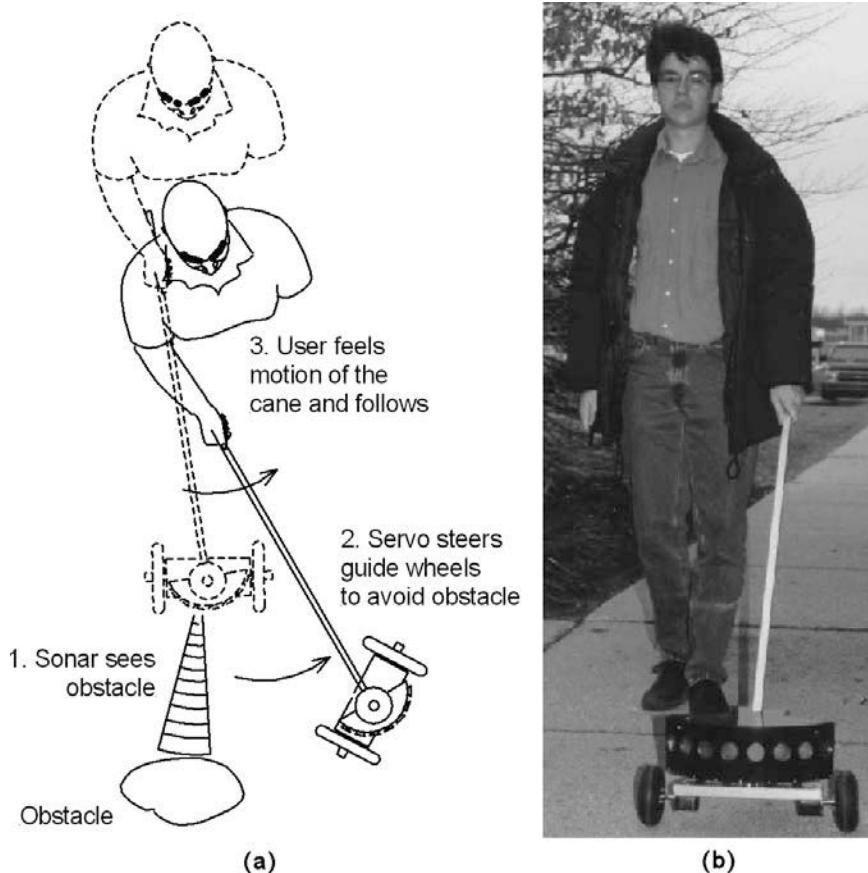
In reality, the beam pattern of the array is modified by the shape of the beam of the individual elements. For that reason small elements with wide beamwidths, rather than the large elements shown in Figure 7-12, are used. This also allows the elements to be packed close together, typically  $< \lambda/2$  apart; otherwise, unwanted grating lobes appear on either side of the main lobe (Brooker, 2008).

### 7.5.1.6 GuideCane

This fairly recently developed device was designed at the University of Michigan to help blind people navigate safely and quickly (Borenstein and Ulrich, 1997). It consists of an array of ultrasonic sensors on a wheeled robotic vehicle that can navigate through a cluttered environment. The operational principles are shown in Figure 7-16.

The GuideCane is used much like a white cane, though it is heavier and rolls on a pair of wheels. A computer-controlled steering servo directs the wheels with respect to the cane. An array of ultrasound sonar transducers mounted on a circular arc senses the environment to the front and sides. In addition, a flux gate compass and wheel encoders provide information about the direction and distance traveled. Finally, a miniature, thumb-controlled joystick is positioned on the handle of the cane.

The joystick commands the initial direction of travel, either straight ahead or to the left or right from the current direction; the steering servo will turn as the GuideCane is pushed forward until it has reached that direction, after which it will follow a straight path. The sonar array images over the 120° sector ahead looking for obstacles as it moves



**FIGURE 7-16 ■**  
Guide cane prototype  
(a) Schematic of the operational principles.  
(b) Photograph of the prototype in operation. [Adapted from (Borenstein and Ulrich 1997) with permission.]

forward. This obstacle map is then used by the onboard computer to calculate the best direction to travel, even through a cluttered environment. If an obstacle is detected in the path, the steering servo drives the angle of the GuideCane wheels in a direction to avoid it. The operator feels the changed resistance on the cane even before the direction change is detected and intuitively follows the direction suggested with almost no hesitation.

The GuideCane solves the problem of stairs in two ways. When a drop-off is reached, the wheels drop off the edge providing a fail-safe indication of the fall. In the other direction, the main forward-looking array sees the bottom step as an obstacle, but the forward-looking up-facing sensor measures a longer range. If the difference between the two ranges is about 300 mm, the object is treated as a flight of stairs. However, if the two distances are almost the same, then the object is treated as a wall and is avoided.

GuideCane incorporates two additional functions: (1) a side-looking sonar for wall following; and (2) a global positioning system (GPS) that can be used for navigation of preprogrammed routes. The latter are generated automatically during a “lead-through” run under the guidance of a sighted person.

For indoor navigation, dead reckoning based on compass readings and wheel-sensor odometry can be used for short distances, but the integrated error grows too large to achieve the required accuracy for distances farther than a few meters. However, algorithms such as simultaneous localization and mapping (Leonard and Durrant-Whyte, 1991), a well-known algorithm used by the robotics fraternity, could be implemented to solve this problem.

### 7.5.2 Issues with Sonar-Based Systems

No visual prostheses are without problems, and all of those here based on sonar discussed have some shortcomings:

- With most of the systems, the user must actively scan the environment to detect obstacles (no scanning is needed with the Sonicguide, but that device doesn't detect obstacles at floor level). This procedure is time-consuming and requires the user's constant activity and conscious effort.
- The user must perform additional measurements when an obstacle is detected to determine the dimensions of the object. A path must then be planned around the obstacle. Again, this is a time-consuming, conscious effort that reduces walking speed.
- One problem with all blind aids based on acoustic feedback is their interference (called *masking*) with the blind person's ability to pick up environmental cues through hearing (Brabyn, 1982; Kay, 1974; Lebedev and Sheiman, 1980). This is a serious concern, as almost all blind people use acoustic information for situational awareness, obstacle detection, and navigation.
- The GuideCane is large and rather unwieldy; it must be lifted up steps and roadside curbs. The wheels can easily wedge in cracks and holes in the pavement.

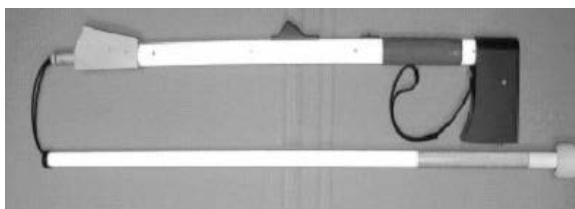
## 7.6 | LASER-BASED SYSTEMS

Introduced in 1973 (Benjamin, Ali et al., 1973), the C-5 laser cane, shown in Figure 7-17, is based on optical triangulation with three laser diodes and three photodiodes as receivers. It can detect obstacles at head height, drop-offs in front of the user, and obstacles up to a range of 1.5 or 3.5 m ahead of the user.

## 7.7 | SENSORY SUBSTITUTION

Sensory substitution is defined by Peter Meijer (2006) as “the replacement of one sensory input (vision, hearing, touch, taste or smell) by another, while preserving some of the key functions of the original sense.” Due to the astonishing plasticity of the brain, sensory substitution devices can help the blind regain some sense of sight through touch or hearing or help a deaf person regain some sense of hearing through vision or touch. It has been shown experimentally that if the brain is deprived of its main form of input, even for quite short periods, it will seek out other ways of gaining this information. Sensory substitution devices are designed to facilitate this process.

**FIGURE 7-17 ■**  
Prototype of the C-5  
laser cane.  
(Benjamin, Ali et al.,  
1973)



In regard to vision replacement, direct cortical stimulation would seem to be the best interface as it does not interfere with other senses. However, the huge interfacing problems and invasive nature of the approach, together with the low-resolution results obtained until now, probably make it a long-term research item.

As an alternative, when considering interference a vibrotactile or electrotactile skin-stimulating system would seem to be a good choice because much of the skin plays only a subordinate role as a communication channel under normal conditions. Unfortunately, that same consideration might indirectly be its major disadvantage, because there is no strong evolutionary reason to expect the touch sensation to provide a high-bandwidth communication channel. This restriction applies not only to the skin interface but also all the way up to the cognitive brain centers.

Experiments in the 1960s conducted by Paul Bach-Y-Rita using a  $20 \times 20$  grid of 1 mm electrotactile stimulators attached to the subject's back and activated by a video camera showed that if the subjects had control of the camera's pan-and-zoom facility they were soon able to discern lines and edges (Bach-y-Rita 1969). With some practice they could even recognize depth, shadows, and shapes and even individual faces. This is an amazing result given the poor resolution of both the camera and the electrotactile stimulator. More importantly, the initial tactile sensations were soon forgotten, and perceived objects appeared to exist in three-dimensional space. However, this heightened perception capability occurred only if subjects were moving the camera themselves, and measurements made using static cameras failed to produce any real perception. It was as if the perception was founded in the sensory-motor loop as a whole and not in the static transfer of an image to the electrotactile array.

It has been argued that reasonable resolution can be achieved by stimulating small areas of skin but that the limited sensory bandwidth available would lead to severe information loss when stimulating a large matrix of skin positions. Experiments have yet to show that this is always the case, but it does suggest that using tactile substitution may not be the ultimate path to high-fidelity sensory substitution.

Knowing the importance of bandwidth in communicating data—in this case detailed environmental information—an alternative would be to exploit the capabilities of the human hearing system. Although it cannot be claimed that this would produce a perfect solution, it is known that the human hearing system is capable of processing and interpreting extremely complicated and rapidly changing acoustic patterns, such as speech or music in a noisy environment. The available effective bandwidth, of the order of 10 kHz, corresponds to a channel capacity of many thousands of bits per second.

As with the tactile option, restrictions may be imposed by the mechanics of the cochlea or with information encoding in the neural architecture. However, in spite of these uncertainties, the known capabilities of the human hearing system in learning and understanding complicated acoustical patterns provide a strong motivation for developing auditory sensory substitution systems.

The following sections address some of the issues involved with both tactile and auditory systems.

### 7.7.1 Auditory Substitution

One of the major problems with auditory sensory substitution is the masking effect, in which sound cues normally used by a blind person are swamped by feedback from the prosthesis. Under laboratory conditions, the interference of an auditory prosthesis with

normal hearing is probably not an issue, and it is possible that it can be overcome outdoors provided the system does not block the hearing system completely. If the system blocks only subtle clues about the environment, normally perceived and used only by blind persons, but replaces them with much more accurate and reliable feedback, it may be an acceptable trade-off.

### 7.7.1.1 Input from Sonar

As the resolution of sonar prosthetics has improved, so has the requirement for higher and higher bandwidth perception. This detailed sensory environment has been championed by Professor Leslie Kay, starting with the Sonicguide in 1974. More recently, he has improved the angular resolution of his binaural sensory aid (BSA) system using a third narrow-beam transmitter for creating an additional monaural signal. This newer system was first called the Trisensor but is now known as the KASPA system (Kay's Advanced Spatial Perception Aid), shown in Figure 7-18. In 1998 this system won Kay the Saatchi and Saatchi Innovation in Communication Award.

**FIGURE 7-18** ■  
Photograph of one  
of the KASPA  
prototypes.  
(Courtesy of  
Zabonne Ltd  
<http://www.zabonne.co.nz/>) with  
permission



Kay is an advocate for supplying blind users with a rich sensory environment, and KASPA is an example of such a prosthesis. It represents object distance by pitch and surface texture through timbre (the sum of a number of harmonically related sounds). To achieve this, use is made of the high-resolution capability of wide bandwidth FMCW signals. The improved, but still modest, resolution positions Kay's system somewhere between obstacle detection and environmental imaging. The best angular resolution is about one degree in the horizontal plane (azimuth detection) for the central beam, which is quite good, but vertical resolution (elevation) is poor. This makes the "view" akin to using binoculars to view the world through a narrow horizontal slit.

It is difficult to settle on a good measure for equivalent image resolution as offered by sonar approaches, because sonar is not just based on mapping a two-dimensional (2-D) projected image to another sensory modality. It can be used to provide a much richer representation of the environment based on ultrasonic interference and acoustic delay patterns. In this regard, Kay's work does demonstrate some features (resolution, texture, parallax) that support its classification as an imaging device for vision substitution.

### 7.7.1.2 Input from Cameras

Concerning the input from the environment, true visual, camera-based input has some major advantages over the use of sonar echo patterns. The possibilities of sonar have been investigated rather extensively, and its short operational range makes it impossible to perceive distant objects that are useful for guidance. Furthermore, visual input matches the most important information providers in our surroundings, such as road signs, books, and television. From a technical perspective, it is difficult to obtain unambiguous high-resolution information using a scanning sonar, whereas any commercially available low-cost camera will perform the function well.

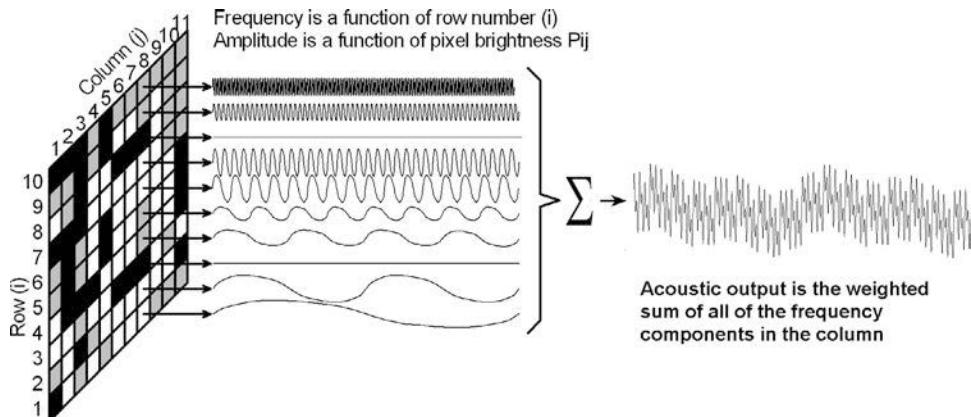
It should be noted that this holds only if the camera input is supplemented by depth clues through changes in perspective to resolve ambiguities in distance. Relative depth information can be obtained from changing relative positions of the viewer and the environment in combination with some knowledge of the sizes of objects in the scene. An equivalent of binocular vision is no strict prerequisite, but there might be advantages in presenting the views from the individual eyes as sound patterns to their associated ears.

Some research groups have tried to simulate vision substitution as accurately as possible. For example, Capelle (Capelle, Trullemans et al., 1998) represents the center of the visual field with more pixels, just as it is in the eye, with each pixel being represented by a particular frequency. Because this system cannot present the visual world continuously, adjacent pixels cannot always be represented by neighboring sinusoids, which makes the representation counterintuitive and potentially more difficult to interpret than a linear mapping.

A different approach was followed by Meijer (2006). To increase the image resolution obtainable via an auditory representation and to decrease the required bandwidth, a time-multiplexed mapping is performed to distribute an image in time. The 2-D spatial brightness map of a visual image is scanned and transformed into a 2-D frequency and time map. It is thought that the subdivision of only one of the two spatial dimensions of the visual image into individual scan lines will be more amenable to a later mental integration than the subdivision in both dimensions into individual pixels.

The human brain is far superior to most, if not all, existing computer systems in rapidly extracting relevant information from blurred and noisy images. Therefore, no attempt is made by Meijer to filter any redundancy out of visual images that are being transformed into

**FIGURE 7-19** ■ Conversion of a  $10 \times 11$  gray-scale image to sound. [Adapted from (Meijer 2006).]



sound patterns. From a theoretical perspective, this means that the available bandwidth is not exploited optimally. However, by keeping the mapping as direct and simple as possible the risk of accidentally filtering out important clues is reduced.

The principles used for spatial encoding of the visual signal to a sound sequence are shown graphically in Figure 7-19 for a gray-scale image. Image processing has reduced the spatial resolution to only 110 pixels ( $10 \times 11$ ) and the intensity to only three gray-scale levels ( $M = 10, N = 11, G = 3$ ). The mapping translates the vertical position of each pixel into a frequency proportional to its position in the column with the amplitude proportional to the pixel brightness. The horizontal position maps to a delay after the start time (denoted by a click) or to an amplitude difference in a stereo system.

As the complete audio signal from each column is output simultaneously, and there are  $N$  columns, the sound is available for  $T/N$  seconds for a total scene update rate of  $T$  seconds. For each column,  $j$ , every pixel is used to excite an associated sinusoidal oscillator within the audible band, with the lowest frequencies at the bottom and the highest at the top. In addition, the oscillators form an orthogonal basis in which they are all integer multiples of some reference frequency. This ensures that all of the information is preserved in this transformation from geometrical to Hilbert space (Meijer, 2006).

The  $M$  oscillator signals from column  $j$  are superimposed with the corresponding sound pattern presented to the ear for  $T/N$  seconds before the sound from column  $j + 1$  is output. This continues until the  $N$ -th column has been converted after  $T$  seconds, when the whole cycle is repeated for a new image. Images are separated by a synchronization click, which is essential to ensure that the subject can reorientate laterally. In addition, the relative amplitude of the sounds differs in the two ears to produce a stereo effect, so that the sound appears to come from the correct direction.

As shown in Chapter 5, the spectrum of a sinusoidal signal with duration  $T/N$  seconds includes other frequency components. These can be determined by convolving the Fourier transform of the rectangular window with that of the sinusoidal signal. One important consideration is that even the lowest frequency is represented by a reasonable number of complete cycles during the observation time; otherwise, it would not be interpreted correctly by the ear.

This processing strategy has been implemented in a sensory substitution prosthesis called vOICe, with which, according to Meijer, it is possible to learn to sense instinctively how the features of a soundscape correspond to objects in the physical world. Pat Fletcher, a proficient user of the vOICe who could see until age 21, describes the sound images



**FIGURE 7-20 ■**  
New generation of the vOICe camera mounted in a glasses frame.  
(Meijer 2006)

in her head as “ghostly” but real. At a meeting of the Cognitive Neuroscience Society in New York, researchers from Harvard Medical School announced that when they viewed the activity in the brains of two vOICe users (one blind at birth, the other who went blind later in life), it was in many respects like that of a sighted person while seeing. However, in tests undertaken by Alex Storer (2006), though his subjects’ identification capabilities improved using vOICe for long periods none felt that the experience was visual in nature.

For obvious reasons, not everyone has the inclination to walk around with a head-mounted camera and a laptop, so Meijer has modified his setup to work using a camera phone. Now, after downloading a simplified version of the software (<http://www.seeingwithsound.com>), practically anyone can point her camera phone at what she wants to see and have a listen to what it looks like.

An alternative for more serious users includes a netbook processor, stereo ear buds, and the installation of the camera into the bridge of a pair of fashion sunglasses shown in Figure 7-20.

So far, the research community has not yet found conclusive answers with respect to the potential of this approach for the blind, and the relevant limitations in human auditory perception and learning abilities for comprehension and development of visual interpretation remain largely anecdotal. In addition, the training effort is expected to be significant while involving perceptual recalibration for accurate sensorimotor feedback. The U.S. National Science Foundation is now funding the first controlled study to look at the benefits of the system while simultaneously attempting to find an optimal training protocol (Trivedi, 2010).

One of the key research questions is to what extent the use of a sensory substitution system can not only provide synthetic vision in a functional sense for extended situational awareness through active sensing but also lead to visual sensations through forms of induced artificial synesthesia.

Other researchers including Stefan Strahl of the UCL Ear Institute in London and Lucy Irving of the University of Brighton have been examining similar aspects of sensory substitution using the vOICe system, with most of their results available on the Web.

The latest functional magnetic resonance imaging (fMRI) results show that in novice users the auditory cortex is involved in the interpretation of the soundscape but, after 10–15 hours of training, activity in the visual cortex and an area known as the occipital tactile-visual (LOtv) region increases. Recent experiments on advanced users including Fletcher involved the use of repetitive transcranial magnetic stimulation (rTMS) to close down the LOtv region. When this occurred they lost the ability to “see” and became confused and frightened (Trivedi, 2010).

### 7.7.2 Electrotactile and Vibrotactile Transducers

Electrotactile stimulation, sometimes called electrocutaneous stimulation, evokes tactile sensations within the skin at the location of the electrode by passing a local electric current through it. In contrast, vibrotactile stimulation evokes tactile sensations using mechanical vibration of the skin, typically at frequencies between 10 Hz and 500 Hz (Kaczmarek, Webster et al., 1991).

Because of the relative simplicity of the technology required, many systems, both commercial and experimental, have been developed over the past half-century to perform these functions. They range from single-element devices to lines of transducers to two-dimensional arrays.

Single stimulation points can convey information through the skin using variations of intensity/amplitude, frequency, or both. The source of information could be temporally varying, such as a simple auditory prosthesis that translates the envelope of inputs from a microphone directly into vibration intensity.

One-dimensional (1-D) arrays consisting of two or more actuators in a line present spatial information more naturally. A good example of this is feedback from a prosthetic arm in which the elbow angle is mapped to the excitation of a linear array mounted elsewhere on the body. From a visual substitution perspective a 1-D array can map the outputs of an array of sonar sensors for devices such as the NavBelt.

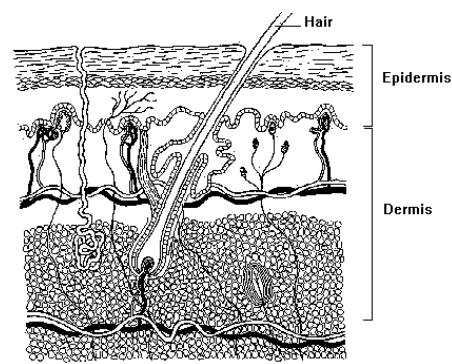
2-D displays, also referred to as tactile vision substitution (TVS) devices, can display spatial information to the skin in a manner similar to the way the lens of the eye presents spatial information to the retina. At its most primitive, each pixel from a camera is mapped directly to an individual vibrotactile or electrotactile element, which is excited in proportion to the pixel intensity. Many researchers including Bach-y-Rita developed systems in the 1970s to study the skin's ability to interpret visual information. They found that subjects could immediately recognize vertical, horizontal, and diagonal lines and that with experience users could recognize common objects and even people's faces (Kaczmarek, Webster et al., 1991).

The commercially available Optacon (optical to tactile converter) and its more modern derivatives convert the outlines of written letters and simple graphics to a vibrotactile array on the user's fingertip. Exceptional users can read ordinary printed text at up to 90 words per minute using such devices.

#### 7.7.2.1 The Skin

Six types of tactile receptors (shown in Figure 7-21) have been identified and characterized in the skin, though all types are not available everywhere across the body with some present beneath hairy surfaces and some in hairless (glabrous) regions. These can be characterized by their response to a step change in the applied pressure to the skin. The response is measured by the change in the receptor's action potential firing rate.

Merkel cells innervate the human fingertip at a density of 100 elements/cm<sup>2</sup> and are mainly responsible for the detection and identification of spatial patterns. Meissner corpuscles are even more densely packed in the human fingertip (150 elements/cm<sup>2</sup>) and are highly sensitive to dynamic skin deformation over a wide receptive field. These corpuscles poorly resolve spatial information but encode for skin motion by detecting low-frequency vibrations indicative of object slip. Pacinian corpuscles are fast-acting receptors with a broad receptive field, are extremely sensitive to the smallest skin motion (in nanometers), and are responsible for the perception of high-frequency stimuli. Their



**FIGURE 7-21 ■**  
Anatomy of the skin and tactile receptors. [Adapted from (Brodel 1969).]

receptor endings	function	location	receptor endings	function	location
	Responds to hair displacement.	Wraps around hair follicle in, of course, hairy skin.		Responds to vibration. Most sensitive in 150–300 Hz range	Deep layers of dermis in both hairy and glabrous skin.
	Responds to pressure on skin.	Dermis of both hairy and glabrous skin.		Responds to vibration. Most sensitive in 20–40 Hz range	Dermis of glabrous skin.
	Responds to pressure.	Lips, tongue, and genitals.		Different types of free nerve endings that respond to mechanical, thermal or noxious stimulation.	Various types are found throughout the skin.

peak sensitivity occurs at stimulation frequencies of 200 to 300 Hz. Ruffini endings have an unknown function and have never been found in the skin of the fingertips (Pasquero, 2006).

Temperature is sensed by two types of free nerve endings, with one type measuring heat and the other type measuring cold. For temperatures above 45 °C and below 10 °C, temperature also stimulates pain endings.

Tests with fine wires and small vibrators have been used to quantify the sensitivity of different areas of skin. The static force required (Weinstein, 1968) varies from 0.05 g on the lips to 0.63 g on the fingertips and belly up to 3.5 g on the sole of the foot. The fingertip threshold corresponds to a skin indentation of about 10 µm.

In regard to vibration, Geldard (1957) used a 1 cm<sup>2</sup> vibrator and found that the fingertips were the most sensitive of all the body locations by at least an order of magnitude. The abdomen was found to be 60 times less sensitive than the fingertips for 200 Hz vibrations.

**TABLE 7-1** ■ Static Simultaneous Two-Point Discrimination Thresholds

Location	Static touch (mm)	Vibrotactile (mm)	Electrotactile (mm)
Fingertip	3	2	<7
Palm	10	—	8
Forehead	17	—	—
Abdomen	36	—	10
Forearm	38	—	9
Back	39	11–18	5–10
Thigh	43	—	10
Upper arm	44	—	9
Calf	46	—	9

Source: Kaczmarek, K. A., J. G. Webster et al., *IEEE Transactions on Biomedical Engineering* 38(1): 1, 1991, with permission.

Threshold amplitude for vibrotactile stimulation also increases after a strong conditioning stimulus, with the reduction in sensitivity being proportional to the amplitude of the conditioning stimulus and its duration.

For vibrotactile and electrotactile displays, the spatial resolution of the skin is extremely important. The basic test is the two-point discrimination threshold (TPDT) and is summarized in Table 7-1. However, other methods to determine tactile spatial resolution include the determination of the minimum width of a deep groove, which can be detected on an otherwise smooth surface. This turns out to be 0.87 mm on the fingertip, much smaller than TPDT. The orientation of grooves in a grating is another test that shows that the fingertip is much more sensitive than expected. In this case the resolution is 0.84 mm.

The skin can also recognize frictionless position shift of a stimulus 10 times smaller than the TPDT can. This indicates that the skin's spatial resolution is much better for some tasks than the TPDT suggests. A good example is the ability of the fingertips to distinguish between different grades of sandpaper with very fine spatial frequencies (Kaczmarek, Webster et al., 1991).

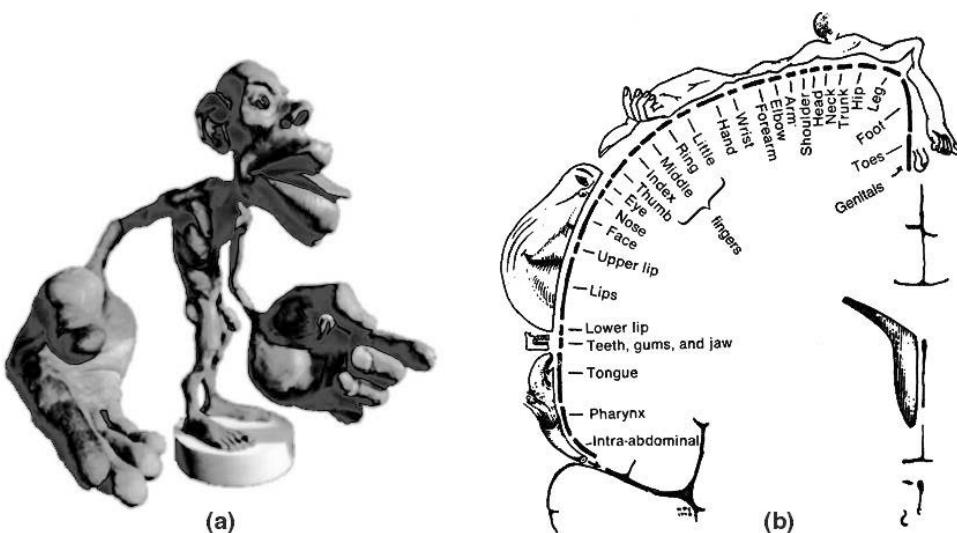
The dynamic range of vibrotactile stimulation for a  $0.78 \text{ mm}^2$  stimulator is about 40 dB if the maximum amplitude of the stimulation is limited to 0.5 mm.

### 7.7.2.2 Vibrotactile Displays

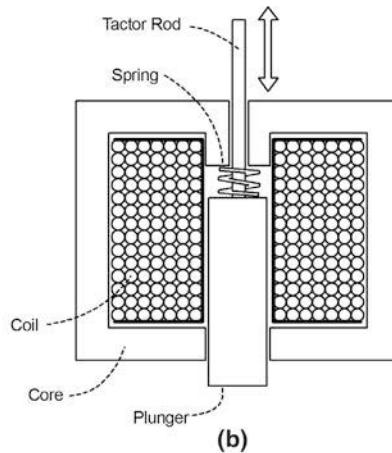
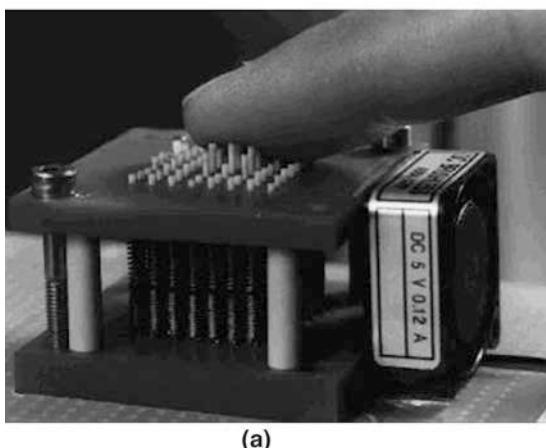
Most attempts at building vibrotactile displays have focused on devices that stimulate the fingertips because of their high tactile acuity. Some researchers have investigated other regions including the tongue and the lips, which are also sensitive, or even the torso and the thighs, which offer the advantage of a large available area. The relative proportion of the sensory cortex associated with these sections is graphically illustrated in Figure 7-22.

Most vibrotactile arrays operate by indenting the skin with arrays of pins that rise out of a surface to create a discrete representation of a texture or pattern, as shown in Figure 7-23.

Other techniques include using vibration, heat and cold, blowing compressed air, or changing shape. Actuators themselves include shape memory alloy (SMA), PZT materials, motors and solenoids, pneumatic valves, rheological fluids, and pistons. Some of these are considered, along with the relevant research reference, in Figure 7-24 (see Pasquero, 2006, for details of individual references).



**FIGURE 7-22** ■ Representations of the sensory cortex as a function of its association with parts of the body  
 (a) Sensory homunculus model.  
 (b) Graphical model.

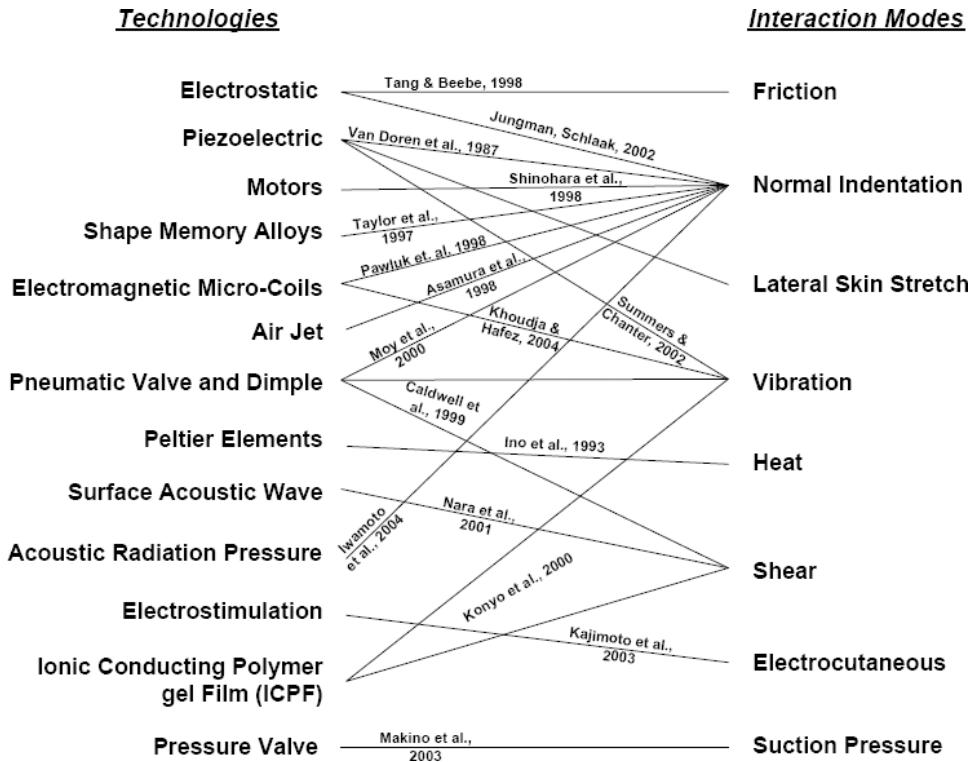


**FIGURE 7-23** ■ Conventional vibrotactile display.  
 (a) Photograph.  
 (b) Diagram of a single solenoid based tactor.  
 (Courtesy of Forschungszentrum Karlsruhe.)

Unfortunately, most of the tactile displays built to date fail to convey meaningful tactile information and to be practical at the same time. The devices are too big; they do not yield enough force or are constrained too low bandwidth; they are limited to a small number of actuators with low-spatial density; they require constant maintenance or are simply too expensive and too complex. These failures are reflected by the relative scarcity of tactile displays that have made it to the commercial world.

Frequency of vibration is a parameter that influences the quality and intensity of perception, with the sensitivity affected by factors like body position, skin temperature, and underlying tissue type (bone, fat, muscle, or a combination). Values between 50 and 300 Hz are generally used as being in the range of frequencies where the tactile corpuscles are most sensitive. Oscillating pressure also adds new degrees of freedom to the design of vibrotactile stimuli. These include waveform shape, for example sinusoidal or square, and amplitude modulations (at different modulation frequencies) of the carrier frequency.

**FIGURE 7-24 ■**  
Tactile display technologies and interaction. [Adapted from (Pasquero 2006)], with permission.

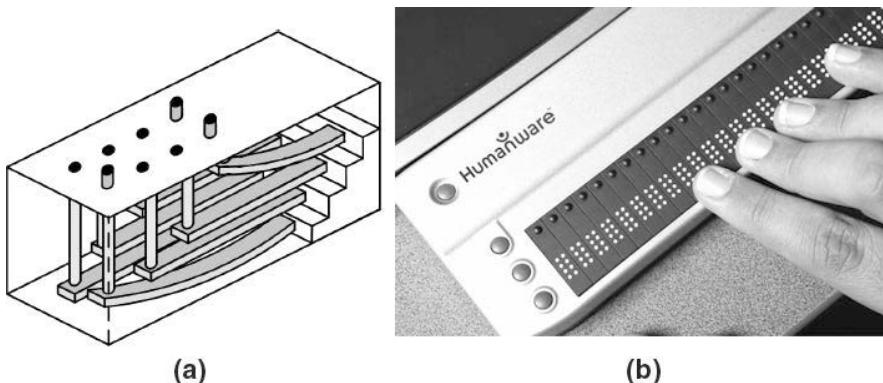


Several features of vibrotactile stimuli can be modulated to convey information over this sensory channel (Cincotti, Kauhanen et al., 2007). The list can be divided into two subsets. The first includes features related to physical perception and are as follows:

- *Frequency*: the main spectral component of the periodic stimulus
- *Intensity*: the amplitude of stimulation (measured either as force applied or as displacement produced)
- *Timbre*: the complexity of the stimulation waveform (i.e., the content of harmonics in the spectral representation)
- *Duty cycle*: the ratio of the durations of the on and off time on an elementary stimulation
- *Spatial location*: the body part or the pattern of parts that are stimulated

Features in the second subset are clearly perceived by an individual but do not rely on any specific property of the receptors. They need to be interpreted on a cognitive level:

- *Rhythm*: the sequences of stimulation and pauses, with specific durations, that compose the current message (i.e., a triplet of stimuli such as a Morse coded SOS)
- *Tempo*: the speed, due to longer or shorter duration of the whole message, given at fixed rhythm
- *Flutter*: an amplitude modulation of the stimulation carrier frequency that can either be perceived as increase and decrease of the intensity (if modulation is slower than 5 Hz) or as “roughness” (if modulation is faster than 10 Hz)



**FIGURE 7-25** ■ Refreshable Braille display showing (a) Diagram of a single cell. (b) Photograph of a complete array made from a number of cells. (Courtesy of HumanWare <http://www.humanware.com/>.) with permission.

Braille, one of the earliest tactile display methods, was invented more than 175 years ago as a means of giving blind people access to the written word. This was originally in the form of a  $2 \times 3$  array of embossed dots on paper, but nowadays it is often displayed on computer peripherals called refreshable Braille displays that use  $2 \times 4$  arrays. These displays have not changed significantly over the past 30 years with typical systems using cantilevered bimorph piezoactuators (reeds) supporting vertical pins at their free end, as shown in Figure 7-25. On actuation, the selected reeds bend upward, pushing the pins above the surrounding surface to generate the selected Braille character.

The individual cells as illustrated in the figure are simple and inexpensive, but the overall cost to generate a line of Braille using 40 to 80 cells becomes significant (Levesque, Pasquero et al., 2005).

As an extension to the Braille units shown previously, one of the most successful applications of vibrotactile displays is the Optacon (OPtical TActile CONverter) manufactured by Telesensory Systems Inc., shown in Figure 7-26. This device consists of a main electronics unit about the size of a large paperback connected by a cable to a small



**FIGURE 7-26** ■ Optacon with the vibrotactile array as an inset. (Courtesy of Wikipedia.)

camera module. The camera module consists of 144 phototransistors arranged in a  $24 \times 6$  configuration along with two small rollers for easy movement and two small lights for illumination. The electronics unit includes a vibrotactile display onto which the user places his finger and then tracks along a line on a printed page with the camera module, which feeds the letter-sized images back to the display. The tactile array contains a  $24 \times 6$  array of 0.08 mm diameter pins, each of which can be independently vibrated at about 230 Hz by a PZT bimorph reed. As the user moves the camera, tactile images of print letters are felt moving across the array of rods under his finger (Efron, 1977). Reading speed is limited to about 100 wpm compared with 250 wpm or more using Braille. But that was not the point, the Optacon offers access to innumerable sources of the written word not previously available to the blind—labels on jars, CD covers, computer screens (with a special lens attachment), ordinary newspapers, and a host more.

A number of derivative products using similar principles have been developed to extend the capability of these vibrotactile systems because they have proven to be so useful to blind people. Devices like the videoTIM have a much larger tactile display and are therefore more amenable to representing simple graphics, handwritten text, and formulas. They are also able to read keys and messages on mobile phone and computer displays if optical character recognition (OCR) software is not available.

Vibrotactile research at Johns Hopkins University has included the development of a 400-element array for fingertip stimulation. This consists of four planes of actuators slightly staggered so that the pins can all pass through a uniform array of holes in a fingertip sized plate. This large number of actuators in a small area is required for research as it is of the same order as the number of tactile corpuscles on the fingertip.

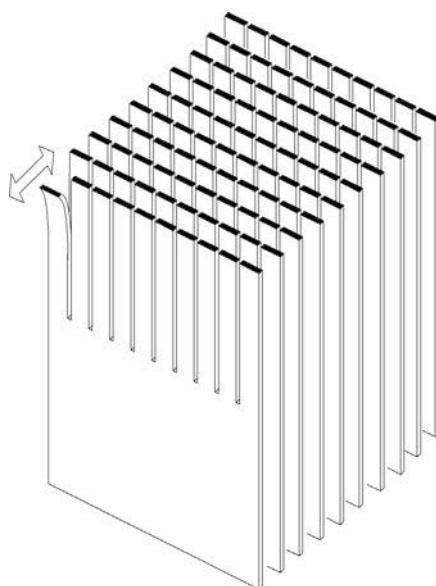
The McGill Haptic lab has been working on the development of an up-to-now unexplored mode of interaction for tactile displays. Their researchers believe that lateral skin stretch is sufficient to give the impression of skin indentation. The objective is therefore to create a device capable of conveying meaningful tactile sensations by lateral skin stretch at the fingertip. The project is based on the simple realization that it might be possible to compensate for both our limited knowledge of touch and the current shortcomings of haptic technology by making use of tactile illusions that are less complex to implement with state-of-the-art actuators.

This research has led to the development of a tactile display called the Stimulator of Tactile Receptors by Skin Stretch (STReSS). Various prototypes of the device have been built based on the principles shown in Figure 7-27 (Pasquero, 2003), and these show potential.

STReSS is a computer-peripheral device that stimulates the fingertip by lateral skin stretch. It is composed of a miniature array of 100 (prototype #1) or 50 (prototype #2) bimorph PZT actuators that induce time-varying programmable strain fields at the fingertip to convey tactile information such as texture or small-scale shape. Tactile signals are generated on a personal computer and then fed to the display via a universal serial bus (USB) port.

Vibrotactile devices are also used for haptic feedback on hand grips, as can be seen in Figure 7-28. Haptic devices are often used in teleremote control applications where visual and audio feedback is insufficient to generate the required immersion. These are discussed in more detail in a later chapter.

From a visual prosthetic perspective, such devices can replace the handle of a cane where they become sensory substitutes and convey visual information to the palm of the hand.



**FIGURE 7-27 ■**  
McGill University  
STReSS tactile  
display. [Adapted  
from (Pasquero  
2003)], with  
permission.



**FIGURE 7-28 ■**  
Vibrotactile haptic  
device. (Becker,  
Jang et al., 2009.)

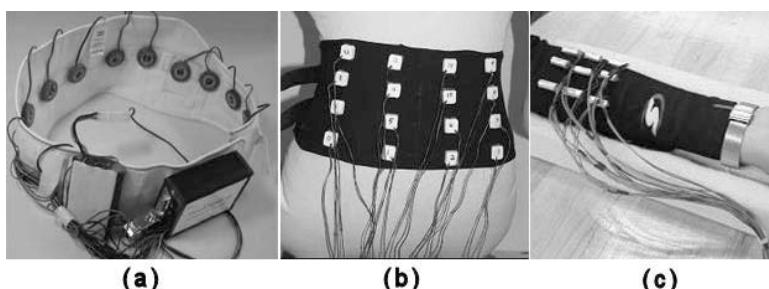
Larger tactile devices are also made for the torso or limbs where the density of sensory corpuscles is much lower. A number of these devices are shown in Figure 7-29. They are often made from micromotors driving an eccentric mass, similar to those commonly found in portable consumer equipment like phones and PDAs. They are therefore very low cost. The vibration frequency is proportional to the applied voltage, and can therefore be controlled. Unfortunately, the response time is poor.

Alternative designs are based on the construction methods for earphones where the diaphragm is replaced by a spring support with a central contactor pin.

### 7.7.2.3 Construction of Vibrotactile Devices

A trade-off is required between the minimum force, and hence displacement, of the pins of a vibrotactile device to ensure that each one registers even after desensitization and a

**FIGURE 7-29 ■**  
**Vibrotactile devices.**  
 (a) Linear array around the waist.  
 (b) 2D array across the back. (c) 2D array around the forearm.



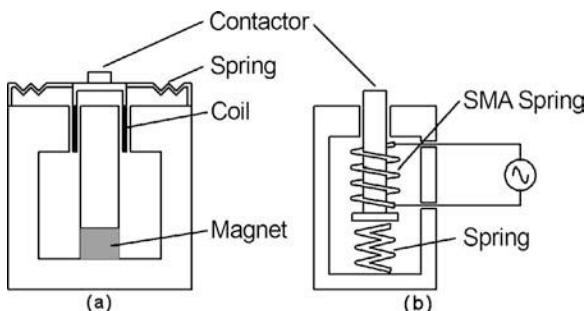
force that results in discomfort or even damage to the skin. This trade-off will vary with the excitation frequency and differ for various parts of the body. For a typical device designed to stimulate Meissner corpuscles in the fingertip, the pin spacing need not exceed the spatial discrimination sensitivity of these corpuscles, which is about 3 mm (Table 7-1). Because they are situated about 0.7 mm below the skin surface, a vertical displacement of at least that distance is required. In addition, each actuator pin should be capable of exerting a force greater than the contact force of the fingers during exploration. This is typically between 50 and 100 mN (Velazquez, Pissaloux et al., 2004).

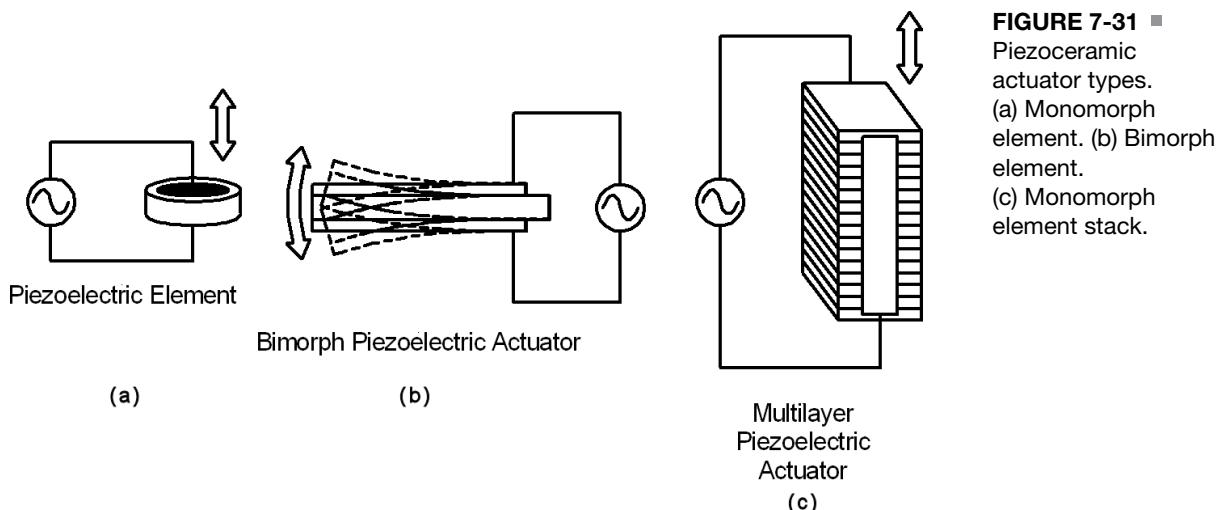
In theory, to minimize the energy expenditure of the device the pin and the drive mechanism should be resonant at the frequency of interest. However, as soon as such a device comes into contact with the skin, the skin's elasticity damps the resonance and the advantage is lost. It is therefore practical to develop an actuator that is energized at the required stimulation frequency with sufficient force to obtain the required displacement with each stroke. Such devices, shown in Figure 7-30 and Figure 7-31, operate using electromagnetic principles, SMAs, and the PZT effect, all of which are described in Chapter 3. However, these are only a few of the many methods to achieve the same result.

The operational frequency of electromagnetic and PZT devices can easily exceed the highest stimulation frequency of 300 Hz required for vibrotactile devices. However, SMA devices, by their nature are limited to a few hertz at most.

In construction, PZT actuators consist of a thin piezoceramic slab with metallized electrodes on the upper and lower surfaces. When a potential difference is applied across the face, the dimensions of the piezoceramic slab are altered. This basic mechanism can be used in isolation if very small displacements are suitable. However, in most cases they are assembled into a stack to increase the displacement for a given voltage, or, alternatively, a pair of elements are bonded on either side of a flexible metal beam to produce the bimorph actuator that flexes, as shown in Figure 7-31.

**FIGURE 7-30 ■**  
**Components of individual vibrotactile elements.**  
 (a) Conventional electromagnetic actuation. (b) Shape memory alloy actuator.





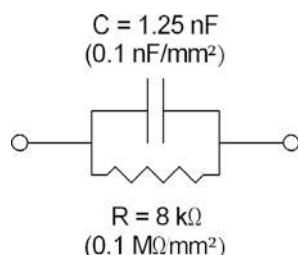
One of the more interesting mechanisms uses magneto-rheological fluids (MRFs), which respond to a magnetic field with a dramatic change in rheological behavior. They can reversibly and instantaneously change from a free-flowing liquid to a semisolid with controllable yield strength, which allows them to be used for high-bandwidth vibrotactile devices.

#### 7.7.2.4 Electrotactile Displays

Most researchers believe that the electric currents passing through the skin from surface electrodes directly stimulate afferent nerve fibers or whole nerve bundles if electrodes are sufficiently large. Depending on the applied voltage, current, waveform electrode size, force and hydration, among other things, subjects describe the sensation as a tingle, itch, vibration, buzz, touch, pressure, pinch, sharp, and ultimately burning pain (Kaczmarek, Webster et al., 1991).

As with the electrodes discussed in Chapter 2, the object is to convert electron flow in the lead wire to ionic flow within the tissue. Electrode requirements are therefore similar to those used for monitoring electrical signals. Ideally, the current density should be kept as uniform and as low as possible, but small variations in the contact pressure and sweat concentration make this impractical.

The impedance of electrodes can be modeled as a capacitor and a resistor in parallel as shown in Figure 7-32. For a monophasic 10 mA pulse with a duration of 10  $\mu$ s



**FIGURE 7-32** ■ Simplified electrical model of a 12 mm<sup>2</sup> electrode. [Adapted from (Kaczmarek, Webster et al., 1991).]

the normalized resistance is about  $0.1 \text{ M}\Omega/\text{mm}^2$  and the normalized capacitance about  $0.1 \text{ nF/mm}^2$ .

The resistive component of the impedance drops sharply with increased current, which results in an extremely nonlinear response. For this reason the electrodes are usually stimulated with constant current rather than constant voltage. Changes in the resistance also alter the effective time constant of the electrode, and this can be a problem for high-frequency stimulation.

The dynamic range of electrotactile stimulation is determined by ratio of the pain threshold level to the sensation threshold level, P/S (for voltage or current). This can vary from a factor of 2 (6 dB) to about 10 (20 dB) at best. This range is very limited compared with the other senses. For example, the ear has a dynamic range of 120 dB and the eye 70 dB.

Dynamic range measurements made by different researchers vary by a large factor because, as are most measurements on human subjects, the thresholds are so subjective. Though the threshold for stimulation is determined with good accuracy, there is no definition for the threshold of *pain*. In addition, pain thresholds vary with a whole raft of psychological factors. Furthermore, skin condition and the position of electrodes also have a large effect on the dynamic range and comfort of electrical stimulation.

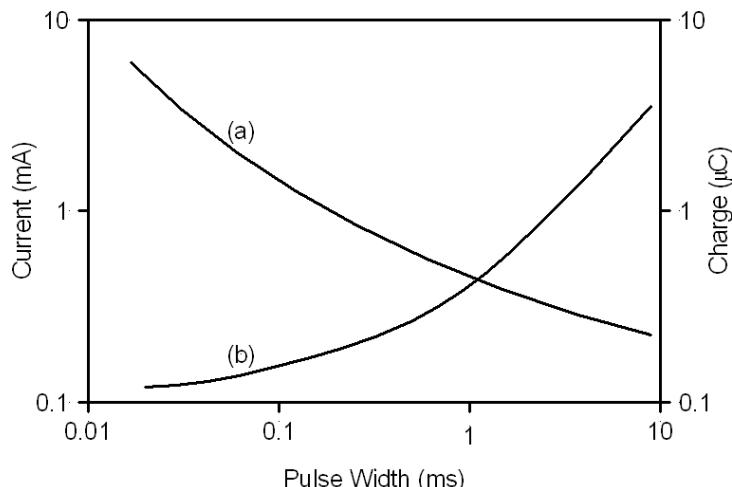
For pulsed stimulation, the sensation threshold increases as the pulse width decreases, suggesting that the threshold is determined by the total charge (current  $\times$  duration), as seen in Figure 7-33.

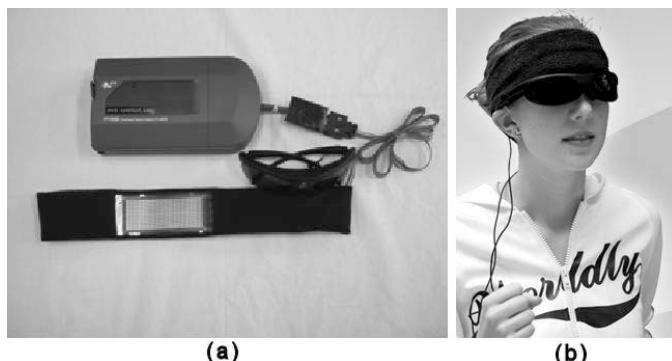
A number of different institutions worldwide are researching various aspects of electrotactile excitation. One of these is the Kajimoto laboratory, which has developed a  $32 \times 16$  element electrotactile display with an electrode spacing of 3 mm. This display is designed to be mounted on the forehead for ease of installation—it is held in place by a sweatband.

Visual images captured by the camera are converted to tactile information through two processes. The first is spatial outline extraction to enhance edges. The second is a temporal band-pass filtering to enhance time-varying information. In essence, the forehead recognition sensory system (FSRS) imitates functions performed in the retina and visual cortex to facilitate image cognition.

As with many other pieces of new technology, the electrotactile ideas developed within the Kajimoto laboratories have been spun off and are being developed for production as

**FIGURE 7-33 ■**  
Electrotactile  
sensation threshold  
as a function of (a)  
current and (b)  
charge. [Adapted  
from (Kaczmarek,  
Webster et al.,  
1991).]



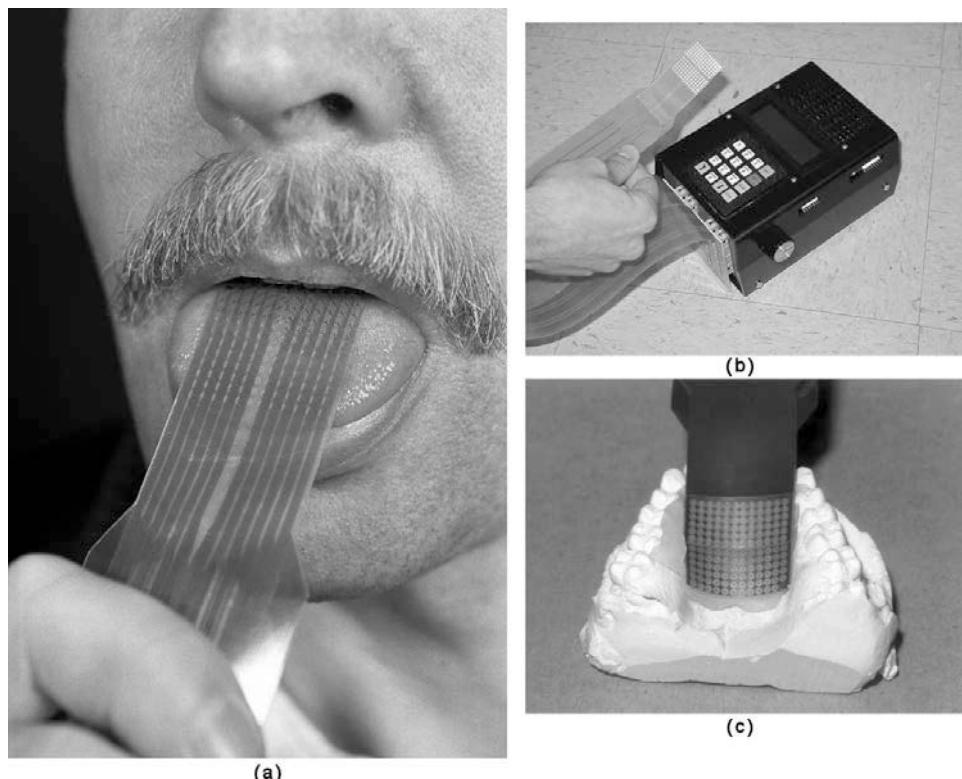


**FIGURE 7-34** ■ Commercial electrotactile display under development  
 (a) Photograph of hardware.  
 (b) Photograph of a model wearing the device. (Courtesy of EyePlusPlus <http://www.eyeplusplus.com/>.)

part of a commercial venture. The product under development by EyePlusPlus is shown in Figure 7-34.

For the past 10 years, researchers in the Department of Rehabilitation Medicine of the University of Wisconsin have been working on making a device that delivers electrotactile input to the tongue via a matrix of electrodes worn inside the mouth, as shown in the photographs in Figure 7-35. Using a camera, a computer, and the input device, individuals who have been blind their whole lives are now able to use this relatively simple and noninvasive device to see basic images.

The research group has developed a generic and flexible way to communicate information to people using these arrays. Initial test results showed that volunteer experimental test subjects could identify very simple geometric patterns such as circles, squares, and



**FIGURE 7-35** ■ Electrotactile display  
 (a) Photograph of electrode matrix stimulating the surface of the tongue.  
 (b) Photograph of the matrix with driver hardware.  
 (c) Photograph of the matrix attached to a plaster cast of the hard palate.  
 (Courtesy of Wisconsin University. <http://kaz.med.wisc.edu/projects.tdu.php>)

triangles and that they identified these figures as accurately on the tongue as on the fingertip.

Volunteers say that the electrical stimulus on the tongue feels like a tingle or vibration or like soda bubbles. The sensation is well controlled and painless unless the user deliberately turns up the excitation level. Occasionally it produces a weak metallic taste sensation, but no tissue irritation has been observed with the gold-plated electrodes.

Applications for the device include the following:

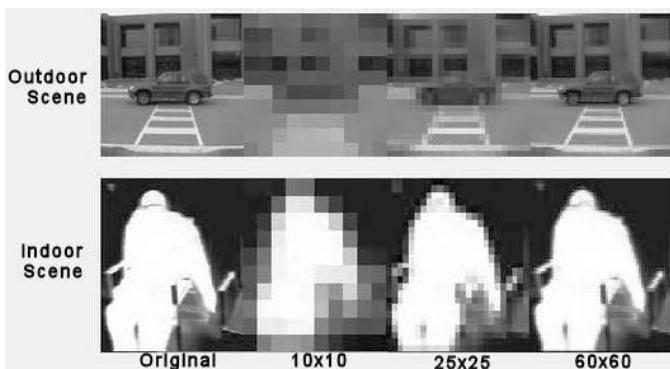
- Provision of vestibular information for people with balance disorders. This is a simple form of sensory substitution, in which the tongue is used to present information from accelerometers or tilt sensors.
- Directional or navigational information for people who operate under central command and control scenarios, such as military and civilian rescue personnel. Providing information via the tongue allows them to fully use their vision and hearing to respond to unforeseen threats or hazards. Experiments have shown that it is possible to navigate a virtual maze using only information received on the tongue.
- Crude visual information provided through the tongue for persons who are completely blind. Eliana Sampaio at the Louis Pasteur University in Strasbourg, France, has used this prosthesis with a small video camera and demonstrated an equivalent visual acuity of about 20–830, which is very poor vision, but possibly useful for certain limited activities with enough practice.
- Provision of tactile feedback to the human operators of robots. For example, Nicola Ferrier of the University of Wisconsin is developing a robot controlled by the tongue of persons with quadriplegia that could incorporate touch sensors into its gripper, relaying the touch information back to the user's tongue (Wisc Edu, 2008).

One of the commercial spinoffs of this technology is the BrainPort vision device under development by Wicab, Inc. This system consists of a postage-stamp-size electrode array for the top surface of the tongue (the tongue array), a base unit, a digital video camera, and a handheld controller for zoom and contrast inversion. Visual information is collected from the user-adjustable head-mounted camera (field of view [FOV] range 3–90 degrees) and sent to the base unit. The base unit translates the visual information into an electrical pattern that is displayed on the tongue.

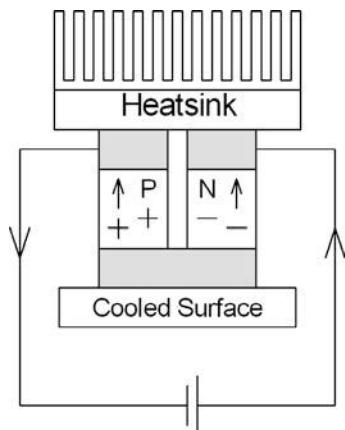
The tactile image is created by presenting white pixels from the camera as strong stimulation, black pixels as no stimulation, and gray levels as medium levels of stimulation, with the ability to invert contrast when appropriate. Users often report the sensation as pictures that are painted on the tongue with champagne bubbles.

With the current prototypes, which have arrays containing between 100 and 600 electro-tactile elements, study participants have been able to recognize high-contrast objects, their location, movement, and some aspects of perspective and depth. The images shown in Figure 7-36 demonstrate how indoor and outdoor information from the video camera is represented on the tongue. The prototypes have  $25 \times 25$  pixel resolution on a  $3 \times 3$  cm tongue display, presented at approximately 30 frames per second, yielding an information rich stream usable by participants (Wicab, 2009).

This technology is beginning to find a much wider market. It is starting to allow divers to “see” using sonar information in murky water and soldiers to obtain 360° night vision using inputs from an infrared camera, among many possible applications (Rosenblum, 2010).



**FIGURE 7-36** ■ Image acuity as a function of the array size [Adapted from (Wicab 2009), with permission.]



**FIGURE 7-37** ■ Schematic diagram showing the construction of a single Peltier element.

### 7.7.2.5 Thermotactile Displays

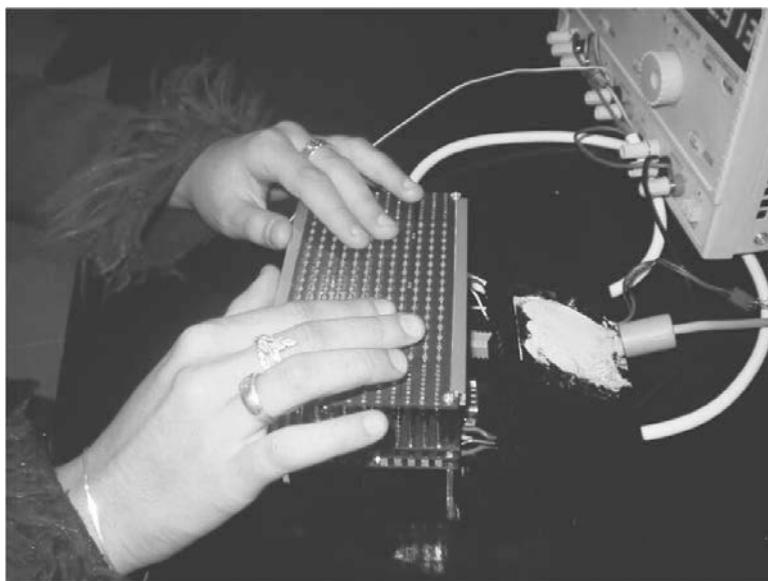
Peltier elements consist of p-doped and n-doped semiconductor material separated by a metallic interconnect, as shown in Figure 7-37. Current flows through the n-type material before crossing a metallic interconnect and passing into the p-type material. Electrons in the n-type material move against the direction of current flow whereas holes in the p-type material move in the direction of current, both remove heat from one side of the device and deposit it on the other. If the polarity is reversed, then the positions of the hot and cold junctions are reversed.

In commercial thermotactile displays, an array of closely spaced Peltier elements can be controlled to generate either hot or cold regions on the upper surface of the array, as shown in Figure 7-38. Control electronics monitors the temperature of each to maintain a reasonably constant element temperature. The main advantages of this form of display are that there are no moving parts and hence nothing to wear out compared with vibrotactile displays. They are also less intrusive than electrotactile displays.

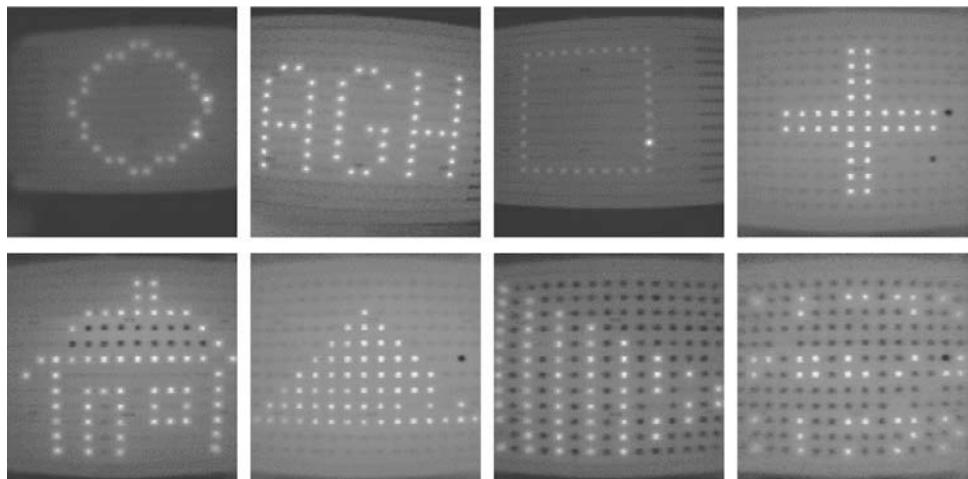
Because the distribution of temperature-sensitive nerve fibers is lower than that of some of the tactile corpuscles, the practical resolution of thermotactile displays will never be as high.

These devices are still in the experimental phase but can be used to produce Braille or standard letters as well as the simple graphics shown in the thermal images of the display visible in Figure 7-39.

**FIGURE 7-38 ■**  
Prototype  
thermotactile  
display. (Monkman  
and Taylor 1993).



**FIGURE 7-39 ■**  
Thermal images of  
the thermotactile  
display outputs.  
(Monkman and  
Taylor 1993).



## 7.8 | GPS-BASED SYSTEMS

Since the introduction of low-cost GPS units, either stand-alone or incorporated into a mobile phone, blind people have had an effective new navigation method to augment their existing capabilities. The genre is well represented by the system for wearable audio navigation (SWAN) developed at Georgia Tech's School of Psychology and College of Computing. The device was designed to help people get from point A to point B and know what's around them as they go.

SWAN uses GPS to locate the person and an electronic compass to detect the direction they are facing. Both feed into a computer that runs custom software to merge the information and to generate a series of sounds. Using recent developments in audio processing and conventional stereo headphones, the sounds generated by SWAN can appear to come from any direction around the user's head.

A series of beeps (beacons) leads users on the path to their destination. Other beacons convey useful situational awareness information (Walker and Lindsay, 2006):

- *Navigation beacon* sounds guide the user along a predetermined path, from a start point, through several waypoints, to arrive at the user's destination.
- *Object sounds* indicate the location and type of objects around the user, such as furniture, fountains, and doorways.
- *Surface transition* sounds signify a change in the walking surface, such as sidewalk to grass, carpet to tile, level corridor to descending stairway, and curb cuts.
- *Locations*, such as offices, classrooms, shops, buildings, and bus stops, are also indicated with sounds.
- *Annotations* are brief speech messages recorded by users that provide additional details about the environment. For example, "Deep puddle here when it rains."

Because the system relies on a priori detailed information about the area, any short-term changes, such as road works, are not registered, and the user needs to use one of the other prostheses discussed in this chapter. Another limitation is that GPS does not work indoors, so the researchers are developing a camera-based system to help in these situations.

## 7.9 | VISUAL NEUROPROSTHESES

---

### 7.9.1 Historical Perspective

The first recorded description of an electrically induced spot of light (later called a phosphene) comes from LeRoy in 1755. He was investigating the potential of electricity to cure various diseases when he discovered that an electrical pulse delivered to the eye produced the sensation of light. Research in the field continued in 1929 when Foerster applied electrical stimulation directly to the occipital cortex and reported the perception of spots of light that he referred to as phosphenes (Lovell, Hallum et al., 2007).

In the 1960s and 1970s, research efforts led by Brindley in England used an array of 80 platinum surface electrodes embedded under the dura membrane, connected to 80 individual receivers, to generate phosphenes using electrical stimulation. Information was conveyed to the receivers using electrical induction. It was found that, though it was possible to stimulate up to 35 individual phosphenes, the currents required were extremely high, typically between 1 and 10 mA. This was completely impractical from a visual prosthetic perspective as the total current from an array of electrodes would reach dangerous levels that could produce seizures.

In the last decade or so, research focus has shifted from the invasive intracranial neurosurgery required for cortical implants to the less invasive retinal options, which are discussed in some detail later.

### 7.9.2 Potential Sites for Visual Neuroprostheses

Of all of the neural structures discussed earlier in this chapter, the photoreceptor layer of the retina, the ganglion cell layer of the retina, the optic nerve, and the cerebral cortical region V1 have been proposed as the most likely sites for visual neuroprostheses. These sites have been proposed because of the relative ease with which they can be accessed



**FIGURE 7-40** ■ Schematic showing the components of a visual neuroprosthesis [Adapted from (Finn and LoPresti 2003).]

from a surgical perspective and because their structures show some correspondence in a visuotopic map. Other possibilities include the supra-choroidal space in the eye and the LGN in the brain.

At this stage, very strong stimulation is required to evoke a useful response (percept) with the result that only a crude sense of vision may be possible. However, considering the rapid progress that has been made with cochlear prostheses over the past decade, there is hope that a second generation of visual prostheses will be more subtle and effective in stimulating a relatively accurate visual response.

### 7.9.3 Components

To make any visual prosthesis acceptable, its components must be integrated into normal systems usually worn by individuals. These include eyeglasses to house the camera and a small processor about the size of a PDA or phone to perform the image processing function. A telemetry device would provide both power and the video signal to an embedded prosthesis containing the stimulation electronics. This would generate the appropriate currents to feed each of the elements of a neural interface array. These components are shown schematically in Figure 7-40.

#### 7.9.3.1 Video Encoder

A video encoder mimics the lost function of the photoreceptors in the retina by transforming the visual image into its electronic form. This could be a photodiode array, a dedicated CCD array, or a miniaturized video camera mounted within the frames of a pair of glasses worn by the patient.

For retinal-based visual neuroprostheses, the encoder would reside within the plane of the retina. This uses existing optics of the eye to form a projection with the result that acquisition of image data would function more naturally without the requirement of head movement to track as the eyes would move normally.

Unfortunately, the spatial resolution of any of these systems will be limited, not by the video encoder but by the limited numbers of electrodes in the neural interface. Temporal resolution is not seen to be a problem as the human visual system is relatively slow (about 30 Hz). Therefore, conventional miniaturized camera technology, already available, would be ideal for use as a video encoder.

#### 7.9.3.2 Signal Processor

The foundation on which all visual prostheses are built is the assumption that the visual system is built into a hierarchical sequence of maps. It is assumed that excitation of neurons close to each other spatially will excite other neurons at various levels in the visual pathways that are also close together. If this mapping were perfectly conformal and linear, then a high-quality neural image of the visual space would be found at each level of the visual pathways. This is not true, unfortunately, and the visuotopic map is conformal

only when considered from a low-resolution perspective. Recent work by Warren and Fernandez (2001) has shown that it is possible to use the relative location of a neuron in the cortex to predict where spots of light projected onto the visual space will excite that neuron. However, this can be done with an accuracy of only about  $0.5^\circ$  of visual angle.

Signal processing must accomplish a number of tasks. First, the electronics must transform the video signal into a set of discrete signals—one for each electrode. Next, the processor must be able to generate the correct amplitude and dynamic range of stimulus levels for the prosthesis, irrespective of the ambient light level. This is an automatic gain function that emulates the adaptive properties of both the photoreceptors and the pupil. Part of this process involves compression of the dynamic range.

The final function of the signal processor is to map the spatial output of the image to the electrode array so that vertical lines are perceived as vertical. The degree of this remapping depends on the degree of apparent randomness in the visuotopic organization at the implant site.

### 7.9.3.3 Telemetry and Power Interface

A visual prosthesis must receive information about the visual scene as well as power and data to run the electronics and subsequently stimulate the retina or cortex to generate the corresponding visual image. Given that it is undesirable for the purposes of long-term implantation to have wires penetrating the body or embedded batteries that have a limited life, it is necessary to send the visual signal and power to the implant wirelessly. It is only with recent advances in microelectronic technology that this has become feasible. If the prosthesis is retinal, then the carrier signal can be either light or radio frequency (RF). However, if it is embedded in the optic nerve or the visual cortex, then RF is generally used. It is envisaged that similar principles to those used for cochlear implants will be used as they have been successful in experimental applications.

The processes involved to accommodate both power and bidirectional telemetry have been discussed in Chapter 5. If the internal prosthesis is retinal, it would be convenient to mount one of the coils within the frame of a pair of eyeglasses and the other coil within the eye. This will ensure good alignment, and small variation of range, with a resultant high-efficiency signal and energy transfer. Radiated power will have to be limited to minimize local heating, and therefore the embedded parts of the prosthesis will have to be extremely efficient. For cortical implants there is no convenient means to mount the external prosthesis really close to the implant. However, behind-the-ear mounting as is used in cochlear implants would still limit the length of the internal wiring and provide a reasonable compromise.

Other issues include the required high reliability, and in the future, as the array sizes of the prosthetics increases, the telemetry will have to support high-bandwidth communications. This will be mitigated to some extent by improvements in image compression along with their associated encoding and decoding techniques.

### 7.9.3.4 Neural Stimulator

Stimulators can be based on existing designs that have been developed for other applications such as those for cochlear implants. These are typically external and have limited bandwidth due to the small number of electrodes involved, so their performance would have to be enhanced significantly to accommodate the requirements of a visual prosthesis.

Current thinking (Finn and LoPresti, 2003) is that stimulators will be digital to facilitate upgrades and to minimize power dissipation. They will include sufficient memory to

map the position of each electrode to its visuotopic map, and these will be dynamically updatable via telemetry as the structure of the image changes. These design issues are already being addressed with prototypes having been implemented on very-large-scale integration (VLSI) chips.

### 7.9.3.5 Neural Interface (Electrode Array)

This should be thought of as an active element of the system that transforms the currents generated by the stimulator electronics into ionic currents that flow in the body. Most materials can achieve this to some degree; however, some, including silver, are toxic to neurons in their vicinity, and platinum, though biocompatible, is not particularly effective. To date, the most effective material that is both biocompatible and an effective transducer from electric to ionic currents is oxidized iridium, and most implantable neural interfaces are made from that material (Robblee and Rose, 1990).

These interfaces are particularly difficult to develop because the human body has developed numerous defenses against intrusions of nonbiological materials. After time most are rendered inert or sealed off by a fibrous tissue capsule, so the challenge is to find materials that the body recognizes as benign in all ways.

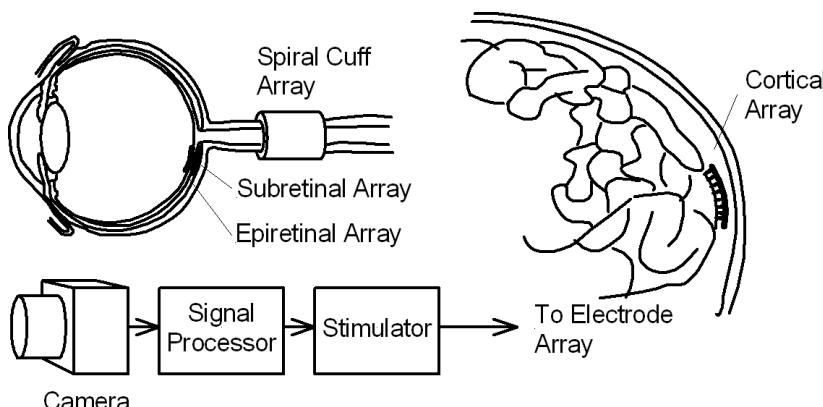
The implant must be chemically benign to avoid reaction, and it must also have a very similar density to that of the surrounding tissue so that it is not differently affected by gravity or kinematic accelerations. Incompatibilities can generate shear forces, which result in micromovements that both displace the electrode from its intended site and also cause chronic inflammation due to the irritation. These are particularly difficult to solve in retinal implants because saccadic eye movements produce large accelerations (Finn and LoPresti, 2003).

Another problem is that because all human cells exchange, for example, nutrients, with extracellular fluid and need to remain in equilibrium, the introduction of a large impermeable structure adjacent to any cells, even if they are biologically invisible, can result in disruption of their equilibrium. Once again, this is a major issue that has not been completely addressed, particularly for retinal implants, where the cell nutrition comes from either the vitreous humor on the one side and the retinal pigment epithelium on the other.

Other issues include the flexibility of the implant. Ideally it should be mechanically compliant to subtle movements of the surrounding tissue caused by blood pressure changes during the cardiac cycle. Typically, polymers of various kinds are superior to more rigid materials in this situation.

The final problem that needs to be addressed is that of tethering. Lead wires to the implant must be conductors and are therefore generally made of metal, which needs to be both insulated and very flexible. As array sizes increase, the numbers of wires must also increase, and the tethering problem will become more acute. Ultimately, flexible image processing electronics integrated directly onto the array will be required to minimize the number of connections to the stimulator.

As discussed earlier in this chapter, four main areas are suitable for the electrode array. The two main sites in the retina are the subretinal and epiretinal regions (below and above the retina) used by most researchers. There are also two main sites within the brain, one around the optic nerve and the second in the visual cortex, as shown in Figure 7-41. Alternatives include the supra-choroidal space behind the eyeball being the site of choice by researchers of the Australian Vision Prosthesis Group (AVPG), and possibly the LGN in the brain, though access to the latter involves a particularly invasive procedure.



**FIGURE 7-41** ■  
Common sites for visual prostheses.

### 7.9.4 Worldwide Research Activity

A number of groups around the world are involved in research and development of neurovisual prostheses of various kinds, with new research institutes entering the field regularly. Many of those now or recently active are listed in Table 7-2.

### 7.9.5 Subretinal Implants

A number of blinding disorders are primarily due to photoreceptor or outer retinal degeneration and destruction. These include but are not exclusive to diseases such as RP and AMD. Most retinal implants are designed to provide some form of vision to this subset of blind patients. Research groups have followed different paths to achieving this goal, with the least invasive being self-powered arrays of phototransducing elements placed in the subretinal space. In this configuration, each element comprises a semiconductor based photodiode-electrode pair. Light striking the photodiode causes a current to flow between its electrode and a reference electrode over the entire back of the array. The result is a voltage gradient that is intended to stimulate the dendrites of the bipolar cells.

The success of this approach depends on three assumptions. The first is that the bipolar cells of the damaged retina still operate in a physiologically normal manner. The second is that the photodiodes can produce sufficient current under normal illumination conditions, and the third is that the electrodes can be placed close enough to the bipolar cells for the induced photocurrent to excite them (Finn and LoPresti, 2003). Whether these assumptions are valid remains controversial; therefore, a number of alternatives have been considered, which will be discussed later in this chapter.

#### 7.9.5.1 Optobionics

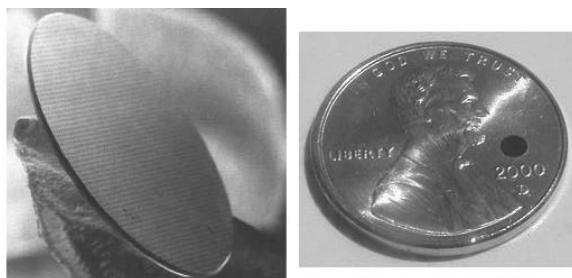
The Optobionics Corporation is a private company started by Dr. Alan Chow and his electrical engineer brother, Vincent Chow, in the 1990s. They developed the artificial silicon retina (ASR) microchip, as described already, which was designed to stimulate damaged retinal cells, allowing them to send visual signals again to the brain.

Initial tests involved a single element containing a photodiode and a large gold electrode ( $36 \text{ mm}^2$  area) that was implanted in the eye of a rabbit. Illumination of the photodiode was registered by increased cortical activity with the activity levels being proportional to the light intensity. With an electrode charge density of  $100 \text{ nC/mm}^2$ , resulting

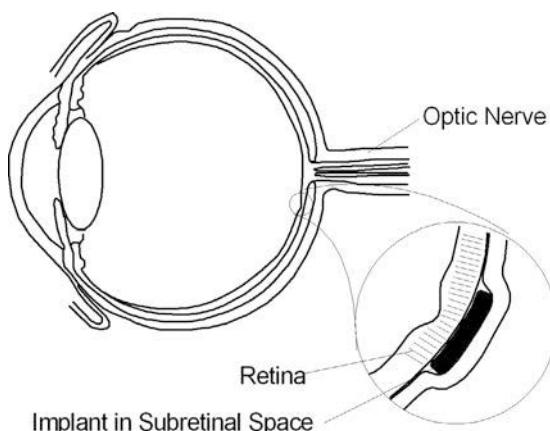
**TABLE 7-2** ■ Worldwide Research Activity in Neurovisual Prostheses

Research Area	Country	Institute	Chief Researcher
Retinal	USA #1	Mass. Eye and Ear Infirmary, Harvard Medical School Circuits and Systems Group of the Research Laboratory of Electronics	Prof. Rizzo Prof. Wyatt
	USA #2	University of Southern California	Prof. Eugene deJuan Prof. Mark Humayun Prof. Weiland
	USA #3 (to 2007)	Second Sight LLC Johns Hopkins University University of California, Santa Cruz	Dr. Robert Greenberg Prof. Dagnelie Prof. Liu
	USA #4	Optobionics Corp The Kresge Eye Institute, Wayne State University School of Medicine	Dr. Chow Prof. Abrams Prof. Iezzi
	USA #5	College of Engineering, Wayne State University	Prof. Auner
	USA #6	Stanford University	Prof. Palanker
	German #1	University of Houston University Eye Hospital Turbingen	Prof. Ignatiev Prof. Eberhart Zrenner
	German #2	University Eye Hospital Regensburg Retina Implant AG	Prof. Gabel
	German #3	Aachen University University of Bonn University of Duisburg	Prof. Walter Prof. Eckmiller
	UK #1	IIP-Technologies	Prof. Toumazou
	UK #2	Imperial College London	Dr. Mathieson
	Japan #1	University of Glasgow	Dr. Morrison
	Japan #2	Tokyo Institute of Technology The Institute of Physical and Chemical Research The Institute for Developmental Research	Dr. Yagi Dr. Mukai Dr. Watanabe
	Japan #3	Tohoku University Dept. of Bioengineering and Robotics Tohoku University Biomedical Engineering Research Organization	Prof. Kurino Dr. Tomita
	Japan #4	Okayama University	Prof. Matsuo Prof. Shimamura
Optic Nerve	Australia	Hayashibara Co., Ltd NIDEK Co., Ltd Osaka University	Prof. Nigel Lovell Prof. Gregg Suaning
	Korea	University of New South Wales	Prof. Kim
	China #1	Seoul National University	Prof. Qiushi Ren
	Belgium	Biomedical Engineering, Shanghai Jiao-Tong University	Dr. Veraart
	China #2	University Catholique de Louvain, Neural Rehabilitation Engineering Lab	Dr. Trulliemans
Cortical	USA #7 (to 2004)	University Catholique de Louvain, Microelectronics Lab	Prof. Qiushi Ren
	USA #8	Biomedical Engineering, Shanghai Jiao-Tong University	Dr. Dobelle
	USA #9	The Dobelle Group	Prof. Richard Normann
	Spain	University of Utah	Prof. Troyk
	Canada	Illinois Institute of Technology	Prof. Fernandez
		University Miguel Hernandez	Prof. Mohamad
		Ecole Polytechnique de Montreal	

Source: Finn, W. and P. LoPresti (Eds.), *Handbook of Neuroprosthetic Methods*, London: CRC Press, 2003, with permission.



**FIGURE 7-42** ■  
Photographs of the Optobionics artificial silicon retina. (Optobionics 2008), reproduced with permission.



**FIGURE 7-43** ■  
Diagram showing the ASR implant in the subretinal space.

from illumination similar to that from a bright sunny day, the invoked activity from the electrode was similar to that measured for the unoperated eye.

Having completed this proof-of-concept design, a prosthesis containing an array of photodiodes and smaller electrodes was developed. The current design of the ASR microchip is a silicon chip 2 mm in diameter and 25 microns thick, as can be seen Figure 7-42. It contains approximately 3,000 microscopic photocells, called microphotodiodes, each with its own stimulating electrode.

The ASR is designed to be implanted in the subretinal space shown in Figure 7-43. In the animal experiments, each element has an area of about  $20 \times 20 \mu\text{m}$ , each separated by an insulating moat with a width of about  $10 \mu\text{m}$  resulting in an array density of about  $1100 \text{ elements/mm}^2$ . The photodiode consists of various layers of doped silicon sandwiched between two layers of gold. The outer gold layer is etched to the same dimensions as the individual photodiode and is sufficiently thin for light to pass through it and enter the silicon where it generates a current. The back layer of gold covers the complete ASR and functions as the return electrode.

The photodiode is responsive to light in the 500 to 1100 nm range, which includes most of the visual spectrum (360 to 830 nm) and extending into the near infrared. One of the advantages of the IR sensitivity is that it can be used to ensure that responses are due to the prosthesis and not to remaining photosensitive cells in the eye.

The results of animal tests showed some cell loss and inflammation, but this was limited because the prosthesis is constructed from well-tolerated materials. General optical stimulation of the implanted eye showed similar electroretinogram (ERG) results to the other eye indicating that the eye itself was still functioning normally. In addition, the ERG response for IR stimulation indicated that the implant was still functioning, and indeed it

did continue to function in one test for 11 months. However, the magnitude of the ERG took up to 2 months to reach a peak and then began to degrade after 4 months.

In the ERG tests, stimulus of positive and negative current-generating photodiodes excited responses with the same polarity as the stimulation, an indication that these were only stimulus artifacts. If the photoreceptor neurons had in fact been stimulated, different sensitivities to the two polarities would have been seen.

When the prosthesis was removed, it was also found that the gold electrode material had been dissolved by the electrical stimulation. This had exposed the underlying chromium layer, which was toxic to the surrounding cells. Alternative electrode materials including platinum and iridium oxide were trialed, but the results are unknown.

Even though no neural response was observed during animal trials except during aggressive stimulation with large electrodes, human safety trials were initiated, and the prosthesis was implanted in the retinas of three people in 2000. In these trials the ASR was also a 2 mm diameter disk 25 microns thick with a total of 3500 elements covering 1.4 mm<sup>2</sup> of the area (Finn and LoPresti, 2003).

Notwithstanding the controversy in the research community regarding the ability of the ASR to generate sufficient current to produce a biologically relevant signal, recent results have proven to be more promising. A team of researchers, some of whom were affiliated with Optobionics and some not, implanted a small (1 mm diameter) ASR in rats and monitored responses in the superior colliculus. The results confirmed retinal stimulation induced neural responses with more robust visible-evoked responses at sites that received implant input compared with those that did not (DeMarco, Yarbrough et al., 2007).

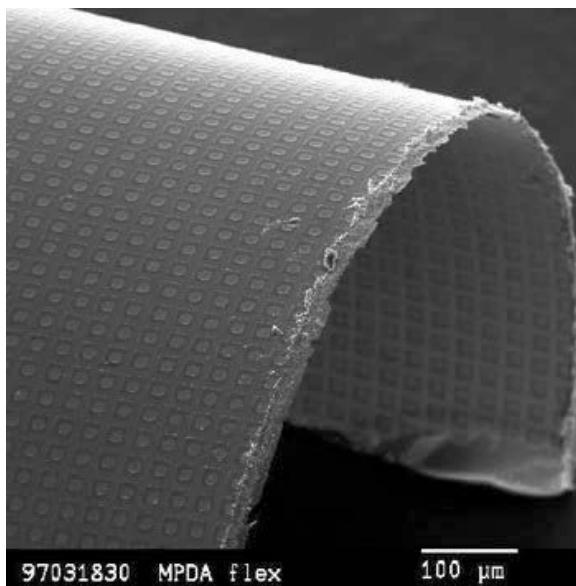
This research was terminated in 2007, and the company closed. Subsequently Dr. Chow acquired the Optobionics name and the ASR implants and is reorganizing a new company under the Optobionics name.

### 7.9.5.2 Research in Germany

A German group based in Tubingen also developed a similar, flexible, multiphotodiode array (MPDA) to that produced by the Optobionics group. It uses the same 20 × 20 μm photodiodes but with only a 5 μm gap between them, which leads to a packing density of 1600 elements/mm<sup>2</sup>, as can be seen in Figure 7-44. The upper electrode size is smaller, only 8 × 8 μm, which allows more light to reach each photodiode and the electrode layer to be thicker. The remainder of the diode is insulated with silicon oxide, which is inert and transparent to light. Later versions of the device also included a hole through the center of each photodiode to ensure that nutrition continued to flow in to retinal cells.

Proof of concept involved inserting the prosthesis into the retinas removed from newly born chicks and rats and then placing both onto a dense array of measurement electrodes, which could measure the response of neurons in the ganglion cell layer closest to it. To ensure that the response was due to stimulation from the MPDA, the photoreceptor layer was destroyed, or a rat strain that had few photoreceptors was chosen.

In both of these species, the activity of ganglion cells could be modulated by light, presumably by the activity of the MPDA. Unfortunately, as the diameter of the light spot was decreased, it became more difficult to evoke a response. A 0.25 mm diameter spot with an intensity of 70,000 lux covering 80 photodiodes could barely elicit a response. From these tests, the group concluded that passive photodiode arrays cannot provide a useful sense of vision under normal lighting conditions. They are currently investigating active prostheses as an alternative (Finn and LoPresti, 2003).



**FIGURE 7-44 ■**  
SEM image of the  
MPDA array  
(Schubert 1999),  
reproduced with  
permission

As expected, *in vivo* implants of the MPDA were no more successful than the ASR in generating neural responses. However, the group did show that the prosthesis using their electrode materials was stable for up to 20 months (the duration of the longest test).

Retina Implant AG was founded to further develop this technology. An operational implant has been produced that is about 3 mm in diameter and subtends a 12° field of view. It contains 1500 pixels, each consisting of two photocells, an amplifier, and a stimulating electrode within an area of about  $70 \mu\text{m} \times 70 \mu\text{m}$ .

The main difference between this and the MPDA array is that additional energy is supplied to the stimulating electrodes through the integrated amplifiers. At present this additional power is provided by an electrical connection to the prosthesis, but a high-frequency electromagnetic link is in development to transfer power wirelessly.

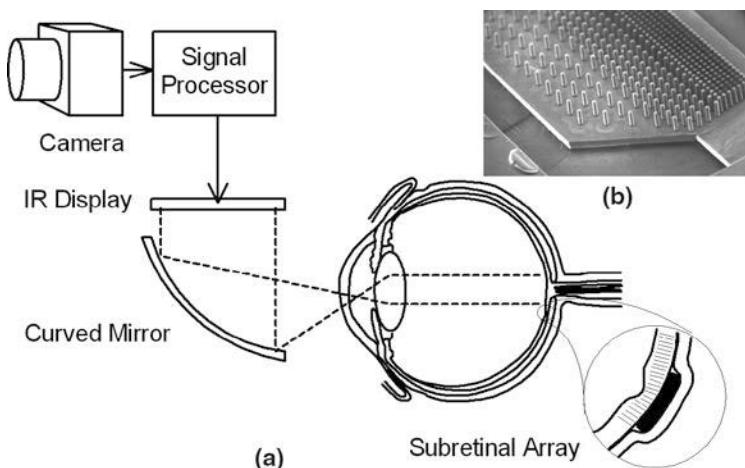
### 7.9.5.3 Stanford University Retinal Implant

An interesting and apparently effective modification to the subretinal implants discussed previously has been developed by researchers from the Hansen Experimental Physics Laboratory and the Department of Neurobiology at Stanford University. Their first innovation was to convert the visual image to a more intense infrared one that is projected into the eye, as shown in Figure 7-45. This image projection device tracks the eye microsaccades essential to proper perception and thus better reproduces what actually occurs in nature. The second innovation is to introduce a subretinal array consisting of photodiode-electrode pairs in which the electrodes are formed as pillars or pores onto which retinal cells are encouraged to migrate. This cell migration results in cell-electrode distance of 7 to 10  $\mu\text{m}$  after only 3 days, with the result that only moderate current densities are required to excite the cell (Levy, 2005).

### 7.9.5.4 Boston Retinal Implant Project

While neuro-ophthalmologist Joseph Rizzo III was researching retinal transplants to restore blind people's vision at the Massachusetts Eye and Ear Infirmary and the Boston

**FIGURE 7-45 ■**  
**Stanford retinal implant.** (a) Conceptual diagram showing the operational principles. (b) SEM image of an experimental subretinal array. [Adapted from (Levy 2005).]



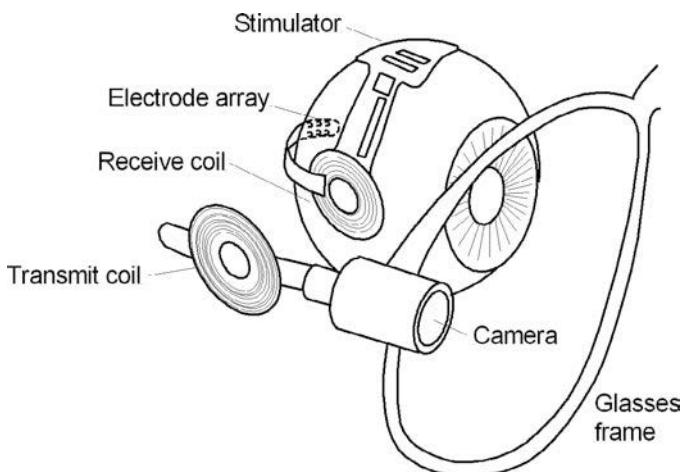
VA Medical Center, he also teamed up with Massachusetts Institute of Technology (MIT) electrical engineer John Wyatt Jr. to pursue the development of electrode-based retinal prosthetics. In 1988, they launched the Boston Retinal Implant Project.

Following a number of neuroscientific studies in which retinal stimulation and electrode life were examined, initial proof-of-concept testing on patients who were blind from advanced RP was undertaken. The surgery was performed while the patients were awake but locally anesthetized, as this allowed them to communicate the percepts that were created to the surgical team during the experiment. A total of six experiments involving the surgical placement of an electrode array into the subretinal layer have been conducted to date. In these experiments, minute electrical currents were delivered through the electrode array, allowing some of the patients who had been legally blind for decades to see relatively small phosphene, about as big as a pea at arm's length and occasionally to distinguish two distinct spots of light. Also, when sufficient electrodes had been stimulated, some patients were able to see a line.

It was decided that internal components of the prosthesis should receive power and video information from outside the body and that its insertion should occur with as little trauma as possible. In the design, the processed image information supplied by a small video camera was transmitted through a coil located on the arm of a specially designed pair of glasses. Both the signal and power were sent to receiving coils located on the prosthesis draped over the eyeball. Current passing through individual subretinal electrodes stimulated bipolar cells in the appropriate areas of the retina corresponding to the main features in the visual scene.

Key to the current prosthesis is a custom designed, low-power mixed-signal integrated circuit (IC) that drives the proper electrodes to stimulate the bipolar cells based on the incoming data. This is connected to a custom microfabricated thin and flexible array made from the best available electrode materials. It is  $10\ \mu\text{m}$  thick, 2 millimeters wide, and three millimeters long and is the only thing that penetrates the eye. The mixed-signal IC, electrode array, communication coils, and other components are assembled onto a flexible circuit substrate that is tacked to the eyeball, as shown in Figure 7-46.

The first prototypes of this new design are working and are undergoing final testing. However, this prosthesis contains an array of only 15 electrodes, each  $400\ \mu\text{m}$  across. Although this will provide only a small area of low-resolution vision, the researchers are



**FIGURE 7-46** ■ The Boston retinal implant flexible prosthesis showing the telemetry and power coil and stimulator draped over the eyeball.

confident that it will help with their first goal: improving blind people's quality of life by allowing them to walk around unfamiliar areas more easily than they can with canes.

The design has a number of special features, including ultralow power consumption and modularity that minimize that amount of hardware placed into the eye. This allows for the use of a minimally invasive surgical method of implantation.

### 7.9.6 Epiretinal Implants

As an alternative to stimulating the bipolar cells, other researchers have proposed placing the stimulating electrodes on the inner surface of the retina and stimulating the remaining retinal ganglion cells. It has been shown that even at the terminal stages of the common pathologies leading to blindness these cells mostly remain intact.

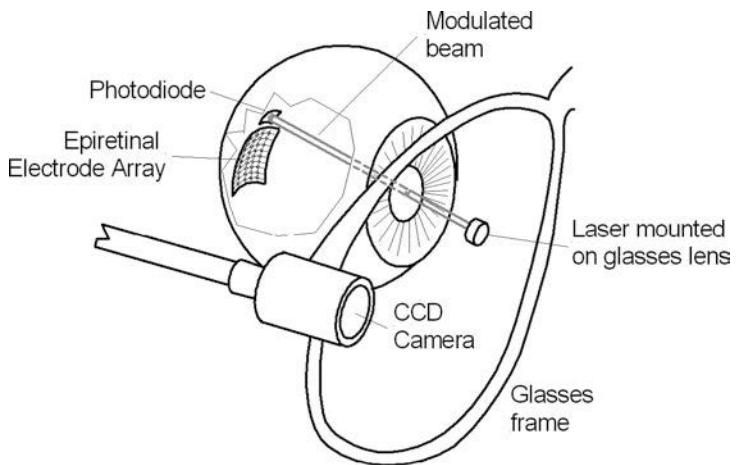
A number of groups have been pursuing similar approaches along these lines. An array of surface electrodes is attached to the inner surface of the retina between the vitreous humor and the inner limiting membrane. Driven by the signal processing electronics, electrodes stimulate ganglion cells to produce a similar pattern of phosphene.

In a clinical device, the signal processing will be placed within the eyeball with control and power supplied by an optical or RF link. The success of this process relies on four assumptions. The first is that the retinal ganglion cells continue to function normally. Second, the neural interface can be attached to the retina without damaging the underlying tissue any further. The third assumption is that useful percepts can be generated at reasonable stimulation levels and, finally, that patterned percepts can be created (Finn and LoPresti, 2003).

#### 7.9.6.1 Harvard Medical School–MIT Collaboration

The Retinal Implant Project began in 1989 as a Massachusetts Eye and Ear Infirmary–Harvard Medical School and MIT collaboration. The chip rests on the inside surface of the retina, opposite the damaged rods and cones, and in contact with the neural ganglion cells. Early research showed that individual percepts could be generated, but that, in most cases, the pattern of the percept did not accurately follow the pattern of the electrode array. In addition, the reliability of the percept was only 66% compared with 82% for the results obtained from the prosthesis fitted to a sighted volunteer. Other problems included the

**FIGURE 7-47** ■ Harvard retinal implant driven using modulated laser light.



high charge densities required to produce a reliable percept. These were found to be as high as  $28 \mu\text{C}/\text{cm}^2$ , which is close to the safe long-term limit.

Studies of the long-term ability to evoke activity in the visual cortex and investigations of the actual retinal excitation using arrays of stimulation and recording electrodes have also been undertaken. The latter investigations used stimulation electrodes with a diameter of  $10 \mu\text{m}$  at  $25 \mu\text{m}$  centers, with recording electrodes of the same diameter placed on  $70 \mu\text{m}$  centers. The retinas were stimulated with biphasic current pulses with amplitudes from  $0.01$  to  $20 \mu\text{A}$  with a total duration of  $800 \mu\text{s}$ . The threshold for observing any extracellular activity ranged between  $0.06$  and  $1.8 \mu\text{A}$  (Finn and LoPresti, 2003).

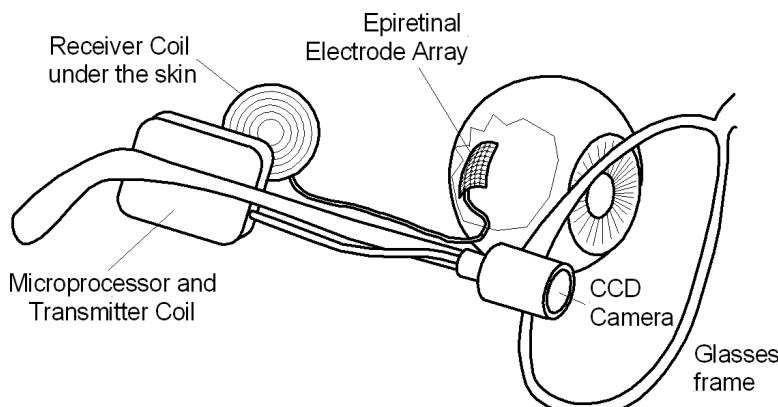
From a hardware perspective, the implant contains two silicon chips, both within a silicone capsule. The top chip receives light entering the eye from a tiny laser attached to a pair of glasses. The chip acts as a photocell and generates current when it is illuminated by the laser, with the voltage depending on the wavelength of the light and its intensity. It measures  $2 \times 2 \text{ mm}$  and consists of 12 photodiodes made of silicon connected in series. Using NIR wavelengths at intensities believed to be safe for the eye, the photodiode array delivers approximately 7 volts, slightly more than needed to power the signal processing chip.

A small CCD camera is mounted on the glasses frame and delivers each visual scene, amplitude modulated onto the same laser that carries power and the image into the eye. The second chip then decodes the laser's visual information and transfers it to an electrode array on the inner surface of the retina as shown in Figure 7-47.

### 7.9.6.2 Doherty Eye Institute

This research group started out as a collaboration between Johns Hopkins University and North Carolina State University, but later Dr. Humayun and Eugene de Juan moved to the University of Southern California and formed the Intraocular Retinal Prosthesis Group of the Doherty Eye Institute. They also have a relationship with a private company called Second Sight.

Early research was undertaken on isolated bullfrog retinas and *in vivo* rabbit retinas with bipolar, biphasic current pulses of between  $50$  and  $300 \mu\text{A}$  with a total duration of  $150 \mu\text{s}$ . Platinum electrodes with surface area of  $0.13 \text{ mm}^2$  spaced  $200 \mu\text{m}$  apart were used for stimulation, and a single electrode was used to record the neural response. Neural responses were observed for current levels as low as  $50 \mu\text{A}$  for the bullfrog and  $150 \mu\text{A}$  for the rabbit retinas, which is equivalent to charge densities of  $3$  and  $9 \mu\text{C}/\text{cm}^2$ , well within safe



**FIGURE 7-48 ■**  
Doherty Eye Institute  
Retinal Implant.

current density levels of  $100 \mu\text{C}/\text{cm}^2$ . Increases in the stimulation current resulted in higher voltages for the neural response, indicating that more ganglion cells are being stimulated.

Longer-term trials with charge densities of less than  $0.1 \mu\text{C}/\text{cm}^2$ , for 12 hours per day over 60 days, did not uncover any gross changes in the function of the retina. In these trials, the array was tacked to the retina, but the group also tried using bioadhesives to perform the same function.

In the first human trials, the question that the researchers wanted to answer was whether bipolar stimulation of the retina would lead to a percept and at what level? It was found that with platinum electrodes charge densities of between  $160$  and  $3200 \mu\text{C}/\text{cm}^2$  would evoke percepts ranging in size from a pinhead to a pea at  $30 \text{ cm}$ . In addition, visuotopic organization of the percepts was verified over a wide region by stimulating widely separated portions of the retina.

Tests with blind volunteers showed that with an electrode spacing of  $435 \mu\text{m}$ , two individual phosphenes were visible separated by about  $1^\circ$ , a spacing consistent with the known magnification of the retina. Later tests using electrode arrays with  $600 \mu\text{m}$  spacing and  $400 \mu\text{m}$  diameters demonstrated the percept of a fused line when a row of electrodes were excited. U- and box-shaped excitations generated H and rectangular percepts. Unfortunately, to generate these patterns required a charge density of about  $300 \mu\text{C}/\text{cm}^2$ , which exceeds the maximum safe level for platinum. New configurations, or the use of iridium electrodes, would allow the current densities to be increased without tissue damage.

These early human trials added to the mounting evidence that people blinded by retinal diseases could get partially restored vision with an implantable microelectronic retinal device. The Doherty system functions in real time in a manner similar to the other prostheses discussed in this chapter. As illustrated in Figure 7-48, the individual components of the system perform the following functions:

- The camera on the glasses captures sequential image frames.
- Signals are sent to an external signal processor the size of a PDA and worn on a belt.
- Processed information is sent back to the glasses and wirelessly transmitted to a receiver under the skin above the ear as with a cochlear implant.
- The internal receiver relays information along a cable to electrodes in the retinal implant.
- Electrodes stimulate the retina with appropriate currents, which trigger neural responses in the ganglion cells that feed visual information to the brain.

Most of the human trials are conducted through Second Sight LLC, and this is where the first, low-resolution devices were implanted into six patients over a 2-year period starting in 2002. Even with only 16 pixels, the results proved to be good. Initially, patients just saw some assembled dots, but because of ongoing brain plasticity they began to report considerably improved acuity with time.

The newest implant, the Argus II, has a higher resolution than the earlier devices. It contains 60 electrodes and is a lot smaller, only  $1 \text{ mm}^2$ , which reduces the amount of surgery that needs to be done to implant the device.

The technology has now been given the go-ahead by the U.S. Food and Drug Administration to be used in an exploratory patient trial that will take place at five centers across America over 2 years, with 50–75 patients aged over 50. If successful, the device could be commercialized soon after, costing around \$30,000.

### 7.9.6.3 EPI-RET Project and IIP Technologies

The main achievements of the EPI-RET project phases I and II at Aachen University were the development of the system concept and the fabrication of a first prototype of an implant featuring full wireless control of 25 platinum electrodes. The group showed that tack fixation of epiretinal devices was safe and sufficient to keep the implant close to the retinal surface and that electrical stimulation achieved specific cortical activation.

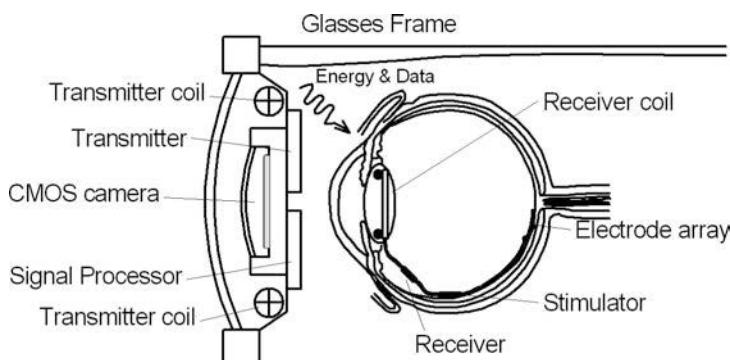
The working principle of this epiretinal implant is shown in Figure 7-49. An image is captured by a camera placed on the eyeglasses frame. The camera signal is preprocessed and data together with energy is transmitted using RF induction to the implant, which consists of a receiver and a stimulator. The receiver unit is inserted into the capsular bag after removal of the lens. The microelectrode array is placed onto the retinal surface and attached with a retinal tack.

IIP Technologies GmbH of Bonn, a subsidiary of IMI Intelligent Medical Implants, also in collaboration with Aachen University, started out investigating the signal processing aspects required for an epiretinal implant and went on to develop the specialized electronics that could mimic some of the image processing functions performed by the retina.

Early proof-of-concept experiments on primates conformed that their neural interface, a multi-microcontact array (MMA) could be secured using retinal tacks and that cortical activation could be generated by electrode stimulation.

Subsequent experiments using cats were aimed at mapping the relationship between the electrode excitation and the cortical response. This has led the group to the conclusion

**FIGURE 7-49** ■  
Schematic diagram  
of the EPI-RET  
implant.



that it is possible to stimulate the desired retinal ganglion cells by the simultaneous control of current in multiple electrode sites (Finn and LoPresti, 2003). This technique is more advanced than those proposed by other researchers in the field and is therefore referred to as the first “intelligent” retinal implant.

IIPs learning retinal implant system replaces some of the signal-processing functions of a healthy retina and provides input to the retinal ganglion cells that, in turn, give input to the optic nerve and the brain. The system comprises three main components:

- The retinal stimulator implant, which is surgically tacked onto the retina of a patient. This is attached via a short pigtail that leads to an antenna just behind the lens.
- Eyeglasses containing an integrated mini-camera and transmitter for wireless signal and energy transmission, called the visual interface. They are also connected via a cable to the pocket processor.
- The pocket processor is attached to the patient’s belt and replaces the information processing function of the formerly healthy retina. The use of a high-speed digital signal processor allows the provision of “intelligent information” to the implant (and the neural ganglion cells) by using adaptive software to approximate the information processing normally carried out by the healthy retina.

The entire process enables patients to optimize their visual perception during the learning phase. Using the patient’s feedback on perception as an input for the tuning of the pocket processor is the unique feature of the system and constitutes the “learning” capability of the learning retinal implant system.

An initial clinical research study in 2003 showed that of the 20 patients suffering from RP who participated in the study, 19 reported that their visual perception had been triggered by electrical stimulations from IIP’s retinal implant. By 2005 trials showed that patients could identify shapes, and a year later wireless transmission of data and energy into an implant in the eye of long-time blind persons had resulted in pattern recognition.

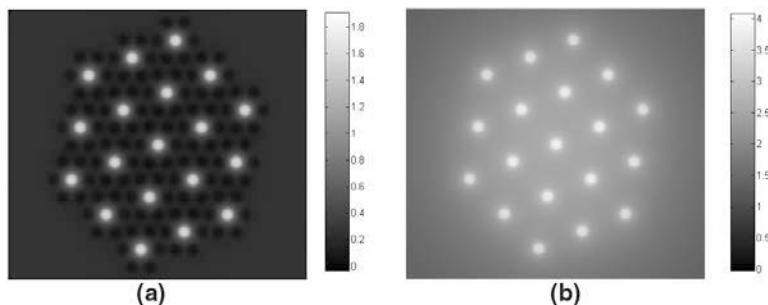
Researchers believe that, in the not-too-distant future, the learning retinal implant system, along with rehabilitation, may allow patients to recognize objects by identifying their size as well as their position, movements, and shapes. In other words, a blind person is expected to be able to move independently in an unfamiliar environment.

#### 7.9.6.4 Australian Vision Prosthesis Group

In 1997, researchers Gregg Suaning and Nigel Lovell from the Graduate School of Biomedical Engineering at the University of New South Wales (UNSW) set out to develop a visual prosthesis for the treatment of disorders causing blindness. As it grew, this association became known as the AVPG.

In 2010, Bionic Vision Australia (BVA) was launched with \$42 million in funding from the Australian Research Council. It is a consortium or research institutions including the AVPG, University of Melbourne, University of Western Sydney, Australian National University, Center for Eye Research Australia, the Bionic Ear Institute, and NICTA. The AVPG wide-view device will be the flagship implant that will take the BVA to human trials by 2013.

Initial work by the AVPG focused on a 100-channel RF stimulator and electrode array divided into a 50-element stimulating block and a similar-sized reference block. One or more electrodes from the stimulating block are driven with respect to one or more



**FIGURE 7-50** ■ Simulated 2-D voltage distribution for 14 parallel current sources. (a) Surrounded by six guard electrodes. (b) One return electrode. (Courtesy of AVPG, with permission).

reference electrodes. A single current source that can provide  $1860\ \mu\text{A}$  is multiplexed into the blocks, allowing for parallel stimuli but limited by the minimum threshold current for effective stimulation.

Inspired by the high-density hexagonal packing of the individual ommatidia making up the compound eyes of *Drosophila*, the array developed at the AVPG is similar. This configuration allows the current injected by a central electrode to be captured by the surrounding six return electrodes. In addition to the optimum packing density, this layout minimizes cross-talk and should result in the generation of finely constrained phosphenes.

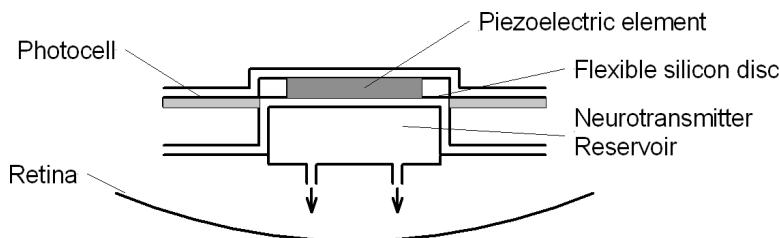
To investigate this possibility, simulations were conducted for a lattice of 98 cylindrical electrodes each with a diameter of 0.4 mm and a height of 0.2 mm immersed in saline solution. The field strength (potential) generated within the saline for two different configurations of ground electrodes is shown in Figure 7-50. In the first case, 1 mA is injected into each of 14 electrodes, each surrounded by six return electrodes at ground potential. It can be seen that these guard electrodes concentrate the field within the ring. In the second case, 1 mA is again injected, but only a single return electrode is provided, with the remaining five left floating. In the second case, 1 mA is again injected, but only a single return electrode is provided resulting in a much more diffuse field.

### 7.9.7 Alternative Implants

The photoreceptors in the retina release neurotransmitters in response to light. These activate the layer of neurons directly above, from where the signals are relayed to the brain via the optic nerve. A potentially viable alternative to electrical stimulation of the retina is chemical stimulation.

Laxman Saggere, an engineer at the University of Illinois at Chicago, has designed an implant that would replace damaged photoreceptors with a set of neurotransmitter pumps that respond to light. The crucial component, a light-powered actuator that flexes in response to the very low-intensity light that strikes the retina, has been built. Multiple actuators on a single chip could be used to pick up the details of the image focused on the retina, allowing some visual information to be passed on to the brain.

The prototype actuator, shown in Figure 7-51, consists of a flexible silicon disk 1.5 mm in diameter and  $15\ \mu\text{m}$  thick. When light strikes a silicon photocell adjacent to the disk, it generates a voltage that is applied to a PZT layer, which flexes, pushing down on the



**FIGURE 7-51** ■ Light-powered retinal implant pump [Adapted from (Biever 2006).]

silicon disk. In future, a reservoir will sit underneath the disk, and this action will squeeze the neurotransmitters out onto retinal cells (Biever, 2006).

### 7.9.8 Optic Nerve Stimulation

The optic nerves provide the only path for visual information from the eye to the visual cortex in the brain. In cases where the visual ganglion cells in the retina have degraded or if patients have lost their eyes in accidents or to cancer, then one alternative is to activate neurons in the optic nerve to restore some sight.

Researchers from the Catholic University of Louvain in Brussels tested this concept in a series of experiments starting in 2000. The prosthesis consists of a self-sizing spiral cuff electrode array that wraps around the optic nerve. In the first human trials, an array consisting of only four surface electrodes was used. Current could be switched in a bipolar manner between any pair of electrodes.

Experiments conducted over a 4-year period have determined the current levels for reliable phosphene generation along with their spatial location. Wide ranges of phosphenes have been excited extending  $85^\circ$  in the horizontal and  $60^\circ$  in the vertical in front of the patient. Their size extends from between 1 and 50 square degrees with the intensity and the size being a function of the stimulus current. Unfortunately, there is a strong cross-correlation between the amplitude of the stimulus and the position of the perceived phosphene with migrations from the edge of the phosphene space to the center, with only a threefold increase in current.

For a single biphasic pulse with a duration of  $213\ \mu\text{s}$ , the average thresholds are about  $350\ \mu\text{A}$ , whereas for pulse trains of 17 pulses delivered at  $160\ \text{Hz}$  the threshold dropped to only  $15\ \mu\text{A}$ . Stimulation rates below  $10\ \text{Hz}$  caused flickering, but above that frequency resulted in continuous phosphene generation with the steady percept fading out after 1 to 3 seconds.

The researchers claim that the phosphene space is sufficiently stable to predict the size, location, and intensity from the stimulation parameters. However, they are able to evoke patterns of only between 4 and 24 phosphenes, which is probably insufficient to generate anything but the simplest shapes.

Questions that remain regarding the usefulness of this technique include the following (Finn and LoPresti, 2003):

- How viable is the population of optic nerve fibers after damage to the ganglion layer in the retina?
- Does electrical stimulation of the optic nerve decrease or increase the rate of degradation?
- How many phosphenes evoked by optic nerve stimulation are required to produce meaningful amounts of visual information?

- Would penetrating arrays provide more focal stimulation of the optic nerve along with better control of phosphene location and intensity?
- Would penetrating arrays allow for significant reductions in the required current stimulation?

More detail concerning penetrating arrays can be found in the following section.

### 7.9.9 Visual Cortex Implants

In the 1960s and 1970s, research efforts lead by Brindley in England and by Dobelle at the University of Utah used platinum surface electrodes embedded under the dura membrane to generate phosphenes using electrical stimulation. It was found that the currents required were extremely high, typically between 1 and 10 mA. This was completely impractical from a visual prosthetic perspective as the total current from an array of electrodes would reach dangerous levels that could produce seizures. The second problem was that the currents that generated phosphenes interacted in a nonlinear manner to displace the phosphenes if multiple electrodes were stimulated. This made generating percepts from multiple phosphenes impossible. However, the research did confirm the visuotopic structure of the human cortex.

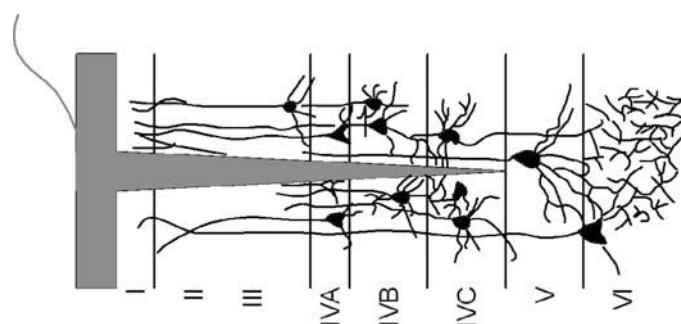
It had been speculated that as the normal input to the visual cortex was layer 4C, electrodes penetrating to this depth (as illustrated in Figure 7-52) would be far more effective in stimulating a more focused response. This hypothesis was confirmed when experiments showed that phosphenes could be evoked with currents of between 1 to 10  $\mu$ A and that two distinct phosphenes could be evoked using electrodes separated by only 500  $\mu$ m. In these experiments, individual electrodes were used, which is impractical if large numbers of electrodes must be inserted for visual stimulation (Finn and LoPresti, 2003).

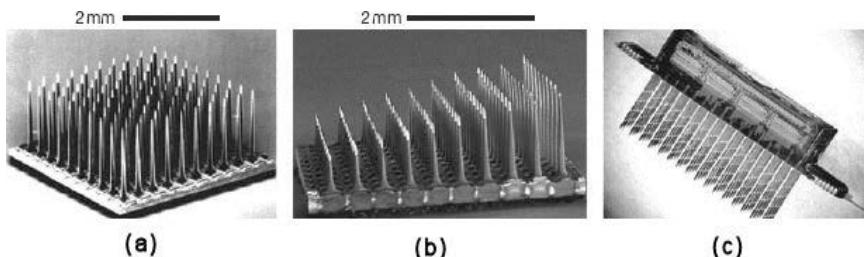
Studies of intracortical stimulation were initiated at Huntington Medical Research Institute (HMRI) in 1979, in which the feasibility of safe, chronic, intracortical stimulation of the cortex was established.

#### 7.9.9.1 Penetrating Electrode Arrays

Starting in 1983, work by Brummer and subsequent work by Robblee, Rose, Cogan, and others at EIC Laboratories eventually resulted in microelectrodes made from activated iridium that could be formed into low-density arrays. Implantation of 38 microelectrodes in a human volunteer at NIH in 1994 provided the motivation for the development of higher-density arrays.

**FIGURE 7-52 ■**  
Graphic showing the penetration of an electrode to cortical layer 4C.





**FIGURE 7-53** ■  
3D microelectrode arrays. (a) Utah electrode array.  
(b) Utah slanted electrode array.  
(c) Umich array.  
(Medscape 2008)

Researchers at the University of Utah–Salt Lake City and the University of Michigan–Ann Arbor have devised methods by which complex three-dimensional microelectrode arrays can be built. The Utah electrode array (UEA) and the Utah slanted electrode array (USEA) are two examples of such neural interfaces and are shown in Figure 7-53a and Figure 7-53b. A three-dimensional electrode array developed by researchers at the University of Michigan was built using more conventional integrated circuit technologies, shown in Figure 7-53c. It has multiple electrode sites that are distributed along a number of electrode shanks, with the planes of the electrode shanks integrated into a single-electrode array.

The UEA consists of 100 1.5 mm long microneedles that were designed to be inserted into the cerebral cortex to a depth of 1.5 mm, the level of normal neural input to the cerebral cortex. The electrodes of the UEA and USEA are built on a square grid with 400  $\mu\text{m}$  spacing. A total of 100 gold bond pads are deposited on the back surface of these arrays, and 100 thin insulated gold wires are bonded to these pads and to a percutaneous connector for connection to external electronics. The tip of each microneedle is metallized with iridium oxide to facilitate electronic to ionic transduction, and the entire array, with the exception of the tip of each microneedle, is insulated with a biocompatible polymer.

The USEA also has up to 100 microneedles, but their lengths are graded from 0.5 mm to 1.5 mm along the length of the array. The graded lengths of the USEA ensure that when it is inserted into a peripheral nerve the electrode tips uniformly populate the nerve, with most nerve fibers being no more than 200  $\mu\text{m}$  away from an active electrode tip (Medscape, 2008).

In the first animal trials, it was found that these arrays containing multiple electrodes could not be pushed into the cortex as they just indent it (like a fakir on a bed of nails). A high-velocity pneumatic insertion tool was developed that could insert the array in 200  $\mu\text{s}$ , corresponding to a velocity of 7 m/s.

### 7.9.9.2 Illinois Institute of Technology Research Group

IIT researchers lead by Philip Troyk started the development of a prototype intracortical visual prosthesis in 2002. The system uses four 256-channel electric stimulators sitting on a patient's skull under the skin to produce the tiny currents needed to drive 1024 miniature electrodes implanted in the visual cortex. Small in size (25  $\times$  25  $\times$  6 mm), each ceramic-encased stimulator module contains the electronics that control the 256 electrodes. The implant is powered and communicated with using transcutaneous magnetic induction.

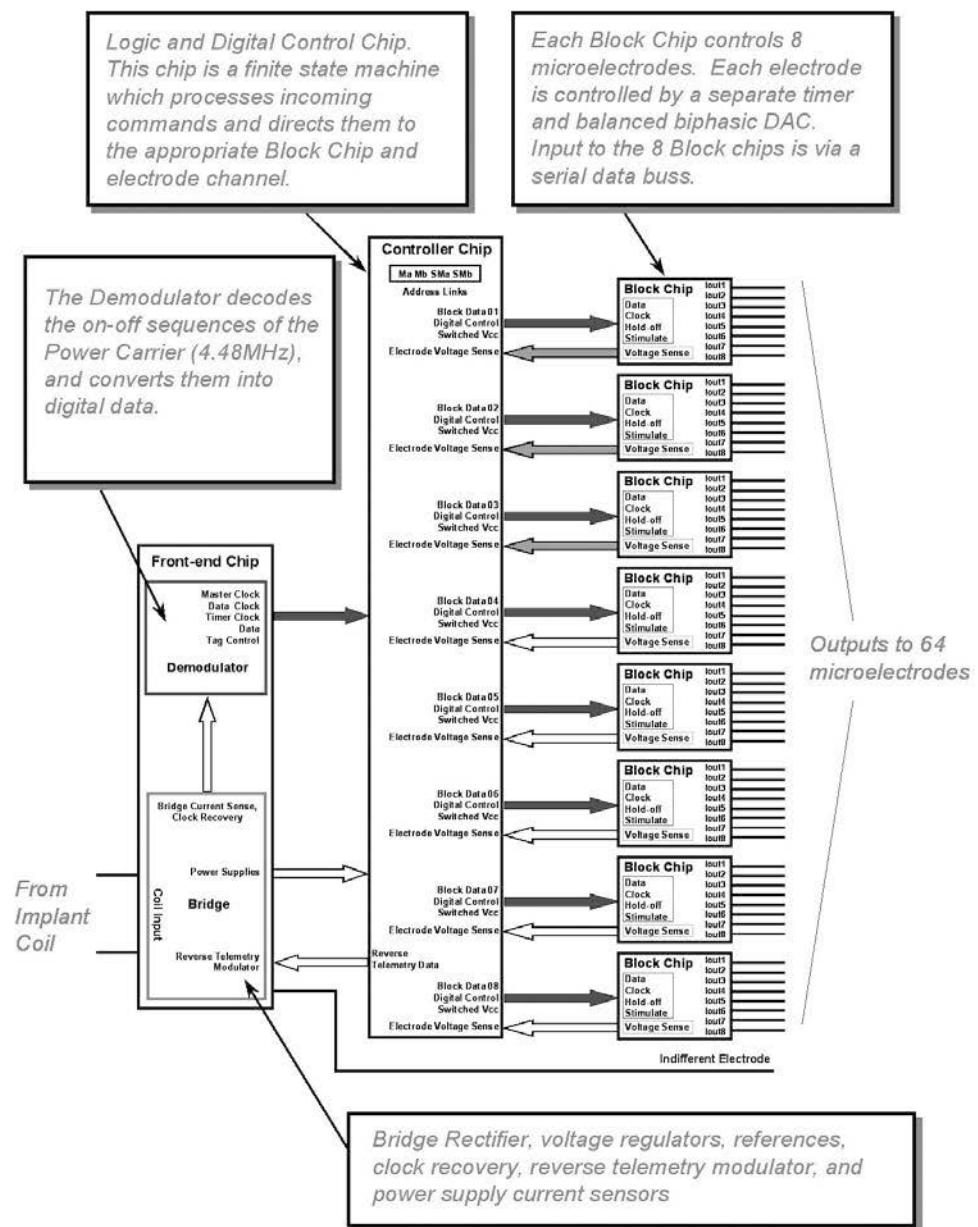
The approach to the design of most implantable devices has been to minimize the functional requirements of the implant itself. This philosophy is based on a desire to maximize the system reliability and operational flexibility. For any implant, the allowable size must be defined relative to its function, and it is not necessarily true that increased

function is associated with larger size. The nature of the power source frequently defines the entire system approach. If the implant requires constant communication, power and communication via an inductive link is often the method of choice. If the implant needs to operate without an external transmitter coil, then an implanted power source is needed.

For an implantable neural stimulator, key questions are as follows:

- Should the stimulator be capable of independent stimulation of the electrodes or should an extracorporeal controller command the implant for each individual stimulation pulse?

**FIGURE 7-54 ■**  
Block diagram of one submodule of the IIT visual prosthesis. (IIT 2002)



- Can the electrodes be multiplexed from a smaller set of current drivers, or should each electrode have its own independent output stage?
- Is true simultaneous stimulation required, or can a time-shared sequential command structure suffice?
- How many and what stimulation rates are required for the electrodes?

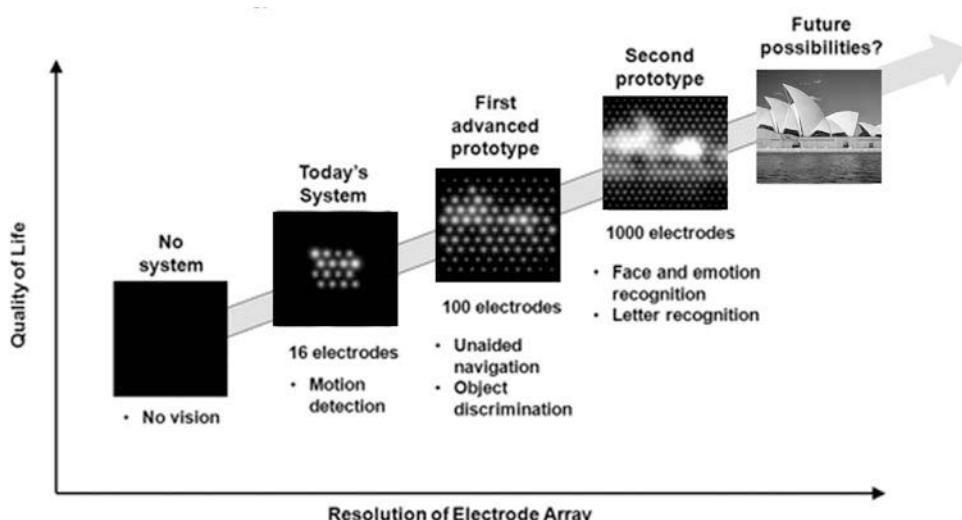
For an implantable neural-signal telemetry device, key questions are as follows:

- Should signal processing and filtering be accomplished within the implant, or should the raw action potentials be telemetered to an external receiver?
- What bandwidth is required for the telemetry, and how many channels are needed per implant?
- What is the allowable size of the implant?

The 1024-channel cortical stimulation system for the visual prosthesis consists of four 256-channel implantable stimulator modules, inductively powered by one external transmitter. Each 256-channel stimulator module contains four individual 64-channel submodules, shown in Figure 7-54.

## 7.10 | THE FUTURE

It is envisaged that as the numbers of electrodes and their density increases, along with improved methods of stimulation, the acuity of the generated image will improve along with the quality of life, as indicated in Figure 7-55. Existing 16-electrode prostheses allow patients to discriminate between light and darkness and to detect motion. However, by the time arrays of 1000 electrodes are available, patients will be able to recognize faces and identify individual letters. Ultimately, it may be possible to restore high-resolution monochrome vision.



**FIGURE 7-55** ■ Improvements in visual acuity that should derive from improvements in electrode numbers and density.  
(Courtesy of AVPG.)

## 7.11 REFERENCES

- Bach-y-Rita, P., C. Collins, F. Saunders, B. White and L. Scadden (1969). "Vision substitution by tactile image projection." *Nature* 221: 963–964.
- Becker, H., T. Jang et al. (2009). "Multi-Channel Vibrotactile Display." Retrieved September 2009 from [http://robotics.bu.edu/dupont/research\\_projects/vibrotactile/index.htm](http://robotics.bu.edu/dupont/research_projects/vibrotactile/index.htm)
- Bellis, M. (2008). "The History of Eye Glasses or Spectacles." Retrieved August 2008 from [http://inventors.about.com/od/gstartinventions/a/glass\\_3.htm](http://inventors.about.com/od/gstartinventions/a/glass_3.htm)
- Benjamin, J., N. Ali, et al. (1973). *A Laser Cane for the Blind*. Proceedings of the San Diego Biomedical Symposium, 12: 53–57.
- Biever, C. (2006). "Solar-Powered Implant Could Restore Vision." *New Scientist* 23, April.
- Bissitt, D. and A. Heyes. (1980). "An Application of Biofeedback in the Rehabilitation of the Blind." *Applied Ergonomics* 11(1): 31–33.
- Borenstein, J. and I. Ulrich. (1997). *The GuideCane—A Computerized Travel Aid for Active Guidance of Blind Pedestrians*. IEEE International Conference on Robotics and Automation, Albuquerque, NM.
- Brabyn, J. (1982). "New Developments in Mobility and Orientation Aids for the Blind." *IEEE Transactions on Biomedical Engineering* BME-29(4): 285–289.
- Brodal, A. (1969). *Neurological Anatomy in Relation to Chemical Medicine*, Oxford, Oxford University Press
- Bronzino, J. (Ed.). (2006). *Medical Devices and Systems*. Boca Raton, FL: CRC Press.
- Brooker, G. (2008). *Introduction to Sensors for Ranging and Imaging*. Raleigh, NC: SciTech.
- Burnside, N. (2004). "A Function that Returns the Atmospheric Attenuation of Sound," MATLAB Central, <http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=6000&objectType=FILE>.
- Capelle, C., C. Trullemans, et al. (1998). "A Real-Time Experimental Prototype for Enhancement of Vision Rehabilitation Using Auditory Substitution." *IEEE Transactions on Biomedical Engineering* 45(10): 1279.
- Cincotti, F., L. Kauhanen, et al. (2007). "Vibrotactile Feedback for Brain-Computer Interface Operation." *PubMed*, Retrieved September 2008 from <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2267023/>
- DeMarco, P., G. Yarbrough, et al. (2007). "Stimulation via a Subretinally Placed Prosthetic Elicits Central Activity and Induces a Tropic Effect on Visual Responses." *Investigative Ophthalmology & Visual Science (IOVS)* 48(2): 916–926.
- Efron, N. (1977). "Optacon—A Replacement for Braille?" *Australian Journal of Optometry* 60(118): 118–129.
- Finn, W. and P. LoPresti (Eds.). (2003). *Handbook of Neuroprosthetic Methods*. London: CRC Press.
- Foley, H. and M. Martin. (2006). "Sensation and Perception." Retrieved September 2009 from <http://www.skidmore.edu/~hfoley/Perc3.htm>
- Geldard, F. (1957). "Adventures in Tactile Literacy." *American Psychologist* 12: 115–124.
- IIT. (2002). "Intracortical Visual Prosthesis." Retrieved September 2008 from <http://neural.iit.edu/intro.html>
- K-Sonar, B. (2008). "Bay Advanced Technologies." Retrieved August 2008 from <http://www.batforblind.co.nz/>
- Kaczmarek, K. A., J. G. Webster, et al. (1991). "Electrotactile and Vibrotactile Displays for Sensory Substitution Systems." *IEEE Transactions on Biomedical Engineering* 38(1): 1.
- Kay, L. (1974). "A Sonar Aid to Enhance Spatial Perception of the Blind: Engineering Design and Evaluation." *Radio and Electronic Engineer* 44: 605–627.
- Knudsen, V. and C. Harris. (1950). *Acoustical Designing in Architecture*. New York: John Wiley & Sons.
- Kolb, H., E. Fernandez, et al. (2009). "Webvision—The Organization of the Retina and Visual System." Retrieved September 2009 from <http://webvision.med.utah.edu/>

- Lebedev, V. and V. Sheiman. (1980). "Assessment of the Possibilities of Building an Echo Locator for the Blind." *Telecommunications and Radio Engineering* 34–35(3): 97–100.
- Leonard, J. and H. Durrant-Whyte. (1991). *Simultaneous Map Building and Localisation for an Autonomous Mobile Robot*. IEEE International Workshop on Intelligent Robots and Systems.
- Levesque, V., J. Pasquier, et al. (2005). "Display of Virtual Braille Dots by Lateral Skin Deformation." *ACM Transactions on Applied Perception* 2(2): 132–139.
- Levy, D. (2005). "Scientists Design 'Bionic Eye' that Could Someday Help the Visually Disabled." *Stanford University News*, March.
- Lovell, N., L. Hallum, et al. (2007). Advances in Retinal Neuroprosthetics. In *Handbook of Neural Engineering*, M. Akay (Ed.). New York: Wiley-IEEE Press.
- Medscape. (2008). "The Neural Interface: The Utah Electrode Arrays." Retrieved September 2008 from [http://www.medscape.com/viewarticle/560817\\_2](http://www.medscape.com/viewarticle/560817_2)
- Meijer, P. (2006). "Seeing with Sound." Retrieved August 2008 from <http://www.seeingwithsound.com>
- Monkman, G. and P. Taylor (1993). "Thermal Tactile Sensing." *IEEE Transactions on Robotics and Automation* 8(3): 313–318.
- Optobionics. (2008). "ASR Device." Retrieved September 2009 from <http://optobionics.com/asrdevice.shtml>
- Pasquier, J. (2003). "STRESS: A Tactile Display Using Lateral Skin Stretch." McGill University, M.Eng.
- Pasquier, J. (2006). "Survey of Communication through Touch: TR-CIM 06.04." Retrieved August 2008 from [http://www.cim.mcgill.ca/~jay/index\\_files/new-research.htm](http://www.cim.mcgill.ca/~jay/index_files/new-research.htm)
- Robblee, L. and T. Rose (1990). *Electrochemical Guidelines for Selection of Protocols and Electrode Materials for Neural Stimulation, in Neural Prostheses: Fundamental Studies*. Eagle-Wood Cliffs, NJ, Prentice Hall.
- Rosenblum, L. (2010). *See What I'm Saying: The Extraordinary Powers of Our Five Senses*. New York: W.W. Norton and Company, Inc.
- Schubert, M. (1999). "Subretinal Implants for the Recovery of Vision." 1999 IEEE International Conference on Systems, Man, and Cybernetics, Tokyo.
- SensComp. (2004a). "40LT16." Retrieved October 2007 from <http://www.senscomp.com/specs/40LT16%20%20spec.pdf>
- SensComp. (2004b). "600 Series Transducer." Retrieved October 2007 from <http://www.senscomp.com/specs/600%20instrument%20spec.pdf>
- Storer, A. (2006). "Visual Perception through Sensory Substitution." Retrieved August 2008 from <http://cns.bu.edu/~storer/sensorysubstitution/Alex%20Storer%20-%20Final%20Paper%20-%20Updated.pdf>
- Trivedi, B. (2010). "Ear Today, Eye Tomorrow." *New Scientist*, August, pp. 43–45.
- Velazquez, R., J. Pissaloux, et al. (2004). "Design and Characterisation of a Shape Memory Alloy Based Micro-Actuator for Tactile Stimulation." *IEEE International Symposium on Industrial Electronics*, 1: 3–8.
- Walker, B. and J. Lindsay. (2006). "Navigation Performance with a Virtual Auditory Display: Effects of Beacon Sound, Capture Radius, and Practice." *Human Factors* 48(2): 265–278.
- Warren, D. and E. Fernandez. (2001). "High-Resolution, Two-Dimensional Spatial Mapping of a Cat Striate Cortex Using a 100-Microelectrode Array." *Neuroscience* 105(1): 19–31.
- Warren, D. and E. Strelow. (1984). "Learning Spatial Dimensions with a Visual Sensory Aid: Molyneux Revisited." *Perception* 13: 331–350.
- Weinstein, S. (Ed.). (1968). *Intensive and Extensive Aspects of Tactile Sensitivity as a Function of Body Part, Sex and Laterality*. Springfield, IL: Charles Thomas.
- Wicab. (2009). "Brainport Technologies—Vision Device." Retrieved September 2009 from <http://wicab.us/technology/vision.php>
- Wisc Edu. (2008). "Tongue Display Unit." Retrieved July 2008 from [http://kaz.med.wisc.edu/projects\\_tdu.php](http://kaz.med.wisc.edu/projects_tdu.php)



# Heart Replacement

## Chapter Outline

8.1	Introduction.....	396
8.2	The Heart as a Pump .....	397
8.3	Heart-Lung Machines.....	403
8.4	Artificial Hearts.....	408
8.5	Ventricular Assist Devices.....	417
8.6	Engineering in Heart Assist Devices.....	446
8.7	Pump Types .....	455
8.8	References .....	466



## 8.1 | INTRODUCTION

More than 2 millennia ago in about 350 BC, Aristotle considered the heart to be the source of all movement, as he considered the heart to link the soul with the organs of life. The Greeks called the pulse *sphygmos*, and therefore sphygmology deals with the theory of the pulse. At about the same time (280 BC) in ancient China, Wang Shu-he wrote ten books about the pulse.

The Greek physician Galen knew that blood vessels carried blood and identified venous (dark red) and arterial (brighter and thinner) blood, each with distinct and separate functions. He believed that growth and energy were derived from venous blood created in the liver, whereas arterial blood gave vitality by containing *pneuma* (air) and originated in the heart. Blood flowed from both of these organs to all parts of the body where it was consumed, and there was no return of blood to the heart or liver (Boylan, 2007).

In 1242, the Arab scholar Ibn Nafis became the first person to accurately describe the circulatory process as illustrated in Figure 8-1a. (Al-Ghazal, 2002). A hundred years later, in 1552, Michael Servetus described the same and Realdo Colombo proved the concept, but it remained largely unknown in Europe. It was another 100 years before William Harvey performed a sequence of experiments and announced, in 1628, the discovery of the human circulatory system as his own. Harvey published an influential book, the *Exercitatio Anatomica de Motu Cordis et Sanguinis in Animalibus*, about it, and the medical world started to take note (Shackleford, 2003).

It is apparent that the beating heart and its external manifestation, the pulse, have played an important role in the roots of Western and Eastern philosophy as well as medicine.

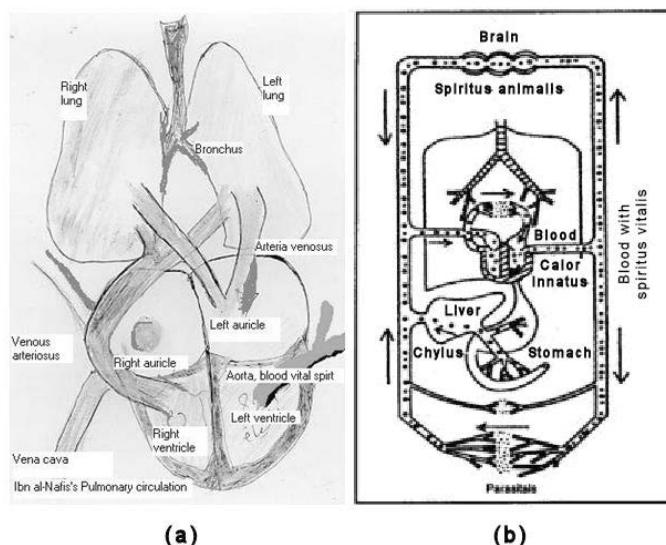
The cardiovascular system is composed of the heart, blood vessels, or vasculature as well as the cells and plasma that make up the blood. The blood vessels represent a closed delivery system, which functions to transport blood around the body, circulating substances such as oxygen, carbon dioxide, nutrients, hormones, and waste products.

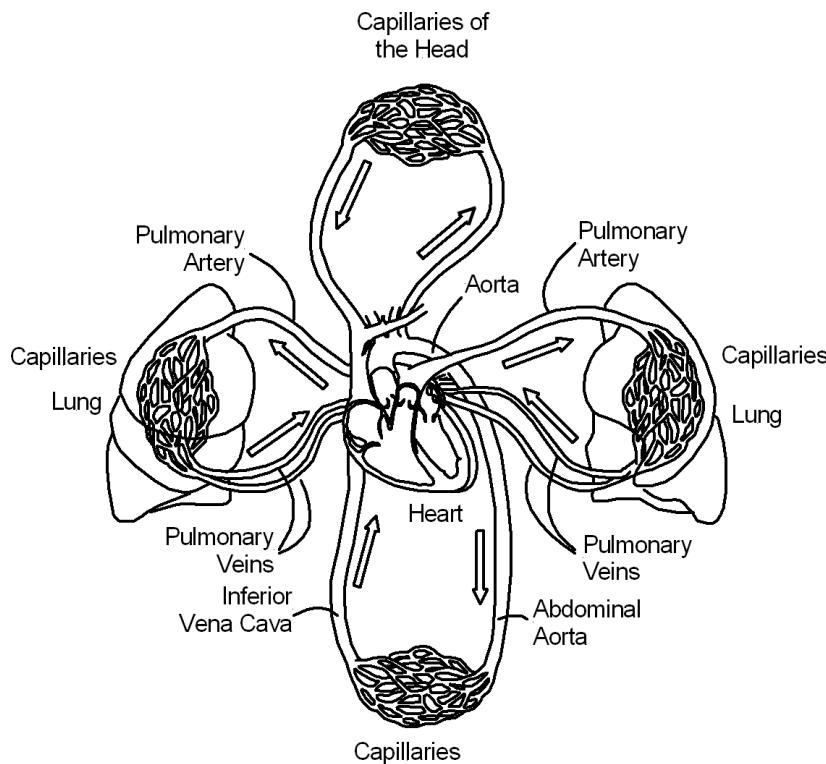
The principal function of the heart is to pump blood continuously around the cardiovascular system. The heart receives signals from both sympathetic and parasympathetic nerve fibers, which can alter the rate of the beat, but they do not initiate contraction. Instead, action potentials generated by autorhythmic cells create waves of depolarization that

**FIGURE 8-1 ■**

Blood circulation models

- (a) According to Ibn Nafis. (b) According to William Harvey. (Al-Ghazal 2002; Shackleford 2003), with permission.





**FIGURE 8-2** ■ The cardiovascular system.

spread to contractile cells via gap junctions. A dense network of blood vessels surrounds the heart and supplies the respiring cardiac tissues with essential oxygen and nutrients and also removes metabolic waste products. The heart's role as a pumping device is considered in more detail later in this chapter.

There are three main types of blood vessels:

- Arteries—the afferent blood vessels that carry blood away from the heart
- Veins—the efferent blood vessels that return blood to the heart
- Capillaries—narrow, thin-walled blood vessels that form networks within the tissues

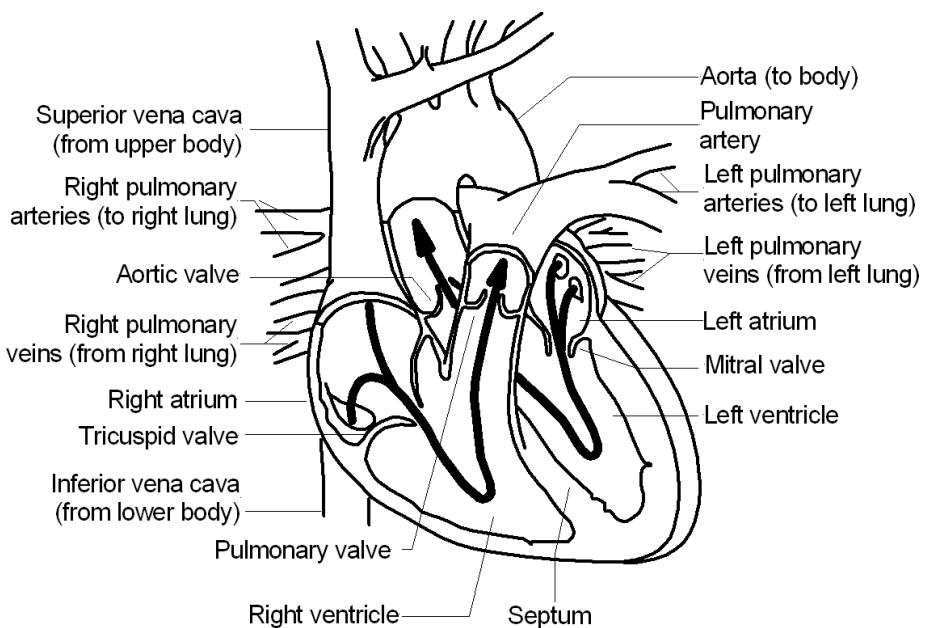
As shown in Figure 8-2, arteries branch and diverge as the distance from the heart increases. Smaller and smaller divisions are formed that eventually terminate in capillaries. By contrast, capillaries merge to form veins and further converge into successively larger blood vessels as the distance back toward the heart decreases. Capillary networks are the site of gas, nutrient, and waste exchange between the blood and the respiring tissues.

## 8.2 | THE HEART AS A PUMP

The heart can be considered a pair of pumps folded together to form a single unit, as shown in Figure 8-3:

- The right half of the heart pumps blood only to the lungs. Deoxygenated blood enters the right atrium through the superior and inferior vena cavae and then out from the right ventricle into the pulmonary arteries at low pressure (25 mm Hg).

**FIGURE 8-3 ■**  
Structure of the heart.



- The left half of the heart pumps blood to the rest of the body. Oxygenated blood enters the left atrium from the lungs through the pulmonary veins and then out from the left ventricle via the aorta at high pressure (120 mm Hg).
- Because the two halves of the heart are attached and connected both physically and electrically, they beat in synchrony.
- The left and right atria receive the incoming blood and pump it into the ventricles through the tricuspid and bicuspid (mitral) valves, respectively.
- The two ventricles produce enough pressure to push the blood out through the semilunar valves and through the pulmonary and systemic circulations.

There are generally no valves where the vena cavae join the right atrium or where the pulmonary veins enter the left atrium, because pressures in the atria are small and valves are not needed. However, in some adults there is a valve between the inferior vena cava and the right atrium.

### 8.2.1 Heart Valves

The heart valves are regulated by the pressure differential between their interior and exterior faces. There are two sets: the atrioventricular (AV) valves between the atria and the ventricles; and the semilunar valves, where the arteries leave the heart.

The AV valves have the following characteristics:

- They are the flap type.
- Chorda tendinae and papillary muscles ensure that they are not inverted by the pressure.
- The left half of the heart has a bicuspid (two-flap), or mitral, valve.
- The right half of the heart has a tricuspid (three-flap) valve.

The semilunar valves close when flow is reversed and blood behind the three cusps forces the lips together to generate a tight seal. The valve in the right half of the heart is called the pulmonary semilunar valve, and the one in the left half is called the aortic semilunar valve.

## 8.2.2 The Pump Cycle

The heart's pumping cycle can be divided into seven different stages, as shown in Figure 8-4.

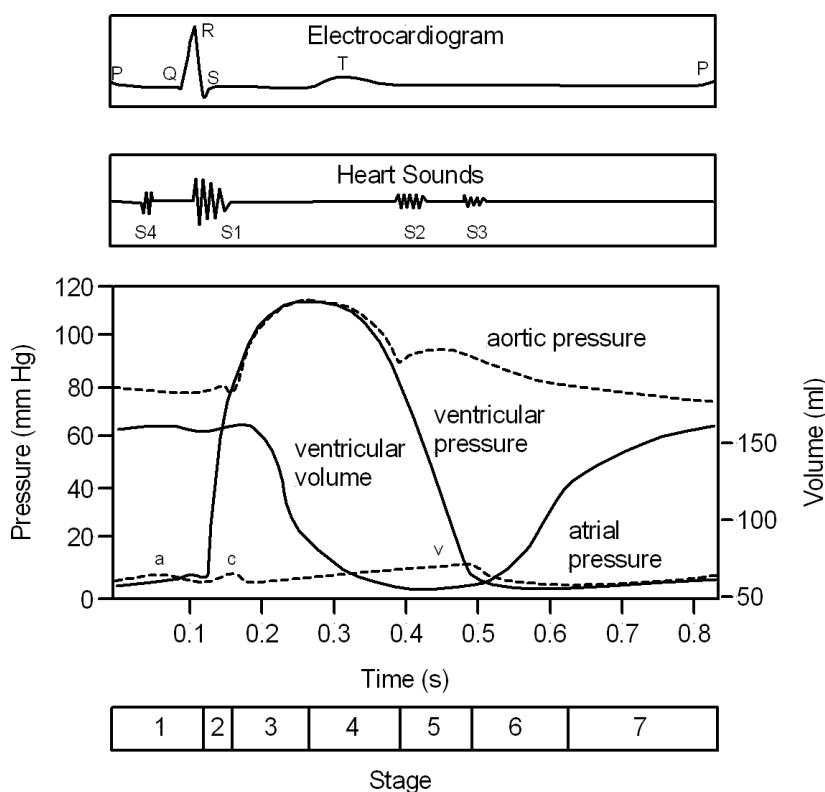
These stages are discussed briefly in the following sections.

### 8.2.2.1 Stage 1: Atrial Systole

Prior to atrial systole, blood has been flowing passively from the atrium into the ventricle through the open AV valve. During arterial systole, the atrium contracts and completes the filling process. The atrial pressure (Figure 8-4a) increases, and this forces blood into the ventricles as well as excess blood back up the jugular vein. This pressure drops when the atria stop contracting.

An impulse from the sinoatrial (SA) node starts the atrial contraction. The P-wave is due to this atrial depolarization.

A fourth heart sound (S4), shown in Figure 8-4, is abnormal and associated with the end of atrial emptying after atrial contraction. It is indicative of hypertrophic congestive heart failure, massive pulmonary embolism, or tricuspid incompetence.



### 8.2.2.2 Stage 2: Isovolumetric Contraction

The AV valves close at the start of this phase when the pressure in the ventricles exceeds the pressure in the atria, and the phase ends with the opening of the semilunar valves. As the ventricles contract, the internal pressure increases, but the volume does not change because blood is incompressible. The pressure increases until it reaches the pressure in the aorta and the pulmonary arteries, respectively.

Meanwhile, the electrical impulse propagates from the AV node through the bundle of His and fibers of Purkinje that cause the ventricles to contract—starting at the apex and moving rapidly toward the base. The QRS complex is caused by ventricular depolarization, and it marks the beginning of ventricular systole. The amplitude of this signal is so large that it masks the underlying atrial repolarization.

The first heart sound, (S1) “lub,” is due to the closing AV valves and associated blood turbulence.

### 8.2.2.3 Stage 3: Rapid Ejection

As the ventricles continue to contract, the pressure exceeds that in the aorta and pulmonary arteries. Both the aortic and pulmonary (i.e., semilunar) valves open, and blood is forced out of the ventricles. Ventricular volume decreases rapidly, but the pressure continues to increase as blood is forced into the arteries; the carotid pulse is thus produced.

Abnormal narrowing of the semilunar valves, called stenosis, causes a high-pitched sound at this time.

### 8.2.2.4 Stage 4: Reduced Ejection

After the pressure peak, blood flow out of the ventricles begins to decrease, and the rate of contraction of the ventricular volume is reduced. When the pressure within the ventricles drops below the arterial pressure, blood begins to flow back into the ventricles, which causes the semilunar valves to close. This marks the end of ventricular systole.

The T-wave marks ventricular repolarization.

### 8.2.2.5 Stage 5: Isovolumetric Relaxation

Throughout this stage and the previous two phases, the atria, in diastole, has been filling with blood on top of the closed AV valves. The V-wave is due to the backflow of blood after it strikes the closed AV valve. It is the second discernible jugular pulse. Meanwhile, the pressure in the ventricles continues to drop, and the ventricular volume reaches a minimum.

The second heart sound, “dub,” occurs when both of the semilunar valves close. This sound is split because the aortic valve closes slightly ahead of the pulmonary valve. Valve leakage results in a swishing sound (heart murmur).

### 8.2.2.6 Stage 6: Rapid Ventricular Filling

The AV valves open, and blood that has accumulated in the atria flows rapidly into the ventricles as their volume increases.

The third heart sound (S3) is usually abnormal and is associated with rapid passive ventricular filling. It occurs in dilated congestive heart failure, severe hypertension, myocardial infarction, or mitral incompetence.

### 8.2.2.7 Stage 7: Reduced ventricular Filling

The rate of ventricular filling decreases as the amount of blood in the atria is reduced.

### 8.2.3 The Cardiac Output

The cardiac output is normally about 5 L/min but can triple during strenuous exercise. This is the product of the stroke volume (the volume of a single output) and the heart rate. These are nominally 70 cm<sup>3</sup> per stroke at 70 beats per minute (BPM).

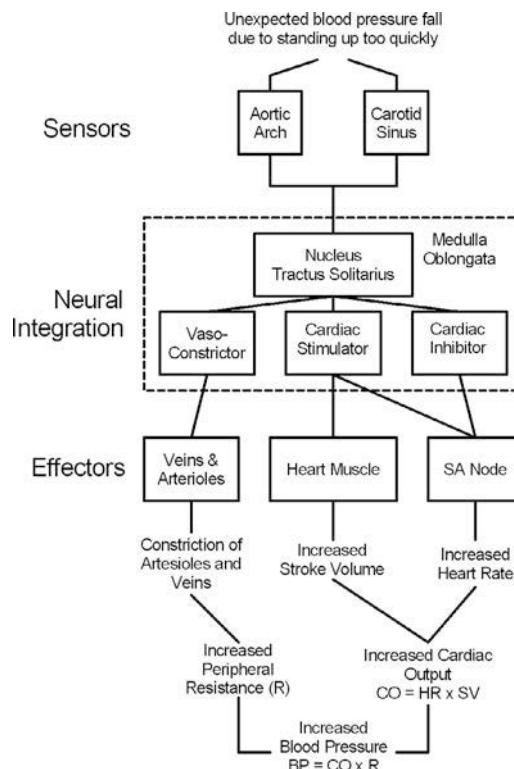
As the heart rate increases, the total output increases proportionally until it reaches about 200 BPM, when it does not have time to fill properly, so the maximum rate is limited to about 15 L/min.

### 8.2.4 Pressure Regulation

The blood pressure (BP) in the aorta alternates between a high pressure (systole) of about 120 mm Hg and low pressure (diastole) of about 80 mm Hg. This lower limit is determined by the elasticity of the ventricular walls. Blood pressure is reported as the systolic pressure over the diastolic pressure, typically 120/80.

Pressure will obviously be a function of the flow rate and the resistance to flow, and the body changes both the cardiac output (CO) and the resistance to maintain the required level, as shown in Figure 8-5. The baroreceptor reflex makes short-term adjustments to the blood pressure (BP) with a time constant of less than 1 minute. An example of this is the increase in heart rate (HR) and stroke volume (SV) that occurs when you stand up suddenly.

Pressure is measured by sensors in the arch of the aorta and the carotid sinus where nerves convey the information to the nucleus tractus solitarius of the medulla oblongata in the brain. Regulation is provided via the Vagus nerve, which slows the heart, and the accelerator nerve, which speeds it up (Orme, 2002).



**FIGURE 8-5 ■**  
Short-term pressure regulation. [Adapted from (Orme 2002).]

Long-term regulation is mediated mostly by the kidneys because they regulate the body's salt and water content, which control blood pressure. Sodium retention is controlled by the Na pump, as regulated by the hormone aldosterone. This is in turn regulated by the hormones rennin and angiotensin. If sodium is retained, the blood osmotic pressure rises, which causes water to be retained. Water reabsorption in the kidney is through water channels in the kidney tubules, and these are controlled by the antidiuretic hormone (ADH). If ADH is present in high concentrations, water absorption will be high and blood pressure will rise.

### 8.2.5 Heart Disease

Cardiovascular disease can be divided into two main categories: coronary heart disease (CHD) and congestive heart failure (CHF).

CHD is caused primarily by hardening of the arteries within the heart. This hardening process is due to deposits of fat and cholesterol (plaque) on the inner walls of the arteries, which reduces their diameter and impairs blood flow. The heart tries to compensate by pumping harder, but sufferers often exhibit symptoms of a lack of oxygen resulting in fatigue as well as severe chest pains.

Congestive heart disease arises when the heart no longer pumps blood efficiently. This leads to an accumulation of blood in the lungs—hence the congestion. Once again, the heart needs to work harder to provide sufficient oxygen to the body, which can lead to excessive wear and tear on the already enfeebled organ. As the disease progresses, patients suffer from shortness of breath and palpitations.

Many sufferers of heart disease are ineligible for heart transplants for the following reasons:

- Age of more than 65 years.
- High pressure in the lung arteries due to permanent changes in the lung blood vessels.
- Irreversible kidney or liver dysfunction not caused by underlying heart failure.
- Symptomatic arterial diseases in the legs, kidneys, neck, or brain.
- Severe chronic lung disease including emphysema, asthma, and chronic bronchitis.
- Chronic infection in the blood, lung, urine, or elsewhere or open wound.
- Insulin-dependent diabetes with evidence of damage to other organs, such as kidney, retina, or nerves.
- Cancer within the past 5 years. Exceptions may be made for some types of early skin cancer or under other very unusual circumstances.
- Other life-threatening diseases likely to severely limit length or quality of life even if the transplant were successful.

In some of these cases, the best alternative is a total artificial heart or a left ventricle assist device (LVAD).

### 8.2.6 Biomechatronic Perspective

From a biomechatronic perspective, the cardiovascular system offers a rich environment for measurement, stimulation, and the development of various prostheses. The state of the cardiovascular system can be determined by measuring the changing blood pressure as the

heart beats, changes in the electrical potential due to this beating, and the concentrations of various gases and nutrients in the blood. If these parameters are out of specification, then various forms of intervention can be applied. For example, internal or external electrical stimulation can be used to restart the heart or to maintain a regular heartbeat. In extreme cases, if the heart has been damaged mechanical pumps can augment its capability or even replace it completely. This chapter documents some of these applications.

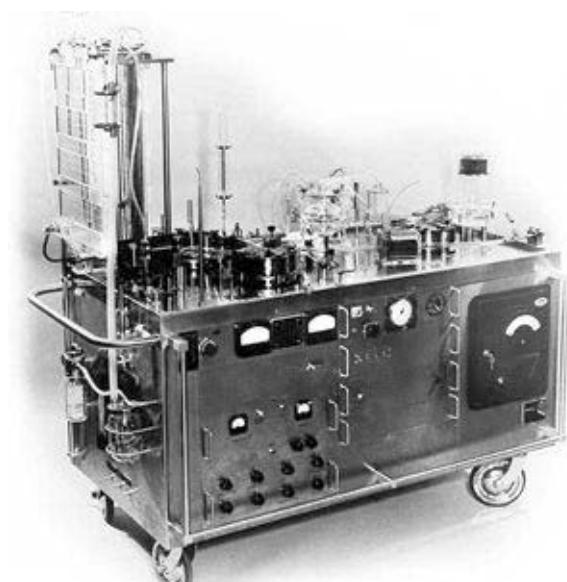
## 8.3 | HEART–LUNG MACHINES

The purpose of a heart–lung machine is to remove oxygen-poor blood from the right side of the heart and return oxygen-rich blood to the left side. This bypasses the heart and lungs and allows the heart to be stopped for short periods, allowing surgery to take place.

### 8.3.1 History

The first heart–lung machine was built by physician John Gibbon in 1935 and is shown in Figure 8-6. He is therefore considered the inventor of the heart–lung or pump oxygenator. This experimental machine used two roller (peristaltic) pumps and had the capacity to replace the heart and lung action of a cat and keep it alive for 26 minutes. Further work on the device was postponed by WWII.

Gibbon joined forces with Thomas Watson in 1946 when Watson, an engineer and the chair of IBM, provided the financial and technical support for Gibbon to further develop his heart–lung machine. Gibbon, Watson, and a number IBM engineers designed an improved machine that minimized hemolysis and prevented air bubbles from entering the circulation. The new device used a refined method of cascading the blood down a thin sheet of film for oxygenation rather than the original whirling technique, which could potentially damage blood cells. Using the new method, 12 dogs were each kept alive for more than 1 hour



**FIGURE 8-6** ■ The first heart–lung machine. (Courtesy of the Mayo Clinic.)

during heart operations. The device was only ever tested on dogs and had a 10% mortality rate.

Some other improvements had been introduced by another engineer in 1945 when Clarence Dennis built a modified Gibbon pump that permitted a complete bypass of the heart and lungs during surgical operations on the heart. Unfortunately, the machine was hard to clean and caused infections, so it never progressed beyond animal trials.

At about the same time, Swedish physician Viking Bjork designed an oxygenator with multiple screen disks that rotated slowly in a shaft, over which a film of blood was injected. Oxygen was passed over the rotating disks and provided sufficient oxygenation for an adult human. Bjork, with help from a few engineers one of whom who was his wife, developed a blood filter and an artificial intima of silicon under the trade name UHB 300. This was applied to all parts of the perfusion machine, particularly the rough red rubber tubes, to delay clotting and save platelets. It was effective, and Bjork took the technology to the human testing phase.

Gibbon performed his first human operation in February 1952, using his machine on a 15-month-old girl with an alleged atrial septal defect. At this time, cardiac catheterization was a major event, and the patient was too small to have a catheterization before surgery to confirm the diagnosis. Unfortunately, the girl died on the operating table because she did not actually have an atrial septal defect but rather a left-to-right shunt.

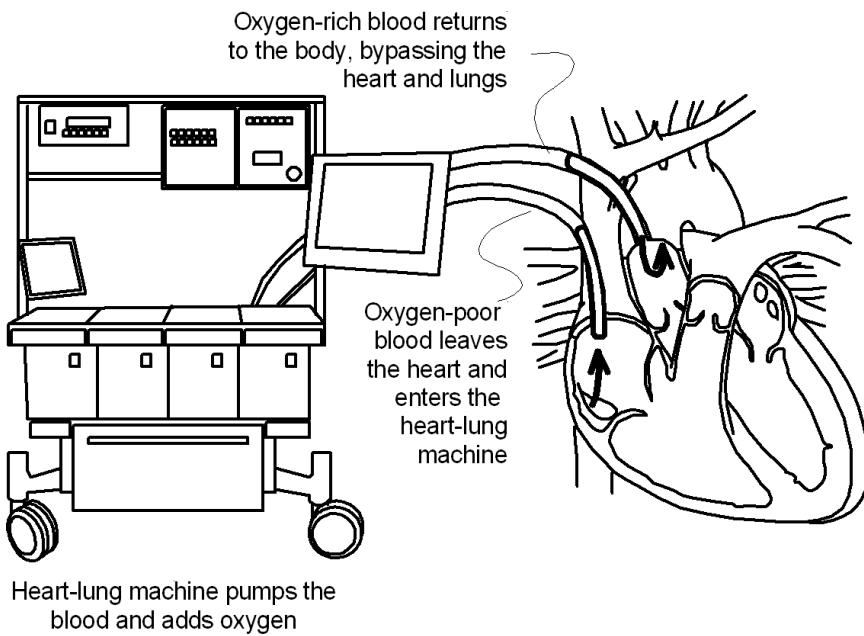
The second operation Gibbon performed was on May 6 of the same year, and it was successful. The patient was 18-year-old Cecelia Bavolek, who had a large left-to-right shunt at the atrial level as revealed by catheterization. After complete heparinization to minimize the likelihood of blood clots, the arterial inflow cannula was placed in the left subclavian artery, and the inferior and superior vena cava were cannulated with plastic tubes. All this was done through a large, bilateral submammary incision, which lifted up the entire upper thorax to expose the heart. The large atrial septal defect was closed with a running cotton suture, and the patient was removed from the heart-lung machine without incident after approximately 26 minutes. She made an uneventful recovery and was discharged 13 days later. Recatheterization 6 months later confirmed that the defect had closed completely. The case made headline news around the world (Cohn, 2003).

On July 3, a heart pump developed for Forest Dodrill by General Motors Research Labs known as the Dodrill-GMR Mechanical Heart was used at the Harper Hospital to keep Henry Opitek alive for 50 minutes while he underwent cardiac surgery. The pump was used to bypass his left ventricle while surgeons repaired the mitral valve in his left atrium (Stephenson, 2002).

The heart measured  $25 \times 30 \times 43$  cm and looked (as you would expect) like a 12-cylinder engine with six separate chambers. It was made from stainless steel, glass, and rubber and was powered by air pressure to circulate blood from the chambers through the patient's body (GMR, 1952).

### 8.3.2 Modern Heart–Lung Machines

One of the problems with heart–lung machines is that foreign surfaces within them activate blood coagulation proteins and platelets that lead to clot formation. As venous and arterial cannulae are inserted, anticoagulants such as heparin are administered that prevent clot formation and allow blood to flow more freely through the machine.



**FIGURE 8-7** ■  
Schematic diagram showing the basic principles of a heart-lung machine.

Large veins and arteries are required to insert the large-bore cannulae that will carry the blood away from the patient to the heart–lung machine and then return the blood from the machine to the patient as shown in Figure 8-7. Sites for venous access can include the inferior and superior vena cavae, the right atrium, the femoral vein (in the groin), or internal jugular vein. Oxygen-rich blood is returned to the aorta, femoral artery, or carotid artery (Enotes, 2002).

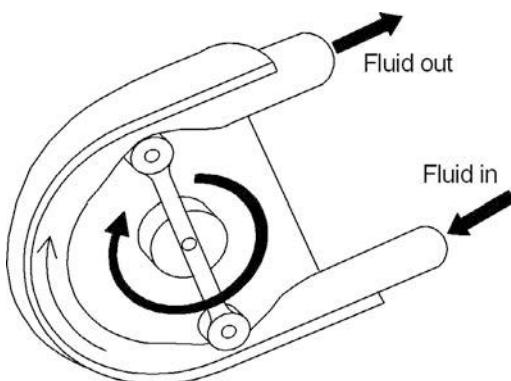
The standard heart–lung machine includes up to five pump assemblies. A centrifugal or peristaltic pump can be used to drive blood circulation. The four remaining pumps are roller pumps that provide fluid, gas, and liquid for delivery or removal to the heart chambers and surgical field:

- Left ventricular blood removal is accomplished by a roller pump that draws blood away from the heart.
- Suction created by a roller pump removes accumulated fluid from the general surgical field.
- A cardioplegia delivery pump is used to deliver a high-potassium solution to the coronary vessels. The potassium stops the heart beating during the surgical procedure.
- Finally, an additional pump is available for emergency backup of the arterial pump in case of mechanical failure.

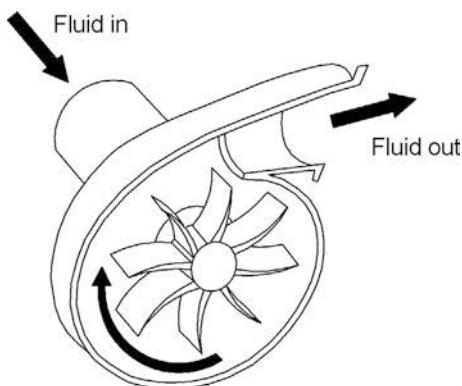
In peristaltic pumps, also known as roller pumps, the roller assembly rotates and engages the tubing, PVC, or silicon, which is then compressed against the pump's housing, propelling blood ahead of the roller head, as shown in Figure 8-8. Rotational frequency and the inner diameter of the tubing determine blood flow. Because of its occlusive nature, the pump can be used to remove blood from the surgical field by creating negative pressure on the inflow side of the pump head.

**FIGURE 8-8 ■**

Schematic diagram of a peristaltic (roller) pump.

**FIGURE 8-9 ■**

Cutaway diagram of a centrifugal pump.



In a centrifugal pump, blood is propelled outward and released to the outflow pipe tangential to the pump housing, which creates a region of low pressure in the center and draws blood into the pump, as illustrated in Figure 8-9. As discussed in more detail later in this chapter, rotational speed determines the blood flow rate, which can be measured by a flowmeter placed at the inlet or outlet. Care must be taken to maintain a reasonable rotational rate, because if it is too low blood may flow in the wrong direction since the system is nonocclusive in nature.

A reservoir collects blood drained from the venous circulation through cannulae. Reservoir designs include open or closed systems. The open system is graduated so that the blood volume in the container can be determined, and, as the name implies, the container is open to atmosphere, allowing blood to interface with atmospheric gases. In contrast, the pliable bag of a closed system eliminates the air–blood interface while still being exposed to atmospheric pressure. Volume is measured by weight or by change in radius of the container. As an additional safety feature, the closed reservoir collapses when emptied.

Gas is bubbled through the blood while it is in the reservoir, and oxygen and carbon dioxide are exchanged across the boundary layer of the blood and gas bubbles. The blood then passes through a filter coated with an antifoam solution, which helps to remove fine bubbles. As blood pools in the reservoir, it has already exchanged carbon dioxide and oxygen and is bright red. From here, tubing carries the blood to the rest of the heart–lung machine.

An alternative to this technique is the membrane oxygenator in which oxygen-poor blood from the reservoir is pumped past a membrane that separates the blood from the

ventilation gases. Oxygen diffuses from the gas mixture into the blood, and carbon dioxide diffuses the other way.

When blood is ready to be returned from the heart–lung machine to the patient, it must first pass through an arterial line filter. This device is used to filter small air bubbles that may have entered or been generated by the heart–lung machine. Finally, the blood is returned to the arterial cannula and back into circulation.

In addition to the previously described flow process, fluid being returned from the left ventricle or via surgical suction requires filtration before being reintroduced into the heart–lung machine. Blood enters a filtered reservoir, called a cardiotomy, which is connected with tubing to the venous reservoir. Other fluids such as blood products and medications are also added into the cardiotomy for filtration of particulates.

A heat exchanger allows body and organ temperatures to be controlled. The simplest heat exchange design is a coiled tube immersed in a water bath. As the blood passes through the coil, the blood temperature will shift toward that of the water. A more sophisticated system separates the blood and water with a metallic barrier. Once the blood is returned to the body, it will alter the temperature of the surrounding tissue. A closed-loop control system adjusts the water temperature to maintain body temperature.

Because respiration is being controlled and a machine is meeting metabolic demand, it is necessary to monitor the patient's blood chemical makeup. Chemical sensors placed in the blood path detect the amount of oxygen bound to hemoglobin, while other, more elaborate sensors constantly measure and plot the blood pH, partial pressure of oxygen and carbon dioxide, and electrolyte levels. This constant trending can quickly analyze the metabolic demands of the body to allow the appropriate adjustments to be made (Encyclopedia of Surgery, 2007).

From this description and the photograph shown in Figure 8-10, it is obvious that heart–lung machines have progressed significantly since their debut in 1952 and are now extremely sophisticated pieces of equipment.



**FIGURE 8-10 ■**  
Modern heart–lung machine. (Courtesy of Jörg Schulze with permission.)

## 8.4 | ARTIFICIAL HEARTS

An artificial heart is an implantable device that replaces all of the pumping actions of a natural heart. It is known as a totally artificial heart (TAH) and includes two pumping chambers that replace both the right and left ventricles. The earliest devices, such as the Liotta-Cooley and Jarvik-7 hearts, were driven by large air pumps through air lines, as shown in Figure 8-11.

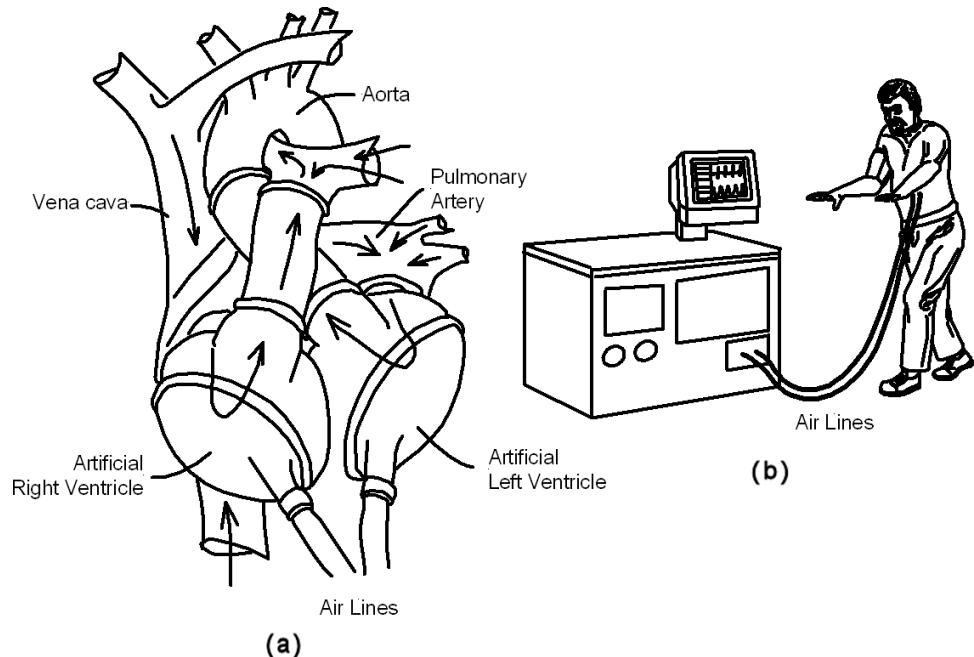
A successful replacement for the natural heart must meet a number of basic requirements. It must be sufficiently small for implantation but still capable of pumping the required blood volume, which is typically between 5 and 10 L/min. It must not produce hemolysis or result in the formation of thromboemboli, and its materials must withstand long-term flexing.

Current pumps use either a flexible sac acted on by a mechanical pusher plate or a rigid chamber divided by a flexible diaphragm separating the blood from either compressed air or hydraulic fluid. The blood contacting surfaces of these devices are commonly made from three types of materials, each with their own advantages and disadvantages: (1) smooth polyurethane; (2) a texturized pseudointima-forming surface; and (3) surfaces coated with biolized or natural tissue coatings (Jaron, 1990).

The choice of energy source is an ongoing design trade-off. It has long been known that pumps using compressed air supplied from outside the body are the simplest to design and control, and all of the early TAHs discussed in this chapter used this technique. However, these are not ideal, as the necessary connecting tubes provide a path for infection and a bulky external compressor restricts patient mobility.

Electrically driven pumps have many advantages for both TAHs and LVADs; therefore, considerable effort has been expended into finding methods of supplying electrical power to implanted pumps. At present, battery technology is not sufficiently advanced to supply

**FIGURE 8-11 ■**  
Pneumatically driven  
artificial heart.  
(a) Schematic  
diagram of the heart.  
(b) Drawing of the  
pump console and  
attachments to a  
human patient.



power from a totally self-contained device for more than 30 minutes or so; thus, external power must be provided at all times. This can be through transcutaneous electrical cables that leave the patient prone to infection or via electromagnetic induction through the skin.

An interesting alternative is the distributed artificial heart (DAH). Instead of inserting a single high-capacity ventricular pump, smaller pumps supply the brain, limbs, and internal organs separately. Because these devices are low power and well separated, heat dissipation is no longer an issue. Disadvantages include a higher component count with its associated reliability concerns and more hardware that must be accommodated in the body (Abe, Ono et al., 2000).

### 8.4.1 History

After emigrating from Holland, Willem Kolff joined the Cleveland Clinic as a research assistant. Within 7 years he and Dr. Tetsuzo Akutsu were testing primitive artificial hearts in animals to identify problems that might be encountered if such a device were to be later implanted in a human patient.

In 1963 the first patented artificial heart was developed by ventriloquist Paul Winchell, with help from Henry Heimlich (for which the Heimlich maneuver is named). Winchell subsequently assigned the patent to the University of Utah, where Robert Jarvik ultimately used it as the model for the Jarvik-7. This generation of artificial hearts were all powered pneumatically and required large external compressors and control systems.

#### 8.4.1.1 Liotta-Cooley Heart

The Liotta-Cooley heart, shown in Figure 8-12, was the first temporary artificial heart implanted in a human being. It was developed by Domingo Liotta and implanted by surgeon Denton Cooley of the Texas Heart Institute on April 4, 1969. The recipient, Haskell Karp, lived for 64 hours with the artificial heart until a human heart became available for transplant (Cooley, 2003).

The heart was a pneumatic, double-ventricle pump with Wada-Cutter hingeless valves to control the direction of blood flow. The two pump chambers (ventricles), the cuff-shaped



**FIGURE 8-12 ■**  
Photograph of  
the original  
Liotta-Cooley heart  
now in the National  
Museum of  
American History.

inflow tracts (atria), and the outflow tracts were lined with a special fabric that promoted the formation of a smooth cellular surface (pseudoneointima-forming surface). The flexible inflow and outflow tracts were made of Dacron fabric, and the pump chambers were made of a combination of Dacron fabric and Silastic.

The pumps were connected to the external power unit with Silastic tubing covered by Dacron. The console, also a major engineering accomplishment at the time, was about the size of a large washing machine. Two pneumatic power units generated the pumping and vacuum actions needed to move blood through the artificial heart. An adjacent control panel could be used to adjust pumping rate and pumping pressure (Cooley, Liotta et al., 1969).

Although Karp died of kidney failure 3 days after receiving a real heart, the procedure demonstrated the viability of artificial hearts as a bridge to transplant (BTT) in cardiac patients. However, some criticized the surgery as unethical because it was performed without formal review by the medical community.

Karp was the only person ever to have a Liotta-Cooley heart implanted, and even though it was a success at sustaining the patient until transplantation, many scientists considered this step of technology too risky. As a result, enough controversy ensued to postpone the use of artificial hearts for the next 20 years. During this hiatus, the development of LVADs was pursued aggressively, and it was only in 1981 that Cooley again demonstrated the potential of a TAH when an Akutsu-III was implanted to provide temporary support until a suitable donor could be located.

#### 8.4.1.2 The Jarvik-7 Artificial Heart

In 1967, Kolff left Cleveland Clinic with engineer Thomas Kessler and surgeon Clifford Kwan-Gett to start the Division of Artificial Organs at the University of Utah. Initial animal trials were not very successful. Even as late as 1970 the best he had managed was a sheep that lived 50 hours with an artificial heart, and it could hardly even lift its head. The Jarvik-3, developed in 1972, was the first of the range developed as a human implant. Cows implanted with these devices lived for up to 4 months with little or no medical intervention. By 1976 the design had been further improved, and the Jarvik-7 had kept a calf alive for 268 days (Hajar, 2005). Three years later, when surgeon William DeVries returned to Utah, animals were walking around, and, according to DeVries, they looked normal except that they were connected to a machine. He was certain that the time was ripe to initiate human transplants.

DeVries spoke to the technician who had made hundreds of excellent artificial hearts for animals, but he was reluctant to move on to human trials, saying, “My God, I can’t do that; these are just for animals. I can’t do that. I mean, I can’t make one good enough.” However, in the end he was persuaded, and in 1982 he and Lyle Joyce implanted the Jarvik-7 in Barney Clark, a Seattle dentist (Hajar 2005).

Clark, who due to his age and severe emphysema had not been a candidate for a transplant operation, was never able to leave the hospital. The system was open to infection, so Clark and subsequent Jarvik-7 recipients got sick. Patients had to be kept on blood thinners to prevent clots and strokes. Although Clark was reported to be in stable condition 48 hours after the implant, his subsequent postoperative condition was not good (Hajar, 2005):

- Day 3: He underwent thoracic exploratory surgery because of subcutaneous emphysema.
- Day 6: He had generalized seizures that left him in a coma.



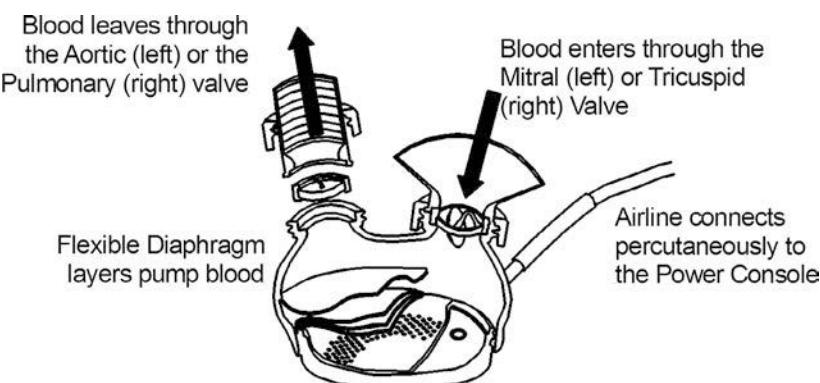
**FIGURE 8-13 ■**  
Photograph of the Jarvik-7 artificial heart. (Courtesy National Institutes of Health [NIH] archives.)

- Day 13: The mitral valve on his heart malfunctioned, and he was taken back into the operating room to replace the left ventricle of the artificial heart.
- Other complications about this time were recurrent pulmonary insufficiency, several episodes of acute renal failure, episodes of fever of unidentified cause, hemorrhagic complications of anticoagulation including continuously bleeding nose, and respiratory failure requiring a tracheostomy.
- Day 92: Diarrhea and vomiting leading to aspiration pneumonia and sepsis.
- Day 112: Death preceded by progressive renal failure and refractory hypotension.

The autopsy revealed extensive pseudo-membranous colitis, acute tubular necrosis, peritoneal and pleural effusion, centrilobular emphysema, and chronic bronchitis with fibrosis and bronchiectasis. However, the Jarvik-7 was still intact and uncontaminated by thrombosis or infection (Dutton, Preston et al., 1988; Hajjar, 2005).

The Jarvik-7, shown in Figure 8-13, is made of polyurethane, polyester, plastic, and aluminum and was designed for permanent implantation. It consists of two identical pneumatically driven pumps that replaced the left and right ventricles of the heart.

Each of the pump sections consists of a chamber divided by a flexible multilayered diaphragm, as shown in Figure 8-14. Direction of blood flow is controlled by tilting-disk valves, held in place by a polycarbonate ring structure. External arterial and venous connections are made using quick-connect cuffs. One of the innovations of the device is the inner coating of rough material, developed by David Gernes. This coating helped the blood to clot and coat the inside of the device, enabling a more natural blood flow. The total weight of the device is about 800 g, and it requires 520 cm<sup>3</sup> of space within the thorax (Jaron, 1990).



**FIGURE 8-14 ■**  
Schematic diagram of one pump of the Jarvik-7. [Adapted from (Hajjar 2005).]

A 2 m long external air line feeds compressed air to the chamber where changes in the pressure flex the diaphragm cyclically to drive blood flow. These air lines are attached to the large compressor and control console, which regulates the pump stroke and dictates the pumping rate.

After Clark's operation and notwithstanding his less than satisfactory quality of life posttransplant, the Jarvik-7 heart was implanted many times. The record for being sustained by this artificial heart is held by William Schroeder, who was hooked to a Jarvik-7 in 1985. He lived for 18 months with a far better quality of life, though he did suffer strokes, sudden hemorrhages, and infections during his final days. A 1985 study of 15 patients with active infections from device implantation showed mortality rates as high as 70%. Most infections developed from the site of the percutaneous tubes passing into the body, though some arose from blood clots that developed from the internal surfaces of the pump (Lemelson-MIT, 2002).

As materials improved, the survival rate improved, but after about 90 people had received the Jarvik device the implantation of artificial hearts was banned for permanent use in patients with heart failure because most of the recipients could not live more than half a year and their quality of life was poor. However, for some time after the ban it was still used as a BTT device.

Hiroaki Harasaki of the Cleveland Clinic developed two important improvements for the artificial heart and other potential artificial organs. The first was a nonclotting surface material that significantly reduced the risk of rejection of the organ by the patient's immune system. The second development, which required the collaboration of many disciplines, was an implantable power source that did not create tissue-damaging heat.

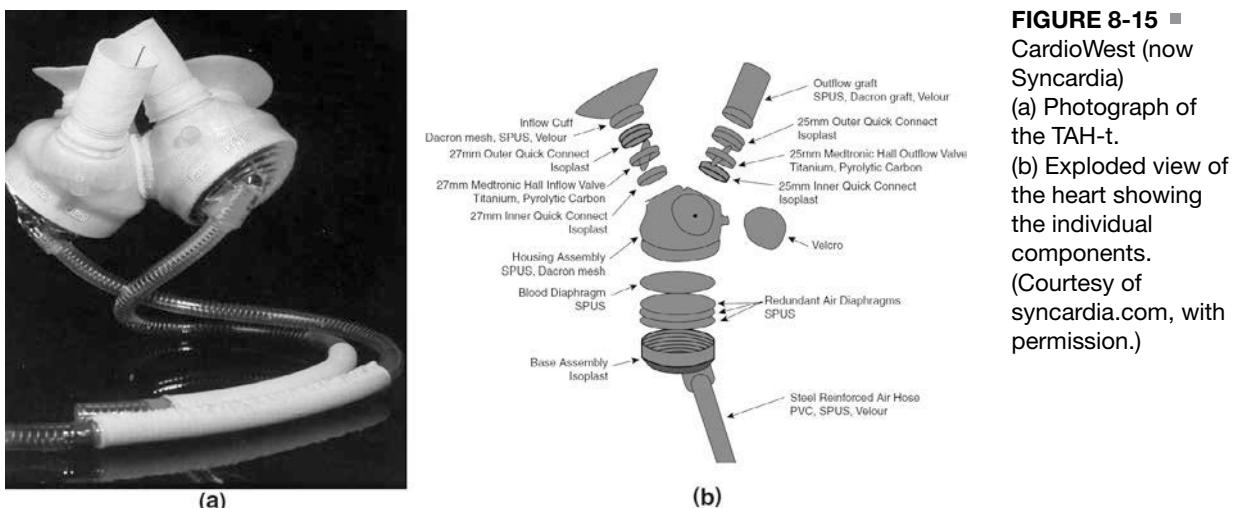
As the technical challenges were overcome, more companies started to manufacture totally implantable artificial hearts. These companies include Thoratec, Medquest, Baxter Novacor, Syncardia Systems, and AbioMed. To date, all of the problems have not been solved, and researchers continue to work on designs for an artificial heart that could provide a realistic, permanent option for survival of patients not considered suitable for heart transplants.

#### 8.4.1.3 Syncardia Systems/CardioWest

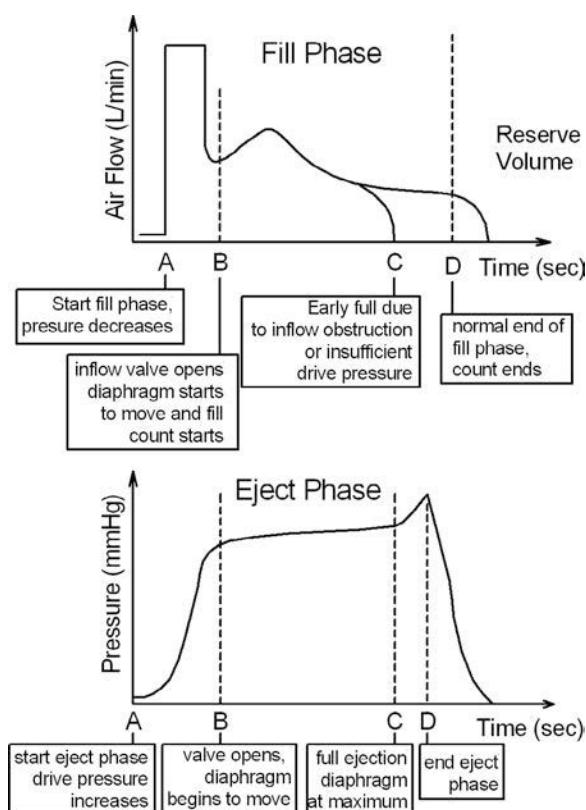
The CardioWest (now Syncardia) temporary total artificial heart (TAHt), shown in Figure 8-15, was developed from the Jarvik-7 by University of Arizona researchers and approved for use in 2004. It is the first implantable artificial heart to be approved by the U.S. Food and Drug Administration (FDA) and has also been approved by the Conformité Européenne (CE) (Europe) and Health Canada (NationMaster Encyclopedia, 2008).

The CardioWest heart is a pulsating biventricular device that is implanted into the chest to replace the patient's left and right ventricles. Compared with the Jarvik-7, it is very light, weighing only 160 g, but in other respects it is very similar to its predecessor. It is lined with polyurethane and has a four-layer pneumatically driven diaphragm. Four Medtronic-Hall mechanical valves ensure that blood flows correctly through the device.

The pneumatic control system is set to fully eject all of the blood from each ventricle with each beat. This is achieved by setting the ejection pressure of the right ventricle to 30 mmHg higher than the pressure in the pulmonary artery and that of the left ventricle 60 mmHg higher than systemic pressure, as shown in Figure 8-16. The ventricles are adjusted to fill to between 50 and 60 ml to allow for some overhead during the fill phase. This produces between 7 and 8 L/min while maintaining the correct Starling law pressure



**FIGURE 8-15 ■**  
CardioWest (now Syncardia)  
(a) Photograph of the TAH-t.  
(b) Exploded view of the heart showing the individual components.  
(Courtesy of syncardia.com, with permission.)



**FIGURE 8-16 ■**  
Details of the CardioWest TAH fill and eject phases.  
[Adapted from (Slepian, Smith et al., 2006).]

differential. At the maximum stroke volume of 70 ml and a rate of 130 BPM, the artificial heart can pump over 9 L/min (Slepian, Smith et al., 2006).

The device is a BTT for patients who do not respond to other treatments and who are at risk of imminent death from nonreversible biventricular failure. The FDA approval was based mainly on the results of a 10-year pivotal study of the artificial heart in 81 patients at high risk of death. The rate of survival after transplantation was 79%, compared with 46%

in a group of control patients who did not receive the artificial heart. The 1-year survival rate among patients who received the artificial heart was 70% compared with 31% among the controls. The 1- and 5-year survival rates among transplant recipients were 86% and 64%, respectively.

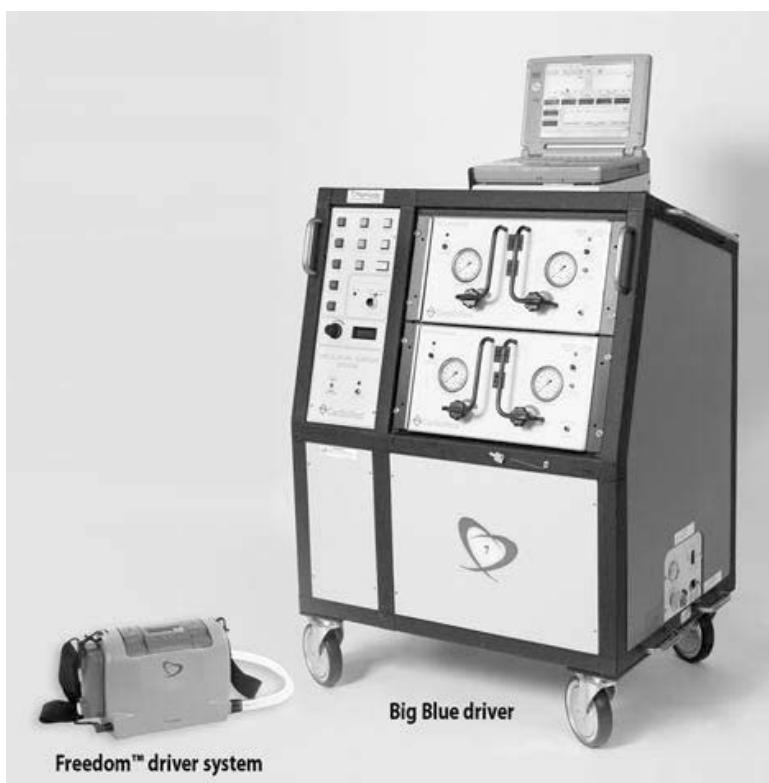
Complications included infection (72% of patients), bleeding (42%), neurological events such as major or minor strokes (25%), and device malfunctions (18%). A total of 17 patients in the study died before a donor heart became available.

These complications notwithstanding, some researchers argue that the alternative of fitting a LVAD to end-stage patients is a worse choice. This is because, in such patients, the right ventricle often fails a while later, which requires more dangerous surgery. In addition, a minimally contracting ventricle is often the source for thrombus formation. They suggest that it is in the patient's interest to remove both ventricles and replace them with a TAH (Slepian, Smith et al., 2006).

In June 2009, the eight hundredth implant of the TAH was performed at the Heart and Diabetes Centre NRW.

Most hospitalized patients in the United States are still connected by tubes from the heart through their chest wall to a large power-generating console called "big blue," which operates and monitors the device, but in Europe the device can be used with portable drivers permitting discharge from hospital (Hajar, 2005). However, this changed in March 2010, when a portable driver weighing only 6 kg was cleared for clinical trials by the FDA. Figure 8-17 shows a comparison between the older "big blue" driver and the new Freedom driver.

**FIGURE 8-17 ■**  
Comparison  
between the original  
"big blue" pneumatic  
driver and the new  
portable Freedom  
driver. (Courtesy of  
[syncardia.com](http://syncardia.com), with  
permission.)



#### 8.4.1.4 AbioMed/AbioCor

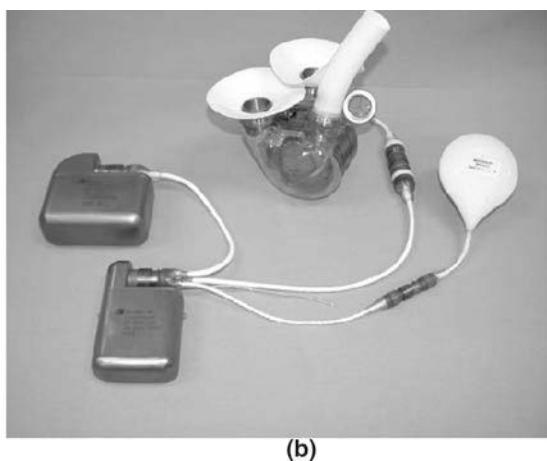
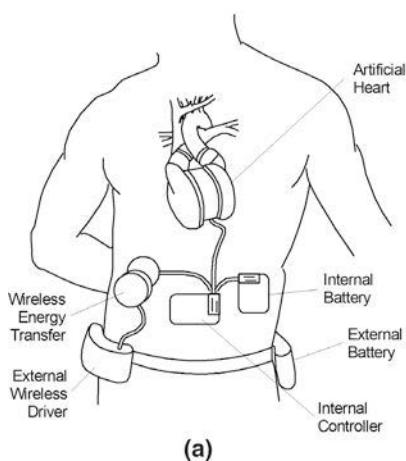
On July 2, 2001, Robert Tools was the first to receive an AbioCor implantable replacement heart produced by AbioMed of Danvers, Massachusetts. It was the first completely self-contained artificial heart. The surgery was performed by University of Louisville doctors at the Jewish Hospital in Louisville, Kentucky. Tools lived for nearly 5 months, and during that time he was sufficiently well to walk about. Later patients fared even better, and Tom Christerson survived for 17 months after an AbioCor transplant (Sherief, 2007).

On September 6, 2006, the FDA approved the first totally implanted permanent artificial heart for patients with advanced biventricular failure. According to the FDA, the AbioCor Implantable Replacement Heart is intended for people who are not eligible for a heart transplant and who are unlikely to live more than a month without intervention.

The AbioCor system consists of a 900 g mechanical heart that replaces the diseased heart, which is removed during the implantation procedure. It also includes a controller and an internal battery, which are implanted in the patient's abdomen, and a transcutaneous power transfer coil that recharges the internal battery. These components are shown in Figure 8-18.

To understand how the system works, consider the various components (Bonsor, 2008):

- **Heart pump:** An efficient dual-cavity hydraulically driven blood pump, which replaces the right and left heart. Each pump is capable of delivering more than 8 L/min. Blood is pumped from the superior and inferior vena cava to the lungs through the pulmonary artery by the right pump and from the pulmonary veins to the rest of the body via the aorta by the left pump.
- **Wireless energy-transfer system:** Also called the transcutaneous energy transfer (TET), it consists of one internal and one external coil that transmit power via electromagnetic coupling from an external battery across the skin without piercing the surface. The internal coil receives the power and sends it to the internal battery and controller device.
- **Internal battery:** A rechargeable battery is implanted inside the patient's abdomen. This gives a patient between 30 to 40 minutes to perform certain activities, such as showering, while disconnected from the main battery pack.



**FIGURE 8-18 ■**  
Abiocor  
(a) Schematic of the TAH installation.  
(b) Photograph of the TAH and peripheral components.  
(Courtesy of AbioMed, with permission.)

- **External battery:** This battery is worn on a Velcro-belt pack around the patient's waist. Each rechargeable battery offers about 4 to 5 hours of power. During sleep and while batteries are being recharged, the system can be plugged into an electrical outlet.
- **Controller:** This small electronic device is implanted in the patient's abdominal wall. It monitors and controls the pumping speed of the heart.

As can be seen in Figure 8-18, the heart is made from clear epoxy and titanium so that it can be visually inspected during the implant process to ensure that it is working correctly and that no air is trapped within the device (Texas Heart Institute, 2008).

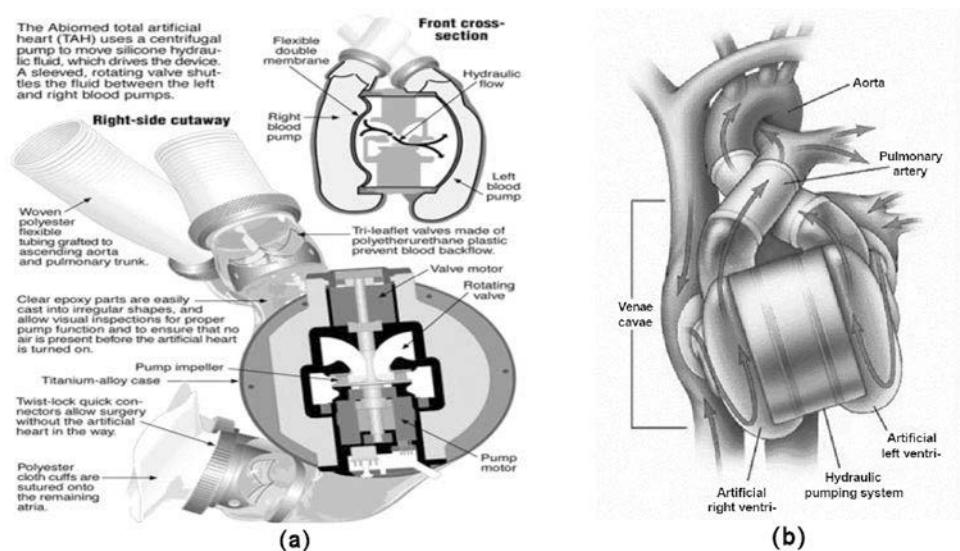
The heart pump consists of the following components, as shown in Figure 8-19:

- **Hydraulic pump:** An efficient electric motor spins the impeller inside the centrifugal pump at 10,000 rpm to create the required hydraulic pressure in a silicone hydraulic fluid.
- **Porting valve:** A separate motor rotates the valve, which opens and closes to let the hydraulic fluid flow from one side of the artificial heart to the other.
- **Artificial ventricle:** When the fluid moves to the right, it compresses the flexible membrane on the inner surface of the right artificial ventricle (pump sac), and blood is pumped to the lungs via a nonreturn valve. When the fluid moves to the left, a similar sac is compressed, and blood is pumped to the rest of the body via a separate nonreturn valve.

AbioMed officials have cautioned against overly optimistic results for the first patients to receive a transplant. While doctors hope for more, the device is designed to only double life expectancy for patients who had only about 30 days to live prior to the operation. The most optimistic predictions are that a patient may be able to live up to 6 months with the AbioCor heart.

**FIGURE 8-19**

Abiocor (a) Cutaway view of the internal workings of the TAH. (b) Graphic showing its connection into a human patient.  
(Courtesy of AbioMed, with permission.)



### 8.4.2 Implanting an Artificial Heart

The implant process is similar for most total artificial hearts. Here is the procedure, as described by University of Louisville surgeon Robert Dowling (Bonsor, 2008):

- Surgeons first implant the energy-transfer coil in the abdomen.
- The breast bone is opened, and the patient is placed on a heart-lung machine.
- Surgeons remove the right and left ventricles of the native heart. They leave the right and left atria, the aorta, and the pulmonary artery in place. This part of the surgery alone takes 2 to 3 hours.
- Atrial cuffs are sewn to the native heart's right and left atria.
- A plastic model is placed in the chest to determine the proper placement and fit of the heart in the patient.
- Grafts are cut to an appropriate length and sewn onto the aorta and pulmonary artery.
- The AbioCor heart is placed in the chest. Surgeons use quick connects to connect the heart to the pulmonary artery, aorta, and left and right atria.
- All of the air in the device is removed.
- The patient is taken off the heart-lung machine.
- The surgical team ensures that the heart is working properly before closing.

## 8.5 | VENTRICULAR ASSIST DEVICES

Ventricular assist devices (VADs) are mechanical pumps that are implanted onto a complete but damaged heart to aid with the pumping process. These can be used as a BTT, bridge to recovery (BTR), or destination devices to be used for the remainder of the patient's life.

According to the United Network for Organ Sharing, adults who need a heart transplant wait an average of 170 days, but nearly 30% are still waiting even after 2 years on the transplant list. At the start of 2009 nearly 3000 Americans of all ages were waiting for heart transplants.

For a heart that is underperforming in regard to the volume of blood that it can transport, the objective of a VAD is to augment the blood flow up to an absolute minimum of 5 L/min to maintain the required blood oxygen level. In addition, the system must be power efficient to allow for battery operation and to minimize heat dissipation within the body. The pump needs to minimize fluid shear stress to minimize damage to red blood cells (hemolysis), and, finally, the pump size and shape must be such that it can be placed in an anatomically compatible location (Paden, Ghosh et al., 2000).

A LVAD consists of an electrically driven mechanical pump, an electronic controller, and a power supply. In cavity-mounted pulsatile devices, the pump weighs about 800 g and is made of titanium with a biologically friendly lining. Placed in the abdominal cavity, the LVAD takes blood from the left ventricle and pumps it into the aorta. Pulsatile pumps operate at a rate of 60–80 beats per minute but can increase to 120 beats per minute with exercise.

The earliest pulsatile assist devices were pneumatic. These included the Thoratec, Cardio West and AbioMed units. The control of these focused on regulating flow, atrial,

or arterial pressures. They were operated in open loop mode, which allowed manual adjustments of the eject rate, duty cycle (systolic/diastolic ratio), and the pneumatic drive pressure.

Pulsatile pumps that are actuated by a positive displacement mechanism include devices such as the Thoratec Heartmate cam drive pump, Novacor spring decoupled solenoid, and the Penn State roller screw pump. These are all controlled by adjusting the stroke volume or the pump rate. They have been shown to be effective for short-term implantation, but they are bulky, have large power requirements, and use complex reciprocating pumping motions.

Other types of ventricular assist devices are dynamic. These include centrifugal, mixed-flow, and axial pumps. Their primary benefit is that they are smaller than the pulsatile units and more energy efficient. Axial pumps, in particular, seem to be quite promising for reasons of simplicity and efficiency. They can be implanted in the patient's abdomen with catheters passing through the diaphragm connecting the pump to the left ventricle, or in some cases they can be attached to the ventricle, inserted within it, or even inserted within the aorta.

Rotary pumps are insensitive to the hydraulic head, which results in their being intolerant of mismatches between physiological demands and operating settings. In addition, ensuring that the pump is operating at an appropriate speed is more critical than it is with pulsatile devices. Rotary pumps in clinical use include the early Medtronic and Biomedicus devices.

Pulsatile pumps still have a role because of their success in stimulating a highly sensitive natural Starling law<sup>1</sup> response to venous pressure. Although nonpulsatile blood flow is now known to be more acceptable than was once believed, it is not yet known with certainty whether all patients will successfully adapt to the unnatural blood shear and flow regimes of dynamic pumps.

In most of these devices, external batteries supply power via a cable through the abdomen or via transcutaneous energy transfer devices using electromagnetic coupling. The batteries are carried in underarm holsters or a waist pack.

To study the efficacy, safety, and cost-effectiveness of an LVAD for permanent use compared with optimal medical management through medication therapy, a clinical trial called the Randomized Evaluation of Mechanical Assistance for the Treatment of Congestive Heart Failure (REMATCH) was undertaken. This trial was conducted on 129 patients at 22 major hospitals across the United States and was financially supported by the National Institutes of Health (NIH). The result of this study showed a 48% decrease in the death rate from all causes with the LVAD over the first 2 years of use. Patients in the LVAD group had a median survival period of 408 compared with 150 days in the medication therapy group. Only 8% (1 of 12) survived 2 years in the optimal medical management group. 23% were alive at 2 years in the LVAD group. The quality of life was also improved in the LVAD group, based on the questionnaire completed by patients from both groups at 1 year. The study was conducted on only the sickest patients with no other options (Jeffrey, 2001).

---

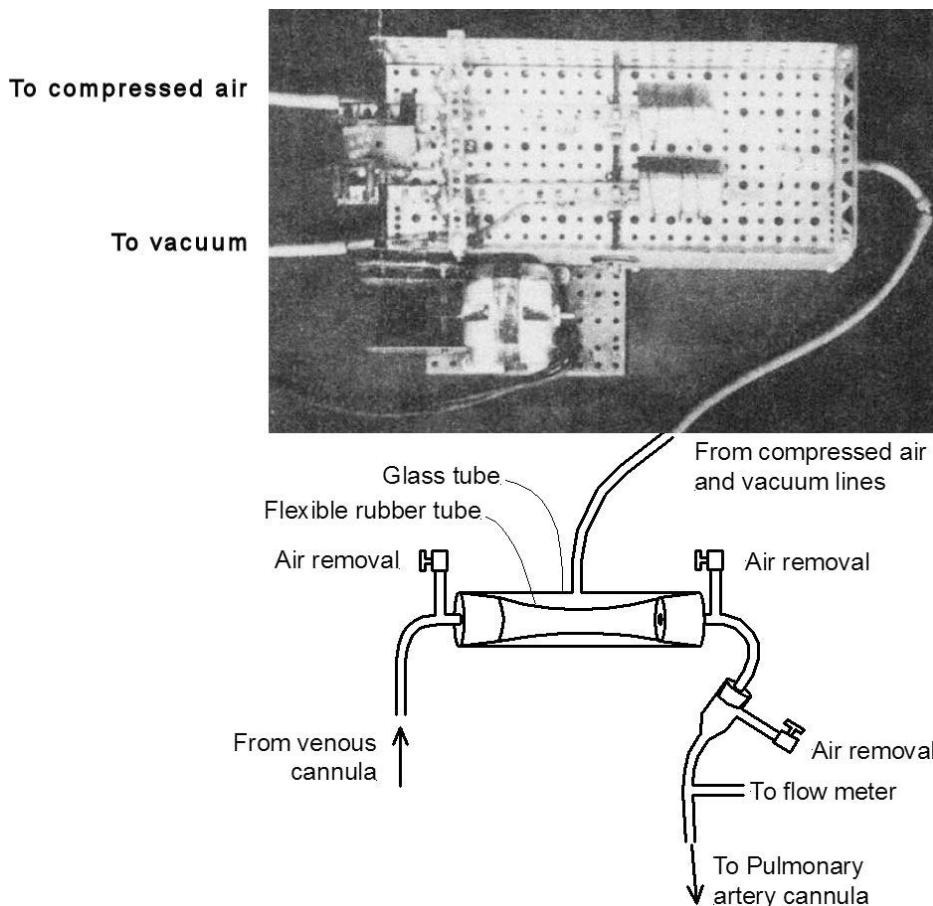
<sup>1</sup>Starling law states that the increased volume of blood entering the heart stretches the ventricular wall, causing cardiac muscle to contract more forcefully.

### 8.5.1 History

Using a toy Erector set, William Sewell Jr. and William Glenn, Yale University medical students, built a section of a heart pump that they used in experimental bypass surgery on dogs. The initial design using a peristaltic (roller) pump driven by the Erector set motor was unsuccessful as the motor was underpowered for the blood pumping function. Instead, Sewell built the device shown in Figure 8-20 to drive eccentric cams that occluded and released small rubber tubes leading to the compressed air and vacuum lines that actually drove the pump (Glenn, 1993).

In 1949 the pump was used to bypass the right heart of dogs in two experiments lasting for 61 and 82 minutes, respectively. During the time of the experiment, the right ventricle was left wide open. After restoration of normal circulation, removal of the pump, and closure of the chest, the dogs made uneventful recoveries.

Search for an optimal circulatory-support device began in 1964 with the NIH artificial heart program. That year, Michael DeBakey inserted the first left ventricular assist device in a human patient using a pump almost identical in design to Sewell's (Glenn, 1993). In 1994, the FDA approved a pneumatically driven LVAD as a bridge to transplant. Four years later, in 1998, it approved a self-contained, vented electric device to replace the pneumatic one that allowed patients to go home while awaiting heart transplant.



**FIGURE 8-20** ■ The first artificial heart pump built by William Sewell and William Glenn [Adapted from (Glenn 1993).]

Recent results have been quite impressive, with some patients showing sufficient improvement while using an LVAD that can be removed. Even so, the overall mortality rate due to end-stage heart failure remains high, mainly due to the limited number of donor heart transplant organs available. Hence, researchers have carefully studied a possible use of the LVAD as a long-term therapy for patients who are not candidates for heart transplant.

Though most LVADS are mounted within the body (intracorporeal), some are mounted outside the body (extracorporeal) with tubes that pass through the skin to convey blood to and from the device.

The following sections of this chapter document some of the VAD devices that have been developed by research institutes and medical companies over the past 2 decades.

### 8.5.2 Extracorporeal Ventricular Assist Devices

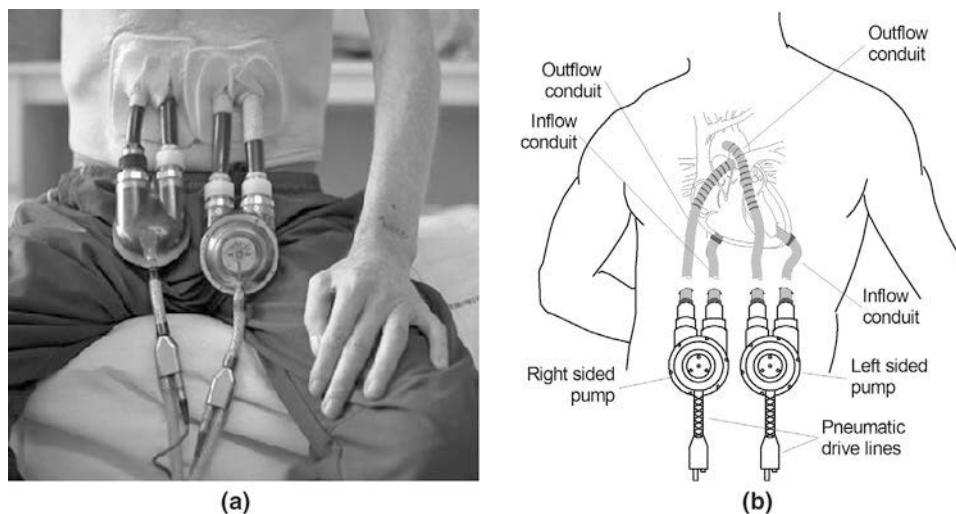
Extracorporeal ventricular devices, as shown in Figure 8-21, consist of an external controller and air pump connected to one or two pneumatically powered heart pumps with cannulae that pass through the abdominal wall and the diaphragm where they connect to the left ventricle and the aorta and to the right ventricle and the pulmonary artery. These devices include the Berlin Heart Excor, BioMedicus BP-80, Medos, Thoratec, and the Abiomed BVS500 devices. They are primarily BTT devices, though occasionally just decreasing the stress to the diseased heart has allowed it to recover sufficiently for the system to be removed, relieving the patient of needing a transplant. This is known as a BTR event.

Most extracorporeal VADS can be tailored to an individual's heart characteristics. Blood volumes ranging from 10 ml required by a child to 80 ml per stroke required by a large adult are available, while the silicon cannulae are available in a number of different diameters to accommodate the different flow rates. Different types of drive units are also available for every conceivable blood pressure and volume, depending on whether the patient will be mainly stationary or mobile. Most of the devices also include features such as a heparin coating, titanium connectors, and multiple membranes.

In the Berlin Heart, the inner surfaces of the blood chambers are extremely smooth and are designed to ensure optimum flow characteristics. Additionally, the heparin coating prevents thrombogenesis. The pump has three membranes separated by a thin graphite

**FIGURE 8-21 ■**

Thoratec extracorporeal biventricular assist device  
 (a) Photograph of installation.  
 (b) Graphic showing internal connections to the heart.  
 (Courtesy of Thoratec Corporation, reproduced with permission.)



lubricant layer, which ensures that they can withstand every type of load. The device is fitted with a deairing stub to remove air pockets easily. The cannulae are made from pure silicon, with the polyester velour jacket forming a bond with the surrounding tissue to provide effective protection against infection.

Human trials conducted at the department of Thoracic and Cardiovascular Surgery, Heart Centre at Ruhr University produced detailed statistics of the survival rates and complications these devices using a wide range of implantable and extracorporeal VADs over a period from 1989 to 1999 (Minami, Arusoglu et al., 2001). For the 85 patients on the Thoratec device (43 biventricular), the mean duration of the BTT support was 49 days (maximum of 386 days). During this time 41% of the group suffered from perioperative bleeding, 24% suffered from cerebral embolism, 2% from driveline infection, and 26% from sepsis. These results were in fact typical of all of the devices trialed. Interestingly, mechanical failure rates for the VADs and the external pumps were low, only 4%.

The Thoratec VAD was originally the only system that offered circulatory support for the left ventricle, the right ventricle, or both. As of July 2006 it had been used in more than 2800 patients ranging in age from 6 to 77 years and in weight from 17 to 144 kg (Deng and Naka, 2007). It is FDA approved for BTT and postcardiotomy recovery of the natural heart.

### 8.5.3 Intracorporeal Left Ventricular Assist Devices

Intracorporeal LVADs can be divided into four generations described in part by the pump mechanism used and in part where they are implanted in the human body, as illustrated in Figure 8-22:

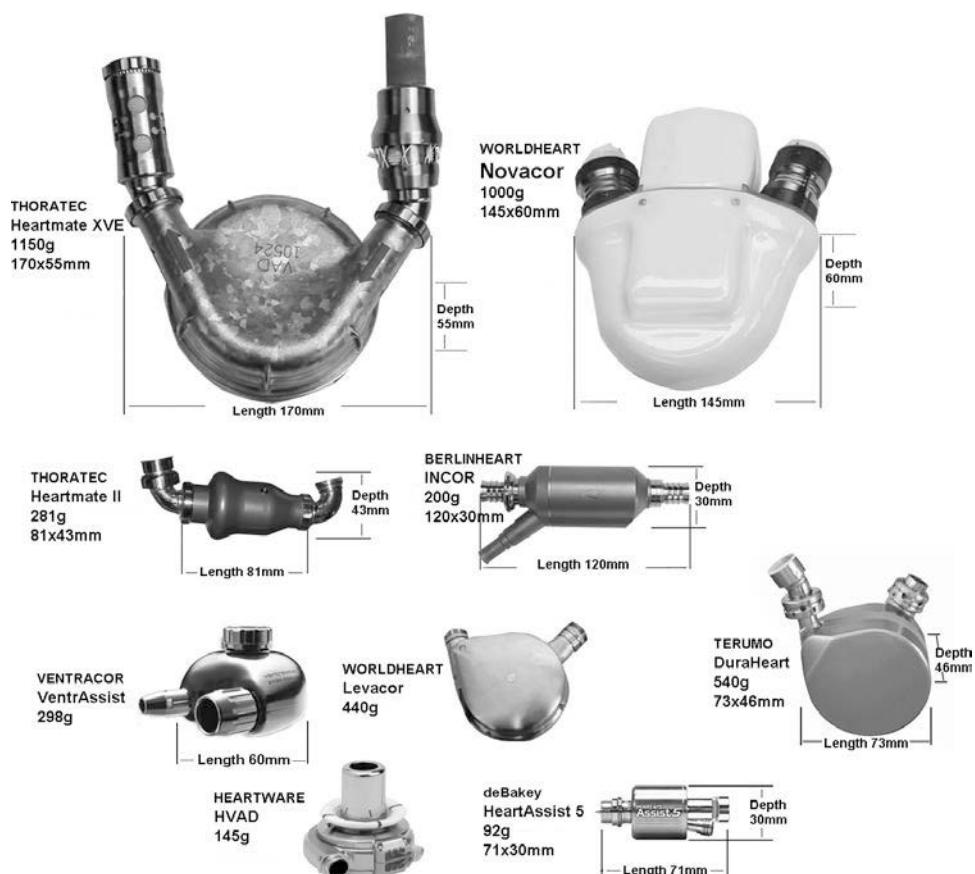
- **Generation 1:** These are the original large pulsatile devices, typically with masses in excess of 1 kg and with volumes of about 1000 cm<sup>3</sup>. They include the Thoratec Heartmate, the WorldHeart Novacor, the Novacor II, and the Arrow LionHeart. They are implanted below the diaphragm.
- **Generation 2:** These VADs include the original lightweight axial-flow devices with masses around 250 g. Because axial-flow pumps are extremely efficient, their volumes are typically between 150 and 200 cm<sup>3</sup>. They include the BerlinHeart Incor and the Thoratec Heartmate II. They are also implanted below the diaphragm.
- **Generation 3:** Improvements in pump design saw a change from axial to centrifugal or mixed flow types. These are typically heavier than the axial-flow types with masses between 300 and 500 g. They include the Ventracor VentrAssist, the Worldheart Levacor, and the Mohawk Technology MiTi Heart. They are also implanted below the diaphragm.
- **Generation 4:** The latest generation of VADs is sufficiently small and light to be implanted above the diaphragm. These have masses ranging from less than 100 g to about 150 g and include the HeartWare HVAD, the DeBakey HeartAssist 5, and the Jarvik-2000.

### 8.5.4 Generation 1 LVADs

#### 8.5.4.1 WorldHeart/Novacor

It was nearly 25 years ago that a Stanford team performed the world's first successful implantation of a Novacor LVAD into Robert St. Laurent to keep him alive long enough to receive a heart transplant. St. Laurent, who had been expected to live only 24 hours

**FIGURE 8-22 ■**  
VADs grouped by generation.

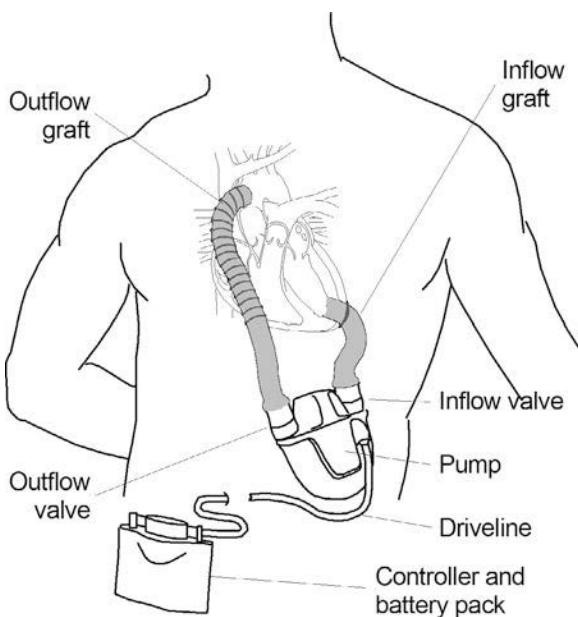


before the surgery, depended on his LVAD for 8 days in 1984 until a donor heart became available (Baker, 2004).

The WorldHeart Novacor consists of an implanted, wearable pump about the size of a human heart as well as a controller and battery pack, as shown in Figure 8-23. It has exhibited good reliability and durability as attested to by the following statistics. Of the more than 1700 recipients, 172 primarily BTT patients have been supported by the device for more than 1 year, and of those 45 have been supported for more than 2 years, 24 for more than 3 years, 11 for more than 4 years, and 1 for more than 6 years. Only 1.4% of the pumps have needed replacement, and no patient deaths have been attributed to pump failure.

Blood enters the Novacor pump via an inflow conduit (graft) through a large opening cored into the patient's left ventricle. The low resistance of the passively filling pump chamber presents a reduced load to the left ventricle, allowing the diseased heart to pump a normal stroke volume with minimal effort. Once the pump chamber is full, the electrically driven pump ejects blood through an outflow conduit into the arterial system, thereby supporting the systemic circulation. The system is completely self-regulating, automatically adjusting its beat rate and stroke volume in response to the patients changing circulatory requirements.

The Novacor design uses two blood-sac pusher plates positioned at opposite sides of the pump body. Electromagnetic actuators coupled to compliant linkages squeeze the



**FIGURE 8-23 ■**  
Schematic diagram showing the installation of the Novacor LVAD. [Adapted from (Cleveland Clinic 2008).]

pusher plates together compressing the blood-sac to eject the blood from the pump in a controlled manner.

The pusher plates are epoxy bonded to a seamless flexible blood sac, consisting of a butyl rubber layer sandwiched between two layers of polyurethane. The polyurethane, which is manufactured by Ethicon Inc. (Somerville, NJ), is used in most existing blood pumps because of its durability and biocompatibility, whereas the butyl rubber layer increases the pump's impermeability.

The blood sac is supported within a cylindrical aluminium annulus that forms the pump housing. The inflow and outflow ports are positioned tangentially on opposite sides of the housing to ensure straight-through blood flow. The ports are formed by an epoxy-impregnated Kevlar fabric shell that is integrated into the housing and support trileaflet inlet and outlet valves made from bovine pericardium tissue.

Tangential inflow into the pump sac initiates a circular flow pattern that is coupled directly to the tangentially mounted outflow to maximize pumping efficiency. It also creates a flow pattern that “washes” all of the internal surfaces of the pump, reducing the potential for blood clot formation.

The external controller is connected to the implanted pump by a percutaneous lead—a small tube that passes control and power wires through the recipient’s skin. The microprocessor-based controller regulates pumping action and monitors system function.

During normal, untethered operation, the controller receives power from two rechargeable power packs. The controller and power packs may be worn on a belt or carried in a shoulder bag, vest, or back pack. The portable nature of the Novacor system facilitates out-of-hospital use and allows recipients to return home and lead near normal lives.

The power pack contains rechargeable nickel metal hydride (NiMH) batteries that last approximately 6 hours per pair (at a flow rate of 6 L/min). A monitor circuit displays charge capacity and alarms for low charge, accidental disconnect, or fault conditions. Users typically have a battery charger that maintains a number of packs on full charge ready to be swapped into place.

**FIGURE 8-24** ■  
Thoratec Heartmate VAD. (Courtesy of Thoratec Corporation, reproduced with permission.)



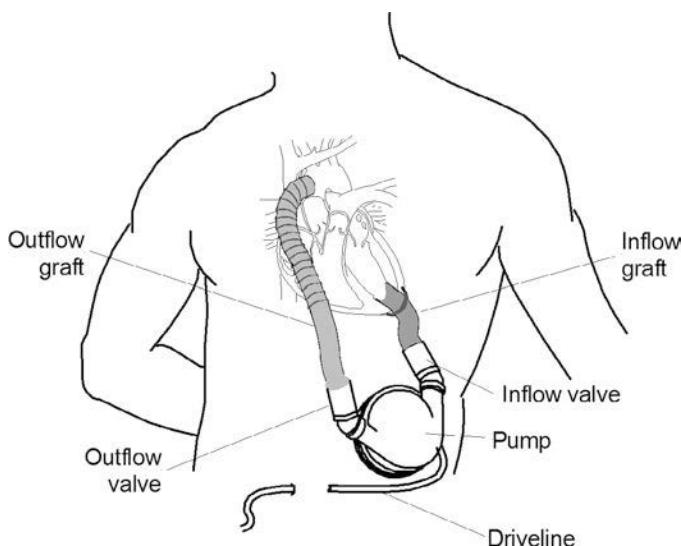
#### 8.5.4.2 Thoratec/Heartmate

The Thoratec Heartmate VAD System includes three major components: blood pump, cannulae, and the portable driver. This system can provide partial or total circulatory assistance when the natural heart is unable to maintain adequate circulation to perfuse vital organs. As with all VADs, the Heartmate receives blood from the ventricle of the natural heart at low pressure and then pumps it back into the arterial system at high pressure.

Manufactured by Thermo Cardiosystems Inc. (Woburn, MA), the device is fully implantable with wearable external components. As shown in Figure 8-24, the pump consists of a flattened titanium cylinder about 50 mm thick and 100 mm in diameter, weighing approximately 1150 g. It is made up of two chambers: the blood chamber and the air chamber. The former contains an electric motor-driven cam mechanism that actuates a pusher plate. This in turn displaces a flexible polyurethane diaphragm that separates the two chambers. A textured interior surface helps prevent blood clots and reduces the need for the patient to take anticoagulants. Each conduit contains a 25 mm porcine valve within a woven Dacron fabric graft. The pump has a maximum stroke volume of 83 ml. It can be operated at up to 120 BPM, resulting in flow rates of up to 10 L/min.

The device is implanted in the upper part of the abdominal wall or in the peritoneal lining, as illustrated in Figure 8-25. At the start of each cycle, blood drains from the left ventricle into to the blood chamber within the LVAD, at which point an external control system triggers pumping. An electric motor-driven pusher plate forces the polyurethane diaphragm upward to pressurize the blood chamber. This motion propels the blood through an outflow conduit that connects to the aorta. The valves in each of the conduits ensure that blood flows only in one direction. A second tube extends outside the body to a battery pack that is carried in a shoulder holster. This external tube also maintains near atmospheric pressure in the air chamber and can also provide pneumatic drive for the LVAD in case of motor failure. The pump is designed to respond to changing flow demands of the body, with variable flow rates up to a maximum of 10 L/min.

As of July 2006, over 4100 patients had been supported by the Heartmate LVAD. It was approved by the FDA as a destination therapy with overall costs of \$160,000 being similar to those for a heart transplant. However, unlike a natural heart, problems associated with the device include infection of the surrounding area, bleeding, and device malfunction. Statistics show that there is a 52% chance of survival after 1 year but also that the device failed 35% of the time after 2 years.



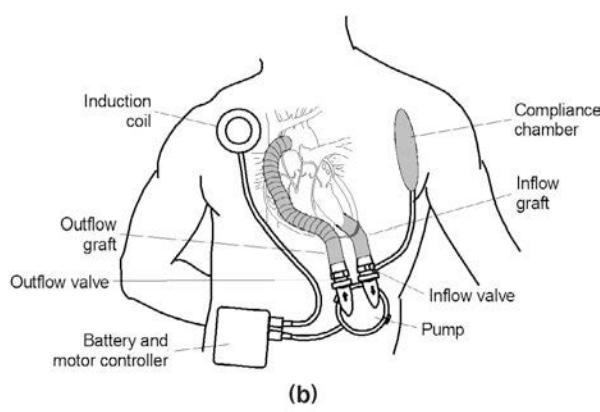
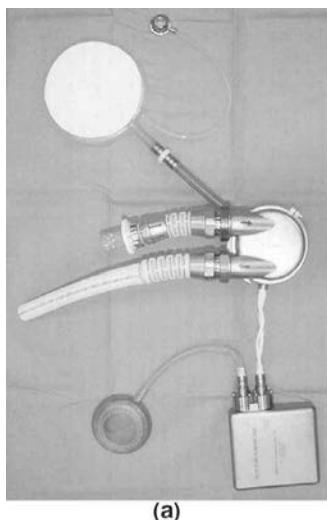
**FIGURE 8-25** ■ Schematic diagram showing the installation of the Thoratec Heartmate VAD. [Adapted from (Cleveland Clinic 2008).]

#### 8.5.4.3 Arrow/LionHeart

The LionHeart development began in 1993 as a collaboration between Arrow International and the Hershey Medical School. Early devices developed at the medical school had used standard externally powered pneumatic pumps, but by the time of the merger a totally implantable pump design had been developed. The first LionHeart was implanted in 1999, and it was totally contained within the body.

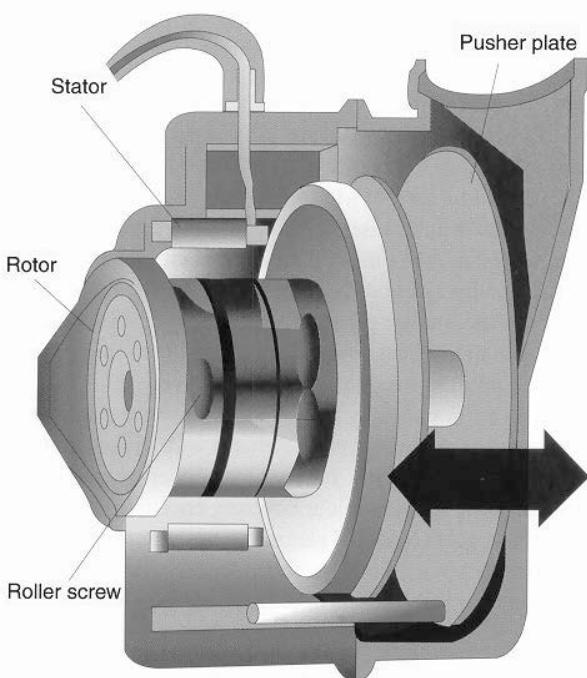
The total weight of the implanted components of the LionHeart is 1.3 kg. It consists of the electrically powered blood pump that is implanted in the abdomen on the left side near the ribs, an internal controller and battery, and a percutaneous inductive power transfer device, as shown in Figure 8-26.

The pump consists of a high-speed compact brushless direct current (DC) motor that spins in one direction before reversing and spinning in the other as shown in Figure 8-27. This drives a roller-screw mechanism that converts the rotary motion into the linear motion



**FIGURE 8-26** ■ LionHeart VAD  
 (a) Photograph of the heart and peripheral components.  
 (b) Diagram showing the implanted device. [Adapted from (Arrow LionHeart 2007), reproduced with permission.]

**FIGURE 8-27 ■**  
 Computer-aided  
 design (CAD) model  
 of the roller-screw  
 energy converter  
 used in the  
 LionHeart VAD.  
 (Snyder, Pae et al.,  
 2001), reproduced  
 with permission.



of a pusher plate. The pusher plate applies a uniform pressure onto a blood sac, compressing it and driving the blood through a tilting-disk outlet valve and around the circulatory system. When the motor reverses, the plates back off, which reduces the pressure and allows blood to flow from the ventricle through another tilting-disk inlet valve to refill the sac.

An intrathoracic compliance chamber 130 mm in diameter and 10 mm thick maintains near thoracic pressure in the energy converter airspace, which allows the pump to fill passively. It is regularly replenished with air via an access port.

The motor controller controls the operation of the blood pump. The blood pump and electronic motor controller are powered by an external source through a high-efficiency (70–80%) inductive coupling or using rechargeable batteries located in the motor controller. Internal power allows the LVAD to function totally free of the external power source for approximately 40 minutes (instructions to the patient say 20 minutes). The motor controller is placed under the abdominal wall on the right side near the ribs, and the internal coil is placed just under the skin of the chest wall (Arrow LionHeart, 2006).

Passive pump filling permits the pump rate to be controlled according to the speed with which it fills. The control system uses no transducers other than the Hall effect sensors needed to commutate the brushless DC motor. The controller analyzes the motor speed and the applied voltage to estimate the load imposed by the pump on the pusher plate, which varies as it moves from end-diastolic to end-systolic position. An abrupt rise in the load indicates that the pusher plate is in contact with the blood sac, beyond which pressure is developed to open the outlet valve. The pusher-plate position at which contact occurs is used to estimate the end-diastolic volume, and this is used to control pump rate.

An end-diastolic volume of greater than 85% of the 70 cm<sup>3</sup> pump capacity is considered to be acceptable. If this threshold is not reached, the pump rate is decreased. Otherwise, the pump seeks an ideal rate as follows. The pump rate increases in small increments so long as the end-diastolic volume remains reasonably constant. As soon as a decrease in volume is noted, the pump rate backs off as long as each change results in improved filling. During constant conditions, this hunting process is minimal, but with changing conditions the pumping rate can increase from about 60 BPM to 120 BPM in less than 1 minute (Snyder, Pae et al., 2001).

A telemetry link permits both adjustment and monitoring of the implant through radio frequency (RF) communication. When the telemetry wand is placed on the skin close to the implanted electronics, the two devices initiate communication automatically and the clinician's monitor displays the device's ID and battery status as well as pump parameters including pumping rate, end-diastolic fill volume, pump volume, and pressure.

The external components consist of the power transmitter coil, its associated oscillator and control electronics, as well as a pair of batteries, shown in Figure 8-28. Also included is a battery charger and external DC supply. Power is preferentially drawn from the external DC source so that when mobility is not desired—for example, when the patient is asleep—the two portable batteries are kept in reserve. Feedback from the control electronics includes the following:

- Status of both of the batteries in the pack.
- Need to change batteries.
- Need to check the coil alignment.
- Whether the implanted battery is completely charged.
- Verification of the presence of an external supply.
- Failure of an external supply.

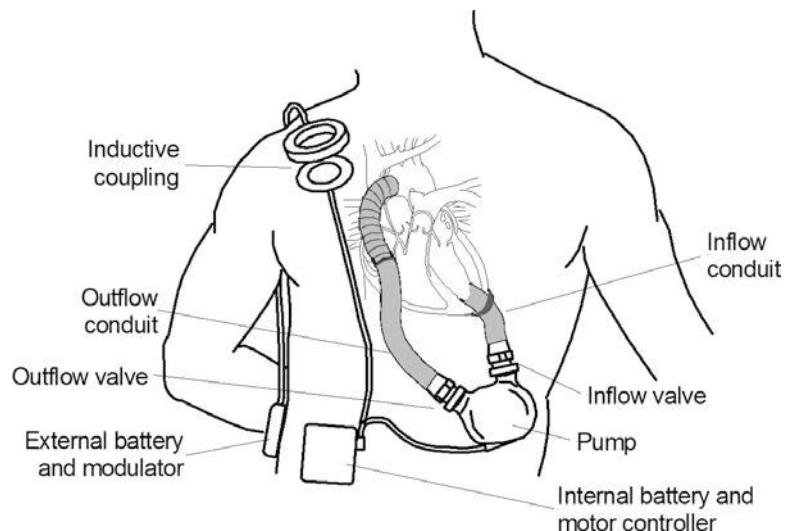
#### 8.5.4.4 WorldHeart/Novacor II

The Novacor II system, currently under development, is the next generation of pulsatile LVADs. It is about half the size of its predecessor and has some unique design features that



**FIGURE 8-28** ■  
External batteries with controller and transcutaneous energy supply coil (Snyder, Pae et al., 2001), reproduced with permission.

**FIGURE 8-29** ■ Schematic diagram showing the installation of the Novacor II [Adapted from (WorldHeart 2008).]



offer many advantages over currently available pulsatile VADs, including the following:

- Smaller size.
- Simpler, quieter operation.
- Reliability and durability, with no bearings or wearing elements.
- Can be fully implantable without a volume compensator.
- Use as LVAD or right ventricular assist device (RVAD).

The system consists of six elements, shown in Figure 8-29:

- Novacor II pulsatile pump with its direct magnetic driver is implanted within the anterior abdominal wall.
- Inflow conduit carries blood from the left ventricle to the pump.
- Outflow conduit carries blood from the pump to the ascending aorta.
- Internal battery and controller regulates the pump action and allows operation for short periods without external power.
- Transcutaneous energy transmission system inductively couples power to the unit.

External battery pack and modulation system conditions the power for transfer.

The pump consists of a pair of chambers separated by a magnetically driven pusher plate containing a transfer valve. When the pusher plate is driven to the right (pumping stroke), the prechamber expands, filling from the left ventricle. Simultaneously, the pumping chamber is compressed, ejecting blood into circulation. When the pusher plate returns to the left (transfer stroke), the prechamber is compressed while the pumping chamber expands; blood transfers from the pre-chamber to the pumping chamber, with no inflow or outflow through the transfer valve in the pusher plate. Because the total volume of the two chambers remains constant as one fills and the other empties, the system can operate without a volume compensator or venting through the skin.

At present development of the Novacor II is on hold while the company focuses on clinical use of the Levacor VAD.

### 8.5.5 Pulsatile Pump Technology

Positive displacement mechanisms in electrically driven pulsatile pumps are commonly actuated by cams, solenoids, and roller-screw devices. However, a number of other experimental technologies are being investigated as the drive to smaller, lighter, and lower-power devices continues. These technologies include linear motors and even shape memory alloy (SMA).

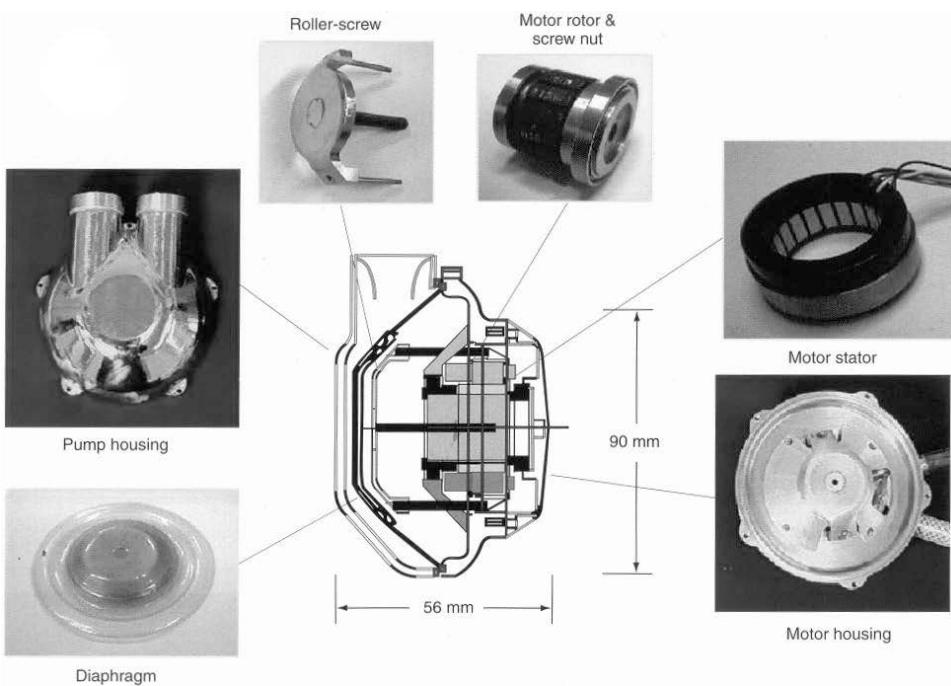
#### 8.5.5.1 Roller-Screw LVAD

A good example of the development of a planetary roller-screw LVAD similar to that used in the Arrow LionHeart device is discussed in Takatani, Ouchi et al. (2001). This device uses a brushless DC motor to drive the roller-screw to produce a stroke length of 12 mm and volume of  $55 \text{ cm}^3$ . The maximum pump output is 8 L/min at an electrical power of 8 W and a 24% electrical-to-hydraulic efficiency. The pump is housed within a titanium alloy shell 90 mm in diameter and 56 mm thick with a total volume of  $285 \text{ cm}^3$  weighing 552 g. Power is transferred to the device transcutaneously by inductive coupling at a frequency of between 100 and 200 kHz.

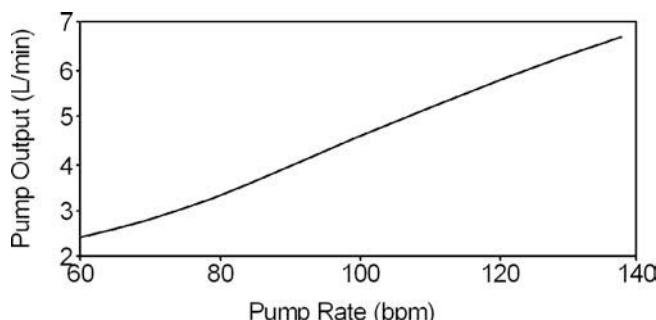
As shown in Figure 8-30, the LVAD consists of a miniature 14-pole Y-wound brushless DC motor from Kollmorgen Inc. and a planetary roller-screw from SKF. Motor rotation is converted into rectilinear motion using the roller-screw attached to a pusher plate, which compresses the diaphragm. It is then reversed after completion of each ejection cycle to allow passive filling of the blood chamber. Hall effect sensors monitor the position of the pusher plate so that the stroke volume and beat rate can be controlled.

The diaphragm is made from polyurethane manufactured by Polymer Technology Inc. using a dip-coating method. The housing is manufactured from a titanium alloy containing 6% aluminium and 7% niobium. It was designed using a computer-aided manufacturing

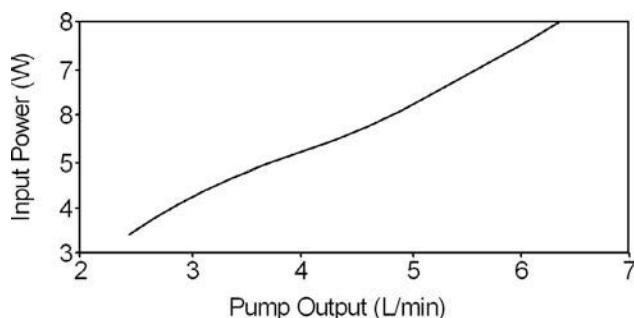
**FIGURE 8-30 ■**  
Schematic diagram of the roller-screw LVAD components (Takatani, Ouchi et al., 2001), reproduced with permission.



**FIGURE 8-31** ■  
Roller-screw pump output as a function of pump rate.



**FIGURE 8-32** ■  
Roller-screw pump input power as a function of pump rate.



(CAM) process and first manufactured using a rapid-prototyping printer before the titanium version was made.

Pump output is a reasonably linear function of pump rate, as shown in Figure 8-31 with the highest pump output of about 7 L/min obtained at 140 BPM. The power requirements shown in Figure 8-32 also bear a reasonably linear relationship to the output flow rate.

Pump efficiencies increased with flow rate up to 140 BPM with the best efficiency of 23% obtained using polyurethane valves. The use of conventional Bjork-Shiley and St. Jude valves resulted in slightly lower efficiencies (about 22% and 20%, respectively).

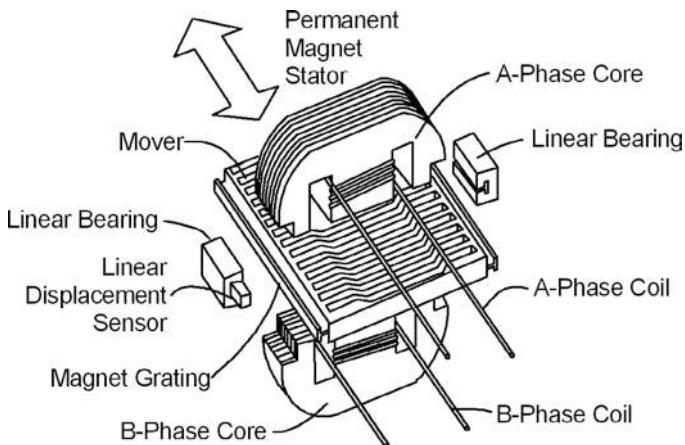
### 8.5.5.2 Linear Motor Driven LVADs and TAHs

Linear motor actuated LVADs have been developed by a number of research institutes in the past decade, including Shinshu University and Tokyo Denki University, both in Japan, and more recently Helmholtz Institute of Aachen University in Germany.

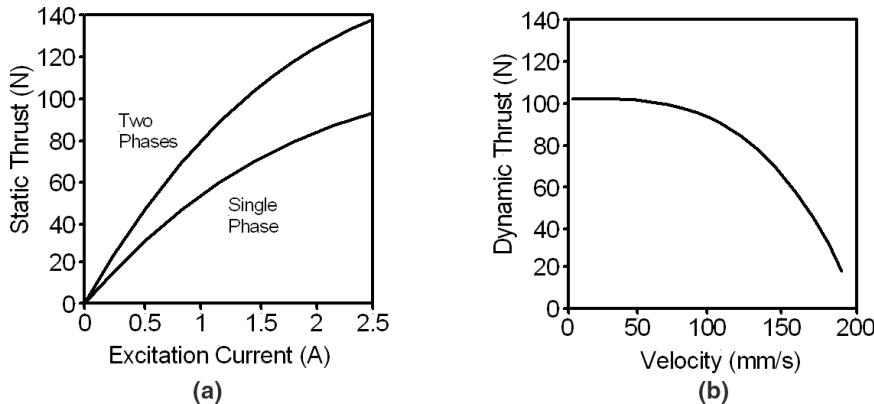
The main advantage of using linear motors compared with rotary-driven devices is a lower component count and hence improved reliability. Using a moving magnet configuration eliminates the need to power a moving coil through flexible cables that may break. On the other hand, a moving coil device can be designed to conduct heat via the pump diaphragm into the flowing blood where it is efficiently transported away from the device.

Most conventional off-the-shelf linear devices are either too weak or too large for implantation. They are also inefficient and therefore generate too much heat when operating within the body cavity. As a rule of thumb, a VAD or TAH cannot dissipate more than 20 W as heat before tissue damage will occur (Finocchiaro, Butschens et al., 2008).

One design for a linear motor system, shown in Figure 8-33, actuates the reciprocating pusher plates directly. In this design it can be seen that the motor consists of upper and lower stators with the mover sandwiched between them and supported on linear bearings.



**FIGURE 8-33 ■**  
Structure of a linear motor to drive a TAH  
[Adapted from (Yamada, Mizuno et al., 1998).]



**FIGURE 8-34 ■**  
Measured characteristics of the linear motor based TAH drive. (a) Static thrust as a function of excitation current. (b) Dynamic thrust as a function of the flow velocity.  
[Adapted from (Yamada, Mizuno et al., 1998).]

Each stator contains 16 teeth with a width of 0.32 mm and a slot width of 0.48 mm, making the total pitch  $\tau = 0.8$  mm. The length of the tooth is 56 mm. The mover tooth pitch is the same as that of the stators and is separated from them by a 40  $\mu\text{m}$  gap. The drive coils are each wound with 110 turns and are supplied with an excitation current of 1.4 A.

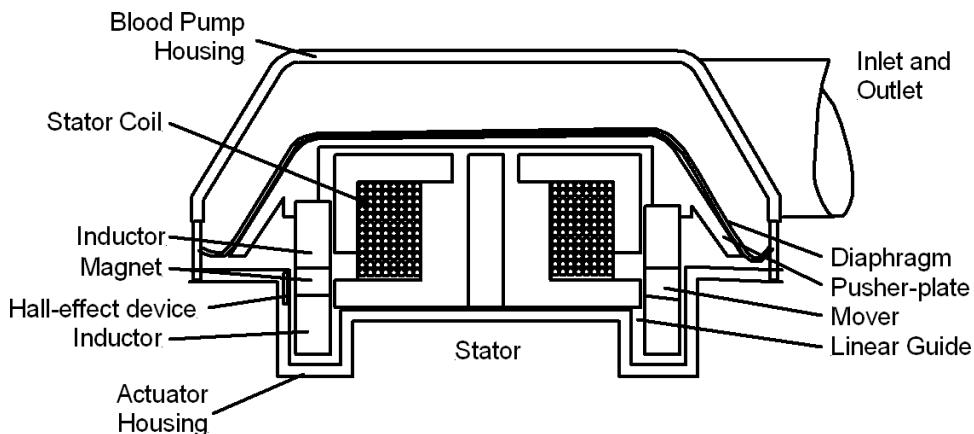
A linear displacement sensor, as discussed in Chapter 3, comprises four magnetoresistive (MR) elements and a magnetic grating embedded within the linear bearing. These together generate a sinusoidal output voltage with a wavelength of 0.8 mm.

A static thrust of just over 100 N was achieved for an excitation current of 1.4 A, but this reduced with mover velocity, as shown in Figure 8-34.

A simpler mechanism is described in Fukui, Funakubo et al. (2004). In this design a moving magnet linear oscillator actuator (LOA) uses conventional voice-coil technology powered by an alternating current (AC) signal to drive pusher plates directly, as shown in Figure 8-35.

The housing is made from epoxy using a rapid-prototyping machine, and the diaphragm is made from segmented polyurethane from PTG medical Co. and formed using a dipping method. All of the blood-contact area within the housing is coated with the same material. The diaphragm, pusher plate, linear guide, and actuator are integrated into a single unit. A Hall effect device is set into the actuator housing to detect the displacement of the mover. A pair of Bjork-Shiley tilting-disk valves is mounted on the inlet and outlet

**FIGURE 8-35** ■ Structure of a linear oscillator actuator-based LVAD.



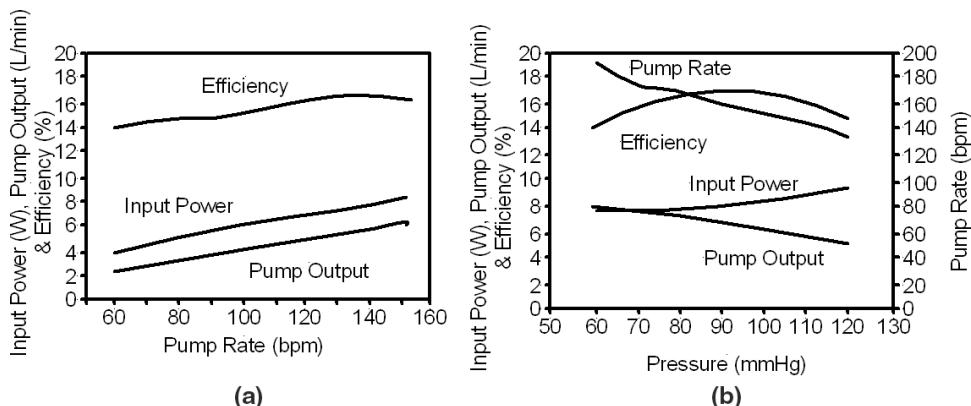
ports of the pump housing. The complete pump has a diameter of 101 mm and a thickness of 49 mm, making its volume  $320 \text{ cm}^3$ . The mass is 770 g, which is significantly heavier than the roller-screw based device.

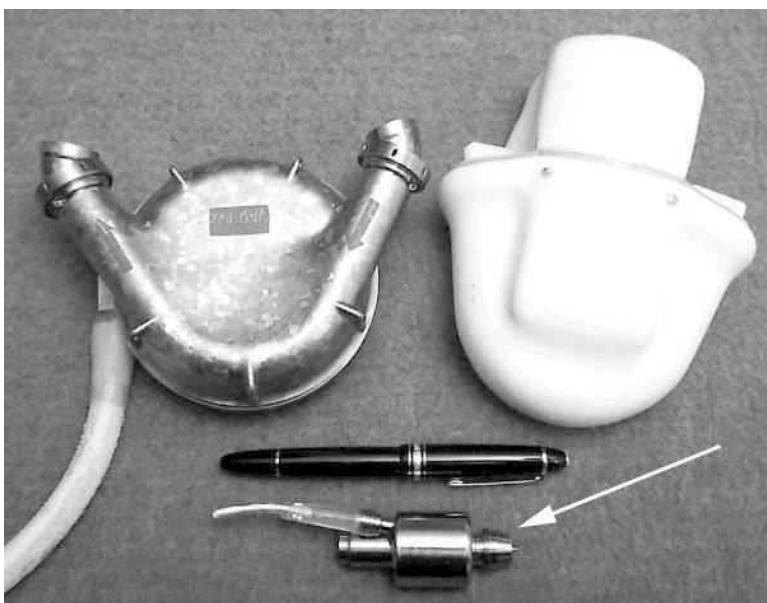
The controller comprises a microprocessor that reads the Hall effect signals and drives a metal-oxide-semiconductor field-effect transistor (MOSFET)-controlled H-bridge that powers the actuator coil from a 9 V DC source. The control unit has two modes: a fixed rate mode where a constant frequency signal is provided to the actuator; and a full fill, full eject (FFFFE) mode that uses the Frank–Starling control mechanism to govern the pump speed.

The measured pump performance is shown in Figure 8-36. For a head of 100 mmHg, the maximum output was 6.1 L/min at 155 BPM, with a power consumption of 8 W. The maximum efficiency was 16.3% at a pump rate of 135 BPM. In the FPPP mode the pump output was 7.9 L/min and 5.1 L/min for loads of 60 and 120 mmHg, respectively, with the pump rate decreasing by 54 BPM over that range. The stroke volume remained between 38 and 43  $\text{cm}^3$ .

If a linear actuator-based system is designed from first principles, then it is possible to use CAD techniques to design the shape of the magnets to optimize the force-displacement curves. In Finocchiaro, Butschens et al. (2008), five different concepts are analyzed. Their conclusions include two viable options. In the first, an ironless drive is used to minimize cogging with moving magnets surrounded by a segmented coil. In the second, a moving

**FIGURE 8-36** ■ Measured performance of the linear oscillator pump.  
(a) Characteristics as a function of the pumping rate.  
(b) Characteristics as a function of the pressure.





**FIGURE 8-37 ■**  
Photograph showing the comparative size difference between the Generation 1 pulsatile VADs and a Generation 2 axial-flow device.

coil is the mover, with magnets and back-iron as the stator. In both cases only the parts of the coil within the high-flux region should be powered to maximize efficiency.

### 8.5.6 Generation 2 VADs

Most second-generation VADs are based on axial pump technology. This was made possible by advances in pump design that minimize damage to blood constituents as well as surface treatments on the surfaces in contact with the blood to minimize clotting. The primary advantages of these designs is the higher pumping efficiency coupled with significant reduction in size and mass, as can be seen in Figure 8-37.

One of the biggest problems with dynamic pump designs is contact-surface wear that reduces the lifetime of the pump. For this reason, much of the design focus over the last decade has been the search for methods to minimize or eliminate wear completely, including single-point contact ceramic bearings as well as magnetically or hydrodynamically suspended impellers.

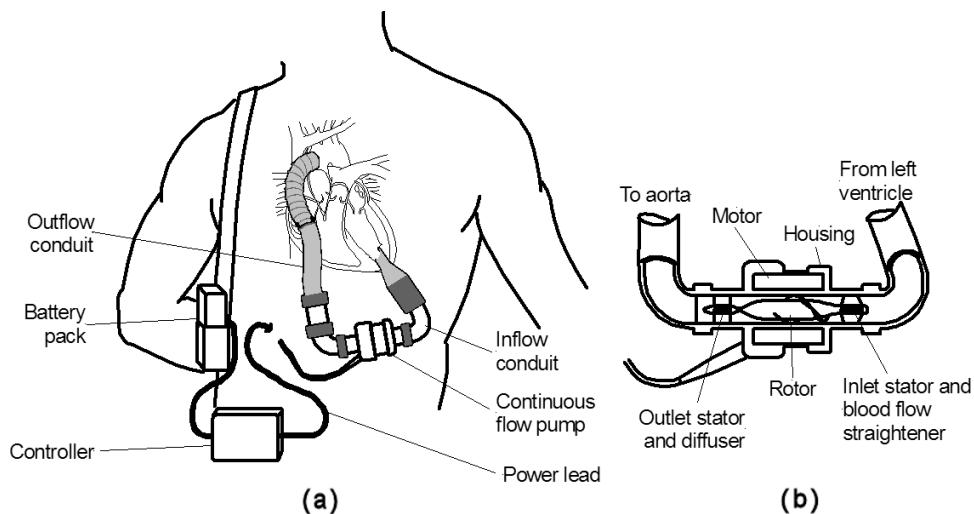
In addition, a wide variety of different design configurations have been developed to try to minimize hemolysis and to maximize efficiency. These range from conventional centrifugal devices to mixed-flow and axial-flow devices.

#### 8.5.6.1 Thoratec/Heartmate II

As can be seen from Figure 8-38, this new-generation LVAD is much smaller than the older devices, with a volume of 124 ml and a length of about 70 mm. This size advantage gives it the potential to help more women, teenagers, and smaller men with end-stage heart failure whose bodies are not big enough for other devices. It is also quieter and has a smaller transcutaneous power cable leading from implanted device to the controller and battery pack that are worn outside the body.

To minimize wear, a pair of cup–socket ruby bearings supports the pump rotor, with the outer boundaries of the bearing's adjacent static and moving surfaces washed directly

**FIGURE 8-38 ■**  
**Thoratec Heartmate II.** (a) Graphic showing installation into a human patient. (b) Detailed cross section of the pump. [Adapted from (Cleveland Clinic 2008).]



by the blood flow. A cylindrical magnet within the rotor is excited by a rotating magnetic field generated by the stator coils.

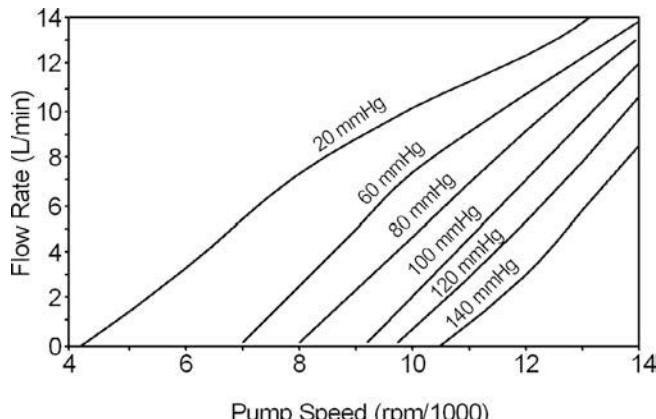
Blood flows from the inlet conduit past three neutral aerofoil-shaped guide vanes that straighten the blood flow before it encounters the rotor. Three curved blades on the rotor impart a radial velocity to the blood before it passes into the outlet stator vanes. These are twisted and convert the radial velocity to an axial one. The exit orifice narrows to convert flow velocity to pressure.

The inlet and outlet conduits are made from woven Dacron and require preclotting, the pump rotor and cowling are made from smooth titanium, and the intraventricular conduits are textured with titanium microspheres.

The performance of continuous flow pumps such as this one is determined primarily by the speed of the rotor and the pressure difference across the pump. As shown in Figure 8-39, flow rate is inversely proportional to the pressure differential across the pump. Pump characteristics are obtained by measuring the pressure differential and the flow rate as outflow resistance is gradually increased until pump shutoff.

Unlike rates of flow in pulsatile devices that are easily evaluated, the pressure-flow characteristics of dynamic pumps require a different interpretation. During the cardiac

**FIGURE 8-39 ■**  
**Characteristics of Heartmate II**  
[Adapted from (Griffith, Kormos et al., 2000).]



cycle the pump differential pressure equals aortic pressure minus left ventricular pressure plus a combined pressure loss across the inlet and outlet conduits. Because the HeartMate II is nonocclusive, it must operate at a sufficiently high speed to avoid pressure differentials that fall below normal expected aortic pressures, as these would result in reverse flow (Griffith, Kormos et al., 2000).

The controller manages the pump's rotational speed to maintain a preset pulsatility during the cardiac cycle. The pulsatility index (*PI*) is defined as

$$PI = \frac{Q_{\max} - Q_{\min}}{Q_{\text{ave}}} \quad (8.1)$$

where  $Q_{\text{ave}}$  is the average flow during a complete cardiac cycle.

Maximum flow occurs during ventricular systole when the inlet-to-outlet pressure differential is the least, and minimum flow occurs during left ventricular diastolic filling when the inlet pressure is lower and the pressure differential is a maximum. Minimal pulsatility occurs in patients with very poor left ventricles or if the pump speed is too high and the ventricle is driven to collapse. For normal operation the *PI* index is set to between 0.3 and 1.0 to ensure safe but responsive auto control.

A study of the performance of the Heartmate II, published in the *New England Journal of Medicine* (Miller 2007), showed that the device is extremely reliable. Of 133 patients at the start of the trial, 68% had not received a transplant and were still relying on the device after 1 year. The same clinical trial included patients who were not eligible to receive a heart transplant and had received the devices as a long-term destination therapy.

At the end of the first 6 months, a total of 100 patients had successful outcomes, with 56 receiving transplants, 43 remaining on the device, and 1 recovering enough to allow the device to be removed. The 25 deaths in the study before 6 months had elapsed, which included 18 patients who died before leaving the hospital, show that serious complications still occur even with the newer-generation devices. Most patients experienced some bleeding related to blood-thinning drugs used with the device, but that complication notwithstanding most experienced cardiac recovery so significant that by the end of 3 months they were moved to a less severe stage of heart failure. There were also significant improvements in quality of life scores and improvements in liver and kidney function.

### 8.5.7 Generation 3 VADs

Improvements in pump design saw a change from axial- to centrifugal- or mixed-flow types as well as the introduction of hydrodynamic or magnetically supported rotors. These devices are typically heavier than the axial-flow types with masses between 300 and 500 g, and because of their size they are also implanted below the diaphragm. They include the Ventracor VentrAssist, the Worldheart Levacor, and the Mohawk Technology MiTi Heart.

#### 8.5.7.1 VentrAssist/VentrAssist

VentrAssist is a third-generation device developed by the Australian company VentraCor. It is designed as a permanent alternative to heart transplant as well as a BTT or a BTR.

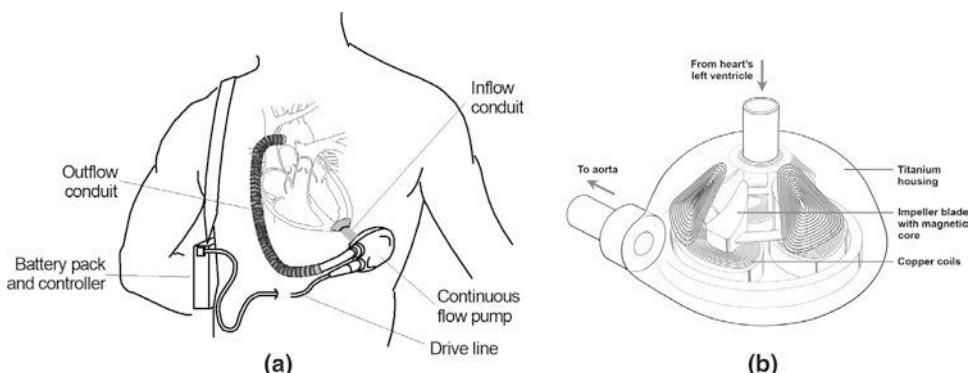
It weighs 298 g and is less than 60 mm in diameter, making it suitable for both children and adults. The device has only one moving part, a hydrodynamically suspended impeller made from a titanium alloy and covered with a diamond-like coating. The device is powered by an external battery pack with each rechargeable battery set lasting about 8 hours for a measured pump efficiency of 19%. It is also mains operable.

**FIGURE 8-40 ■**

VentrAssist VAD.

(a) Schematic diagram showing installation into a human patient. (b) Schematic diagram of the pump.

[Adapted from (Gosline 2004).]



Its strengths, compared with other similar designs, are that because of the large impeller size it rotates relatively slowly and the impeller has no shaft seals or bearings and has clean flow lines with no stagnant zones. Initial tests made at a flow rate of 5 L/min and 100 mmHg differential pressure showed a normalized index of hemolysis of 0.002–0.005 g/100L (Watterson, Woodard et al., 1999). It was found that improving the surface finish to 0.2  $\mu\text{m}$  improved this significantly, with a measured normalized index of hemolysis of 0.000167  $\pm$  0.00007 g/100L in whole human blood (James, Wilkinson et al., 2003).

The system consists of five components, shown in Figure 8-40:

- The natural heart remains in place.
- A short inflow conduit is attached to left ventricle, which delivers blood from the heart to the device.
- The outflow conduit from the pump delivers blood from the device to the ascending aorta.
- The device is implanted in the “pump pocket,” which is located on the left side of the body, behind the muscles of the abdominal wall and below the rib cage. The drive line from the pump exits from the right side of the abdomen below the ribs.
- This connects the pump to the controller and batteries worn on an external belt or backpack.

Rotating magnetic fields generated by the six copper coils in the base and walls of the unit interact with the permanent magnets mounted within the rotor and cause it to spin rapidly. Hydrodynamic forces, which result from the small clearances between the outside surfaces of the impeller and the pump walls, support it. These small clearances range from approximately 50 to 230  $\mu\text{m}$  (Chung, Zhang et al., 2004). Blood enters the center of the pump and is spun up by the rotors and forced outward by centrifugal force, where it exits through a pipe on the outside edge. A photograph of the device is shown in Figure 8-41.

Ideally, pump operating parameters are measured directly with pressure and flow sensors, but in the case of LVADs the rationale is that additional complexity leads to additional points of failure and a lower overall system reliability. Therefore, the controllers of the VentrAssist and most other LVADs estimate them using the impeller speed and input power for an assumed blood viscosity.

The VentrAssist device was first patented in 2001. The first human implant was made in June 2003 at The Alfred Hospital in Melbourne. By 2008, more than 250 implants had been made in 36 centers across 10 countries. At this time the product had achieved CE



**FIGURE 8-41 ■**  
Photograph of the  
VentrAssist LVAD.

market approval in Europe and Therapeutic Goods Administration (TGA) market approval in Australia and was undergoing FDA trials in the United States. Unfortunately, early in 2009 the company folded.

#### 8.5.7.2 Mohawk Innovative Technology/*MiTHeart*

Another third-generation blood pump has been under development for use as an LVAD by Mohawk Innovative Technology, Inc. The innovation of this pump design is a novel, patented, hybrid passive–active magnetic bearing system. The *MiTHeart* LVAD is a high-efficiency centrifugal pump that exhibits extremely low power loss, low vibration, low hemolysis, and high reliability under transient conditions and varying pump orientations.

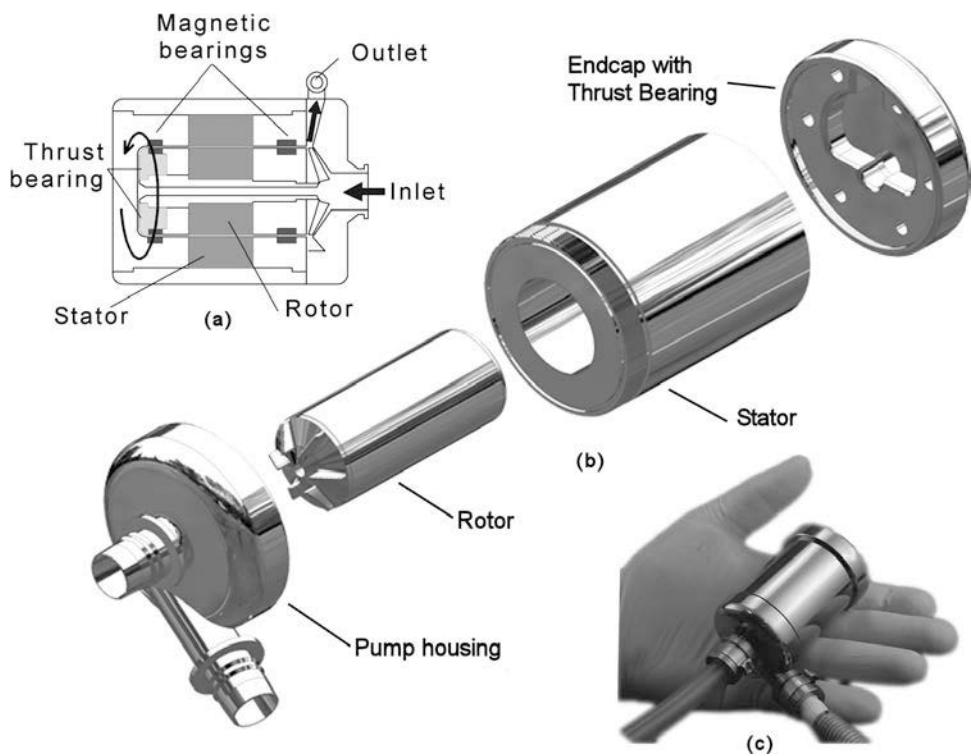
The original magnetic levitation bearing was developed for a liquid oxygen pump for the space shuttle. Subsequent research was internally funded until 1996, when the company received a Small Business Innovative Research (SBIR) grant from the NIH. The development of the LVAD took until 2005 and cost about \$15 million before the device was evaluated in comprehensive *in vitro* and *in vivo* animal tests.

Magnetic suspension has advantages from the viewpoints of power loss, wear life, and blood damage. Power lost due to bearing friction is extremely low, especially when compared with hydrodynamic rotor support systems. Most active magnetic bearing supported pumps use a magnetic bearing system with five active axes (one axial, two radial, and two tilt) to provide complete control of the pump rotor during operation. The *MiTHeart* design is different and uses a hybrid passive–active magnetic bearing system that requires only one actively controlled axis. This reduces the power required to operate the bearings and increases operating time before a battery change or recharge is required (*MiTHeart*, 2006; Jahanmir, Hunsberger et al., 2008).

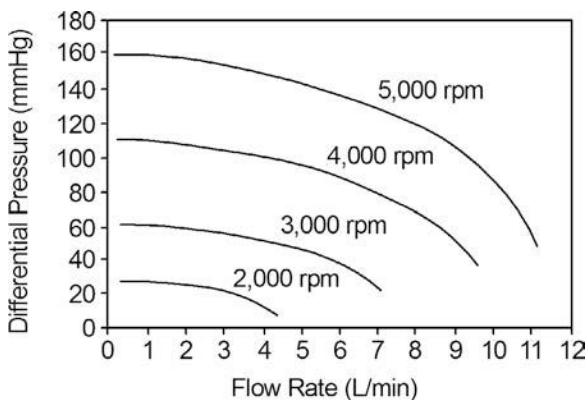
As shown in Figure 8-42, the cylindrical pump consists of four components: (1) the pump housing; (2) the stator; (3) the rotor with integrated vanes; and (4) an end cap. It is 80 mm long with a diameter of 50 mm and a total mass of 640 g. The pump is designed for a flow rate of 5 L/min at 100 mmHg pressure rise. In initial tests, flow rates from 2 to 7 L/min and pressure rises from 50 to 150 mmHg were measured. The nominal design flow of 5 L/min at 100 mmHg pressure rise was successfully achieved at a speed of 3,000 rpm.

**FIGURE 8-42 ■**

The MiTiHeart design. (a) Cross section through the integrated motor and pump. (b) Model of the pump assembly. (c) Photo of the pump. (Courtesy of MiTiHeart Corporation, reproduced with permission.)

**FIGURE 8-43 ■**

Measured performance of the latest MiTiHeart LVAD design  
[Adapted from (MiTiHeart 2006).]



However, a subsequent reduction in the size of the device has altered these characteristics somewhat, with the nominal operating point of 5 L/min and 100 mmHg occurring at 4000 rpm, as shown in Figure 8-43.

Another unique feature of the MiTiHeart design configuration is the very simple and direct flow path for both main and washing blood flows. The noncontact nature of the magnetic bearing allows for fully washed flow paths to avoid stagnation points that might promote thrombi formation. The design is sufficiently flexible to avoid stagnation points while at the same time maintaining large enough clearances throughout the flow path to also eliminate regions of high shear stress. The low shear associated with the relatively large gaps between the rotating surfaces and lower speeds (4,000 rpm compared with 10,000 rpm

**TABLE 8-1** ■ Measured Parameters for in vivo Tests with the MiTiHeart

Parameter	Test 1		Test 2	
	Mean	Std Dev	Mean	Std Dev
Speed (rpm)	2560	28	2709	48
Flow (L/min)	5.8	2.7	5.2	3.9
Aortic pressure (mmHg)	91.1	7.9	110.9	9.5
Motor power (W)	8.2	0.3	9	0.4
Bearing current (mA)	34.9	26.8	46	65

Source: Jahanmir, S., A. Hunsberger et al., *Artificial Organs* 32(5): 366–375, 2008, with permission.

in axial pumps) ensures low blood damage. Measured normalized index of hemolysis was 0.0001 g/100L. Equally important is using of noncontact bearings that eliminate the lifetime reducing wear and tribocompatibility issues present in rolling element or point contact bearing systems (Jahanmir, Hunsberger et al., 2008).

Two in vivo tests were conducted on 90 kg Holstein calves. It was found that at speeds close to 3000 rpm suction was observed in both studies but that the optimum speed to avoid suction or regurgitation was different in the two cases, as shown in Table 8-1. It can be seen that the pump speed fluctuated by about 3% around the mean and that the instantaneous flow rate varied from 3 to 10 L/min due to the pulsatility of the calves hearts. Aortic pressure fluctuated between about 80 and 110 mmHg with a mean motor power of less than 9 W.

### 8.5.7.3 Other Generation 3 VADs

These include the Terumo Duraheart, CorAide, and the EvaHeart VADS (Deng and Naka, 2007).

Terumo Heart, Inc. is a wholly owned U.S. subsidiary of Terumo Corp (Tokyo) that produces the third-generation centrifugal Duraheart. The pump is made from a titanium shell with separate compartments housing the pump system and the brushless DC motor. The magnetically levitated motor is also magnetically coupled to the impeller to be completely wear-free. Its blood-contacting surfaces are modified with a heparin immobilization technique to enhance blood compatibility.

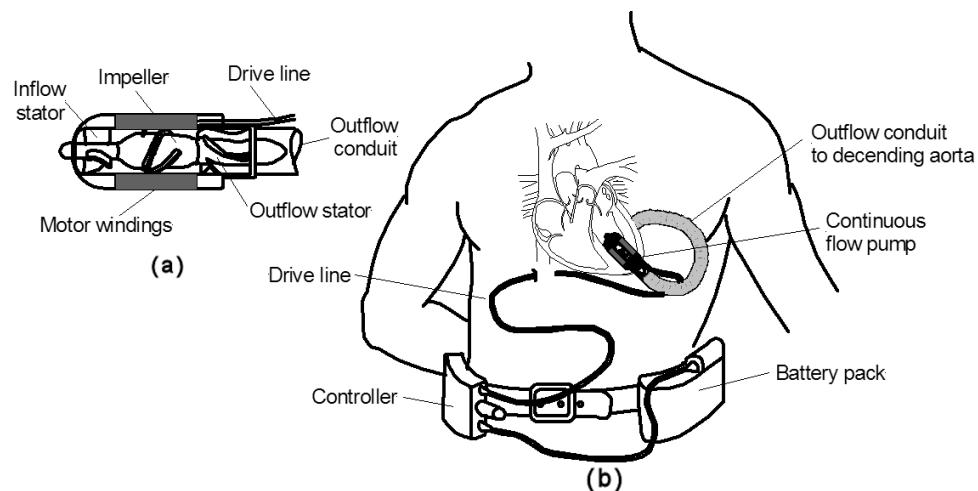
The CorAide is also a centrifugal pump made of titanium with a blood-lubricated journal-bearing supporting magnetically driven impellers. It has been plagued by rotor balance issues and unacceptably high hemolysis levels caused by incorrect journal-bearing clearance.

The EvaHeart is a centrifugal pump covered with a unique thromboresistant coating (2-methacryloyloxyethyl phosphorylcholine). The pump is driven by a stator-generated rotating magnetic field coupled to a permanent magnet within the rotor shaft. The unit is cooled by a fluid pumped by way of a percutaneous tube from an external unit.

### 8.5.8 Generation 4 VADs

The latest generation of VADs is sufficiently small and light to be implanted above the diaphragm. These have masses ranging from less than 100 g to about 150 g and include the HeartWare HVAD, the DeBakey HeartAssist 5, and the Jarvik-2000.

**FIGURE 8-44 ■**  
**Jarvik-2000 LVAD.**  
 (a) Detailed cross section of the pump mechanism.  
 (b) Graphic showing installation in a human patient.  
 Schematic diagram of [Adapted from (Jarvik\_Heart 2008).]



### 8.5.8.1 Jarvik-2000

Jarvik Heart, Inc. and the Texas Heart Institute began developing the Jarvik-2000 in 1988. This device is possibly the smallest and simplest LVAD available today—small enough to be implanted in small adults and even children. Its battery-powered flow pump is about the size of a thumb that fits inside the left ventricle, as shown in Figure 8-44.

Although recipients must take blood thinners, this device was designed to minimize the risk of blood clotting and infection. Actual hemolysis levels are not quoted by the manufacturer but are said to be “negligible,” while a small study undertaken in 2002 indicate mild hemolysis after implantation (Siegenthaler, Martin et al., 2002).

The Jarvik-2000 is FDA approved only for experimental study as a BTT device. However, the manufacturers hope that the pump will be safe and effective as a permanent assist device for a failing heart (it has a design lifetime of 10 years) and as a temporary implant to facilitate recovery of hearts treated with new medications or gene therapy approaches.

The Jarvik-2000 is an electrically powered axial-flow blood pump 25 mm wide and 55 mm long, as can be seen in Figure 8-45. It weighs 85 g. The titanium impeller houses a neodymium-iron-boron magnet, which is surrounded by stator coils that generate a rotating

**FIGURE 8-45 ■**  
 Photographs of the Jarvik-2000 pump along with prototype child and infant versions. (Courtesy of the University of Maryland.)



magnetic field that causes it to spin at high speed. The impeller is supported by blood-immersed ceramic bearings, and the whole construction is housed within a titanium shell. The normal operating speed is 8000 to 12,000 rpm, which will generate an average pump flow rate of 5 L/min. Smaller versions of the pump, suitable for children (15–25 kg) and infants (3–15 kg), are under development at the University of Maryland Medical School.

A small transcutaneous cable delivers power to the impeller. In the BTT implants the cable exits through the abdomen as shown, but in the latest permanent implants the cable exit point is via a skull-mounted pedestal similar to those used for cochlear implants.

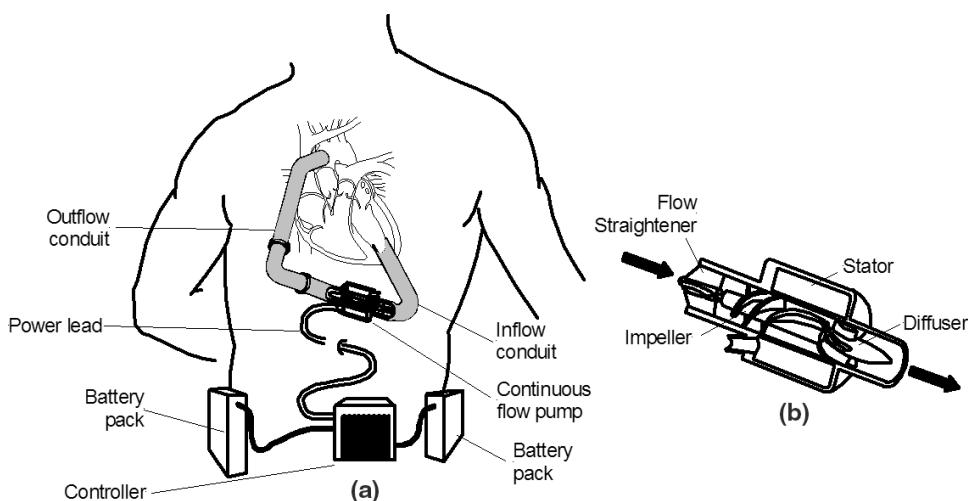
The pump speed is controlled by an analog controller, which can adjust the speed from 8000 to 12,000 rpm in increments of 1000 rpm. The control unit also monitors the pump function and the remaining power in the batteries. Audible and visual alerts notify the user of any problems.

Preliminary human trials have been positive, with few complications and only slightly increased hemolysis levels compared with preoperative levels (Siegenthaler, Martin et al., 2002). So far, the Jarvik-2000 has been used to treat more than 200 patients in the United States, Europe, and Asia. Of those, roughly 79% received the Jarvik-2000 as a BTT and 21% as a permanent implant, with a number of patients in each group being very ill, near-death cases. Nearly 70% of those patients treated as BTT have either undergone transplantation already or are currently being supported by the Jarvik-2000 (Jarvik Heart, 2008).

### 8.5.8.2 Micromed/DeBakey LVAD

The DeBakey VAD HeartAssist-5 pump system consists of a lightweight titanium pump (92 g), as shown in Figure 8-46. It is based on technology developed by NASA and licensed for use by the company for cardiovascular applications. As with all the other LVADs discussed so far, the pump is attached to an inlet cannula that is placed into the left ventricle. A graft is connected to the pump outlet and attached to the ascending aorta. In the latest versions of the system, a noninvasive probe fits around the outflow to generate an accurate measure of flow rate.

The inducer/impeller (rotor) is the only moving part of the pump. It has six blades with eight magnets hermetically sealed in each blade. The rotating magnetic field generated by



**FIGURE 8-46 ■**  
DeBakey LVAD.  
(a) Graphic showing installation of pump in a human patient.  
(b) Detailed cross section of the pump mechanism.  
[Adapted from (MedGadget 2006).]

the brushless DC motor stator induces a rotation rate of between 2500 and 7500 rpm in the impeller, which is capable of generating flow rates in excess of 10 L/min. The components are fully enclosed in a titanium flow tube that has been hermetically sealed.

The DeBakey VAD clinical data acquisition system (CDAS) is used to monitor the patient and adjust the operating parameters of the LVAD. It is designed for use during surgery and while the patient is in intensive care. The system provides power to the unit and is used to adjust the speed of the pump to increase or decrease blood flow. The CDAS also monitors critical patient data, including blood flow rate, speed of pump, and power usage.

### 8.5.8.3 HeartWare LVAD

The HeartWare LVAD with a diameter of 40 mm and a height of only 20 mm is sufficiently small for intrapericardial placement. Similar to the VentrAssist device, the HeartWare includes magnets integrated within the impeller blades driven by a rotating magnetic field generated by stator coils. The rotor is suspended by a proprietary hybrid magnetic and hydrodynamic bearing system and is capable of pumping up to 10 L/min (Deng and Naka, 2007).

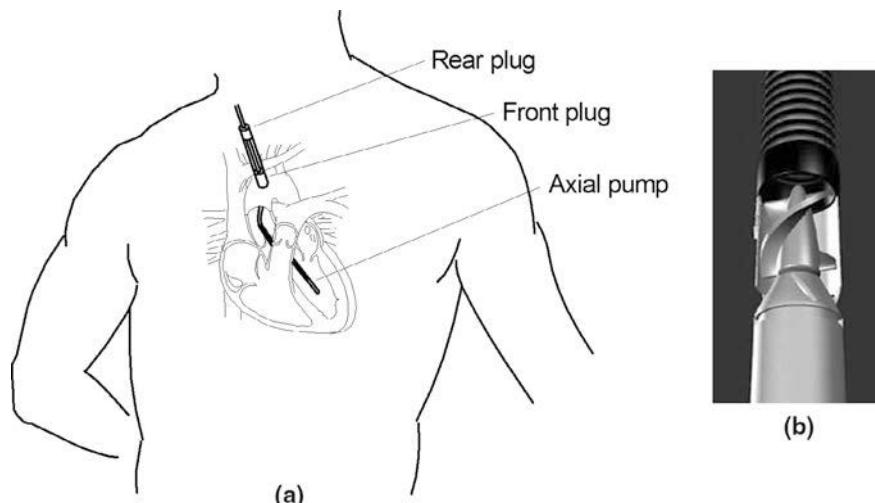
### 8.5.8.4 AbioMed/Impella

The Impella is a BTR circulatory support system that includes three different pump configurations and sizes and has been implanted in more than 1000 patients in over 200 centers worldwide. The Impella Recover system, shown in Figure 8-47, is a miniaturized impeller pump located within a catheter. The device can provide support for the left side of the heart using either the Recover LD 5.0 (implanted via direct placement into the left ventricle) or the Recover LP 5.0 LV (placed percutaneously through the groin and positioned in the left ventricle). The microaxial pump used in these systems can pump up to 4.5 L/min at a speed of 33,000 rpm. The pump is located at the distal end of a catheter that contains the electrical connections for the pump motor and sensor as well as for a separate tube used for transfer of purged fluid. Blood is pumped from the left ventricle, across the aortic valve, and into the aortic root.

This configuration mimics the natural blood flow and thus provides both an active flow and systemic pressure contribution leading to increased cardiac power output. In addition,

**FIGURE 8-47 ■**

Abiomed Impella LVAD. (a) Graphic showing the installation of the pump into the left ventricle of a human patient. (b) Photograph of the pump with the housing cut away to expose impellor. (Courtesy of AbioMed, with permission.)



by drawing blood from the left ventricle it reduces ventricular end-diastolic volume and pressure. This reduces the amount of work performed by the heart and myocardial wall tension, both of which reduce myocardial oxygen demand. It also increases the overall flow rate, which increases oxygen supply to the heart (Weber, Raess et al., 2009).

The Impella 5.0 is a variant of the directly inserted Impella LD. It is not placed surgically like the LD but rather via a peripheral blood vessel such as the femoral artery to provide the heart with left ventricular support. It is the smallest device available for left ventricular pumping; about the diameter of a pencil and weighing only 8 g, it can still pump up to 5 L/min.

This system was developed to provide ventricular support in patients who develop heart failure after heart surgery (called cardiogenic shock) and who do not respond to standard medical therapy. It provides immediate support and restores hemodynamic stability for a period of up to 7 days. Used as a bridge to therapy, it allows doctors time to develop a definitive treatment strategy or for the heart to regain its function naturally.

### 8.5.9 Toward an Ideal Replacement Heart

The ideal mechanical replacement for a human heart would be small and easy to implant. It should be made from materials that are completely biocompatible and should operate using principles to minimize hemolysis so that no medication would be required. The device should be sufficiently reliable that it will not need replacement during the lifetime of the patient while still being affordable.

Table 8-2, reproduced from Deng and Naka (2007), gives an indication of where the current generation of TAHs and LVADs is with respect to the ideal.

### 8.5.10 Other Pump Types

A number of other types of pumps for LVADs have been developed, including a shape memory alloy concept, a high-frequency (10–20 Hz) pulsatile type, rotary pulsatile devices, and a counterpulsation cuff.

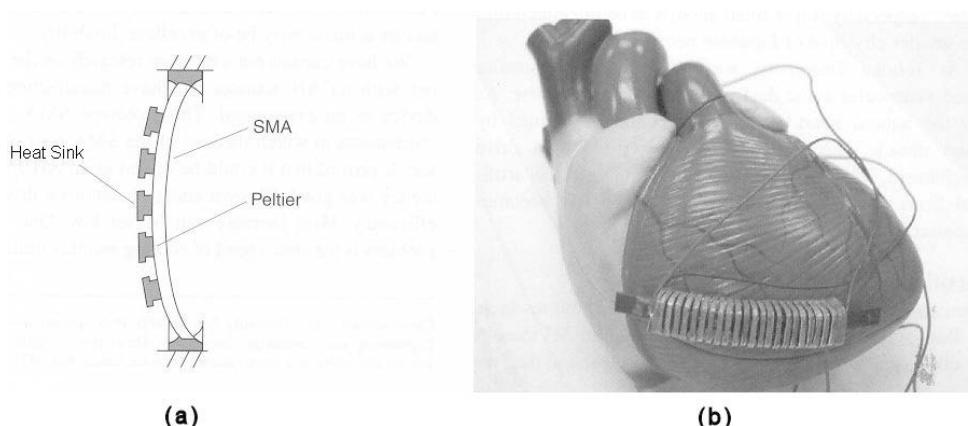
**TABLE 8-2** ■ Characteristics of Currently Available TAHs and LVADs Compared with the Ideal

Type	Size	Reliability	Implantability	Biocompatibility	Cost
<b>Ideal</b>	***	***	***	***	***
Abiocor	*		***		*
CardioWest	*	**	*		*
WorldHeart/Novacor		***		*	*
Thoratec/Heartmate	*	*	*	**	*
Arrow/LionHeart	*		***		*
WorldHeart/Novacor II	***	***	***	*	*
Thoratec/Heartmate II	**				
Ventracor	**		*	*	
Terumo/Duraheart	**		*		
CorAide	**				
EvaHeart	*				
Jarvik-2000	**	***	*		*
Micromed/DeBakey	**				

Source: Deng, M. and Y. Naka, *Mechanical Circulatory Support Therapy in Advanced Heart Failure*, London: Imperial College Press, 2007, with permission.

**FIGURE 8-48 ■**

Novel heart assist device. (a) Graphic showing the combined SMA–Peltier actuator. (b) Photograph of actuator on a plastic model of the heart. (Yambe, Maruyama et al., 2001), reproduced with permission



### 8.5.10.1 Shape Memory Alloy Driven VAD

An alternative to having an external pump is to retain the left ventricle as the pumping chamber but to develop a method to assist the heart muscle to contract. The contractile properties of SMA can be used to this effect. It can easily be heated quickly enough to change state and apply the required force to the ventricle, but natural cooling is generally much slower. A solution to this problem is to use an array of Peltier elements to both heat and cool the SMA mechanism much more quickly (Yambe, Maruyama et al., 2001). A model of the device is shown in Figure 8-48.

### 8.5.10.2 Pulsatile Rotary Pumps

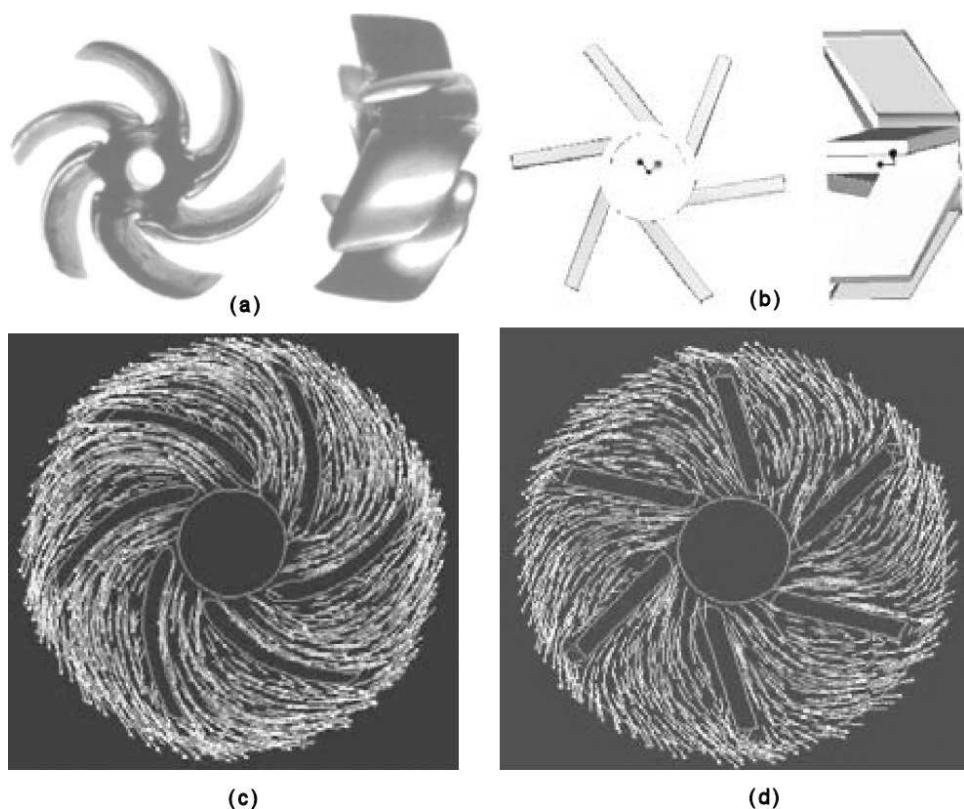
Pulsatile rotary pumps operate by cyclically varying the rotation speed of a centrifugal pump to impose a synthetic cardiac rhythm. To minimize increases in Reynolds's shear stress during acceleration and deceleration, the impeller vanes are curved (Kunxi and Ying, 2008). Figure 8-49 shows a comparison between the flow lines of a centrifugal pump with curved impeller vanes and a conventional one with straight vanes. It can be seen that in the former the blood flow follows the curved vanes with minimal impact or separation, which results in a lower Reynolds's shear stress and lower hemolysis.

### 8.5.10.3 Sunshine Heart/C-Pulse

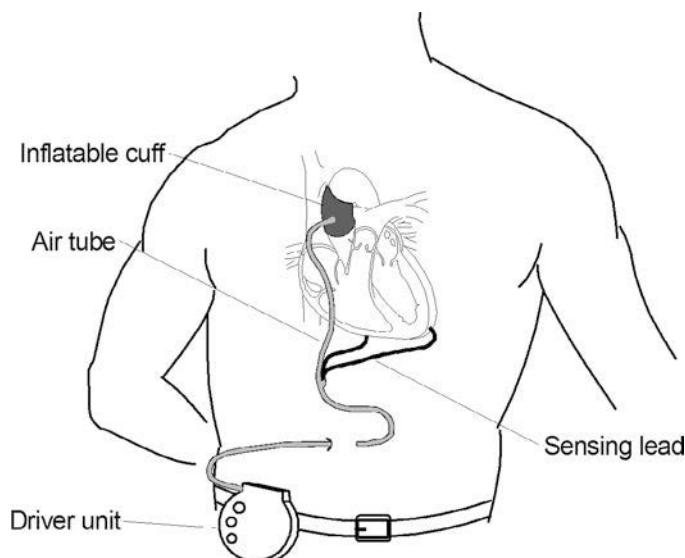
The C-pulse is a pliable, inflatable cuff that encircles the patient's aorta, as shown in Figure 8-50. It inflates and deflates in time with the patient's heart to perform a function known as counterpulsation, which enhances blood flow and reduces the workload of the left ventricle.

When the heart is filling, the C-pulse inflates, pumping blood from the aorta into circulation. Just before the left ventricle begins its contraction, the cuff deflates, increasing the volume available in the aorta, decreasing the aortic pressure, and therefore decreasing the heart workload.

Because there is no blood contact and the C-pulse is inserted with minimum trauma, it is an effective and safe method to counteract the effects of moderate heart failure.



**FIGURE 8-49** ■ Centrifugal pump performance (a) CAD model of curved bladed impellor. (b) CAD model of straight bladed impellor. (c) CFD model showing fluid flow for pump with curved impellor blades. (d) CFD model showing fluid flow for pump with straight impellor blades. (Kun-xi and Ying 2008)



**FIGURE 8-50** ■ The C-pulse counterpulsation LVAD.

## 8.6 | ENGINEERING IN HEART ASSIST DEVICES

### 8.6.1 Fluid Dynamics in Pulsatile LVADs

Thrombosis formation is determined by the physiological condition of the patient, the biocompatibility of the materials in contact with the blood, and fluid dynamic characteristics of the blood pump. Adverse flow factors that can induce thrombosis include turbulence, recirculation stasis (stagnation), and high shear stresses. Pump performance can be determined after manufacture, but that is expensive if major redesign is required. Therefore, it is more cost-effective to produce accurate models of the device and to use computational fluid dynamic (CFD) analysis to predict flow prior to manufacture.

Detailed analysis has been performed for a typical pusher-plate LVAD (Okamoto, Fukuoka et al., 2001) under the following conditions:

- Inflow pressure 10 mmHg
- Outflow pressure 100 mmHg
- Pusher-plate velocity 27 mm/s
- Pump stroke volume 65 cm<sup>3</sup>
- Pump stroke 12 mm
- Blood density  $1.06 \times 10^{-3}$  g/mm<sup>3</sup>
- Blood viscosity  $6 \times 10^{-3}$  g/mm · s

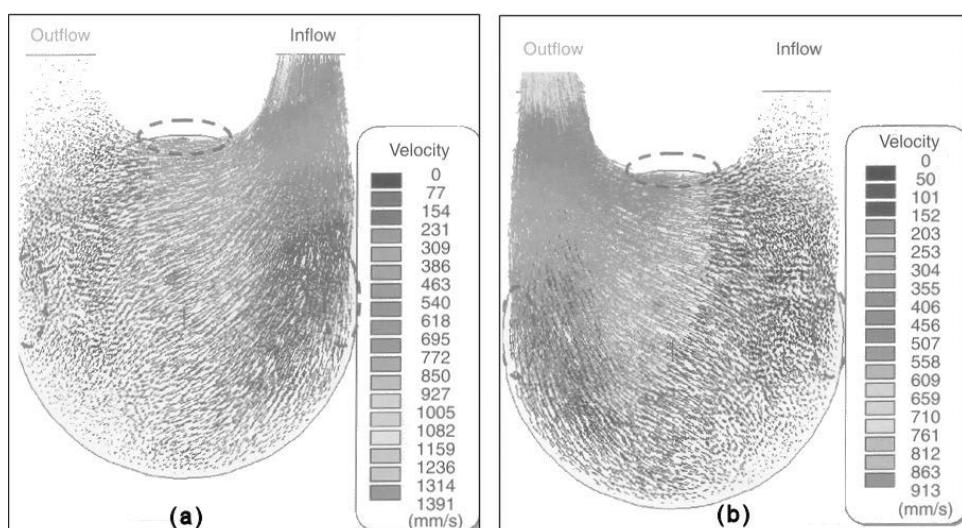
The analysis can be used to estimate the NIH normalized index of hemolysis for each of the different pump configurations as follows.

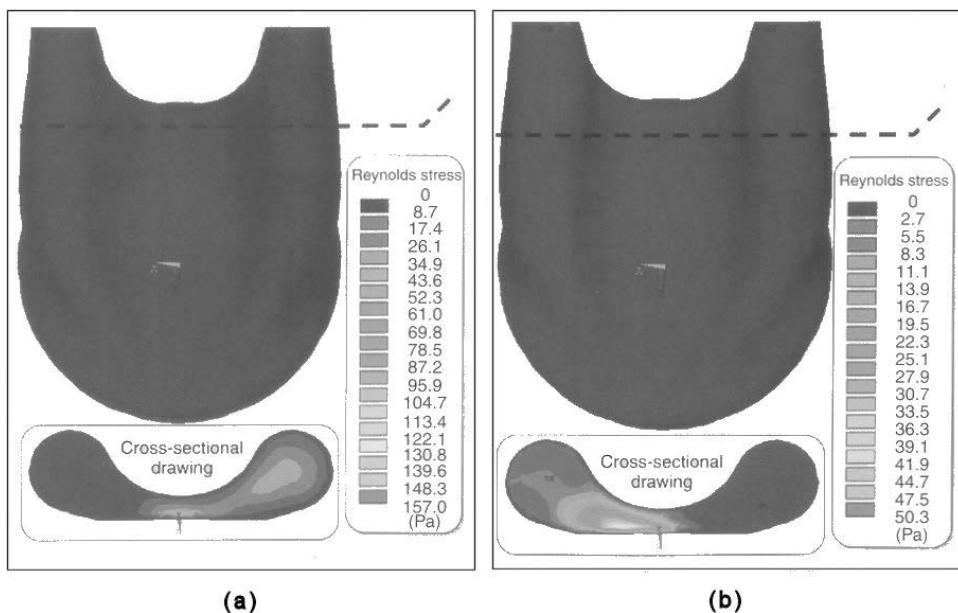
It is first necessary to use CFD to determine the distribution of the flow velocity during the filling and ejection cycles. This is shown in Figure 8-51.

The turbulent kinetic energy,  $k$ , can then be determined from the time-averaged fluctuating component of the fluid velocity,  $\bar{u}_i$ ,

$$k = \frac{1}{2} \bar{u}_i^2 \quad (8.2)$$

**FIGURE 8-51 ■**  
Distribution of flow velocity for a pulsatile pump.  
(a) During filling.  
(b) During ejection.  
(Okamoto, Fukuoka et al., 2001), copyright Informa Healthcare, reproduced with permission





**FIGURE 8-52** ■ Distribution of turbulent kinetic energy in a pulsatile pump. (a) During filling. (b) During ejection. (Okamoto, Fukuoka et al., 2001), copyright Informa Healthcare, reproduced with permission.

The Reynolds's shear stress,  $\tau$ , is defined as

$$\tau = -\rho \bar{u}_i \bar{v}_j \quad (8.3)$$

where  $\rho$  is the density,  $u_i$  is the fluctuating component of the velocity in one direction, and  $v_i$  is the fluctuating component at right angles to  $u_i$ .

Using (8.2) and (8.3), the shear stress can be approximated by

$$\tau = -2\rho k \quad (8.4)$$

A snapshot of the distribution of the Reynolds's shear stress through the dotted cross sections is shown in Figure 8-52.

It has been experimentally observed that red blood cell damage in shear flow is due to two factors acting at the same time: (1) the level of shear stress; and (2) the exposure time of the blood cell membrane to these stresses. A basic model, developed by Wurzinger, which relates the damage to the effects of shear stress on blood corpuscles, is

$$L_{RBC}(\%) = \frac{\Delta Hb}{Hb} = 3.62 \times 10^{-5} \times t^{0.785} \times \tau^{2.416} \quad (8.5)$$

where  $L_{RBC}(\%)$  is the lysis rate for red blood cells,  $Hb$  is the haemoglobin content,  $\Delta Hb$  is the damaged hemoglobin content, and  $t$  is the exposure time.

To obtain accurate results, the incremental lysis rate as a blood cell travels through the pump is determined. This is achieved by calculating a number of flow lines from the input to the output of the pump and determining the integrated lysis rate for each.

This lysis rate, along with the total blood volume, the flow rate, and the hematocrit (%), is then used to determine the free hemoglobin volume and ultimately the normalized index of hemolysis.

The authors show that small changes to the shape of the pump were able to decrease the Reynolds's stress by a significant margin with a result that the normalized index of hemolysis was reduced significantly. The results are summarized in Table 8-3.

**TABLE 8-3** ■ Calculated Results of Normalized Index of Hemolysis

Parameter	Model #1	Model #2
Reynolds's stress (filling)	28.6 Pa	7.3 Pa
Reynolds's stress (ejection)	14.5 Pa	3.1 Pa
Lysis rate (filling)	0.03311%	0.0012%
Lysis rate (ejection)	0.0064%	0.00015%
Normalized index of hemolysis	2.72 g/100L	0.098 g/100L

Source: Okamoto, E., S. Fukuoka et al., *Journal of Congestive Heart Failure and Circulatory Support* 1(4): 391–398, 2001, with permission.

Because pulsatile pumps have slow-moving components, they rely on low Reynolds's shear stresses to minimize hemolysis levels. CFD-based design is critically important to make sure that this does in fact occur.

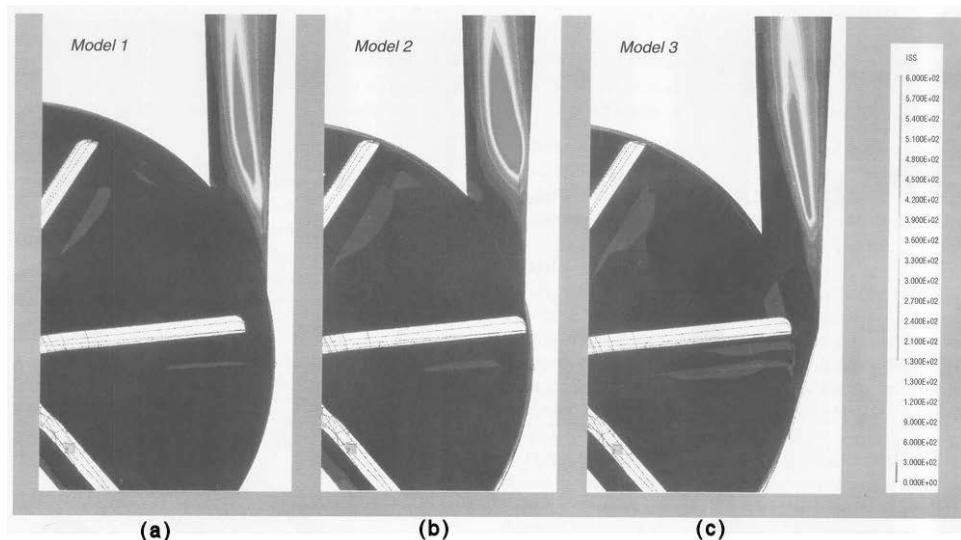
### 8.6.2 Fluid Dynamics in Centrifugal and Axial LVADs

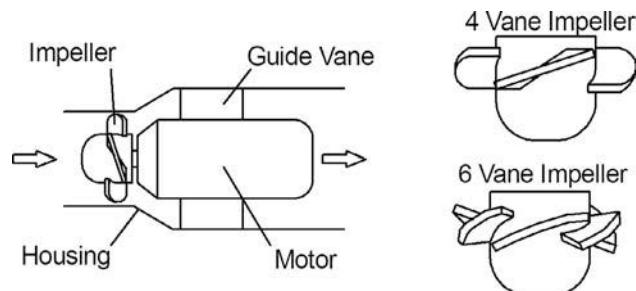
CFD analysis of centrifugal and other rotary pumps has led to improved designs that reduce both shear stresses and stagnation. As an example of the analysis now possible, the following section discusses a CFD-based investigation of a reasonably conventional centrifugal design (Tsukamoto, Ito et al., 2001). A number of different impeller lengths and outlet diffuser configurations were compared from a shear stress perspective, as shown in Figure 8-53.

Figure 8-53b shows a high shear stress region expanding from the outlet diffuser. It is larger than those produced by either Figure 8-53a or Figure 8-53c. Hemolysis calculations showed that Figure 8-53b was worse by a factor of two than the other models. Further analysis showed that wash-out holes effectively reduced flow stagnation behind the impeller and that the size and location of the holes is very important.

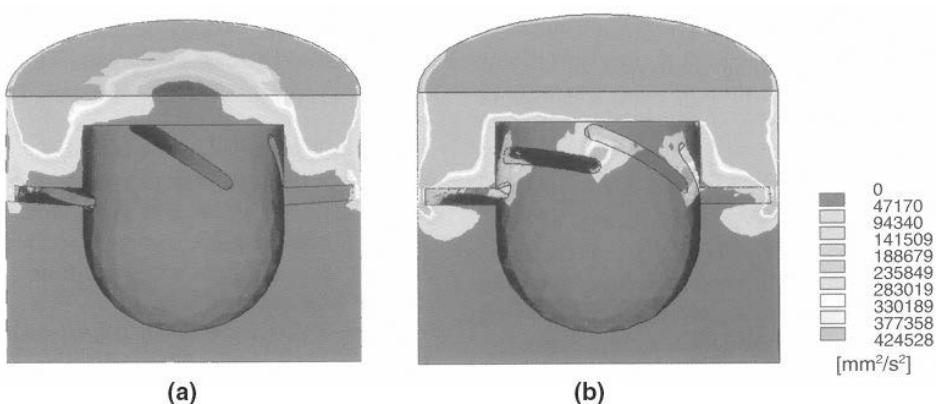
Modeling of axial pumps has also been undertaken (Mitamura, Nakamura et al., 2001). As with the other studies, the CFD analysis was used to determine the turbulent

**FIGURE 8-53** ■ Comparison of the shear stress distribution in three different pump designs. (a) Original design. (b) Longer impeller design. (c) Altered output diffuser design. (Tsukamoto, Ito et al., 2001), reproduced with permission.





**FIGURE 8-54** ■ Structure of an axial-flow pump. [Adapted from (Mitamura, Nakamura et al., 2001).]



**FIGURE 8-55** ■ Turbulent Kinetic energy in an axial pump. (a) Four vane impeller. (b) Six vane impeller. (Mitamura, Nakamura et al., 2001), copyright Informa Healthcare, reproduced with permission.

kinetic energy along streamlines through the pumps for a 100 mmHg head and a flow rate of 5 L/min. Analysis was conducted for pumps with four and six vanes, as shown in Figure 8-54.

From the turbulent kinetic energy results shown in Figure 8-55, the Reynolds's shear stress was then calculated, which allowed hemolysis to be estimated. To achieve this, a similar analysis to that discussed in the section on fluid dynamics of pulsatile devices was performed.

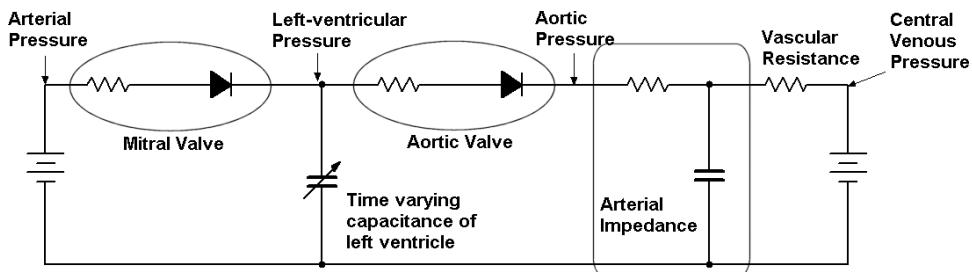
The predicted hemolysis based on the integrated shear stress along 30 stream lines flowing through each of the pumps is compared with measurements made using animal blood. A correlation coefficient of 0.87 was obtained, indicating that the modeling process is reasonably accurate.

Because the hemolysis level is a function of both the Reynolds's shear stress and the exposure time of the red blood cell membranes to these forces, it is possible to achieve the required low level even from a fast-operating axial pump if the exposure time is short enough.

Care should be taken when using Wurtzinger's formula, equation (8.5), because it has been found to be incorrect for high shear but short episodes ( $\tau > 100 \text{ Pa}$   $t < 5 \text{ ms}$ ) or in low shear but long episodes ( $\tau < 5 \text{ Pa}$   $t > 2 \text{ s}$ ). In these cases, the hemoglobin release is smaller than predicted by two or three orders of magnitude.

It is clear from this brief introduction that off-the-shelf designs are not suitable for use as blood pumps and that comprehensive analysis using CFD is essential if a biocompatible device is required.

**FIGURE 8-56** ■ Circuit model for the left ventricle and its load [Adapted from (Paden, Ghosh et al., 2000).]



### 8.6.3 Estimation and Control of Blood Flow

Most LVAD systems receive blood through an inflow cannula that is inserted into the apex of the left ventricle through a “cored” hole. The device then pumps the blood back into circulation through an outflow cannula grafted into the ascending aorta. Thus, the LVAD operates in parallel with the left ventricle. Since the pressure in the left ventricle varies as the heart beats, the LVAD is subjected to cyclical variations in loading that must be accommodated.

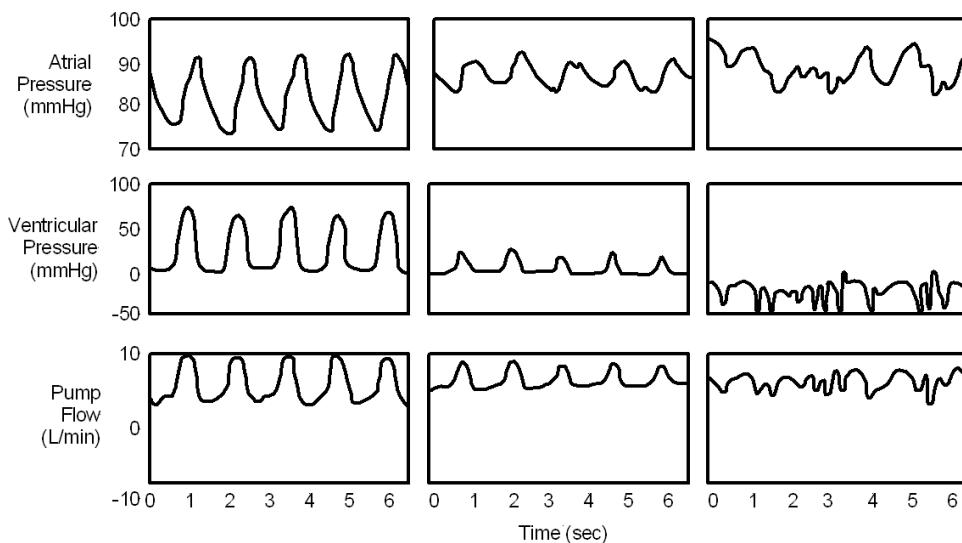
System models for simulation often replace the mechanical components with their electrical equivalents to form a circuit in which current corresponds to blood flow and voltage corresponds to blood pressure. An example of such a model for the left ventricle is shown in Figure 8-56.

The electrical model of the LVAD is placed in parallel with the two series resistor-diode sections.

In general, pulsatile systems concentrate on regulating flow or arterial pressure synchronously with the natural beating of the heart by using the passive filling of the LVAD pumping chamber to regulate stroke volume and stroke rate. In centrifugal and axial-flow pumps typical of the newer generation of LVADs, only the rotation speed can be controlled to maintain the required pressure and flow. In these cases flow is mostly governed by supply from the venous system, and innate feedback is provided by resistance to blood flow. This open-loop control is adequate over a small operating range, but as soon as the patient becomes rehabilitated and starts to resume a normal lifestyle LVADs must be capable of responding to demand.

Axial-flow pumps do not use valves, so it is possible for blood to flow backward through the pump if it is run too slowly. Conversely, if it runs too quickly it may attempt to pump more blood from the ventricle than is available, causing ventricular collapse or kinking in the inlet cannula. This is illustrated *in vivo* using a Thoratec Heartmate implanted in a calf in Figure 8-57. At 9000 rpm the minimum pump flow is close to zero and the ventricular pressure is always positive, reaching about 70 mmHg at its peak. At 10,000 rpm the flow is much less pulsatile with a minimum value significantly greater than zero. The pressures are also less pulsatile with the peak ventricular pressure reaching only about 30 mmHg. Finally, if the speed is increased to 11,000 rpm, the pulsatile pattern is lost and the pump flow oscillates erratically. This last condition corresponds to ventricular suction and can result in damage to the heart muscle if it is not identified and corrected quickly (Antaki, Boston et al., 2003).

The normal heart is controlled via the nervous system to maintain a mean arterial blood pressure (MABP) of 90 mmHg. As discussed earlier in this chapter, in the natural heart this set point is maintained by feedback from baroreceptors in the atria.



**FIGURE 8-57 ■**  
Waveform changes for different rotation speeds of a Thoratec Heartmate.  
(a) 9,000 rpm.  
(b) 10,000 rpm.  
(c) 11,000 rpm.  
(Adapted from [Antaki, Boston et al., 2003].)

Flow rate and pressure difference (or head) are key variables needed in the control of implantable rotary blood pumps. However, use of flow or pressure probes can decrease reliability and increase system power consumption and expense. For a given fluid viscosity, the flow state is determined by any two of the four pump variables: flow, pressure difference, speed, and motor input power. Thus, if viscosity is known or if its influence is sufficiently small, flow rate and pressure difference can be estimated from the motor speed and motor input power (Tanaka, Yoshizawa et al., 2001).

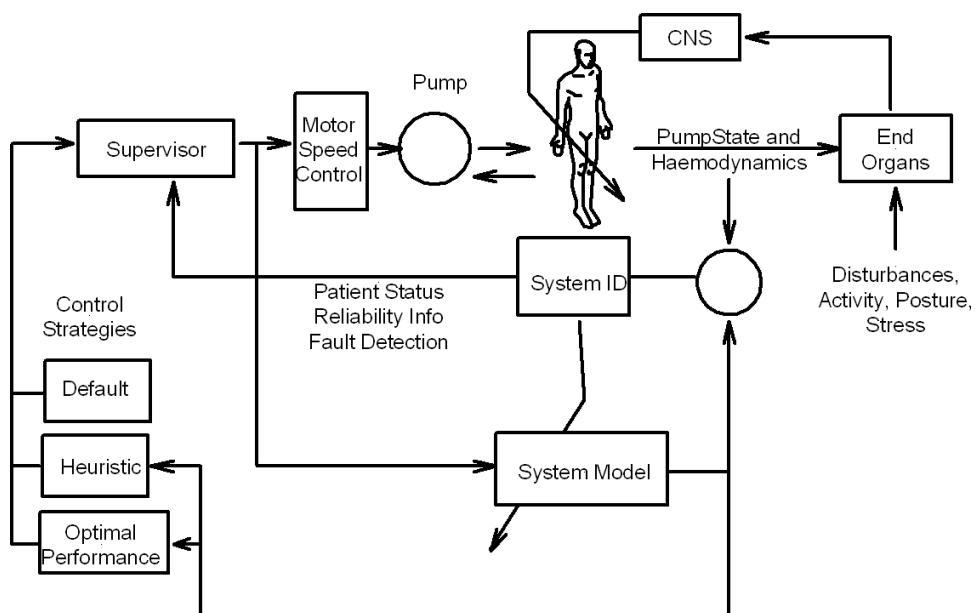
As discussed in Paden, Ghosh et al., (2000) and Antaki, Boston et al., (2003), any cardiac augmentation system must function within the following constraints:

- Cardiac output should be above a minimum value. This is nominally 5 L/min but will vary between 3 and 6 L/min depending on the size of the patient.
- Left atrial pressure should be maintained below 10 to 15 mmHg to avoid pulmonary edema and above 0 mmHg to avoid suction.
- Systolic arterial pressure should be maintained within specific limits to ensure an adequate oxygen supply while avoiding risks associated with hypertension.
- System should be maximally efficient in terms of blood flow and pump power.

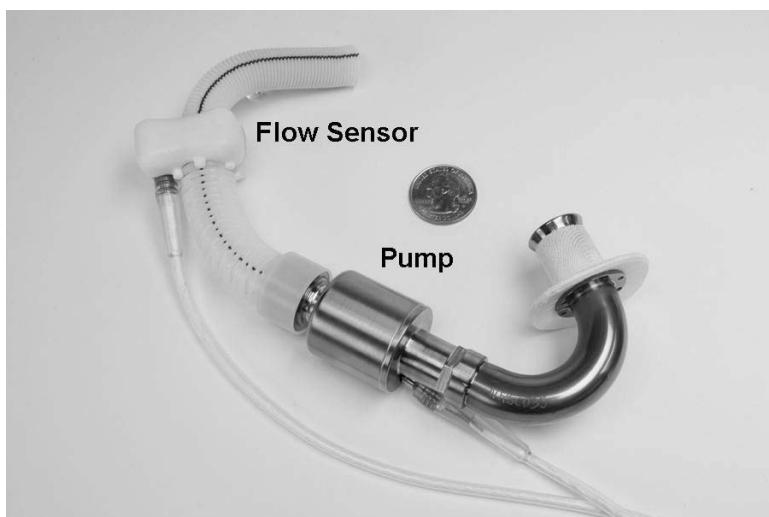
In general, it is not possible to minimize all of these simultaneously, so control systems are designed that optimize performance based on cost functions associated with deviation from the constraint. These cost functions are normally asymmetrical because of hard minima below which the patient cannot function. An intelligent controller based on multiobjective optimization of these parameters as well as information about the patient's activity level is used to control the pump motor speed, as shown in Figure 8-58.

Other control architectures have been developed. For example, Vollkron discusses a closed-loop controller for a DeBakey VAD that uses venous return based on flow pulsatility (peak-to-peak flow variation) as well as the available return derived from the patient's own heart rate (desired flow) along with power use and minimal flow as inputs. These are analyzed on a beat-to-beat basis within a 10-second moving window before the motor speed is adjusted (Vollkron, Schima et al., 2006).

**FIGURE 8-58** ■ Control architecture for an axial pump based LVAD [Adapted from (Antaki, Boston et al., 2003).]



**FIGURE 8-59** ■ DeBakey pediatric LVAD with a noncontact flow sensor. (Courtesy of micromedcv.com, reproduced with permission.)

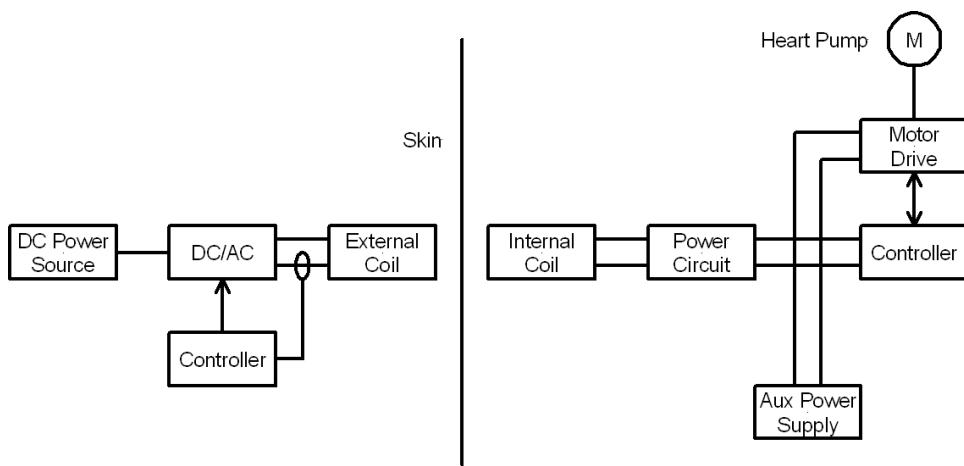


The latest DeBakey LVAD includes a noncontact flow measurement device based on magnetohydrodynamic principles, as shown in Figure 8-59. Measuring rather than inferring flow simplifies the control process significantly.

#### 8.6.4 Transcutaneous Energy Transfer

A transcutaneous energy transfer (TET) system to power an artificial heart consists of components both external and within the body of the patient, as shown in Figure 8-60. The external components include the following:

- DC power source: This is a battery pack with sufficient capability to provide 10 W to power a VAD and 20 W to power a TAH for between 6 and 8 hours. It uses



**FIGURE 8-60** ■  
Block diagram showing the components of a TET system.

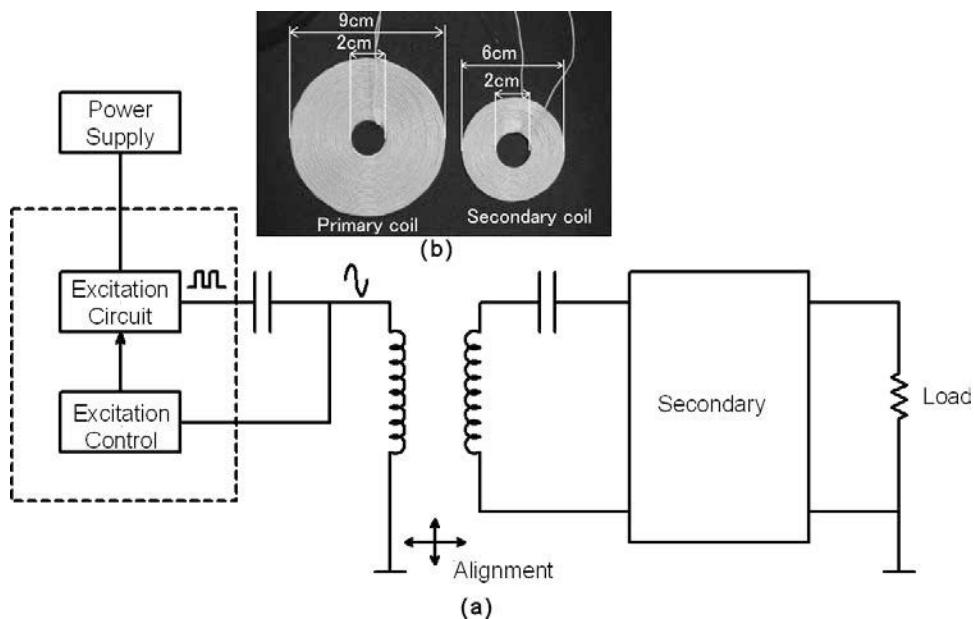
rechargeable cells with the highest power density possible to minimize the mass that the patient needs to carry.

- DC/AC converter: This consists of an inverter or power oscillator that generates an adjustable square wave signal at a frequency of between 100 kHz and 1 MHz typically.
- Controller: The energy transfer efficiency depends on the alignment of the internal and external coils and their mutual resonant frequency. The controller monitors the current to the coil and adjusts the DC/AC converter parameters to maintain optimum conditions for transfer. It also detects faults and closes down the circuit if the coils are completely out of alignment.
- External coil network: This is the primary coil, between 5 and 10 cm in diameter, that is excited by the AC signal and generates a changing magnetic field that couples through the skin into the subcutaneous secondary coil. It is generally part of a tuned LC resonant circuit.

The internal components include the following:

- Internal coil network: This coil is hermetically sealed and inserted just below the skin. It is typically a little smaller than the primary to make alignment as simple as possible. It is also part of the tuned LC circuit so that it captures the highest possible proportion of the radiated magnetic energy.
- Power circuit: Consists of a rectifier and regulator to produce the required DC voltages to power the control electronics and the pump.
- Auxiliary power supply: An internal battery that allows the device to operate for a reasonable period (typically 30 minutes) without external power. It also includes the electronics to recharge the internal battery when external power is available.
- Controller: Most heart pumps consist of brushless DC motors that must be controlled to suit the patient's requirements. The controller monitors the rotation rate and the current requirements and adjusts the supply voltage to suit.
- Drive circuit: Depending on the motor types, the drive circuit is usually a half- or full-bridge pulse-width modulated MOSFET circuit that chops the DC supply to drive the heart pump.

**FIGURE 8-61** ■ Transcutaneous energy transfer device. (a) Simplified schematic diagram showing the coupling method. (b) Photograph of typical primary and secondary coils.



Though all of the components of this system are critical, the energy transfer efficiency is governed by the coupling between the primary and secondary coils. To maximize this, the coupling system is generally resonant, with the excitation frequency adjustable to maintain this resonance with changing load and coil alignment, as shown in Figure 8-61.

The excitation circuit generates a pulse train with the required frequency and duty cycle to excite a series resonant LC circuit where the inductance includes the inductance of the primary and the inductance of the secondary as seen by the primary.

Experiments have shown that energy transmission efficiencies AC-to-AC across the transformer can be as high as 94.5% at frequencies between 500 and 800 kHz for a power output of 20 W (Shiba, Nukaya et al., 2006).

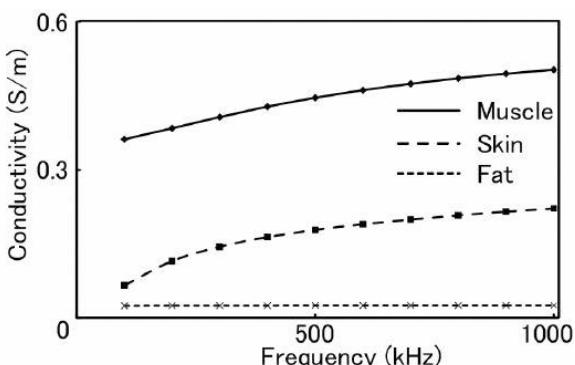
The specific absorption rate (SAR) (W/kg) is a good indication of the temperature increase that can be expected in the skin, fat, and muscle in the region of the transformer. This can be estimated by

$$\text{SAR} = \frac{\sigma E^2}{\rho} \quad (8.6)$$

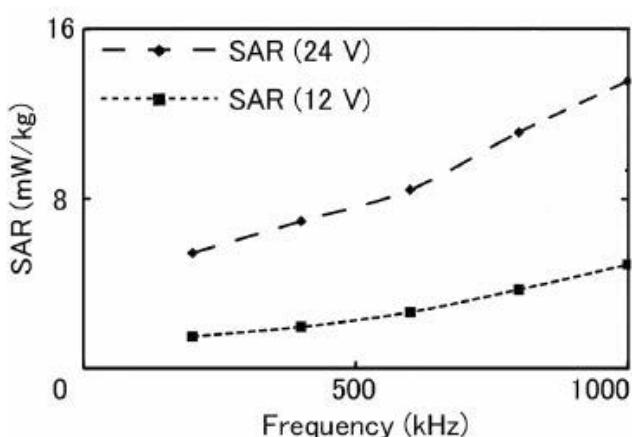
where  $E$  (V/m) is the root mean square (RMS) electric field strength,  $\sigma$  (S/m) is the biological tissue conductivity, and  $\rho$  (kg/m<sup>3</sup>) is the tissue density.

The density can be taken as a constant 1000 kg/m<sup>3</sup>, but the conductivity is dependent on the tissue type and the measurement frequency, as shown in Figure 8-62.

For a fixed power transmission of 20 W across the transformer, with an output voltage of either 12 or 24 V, the maximum recorded SAR is 14 mW/kg, which is well within the 2 W/kg maximum allowed for general public exposure. It can be seen from Figure 8-63 that the SAR increases with frequency.



**FIGURE 8-62** ■ Conductivity of human tissue as a function of frequency. [Adapted from (Shiba, Nukaya et al., 2006).]



**FIGURE 8-63** ■ Maximum SAR as a function of frequency [Adapted from (Shiba, Nukaya et al., 2006).]

## 8.7 | PUMP TYPES

Turbo pumps, also known as dynamic or roto-dynamic pumps, produce a head and a flow by increasing the velocity of the liquid through the device with the help of a rotating vane impeller. In broad terms there are two classes of turbo pumps. In the first class, there is a pronounced change in radius from the inlet to the discharge. These are known as centrifugal pumps. The other main class, axial pumps, creates a flow that is mostly parallel to the axis of rotation. Mixed-flow pumps lie somewhere between the two extremes, with the flow proceeding along a conical surface of revolution.

Positive displacement pumps operate by alternately filling a cavity and then displacing a given volume of liquid. The positive displacement pump delivers a constant volume of liquid for each cycle irrespective of the discharge pressure. They can be classified as follows:

- Rotary pumps: gear, lobe, screw, vane, regenerative (peripheral), and progressive cavity.
- Reciprocating pumps: piston, plunger, and diaphragm.

### 8.7.1 Centrifugal and Axial Pump Characteristics

Fluid enters the central portion, called the eye, flows radially outward, and is discharged around the entire circumference into a casing. During flow through the rotating impeller, the fluid receives energy from the vanes, resulting in an increase in both pressure and absolute velocity. Since a large portion of the energy leaving the impeller is kinetic, it is necessary to reduce the absolute velocity and transform much of this energy to pressure head. This is accomplished in the volute casing surrounding the impeller or in flow through diffuser vanes. If a higher capacity is required without an increase in the diameter, the pump dimensions in the direction parallel to the shaft must be increased (Daugherty and Franzini, 1977). This becomes a mixed-flow pump and is shown in Figure 8-64 along with a number of examples of centrifugal pumps.

The configuration of the axial pump, as shown in Figure 8-65, occurs at the limit where no radius change exists between the streamlines moving from the inlet to the outlet, and centrifugal action plays no part.

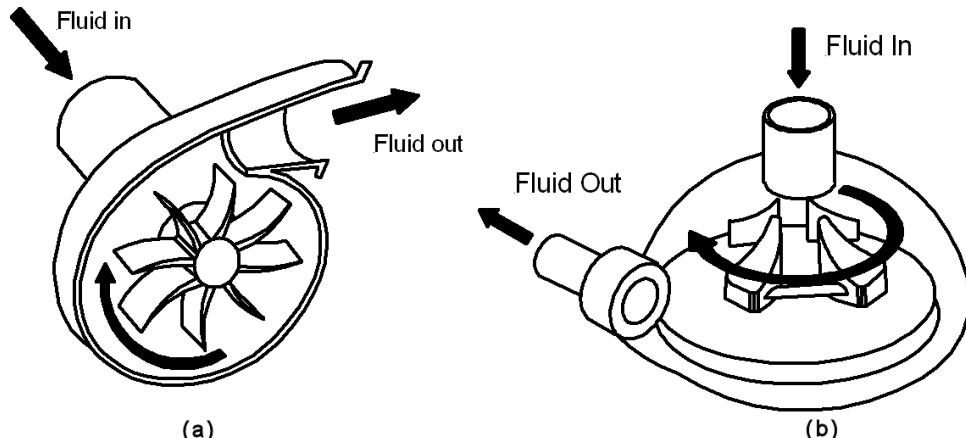
#### 8.7.1.1 Torque, Energy, and Power

The required torque,  $\tau$  (Nm), from the motor can be expressed in terms of the centrifugal pump velocity,  $n$  (rpm), as

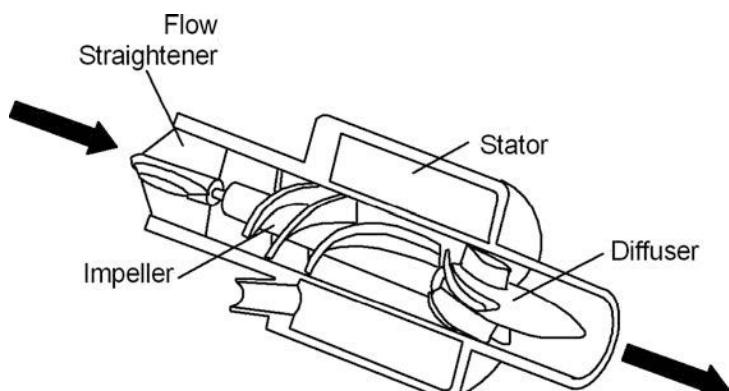
$$\tau = kn^2 \quad (8.7)$$

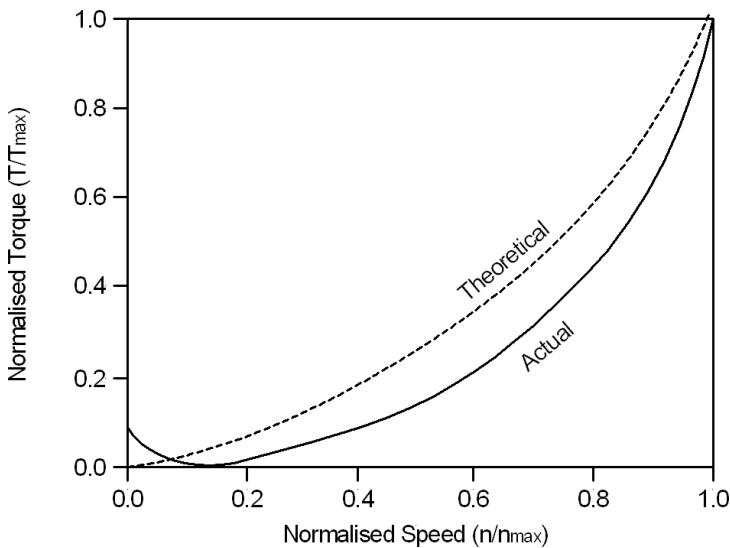
where  $k$  is a constant.

**FIGURE 8-64 ■**  
Common turbo pump configurations.  
(a) Centrifugal.  
(b) Mixed-flow.



**FIGURE 8-65 ■**  
Axial pump type.





**FIGURE 8-66** ■  
Theoretical and measured characteristics of a centrifugal pump.

The theoretical characteristic of a centrifugal pump torque is therefore a parabola starting from the origin and proportional to the square of the speed, as shown in Figure 8-66.

With the discharge valve closed, the required torque amounts to between 30 and 50% of the nominal torque at full speed.

The relationship between the power and the torque is

$$P = \omega\tau \quad (8.8)$$

where  $\tau$  (Nm) is the torque,  $P$  (W) is the power, and  $\omega$  (rad/s) is the angular rotation rate. This can be rewritten in terms of the pump speed,  $n$  (rpm),

$$P = \frac{\pi n}{30} \tau \quad (8.9)$$

In a volute pump in which fluid enters the pump through the eye of the impeller and is accelerated radially outward from the pump casing, a partial vacuum is created at the impeller's eye, continuously drawing fluid into the pump.

The kinetic energy of the rotating fluid is determined according the Bernoulli equation. The energy transferred corresponds to the velocity at the edge or vane tip of the impeller. Therefore, the faster the impeller revolves or the larger its diameter, the higher will be the velocity of the liquid and the larger its kinetic energy.

### 8.7.1.2 Pressure and Head

The kinetic energy of a liquid coming out of an impeller is reduced by creating resistance in the flow. The first resistance is created by the pump casing. From the conservation of energy, when the liquid slows down the kinetic energy is converted to pressure and some heat. Stated another way, a pump creates not pressure but only flow, and pressure is a measurement of the resistance to flow. It is the resistance to the pump's flow that is read on a pressure gauge attached to the output.

Pump pressure specifications are often given in non-SI units such as mmHg, and these must be converted to Pascals or N/m<sup>2</sup> for calculation purposes. Table 8-4 lists the conversion factor among a number of units for a pressure of one atmosphere.

**TABLE 8-4** ■ Common Units of Pressure

Units	Atmospheric Pressure
Atmospheres (atm)	1
Bar	1.01325
Pascal (Pa) ( $\text{N}/\text{m}^2$ )	$1.01325 \times 10^5$
Pounds/in <sup>2</sup> (psi)	14.6960
Inches Hg	29.9213
mm of Hg (torr)	760
Inches of $\text{H}_2\text{O}$	406.8
cm of $\text{H}_2\text{O}$	1033.3
Dynes/cm <sup>2</sup>	$1.01325 \times 10^6$

The conversion of mmHg to Pa is therefore  $1.01325 \times 10^5 / 760 = 133.32$ .

If the discharge of a centrifugal pump is pointed straight up into the air, the fluid will be pumped to a certain height, which known as the shutoff head. This maximum head is mainly determined by the outside diameter of the pump's impeller and the speed of the rotating shaft. In Newtonian fluids (nonviscous liquids like water or petrol) the term *head* is used to measure the kinetic energy of the fluid because pressure changes with changes in the specific gravity of the liquid but the head does not.

Centrifugal pumps are *constant head machines*, but not constant pressure machines since pressure is a function of head and density. The head is constant, even if the density (and therefore pressure) changes.

Using the energy equation, the head rise through a pump can be expressed as

$$h_a = \frac{p_2 - p_1}{\rho g} + h_2 - h_1 + \frac{v_2^2 - v_1^2}{2g} \quad (8.10)$$

where  $h_a$  (m) is the total head rise developed,  $p_2$  ( $\text{N}/\text{m}^2$ ) is the pressure at the outlet,  $p_1$  ( $\text{N}/\text{m}^2$ ) is the pressure at the inlet,  $\rho$  ( $\text{kg}/\text{m}^3$ ) is the density,  $g$  ( $9.81 \text{ m}/\text{s}^2$ ) is the acceleration due to gravity,  $v_1$ (m/s) is the velocity at the inlet, and  $v_2$ (m/s) is the velocity at the outlet.

Outlet velocity is reduced because of the following losses through the machine:

- Skin friction in the blade passages
- Flow separation
- Impeller blade casing clearance flows
- Other three-dimensional flow effects

The flow rate,  $Q$  ( $\text{m}^3/\text{s}$ ), is the product of average velocity of the fluid at the outlet, and  $A$  ( $\text{m}^3$ ), the cross sectional area of the outlet pipe.

$$Q = v_2 A \quad (8.11)$$

For a very common medical installation, the inline pump, where the inlet velocity and the outlet velocity are equal ( $v_2 = v_1$ ), and the inlet and outlet elevation are also equal ( $h_2 = h_1$ ), then (8.10) can be modified to

$$h_a = \frac{p_2 - p_1}{\rho g} \quad (8.12)$$

The density of blood is slightly higher than that of water, being about  $1062.1 \text{ kg/m}^3$ . If the outlet and inlet pressures of a LVAD are, respectively,  $120 \text{ mmHg}$  and  $20 \text{ mmHg}$ , the head rise provided by the pump is

$$h_a = \frac{(120 - 20) \times 133.32}{1062.1 \times 9.81} \\ = 1.27 \text{ m}$$

From this result, it can be seen that if a major artery is severed the heart would be capable of pumping blood even higher than this because the full  $120 \text{ mmHg}$  head would be available into the atmosphere.

### 8.7.1.3 Specific Speed

The specific speed determines the general shape of a centrifugal pump impeller. As the specific speed increases, the ratio of the impeller outlet diameter to the inlet of the eye diameter decreases. This ratio becomes unity for an axial-flow pump.

Radial flow impellers develop head through centrifugal force and are low-flow high-head designs, whereas pumps with higher specific speeds develop head partly by centrifugal force and partly by axial force. In the limit, an axial pump develops head using axial forces only. Specific speed,  $n_s$ , can be determined using

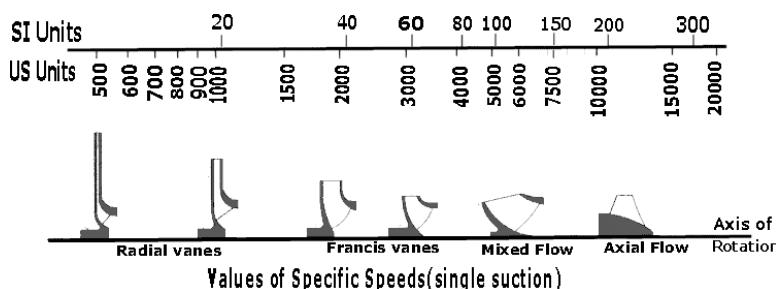
$$n_s = \frac{n_e \sqrt{Q}}{h^{3/4}} \quad (8.13)$$

where  $n_e$  (rpm) is the rotation speed for maximum efficiency,  $Q$  ( $\text{m}^3/\text{s}$ ) is the flow rate, and  $h$  (m) is the head delivered by the pump at the point of maximum efficiency.

As shown in Figure 8-67, high  $n_s$  pump impellers have inlet diameters that approach or equal the outlet diameter and relatively large open flow passages. Low  $n_s$  pump impellers have outlet diameters that are much larger than the inlet diameters and relatively narrow flow passages.

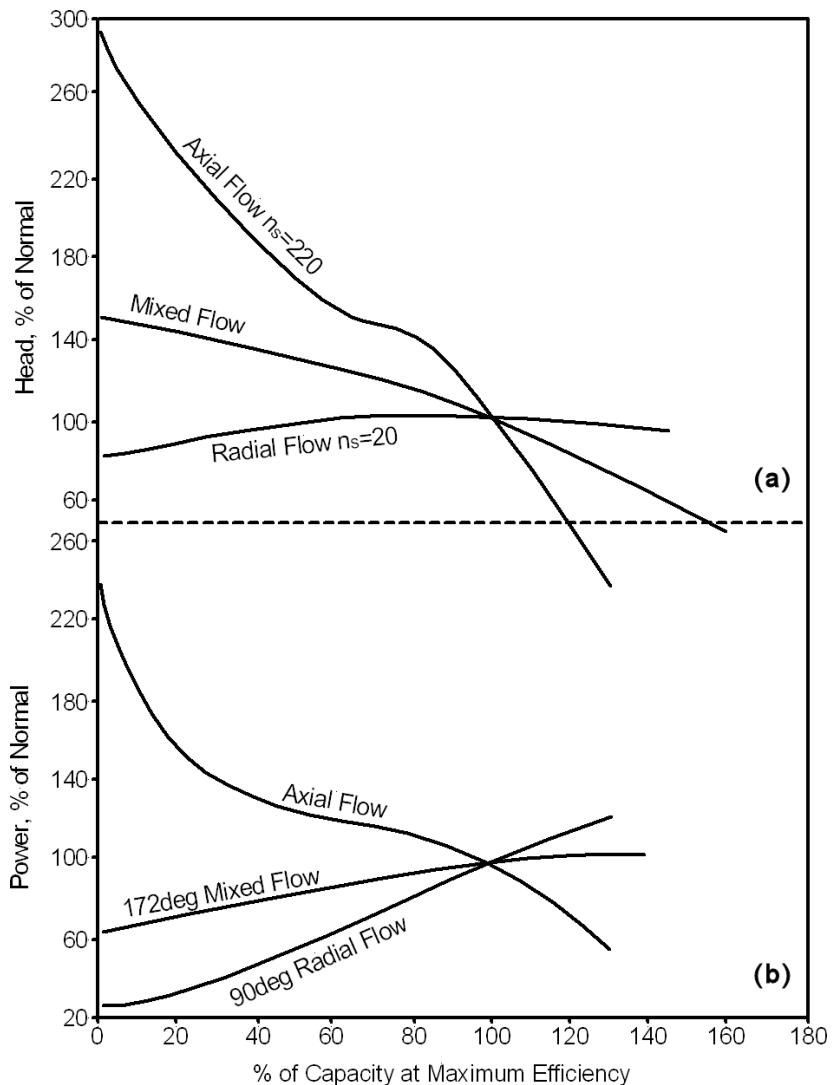
By using different impeller and casing designs, it is possible to vary the pump characteristics over a wide margin, as shown in Figure 8-68. A flat characteristic allows considerable variation in the flow rate with little change in the head, whereas a steep characteristic gives a small variation in flow for a large change in head.

It can be seen that the axial-flow pumps have a much steeper head capacity curve than centrifugal pumps, and instead of power at shutoff being a minimum, as it is for centrifugal pumps, it is not only a maximum but is also much larger than the power required at the point of maximum efficiency. This is a disadvantage both in starting and also running at low capacity. This consideration is reflected in the narrow operating speed for axial pumps used in LVADs.



**FIGURE 8-67** ■ Relationship between the pump configuration and specific speed.

**FIGURE 8-68 ■**  
 Head capacity characteristics for different types of centrifugal and axial-flow pumps at constant speed  
 [Adapted from (Daugherty and Franzini 1977).]



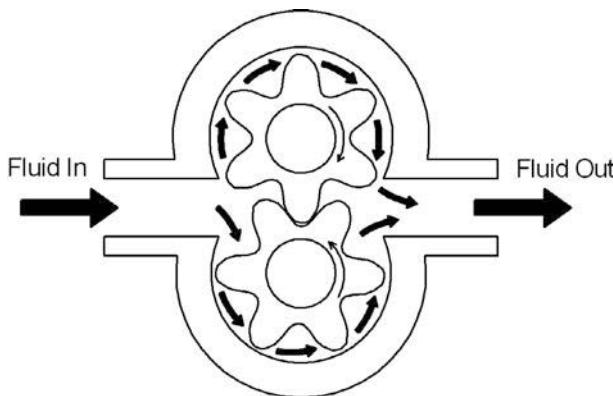
### 8.7.2 Rotary Pump Characteristics

Rotary pumps operate by filling a cavity and then displacing a given volume of liquid. They deliver a constant volume of liquid for each cycle irrespective of the discharge pressure.

#### 8.7.2.1 Gear Pumps

The pump includes two gears, of which one is driven by a motor, as illustrated in Figure 8-69. The most popular gear types are straight spur, though these can be noisy and subject to vibration if they are not manufactured to high standards. Helical gears can be used to minimize vibration, but high side loads result and the use of double helical gears to eliminate side loads results in expensive units.

When operating with straight spur gears, these pumps are reliable low-cost units that can be run for long periods if operated correctly. They have good high pressure operating characteristics. However, close tolerances are required between the internal components for the pump to operate effectively.



**FIGURE 8-69 ■**  
Schematic diagram showing gear pump construction.

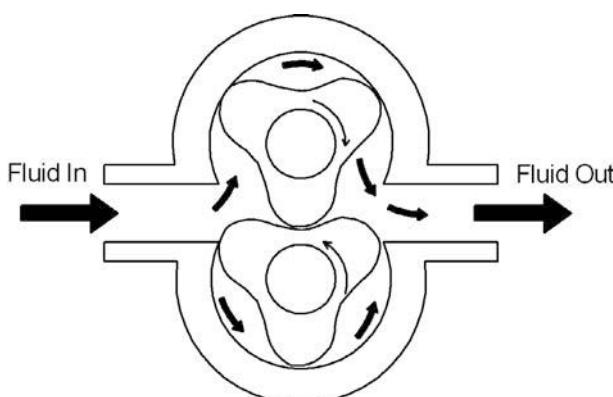
The gear pump has moderate efficiency, and it is not recommended for handling suspended solids. Because the gears are in contact, the fluid can be highly sheared as it is transferred. These pumps can transfer fluids at reasonable flow rates at developed heads of up to 200 bar if designed correctly.

### 8.7.2.2 Lobe Pumps

Lobe pumps are based on two parallel rotors located within a shaped case, as illustrated in Figure 8-70. The rotors include a number of lobes that are arranged so that as the rotors turn they contain spaces that increase and reduce in volume. Fluid enters these spaces through the inlet connection and is trapped as the rotors turn until it is discharged through the outlet port. The rotors are synchronized by external gears, and therefore the internal contact between the lobes is a sealing contact and not a driving contact.

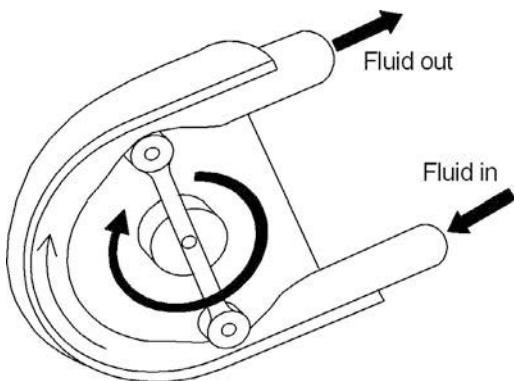
Various shapes of rotor are used, with the trilobe rotor being the most popular. The lower the number of lobes the better the pump is for handling viscous and solids-laden fluids. The rotor can be made from a wide range of materials including steel and reinforced rubber. When soft rotors are used this type of pump can achieve high levels of volumetric efficiency.

Lobe pumps generate relatively low internal fluid velocities with low level of shear and are therefore suitable for blood pumps. The resulting flow includes some pulsation. The pump can also run dry without damage, subject to the design of the bearings. It is also self-priming if the rotors are wetted. As the pump has clean internal surface with few crevices it can be used for hygiene-related applications.



**FIGURE 8-70 ■**  
Lobe pump schematic diagram.

**FIGURE 8-71** ■  
Schematic diagram  
of a peristaltic pump.



### 8.7.2.3 Peristaltic Pumps

Peristaltic pumps are based on an elastomeric tube through which the fluid is forced by the action of a number of lobes or rollers that progressively squeeze along its length, as illustrated in Figure 8-71. The tube should be closed by at least one lobe/roller at all times throughout the pumping cycle. The squeezing items are generally located on the rotating support, which is driven by a variable speed drive. This mechanism includes no glands and is very smooth operating.

The flow rate of the pump is related directly to the diameter of the tube and the speed of rotation of the drive. The pump duty is limited by the tube material used in construction. The suction capability is related to the tube's ability to rapidly expand after the compression cycle. This pump type can generate heads of up to 5 m at flows of up to  $10 \text{ m}^3/\text{hr}$ .

### 8.7.3 Reciprocating Pump Characteristics

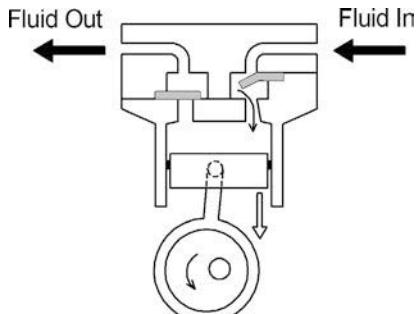
The two primary forms of reciprocating pumps used in medical applications are piston and diaphragm pumps.

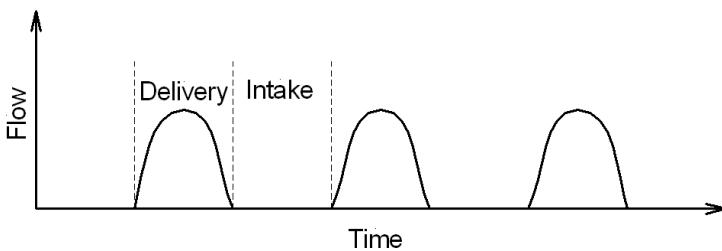
#### 8.7.3.1 Reciprocating Piston Pumps

In a reciprocating piston pump, a motor-driven cam pulls the piston back and forth in the pump head, as shown in Figure 8-72. A flexible seal around the periphery of the piston prevents leakage of the liquid from the back of the pump. Check valves mounted in the head open and close in response to small changes in pressure to maintain a one-way flow of the liquid.

During the intake stroke, reciprocating cylinder pumps increase the volume of the pump cavity, which reduces the pressure and draws the liquid in through an inlet valve.

**FIGURE 8-72** ■  
Schematic diagram  
of a reciprocating  
piston pump.





**FIGURE 8-73** ■  
Pump flow as a function of time.

Once this stroke is complete, the pump starts to reduce the volume of the pump cavity. This results in an increase in pressure, which closes the inlet valve and opens the outlet valve to allow the liquid to escape.

During the delivery stroke, flow increases from zero up to a maximum and then decreases back to zero during the intake stroke, when flow is zero. The pressure inside the pump changes in the same way as flow, going from zero to a maximum value before reverting to zero during the intake stroke, as shown in Figure 8-73.

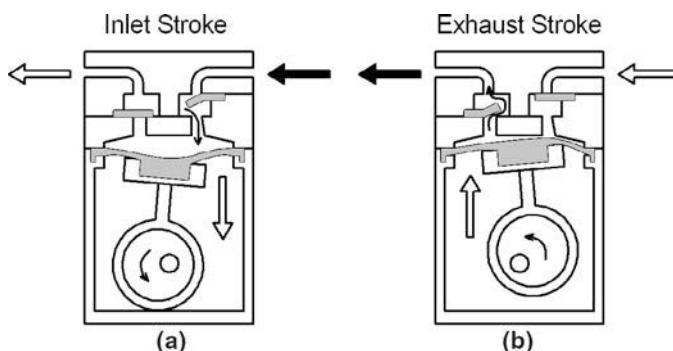
For some applications these pressure pulses are undesirable, but because the cardiovascular system is designed to accommodate such variations pulsatile pumps are commonly used for artificial hearts and LVADs.

There are many ways to generate a more uniform flow, if that is required. One approach is to keep the single-piston design but to vary the shape of the cam or the speed of the motor. The shape of the cam leads to a flatter flow curve at the middle of the delivery stroke. In addition, the motor can be made to speed up during the intake stroke and slow down during the delivery stroke. Some pulsation still remains, however, and these pumps often use some form of pulse dampening to further reduce the flow fluctuations if required.

Another common approach combines the output flow from two heads operating 180 degrees out of phase, such that the intake stroke from one head coincides with the delivery stroke from the other. This means that while one cylinder is filling the pump cylinder the second cylinder is delivering. Then, when the second refilling, the first cylinder delivers. These two flows can be combined by feeding each pump output into a tee that connects a common outlet. The inlet line from the reservoir likewise is fed to a tee, which branches to feed both cylinders of the pump.

### 8.7.3.2 Diaphragm Pumps

The operating principle of a diaphragm pump and its construction are very simple, as can be seen in Figure 8-74. The diaphragm is clamped at its circumference between the pump housing and the pump head. An eccentric cam displaces the connecting rod, which in turn moves the diaphragm back and forth. This produces a periodic change in volume of the



**FIGURE 8-74** ■  
Schematic diagram of a diaphragm pump. (a) Inlet stroke. (b) Exhaust stroke.

working chamber, similarly to a piston pump. In combination with automatic inlet and exhaust valves, this change in volume produces a pumping action. To prevent excessive stretching of the elastic diaphragm, a support is arranged below the diaphragm.

Simple diaphragm pumps can also be driven by an electromagnet or by a voice-coil type of arrangement excited by an AC supply. They are less efficient than the cam driven variety as they rely on the stiffness of the diaphragm material or a spring to perform the pump stroke.

Depending on the type of drive used, diaphragm pumps can be classified as high speed and low speed. Low-speed pumps operate up to about 300 strokes/min. The flow rate,  $Q$  ( $\text{cm}^3/\text{s}$ ), of such pumps exhibits a reasonably linear relationship with the speed of rotation of the drive motor. This makes such low-speed diaphragm pumps especially suitable for metering in medical applications. The flow rate can be varied by changing the stroke of the diaphragm and changing the rotation speed of the motor.

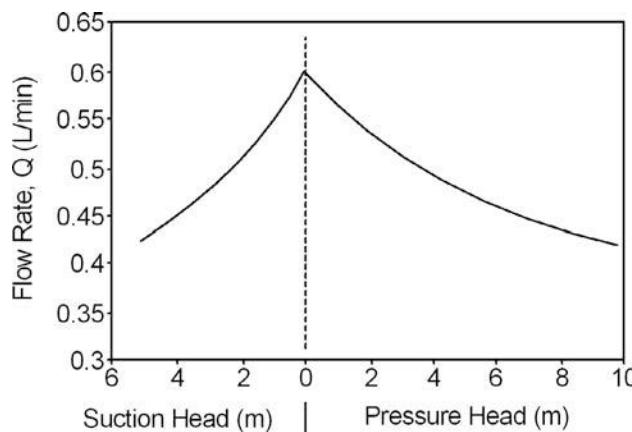
High-speed diaphragm pumps operate at between 2500 and 3000 strokes/min, about 10 times as fast as their slow-running counterparts. Their size is generally in inverse proportion to their speed, which allows very compact pumps to be manufactured.

The important characteristics of these pumps include the following:

- Reliability, thanks to their simple and sturdy construction.
- Chemical resistance; all liquid-contact parts can be made of chemically resistant materials such as polytetrafluoroethylene (PTFE), perfluorinated elastomer (FFPM), or polyvinylidene fluoride (PVDF).
- Can operate in any position.
- Very compact.
- Quiet.
- Self-priming because they can also handle gases and gas–liquid mixtures.
- Can run dry.
- Maintenance-free.
- Have long working lives.

A typical high-speed diaphragm pump running at 3000 strokes/min has the flow rate versus head characteristic shown in Figure 8-75. Note that deformation of the diaphragm

**FIGURE 8-75 ■**  
Performance curves  
for a small  
high-speed  
diaphragm pump.



decreases the displaced volume slightly, and therefore the flow rate reduces slightly with increasing head.

Most total artificial hearts and pulsatile ventricle assist devices use variations of the diaphragm pump operating at very low speed.

### WORKED EXAMPLE

---

Consider a conventional pulsatile LVAD like the LionHeart that consists of a brushless DC motor driving a pusher plate using a ball-screw mechanism. The pusher-plate diameter is 8 cm, and the stroke volume is  $V = 70 \text{ cm}^3$ , which makes the stroke length approximately

$$\begin{aligned}\Delta s &= \frac{4V}{\pi d^2} \\ &= \frac{4 \times 70}{\pi \times 8^2} \\ &= 1.39 \text{ cm}\end{aligned}$$

The area,  $A$  ( $\text{cm}^2$ ), of the pusher plate that applies a force to the diaphragm is

$$\begin{aligned}A &= \frac{\pi d^2}{4} \\ &= \frac{\pi \times 8^2}{4} \\ &= 50 \text{ cm}^2\end{aligned}$$

Assume that the pressure against which the plate must push is constant and equal to the highest arterial pressure, which is 120 mmHg. As the conversion of mmHg to Pa is 133.32, this equates to 16 kPa (16,000 N/m<sup>2</sup>).

The force,  $F$  (N), is the product of the pressure,  $P$  (N/m<sup>2</sup>), and the area,  $A$  (m<sup>2</sup>),

$$\begin{aligned}F &= PA \\ &= 16 \times 10^3 \times 50 \times 10^{-4} \\ &= 80 \text{ N}\end{aligned}$$

The energy, or work done  $E$  (J), is the product of the force,  $F$  (N), and the distance, or pump stroke,  $\Delta s$  (m),

$$\begin{aligned}E &= F\Delta s \\ &= 80 \times 1.39 \times 10^{-2} \\ &= 1.1 \text{ J}\end{aligned}$$

Assuming that the LVAD beats at 72 BPM and that 60% of the period is used for filling the pump and 40% for ejection, the amount of time available for ejection is

$$\begin{aligned}t_{ej} &= \frac{60}{72} \times 0.4 \\ &= 0.33 \text{ s}\end{aligned}$$

The filling time,  $t_{fill} = 0.5$  s.

The mechanical power,  $P_{mech}$  (W), used to perform the pumping function is the work done per unit time

$$\begin{aligned} P_{mech} &= \frac{E}{t_{ej}} \\ &= \frac{1.1}{0.33} \\ &= 3.33 \text{ W} \end{aligned}$$

Assuming that no energy is expended in reversing the pump stroke, then the average power used over a full pump cycle is

$$\begin{aligned} P_{ave} &= \frac{t_{ej}}{t_{ej} + t_{fill}} P_{mech} \\ &= \frac{0.33}{0.33 + 0.5} \times 3.33 \\ &= 1.32 \text{ W} \end{aligned}$$

In previous sections of the chapter, pulsatile pump efficiencies of about 20% ( $\eta = 0.2$ ) are quoted; therefore, the average electric power required to drive the pump is

$$\begin{aligned} P_{elec} &= \frac{P_{ave}}{\eta} \\ &= \frac{1.32}{0.2} \\ &= 6.6 \text{ W} \end{aligned}$$

This equates quite well with the 8 W power requirement quoted for most pulsatile devices.

---

### 8.7.4 Bearings

Impellers in rotary blood pumps must be supported by bearings. These can include conventional sealed bearings made from stainless steel or ceramics, blood-immersed bearings, magnetically supported bearings, hydrodynamically supported impellers, or combinations of these.

It is believed that the high shear stresses formed around blood-immersed bearings may lead to increased hemolysis, with the result that most new developments aim to eliminate this and use magnetic or hydrodynamic bearings with quite large gaps.

Magnetic bearings are often supported by permanent magnets augmented by a number of electromagnets used to maintain stability and drive the impeller. Such support is used by the MiTiHeart.

## 8.8 | REFERENCES

---

- Abe, Y., T. Ono, T. Isoyama, S. Mochizuki, K. Iwasaki, T. Chinzei, I. Saito, A. Kouno and K. Imachi. (2000). "Development of a Miniature Undulation Pump for the Distributed Artificial Heart." *Artificial Organs* 24(8): 656–658.
- Al-Ghazal, S. (2002). "Ibn Al-Nafis and the Discovery of Pulmonary Circulation." Retrieved September 2008 from <http://www.islamonline.net/english/Science/2002/08/article06.shtml>

- Antaki, J. F., J. R. Boston, and M. A. Simaan. (2003). "Control of Heart Assist Devices." *Proceedings of the 42nd IEEE Conference on Decision and Control, 2003.*
- Arrow LionHeart. (2006). "LionHeart LVAS." Retrieved September 2008 from <http://www.hmc.psu.edu/lionheart/>
- Baker, M. (2004). "Heart Transplant Reunion Party Celebrates Lifesaving Milestone." Retrieved September 2008 from <http://news-service.stanford.edu/news/medical/2004/april14/transplant.html>
- Bonsor, K. (2008). "HowStuffWorks: How Artificial Hearts Work." Retrieved September 2008 from <http://health.howstuffworks.com/artificial-heart.htm/printable>
- Boylan, M. (2007). "Galen: On Blood, the Pulse and the Arteries." *Journal of the History of Biology* 40: 207–230.
- Chung, M., N. Zhang, G. Tansley and Y. Quin. (2004). "Experimental Determination of Dynamic Characteristics of the VentrAssist Implantable Rotary Blood Pump." *Artificial Organs* 28(12): 1089–1094.
- Cleveland Clinic. (2008). "Ventricular Assist devices (VAD)." Retrieved September 2008 from [http://my.clevelandclinic.org/heart/disorders/heartfailure/lvad\\_devices.aspx](http://my.clevelandclinic.org/heart/disorders/heartfailure/lvad_devices.aspx)
- Cohn, L. (2003). "Fifty Years of Open-Heart Surgery." *Circulation* 2003(107): 2168–2170.
- Cooley, D. (2003). "The Total Artificial Heart." *Nature Medicine* 9: 108–111.
- Cooley, D., D. Liotta, G. Hallman, R. Bloodwell, R. Leachnam and J. Milam. (1969). "Orthotopic Cardiac Prosthesis for Two-Staged Cardiac Replacement." *American Journal of Cardiology* 24(5): 723–730.
- Daugherty, R. and J. Franzini. (1977). *Fluid Mechanics with Engineering Applications*. McGraw Hill Kogakusha.
- Deng, M. and Y. Naka. (2007). *Mechanical Circulatory Support Therapy in Advanced Heart Failure*. London: Imperial College Press.
- Dutton, D., T. Preston, and N. Pfund. (1988). *Worse Than The Disease: Pitfalls of Medical Progress*. Cambridge, UK: Cambridge University Press.
- Encyclopedia of Surgery. (2007). "Heart-Lung Machines." Retrieved September 2007 from <http://www.surgeryencyclopedia.com/Fi-La/Heart-Lung-Machines.html>
- Enotes. (2002). "Heart-Lung Machines—Encyclopedia of Nursing & Allied Health." Retrieved September 2008 from <http://www.enotes.com/nursing-encyclopedia/heart-lung-machines>
- Finocchiaro, T., T. Butschen, P. Kwant, U. Steinseifer, T. Schmitz-Rode, K. Hameyer and M. Lessmann. (2008). "New Linear Motor Concepts for Artificial Hearts." *IEEE Transactions on Magnetics*, 44(6): 678.
- Fukui, Y., A. Funakubo, and K. Fukunaga. (2004). *Development of the Assisted Artificial Heart with Linear Motor Actuator*. SICE 2004 Annual Conference.
- Glenn, W. (1993). "Seawell's Pump." *Guthrie Journal* 63(1).
- GMR. (1952). "Generations of GM—History: 1952, The First Mechanical Heart Pump." Retrieved May 2010 from [http://wiki.gmnext.com/wiki/index.php/1952,\\_The\\_First\\_Mechanical\\_Heart\\_Pump](http://wiki.gmnext.com/wiki/index.php/1952,_The_First_Mechanical_Heart_Pump)
- Gosline, A. (2004). "Simpler Pump Boosts Failing Hearts." *New Scientist*, July 28.
- Griffith, P., R. Kormos, H. Borovetz, K. Litwak, J. Antaki, V. Poirier and K. Butler. (2000). "Heart-Mate II Left Ventricular Assist System: From Concept to First Clinical Use." Fifth International Conference on Circulatory Support Devices for Severe Cardiac Failure, New York.
- Hajar, R. (2005). "The Artificial Heart." *Heart Views* 8(2): 70–76.
- Jahanmir, S., A. Hunsberger, H. Henshmat, M. Tomaszewski, J. Walton, W. Weiss, B. Lukic, W. Pae, C. Zapanta and T. Khalapyan. (2008). "Performance Characterization of a Rotary Centrifugal Left Ventricular Assist Device with Magnetic Suspension." *Artificial Organs* 32(5): 366–375.
- James, N., C. Wilkinson, N. Linqard, A. vanderMeer and J. Woodard. (2003). "Evaluation of Hemolysis in VentrAssist Implantable Rotary Blood Pump." *Artificial Organs* 27(1): 108–113.
- Jaron, D. (1990). "Current Status of the Artificial Heart and Cardiovascular Assist Devices." *Proceedings of the 1990 IEEE Colloquium in South America, 1990*.

- Jarvik Heart. (2008). Retrieved September 2008 from <http://www.jarvikheart.com/home.asp>
- Jeffrey, S. (2001). “REMACH: LVAD Support Superior to Medical Management for Transplant-Ineligible CHF Patients.” Retrieved September 2008 from <http://www.theheart.org/viewEntityDispatcherAction.do?legacyId=26851>
- Kun-xi, Q. and J. Ying. (2008). “Artificial Heart Rejects High Tech? Lessons Learned from Non-pulsatile VAD with Straight Impeller Vanes.” *International Conference on BioMedical Engineering and Informatics, 2008 (BMEI 2008)*.
- Lemelson-MIT. (2002). “Inventor of the Week: Robert Jarvik.” Retrieved September 2008 from <http://web.mit.edu/invent/iow/jarvik.html>
- MedGadget. (2006). “Apple Solution for DeBakey VAD.” Retrieved September 2008 from [http://medgadget.com/archives/2006/10/apple\\_solution.html](http://medgadget.com/archives/2006/10/apple_solution.html)
- MedGadget. (2008). “Impella 2.5 Heart Pump Given Green Light in US.” Retrieved September 2008 from [http://medgadget.com/archives/2008/06/impella\\_25\\_heart\\_pump\\_given\\_green\\_light\\_in\\_us.html](http://medgadget.com/archives/2008/06/impella_25_heart_pump_given_green_light_in_us.html)
- Miller, L., F. Pagani, S. Russell, R. John, A. Boyle, K. Aaronson and J. Conte (2007). “Use of a Continuous-Flow Device in Patients Awaiting Heart Transplantation.” *New England Journal of Medicine (NEJM)* 357: 885-896.
- Minami, M., L. Arusoglu, A. El-Banayosy and A. Sezai. (2001). “Bridging to Heart Transplantation using Paracorporeal and Implantable Ventricular Assist Devices.” *Journal of Congestive Heart Failure and Circulatory Support* 1(4): 179–184.
- Mitamura, Y., H. Nakamura, K. Sekine, D. Kim and R. Yozu. (2001). “Prediction of Hemolysis in Rotary Blood Pumps with Computational Fluid Dynamics Ananysis.” *Journal of Congestive Heart Failure and Circulatory Support* 1(4): 331–336.
- MiTiHeart. (2006). Retrieved September 2008 from <http://www.mitiheart.com/Technology.html>
- NationMaster Encyclopedia. (2008). “Artificial Heart.” Retrieved September 2008 from <http://www.nationmaster.com/encyclopedia/Artificial-heart>
- Okamoto, E., S. Fukuoka, M. Momoi and E. Iwasawa. (2001). “FEM and CAD/CAM Technology Applied for the Implantable LVAD.” *Journal of Congestive Heart Failure and Circulatory Support* 1(4): 391–398.
- Orme, F. (2002). “Human Physiology—Lecture Notes.” Retrieved September 2008 from <http://members.aol.com/Bio50/index.html>
- Paden, B., J. Ghosh, and J. Antaki. (2000). “Control System Architecture for Mechanical Cardiac Assist Devices.” *Proceedings of the 2000 American Control Conference, 2000*.
- ScienceDaily. (2006). “VCU Medical Team Implants Total Artificial Heart.” Retrieved September 2008 from <http://www.sciencedaily.com/releases/2006/04/060405022627.htm>
- Shackleford, J. (2003). *William Harvey and the Mechanics of the Heart (Oxford Portraits in Science)*. Oxford, UK: Oxford University Press.
- Sherief, H. (2007). “Biomedical Engineering: Artificial Heart & Heart Assist Devices.” Retrieved September 2008 from <http://mybiomedical.blogspot.com/2007/08/artificial-heart-heart-assist-devices.html>
- Shiba, K., M. Nukaya, T. Tsuji and K. Koshiji. (2006). “Analysis of Current Density and Specific Absorption Rate in Biological Tissue Surrounding an Air-core Type of Transcutaneous Transformer for an Artificial Heart.” *28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2006 (EMBS '06)*.
- Siegenthaler, M., J. Martin, A. Van de Loo, T. Doenst, W. Bothe and F. Beyersdorf. (2002). “Implantation of the Permanent Jarvik-2000 Left Ventricular Assist Device: A Single Center Experience.” *Journal of the American College of Cardiology* 2002(39): 1764–1772.
- Slepian, M., R. Smith, and J. Copeland. (2006). The Syncardia CardioWest Total Artificial Heart. In *Treatment of Advanced Heart Disease*, K. L. Baughman and W. M. Baumgartner (Eds.). Boca Raton, FL: Taylor and Francis, pp. 473–490.

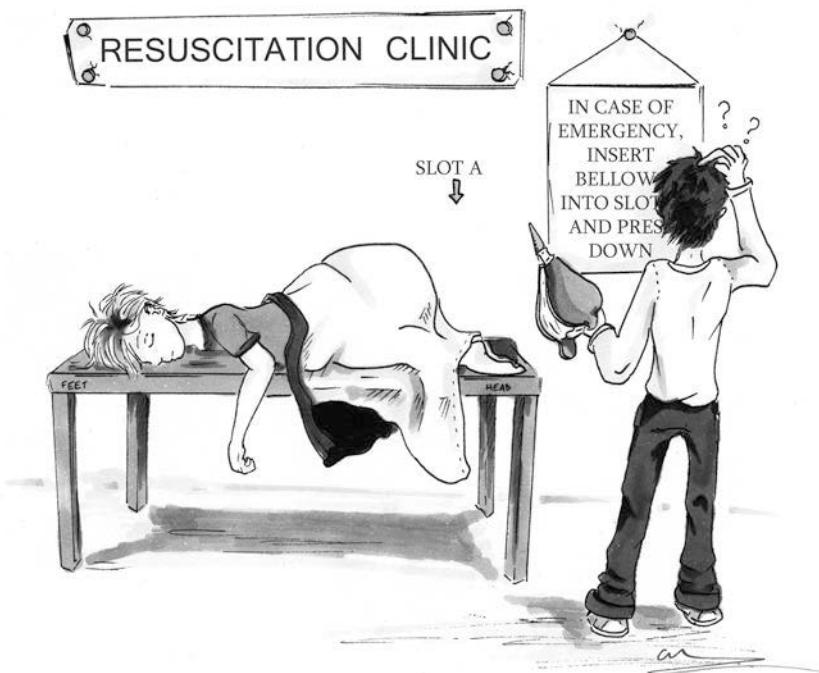
- Snyder, A., W. Pae, J. Boehmer and G. Rosenberg. (2001). "First Clinical Trials of a Totally Implantable Destination Therapy Ventricular Assist System." *Journal of Congestive Heart Failure and Circulatory Support* 1(4): 185–192.
- Stephenson, L. (2002). "The Michigan Heart: The World's First Successful Open Heart Operation?" *Journal of Cardiac Surgery* 17(3): 238–246.
- Takatani, D., K. Ouchi, M. Nakamura and T. Sakamoto. (2001). "Ultracompact, Totally Implantable, Permanent, Pulsatile VAD System." *Journal of Congestive Heart Failure and Circulatory Support* 1(4): 407–412.
- Tanaka, A., M. Yoshizawa, T. Yamada, K. Abe, H. Takeda, T. Yambe and S. Nitta. (2001). "In Vivo Evaluation of Pressure Head and Flow Rate Estimation in a Continuous-Flow Artificial Heart." *Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2001*.
- Texas Heart Institute. (2008). "Heart Assist Devices: AbioCor Implantable." Retrieved September 2008 from <http://www.texasheart.org/Research/Devices/abiocor.cfm>
- Tsukamoto, Y., K. Ito, Y. Konishi, T. Masuzawa, T. Yamane, M. Nishida, A. Aouidef, T. Tsukiya and Y. Taenaka. (2001). "Computational Fluid Dynamics Analysis for Centrifugal Blood Pumps." *Journal of Congestive Heart Failure and Circulatory Support* 1(4): 337–343.
- Vollkron, M., H. Schima, L. Huber, B. Benkowski, G. Morello and G. Wieselthaler. (2006). "Control of Implantable Axial Blood Pumps Based on Physiological Demand." *American Control Conference, 2006*.
- Watterson, P., J. Woodard, V. Ramsden and J. Reizes. (1999). "VentrAssist Hydrodynamically Suspended, Open, Centrifugal Pump." *Artificial Organs* 24(6): 475–477.
- Weber, D., D. Raess, J. Henriques and T. Siess. (2009). "Principles of Impella Cardiac Support." *Principles of Hemodynamics: Suppliment to Cardiac Interventions Today*, August–September, 4–18.
- WorldHeart. (2008). "Novacor II LVAS." Retrieved September 2008 from [http://www.worldheart.com/products/novacor\\_2\\_lvias.cfm](http://www.worldheart.com/products/novacor_2_lvias.cfm)
- Yamada, H., T. Mizuno, H. Wakiwaka, Y. Izumi, Y. Kataoka, M. Karita, M. Maeda and Y. Kikuchi. (1998). "Drive Control of Linear Pulse Motor for Artificial Heart." *Proceedings of the 1998 International Conference on Power Electronic Drives and Energy Systems for Industrial Growth*.
- Yambe, T., S. Maruyama, T. Takagi and M. Yoshizawa. (2001). "Smallest Ventricular Assist System by use of Peltier Elements with Shape Memory Alloy." *Journal of Congestive Heart Failure and Circulatory Support* 1(4): 403–405.



# Respiratory Aids

## Chapter Outline

9.1	Introduction.....	472
9.2	Construction.....	473
9.3	The Mechanics of Respiration.....	476
9.4	Energy Required for Breathing .....	485
9.5	Measuring Lung Characteristics.....	488
9.6	Mechanical Ventilation.....	494
9.7	The Physics of External Negative-Pressure Ventilation.....	508
9.8	Positive-Pressure Ventilators.....	511
9.9	References .....	520

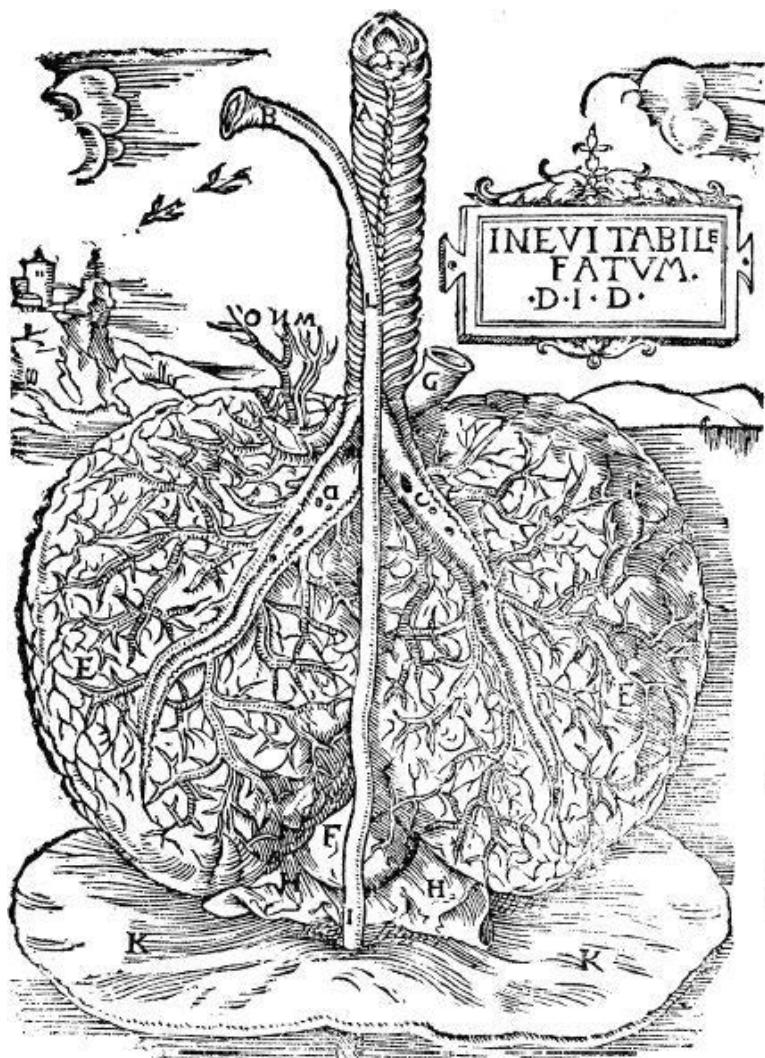


## 9.1 | INTRODUCTION

Knowledge of the structure and function of the lungs has depended on the development of the microscope. For almost 2000 years prior to that it was thought that the lungs were similar in structure to the other internal organs, with the only important difference being that air could enter the lungs through the trachea to mix physically with blood to cool it as illustrated in Figure 9-1. The anatomical drawings of Leonardo da Vinci (1452–1519) and Versalius (1514–1564) were regarded as consistent with this view, and it was only the use of the microscope by Malpighi (1628–1694) that demonstrated for the first time the air-filled alveoli—the blind ends of the air passages into the lungs. He described them as “an almost infinite number of orbicular bladders.” Malpighi also showed that the blood capillaries in the lungs were separate from the gas and allowed the passage of blood through the lungs, as deduced but not demonstrated by William Harvey. Malpighi knew nothing of gas exchange and thought the function of alveolar ventilation was to stir and mix the blood in the capillaries (Wilson 1960).

**FIGURE 9-1 ■**

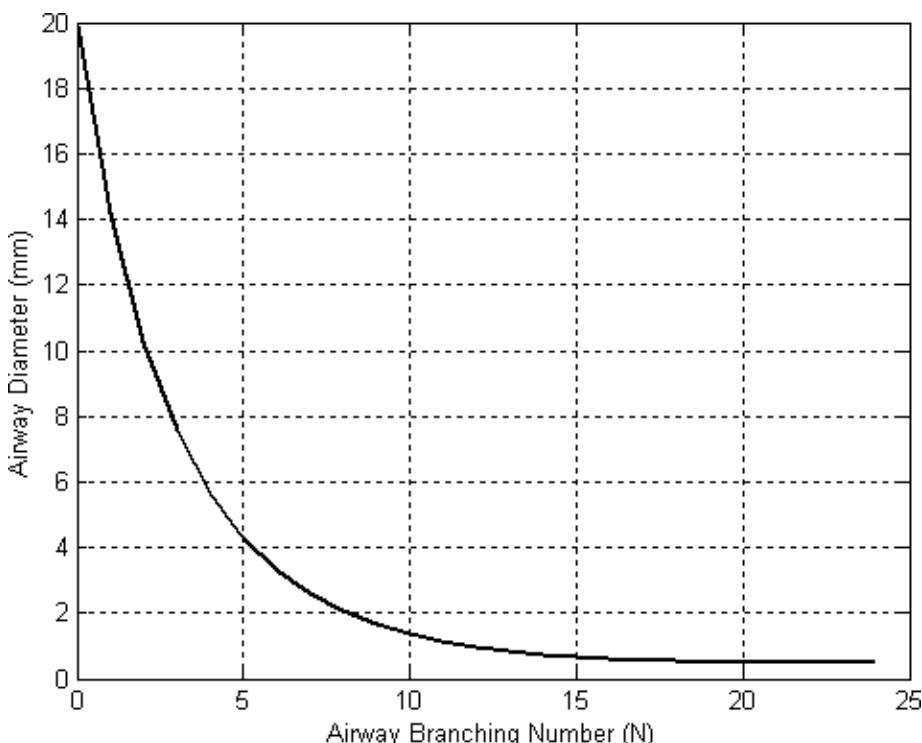
Woodcut from  
*Mundinus*  
*Anatomica* (Marburg  
1541)



## 9.2 CONSTRUCTION

Air passes through the nose and mouth farther into the airways, where it is warmed, humidified, and filtered. From the trachea to the alveoli, there are about 23 branching generations of airways of ever decreasing diameter. For example, a single trachea begins with a diameter of about 20 mm in an adult; by the tenth generation there are 1000 bronchi, each with a diameter of about 1.4 mm, and by the twentieth generation there are 1 million bronchioles, each with a diameter of about 0.5 mm as illustrated in Figure 9-2. The first 16 branches or so constitute the conducting zone, which is an anatomic dead space because no gas exchange takes place there. This dead space volume is about 150 ml in a healthy adult.

N	$2^N$	Diameter (mm)
0	1	20
1	2	14
2	4	10
3	8	7.5
4	16	5.5
5	32	4.25
10	1024	1.4
20	1,048,576	0.5

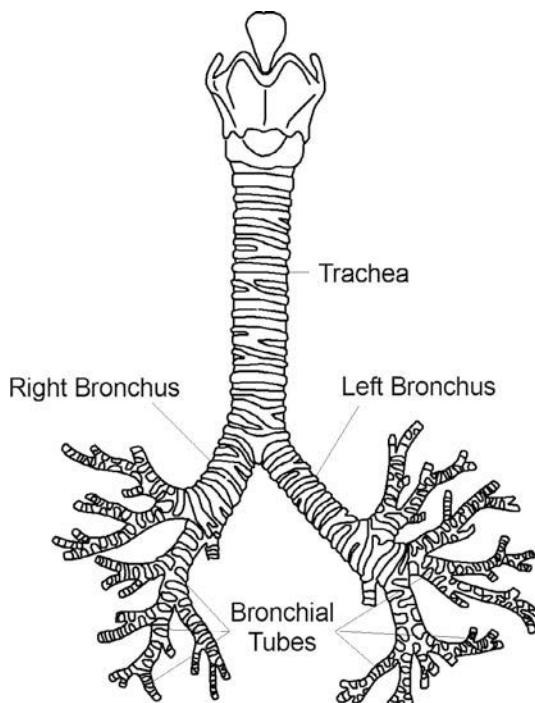


**FIGURE 9-2 ■**  
Airway diameter as a function of the division level.

If an average tidal volume of 500 ml is inhaled, at the end of inspiration only 350 ml will have entered the alveoli and 150 ml will remain in the airways. The ventilation used for gas exchange will be only 350/500 ths, or 70%, of the total ventilation. The remaining 30% dead space is sometimes called wasted ventilation but does perform the vital function of conditioning the air before it reaches the alveoli. If inspired air reached the alveoli directly, if it were cold it would cool the tissue, if hot it would heat it, and if dry it would parch and destroy the alveolar walls. As cold, dry air is inspired, it is raised to body temperature and 100% humidity after passing through the first few generations of bronchi. In this process the airway lining itself (the mucosa) is cooled and dried, but it has performed its function of protecting the alveoli. On breathing out, the mucosa take up some of the heat and water vapor from the expired air, restoring it to normal. In this way, not only are the alveoli protected, but also loss of heat and water from the body as a whole is reduced.

The walls of the trachea and bronchi, shown in Figure 9-3, consist of several layers. On the inner lining surface, the *luminal* side, there is an epithelial layer the cells of which are covered with microscopic cilia (hairs) that continuously sweep any surface material toward the larynx, where it is coughed up or swallowed. Other cells secrete mucus. Just under the epithelium is a dense blood capillary network that provides nutrition for the epithelium and glands and may be the site of uptake of inhaled pollutants and drugs. Deeper in the wall are the submucosal glands, the main source of the mucus that lines the airways. The glands are stimulated to secrete by many factors, the most important being pollutants (e.g., cigarette smoke) and viral or bacterial infections of the airways. The secreted mucus creates a barrier and takes up soluble pollutants and smoke particles, slowing down their entry into the body and protecting the epithelium from their harmful effects. It can also stimulate coughing as an even more rapid means of their removal.

**FIGURE 9-3 ■**  
The trachea and bronchi.



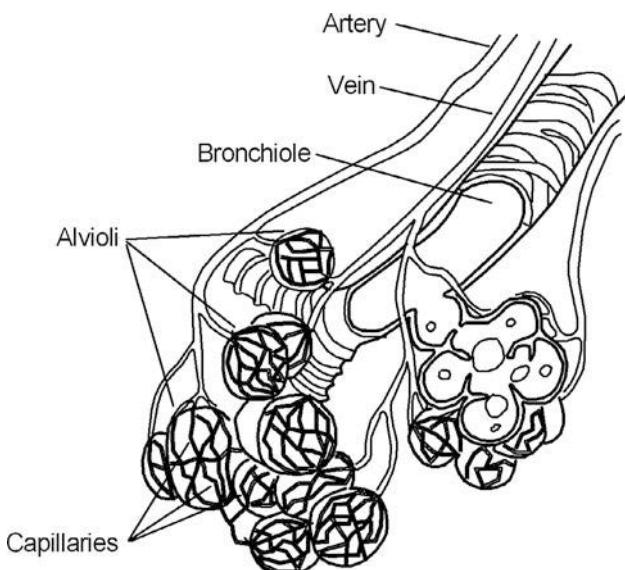
Even deeper in the airway wall are layers of cartilage and smooth muscle. The cartilage stabilizes the airways and prevents their collapse during vigorous acts of breathing, while the smooth muscle has not been shown to have a physiological role, though it may adjust the diameter of the airways to make them optimally efficient for gas transfer to the alveoli.

The trachea and bronchi contain sensory nerves of two types. In the smooth muscle are receptors that signal the degree of stretch and therefore of inflation of the airways and lungs. These signals are used to control the pattern of breathing. If the Vagus nerves, which carry sensory information from the bronchi, are cut, breathing becomes slow, deep, and mechanically inefficient. The second type is the fine nerve fibers in the epithelium, with finger-like projections reaching almost to the airway lumen. These respond to inhaled pollutants and irritants and set up a range of reflex responses. The most striking is the cough, but there is also reflex mucus secretion and smooth muscle contraction.

The respiratory zone includes branches 17 through 23 and consists of the respiratory bronchioles, alveolar ducts, and alveolar sacs. The smallest air-conducting vessels, the bronchioles, have no cartilage in their walls and no mucus cells in their epithelium. When the lungs inflate they probably distend equally with the alveoli, but there is little gas exchange in them because they are not well vascularized.

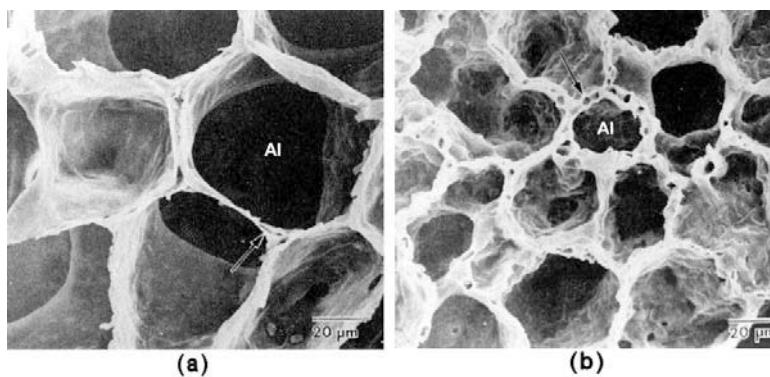
There are about 15 million alveolar ducts, each giving rise to about 20 alveolar air sacs and making up a total of approximately 300 million alveoli, each with an inflated diameter of 100–300  $\mu\text{m}$  and surrounded by approximately 1000 capillaries, as illustrated in Figure 9-4. The total contact area between pulmonary capillary blood and alveolar air ranges from 70 to 140  $\text{m}^2$  in adult human beings, and this surface area increases during exercise through recruitment of new capillaries in particular in the apical parts of the lungs (Widdicombe, 2008).

Since the rate of diffusion of a gas depends on the surface area, the thinness of the diffusion barrier, and the solubility of the gas (Fick's law), the alveolar wall is extremely thin, from 0.2 to 0.5  $\mu\text{m}$  depending on the degree of inflation of the lungs. The change in alveolar volume and wall thickness with respiration is shown in Figure 9-5.



**FIGURE 9-4** ■  
Schematic diagram showing an alveolar cluster.

**FIGURE 9-5 ■**  
 Scanning electron microscope view of lung alveoli, magnified  $\times 750$ .  
 (a) Full capacity.  
 (b) After normal expiration. (Murray and Nadel 2008)  
 copyright W. B. Saunders (Elsevier) reproduced with permission



The barrier to diffusion has three components. On the inside surface of the alveoli is a thin layer of secretion about  $0.15\text{ }\mu\text{m}$  thick, containing a phospholipid surfactant that lowers the surface tension of the lungs and allows them to be stretched by relatively low pressure. Then there is the epithelial cell layer of the alveoli and finally the capillary endothelium. Cells of a different type, the alveolar macrophages, are found within the cavities of the alveoli; their function is to ingest and remove solid particles, such as those of smoke.

### 9.3 | THE MECHANICS OF RESPIRATION

Air flows from a region of higher to lower pressure by bulk flow. When total gas pressure in the alveoli,  $P_A$  (Pa), is the same as the atmospheric pressure  $P_B$  (Pa), no air flow occurs because no pressure gradient exists. To initiate air flow into gas exchange sites of the lung,  $P_A$  must decrease below  $P_B$ , or  $P_B$  must increase above  $P_A$ , as occurs during mechanical ventilation.

During respiratory muscle contraction, the volume of the thoracic cavity and lungs increases, which decreases  $P_A$ . Air then moves into the alveoli by bulk flow until the pressure equalizes at the end of inspiration. The lung is expanded elastically during inspiration, so when the respiratory muscles relax it recoils to compress the alveolar gas volume. This elevates  $P_A$  above  $P_B$ , so air is expelled until equilibrium is again achieved at the end of expiration.

When the respiratory muscles contract, several opposing forces must be overcome to bring about enlargement of the lungs. These are related to the physical characteristics of the lungs, friction resistance of moving air, and inertia of both the lungs and the air. Nerve impulses initiating contraction of the respiratory muscles originate in the medulla of the brain stem. The force generated by the respiratory muscles can be increased by increasing the frequency of discharge in individual motor units, activating additional motor units, or calling on the accessory muscles of respiration. These are muscles not normally used during normal breathing.

During inspiration, contraction of the diaphragm and external intercostal muscles enlarges the chest cavity, which leads to an expansion and stretching of the lung. As the lung inflates, potential energy is stored in the stretched elastic structures of the lung so that with relaxation of the respiratory muscles expiration occurs from elastic recoil of the lung–chest cage apparatus and no muscle contraction is required.

During contraction, the central portion of the diaphragm becomes flattened by about 1.5 cm during a normal breath and up to 10 cm for a deep breath. This increases the volume

of the thoracic cavity and hence the lungs. Normally about two-thirds of the tidal volume is directly attributable to diaphragmatic contraction.

In human beings, the ribs can rotate at their interface with the thoracic vertebrae. When the external intercostal muscles contract they lift the anterior end of each rib, pulling it upward toward the horizontal. This increases the volume of the thoracic cage as well as prevents the rib cage from being pulled downward and inward as the diaphragm descends during inspiration. While the external intercostals can maintain a considerable level of breathing on their own, their paralysis does not prevent ventilation by the diaphragm alone.

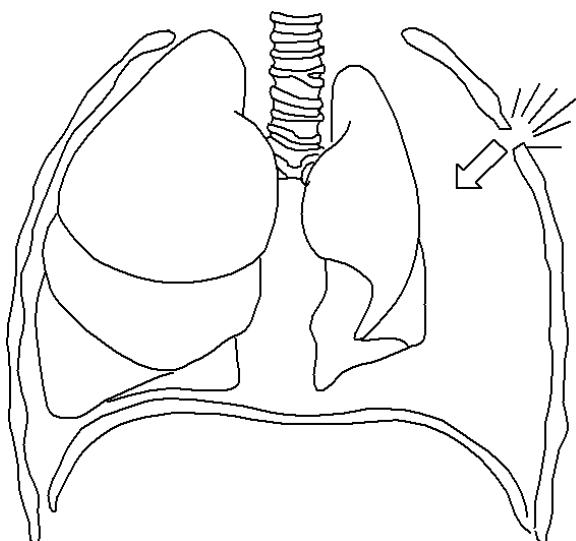
The accessory muscles of respiration include the scalenes, which elevate the first ribs, and the sternomastoids, which raise the sternum and the back muscles. The accessory muscles are usually inactive until  $Q_E$  reaches a fairly high rate of 50 to 100 L/min in an adult during heavy exercise.

In addition, with high ventilatory rates the muscles of expiration comprising the abdominal group and the internal intercostals are important. They are also the principal muscles responsible for coughing and for forced lung expiratory volume measurements, like the vital capacity. The internal intercostals are attached between adjacent ribs and act antagonistically to the external intercostals. Their contraction compresses the rib cage to decrease the volume of the thoracic cage.

The diaphragm is also a participant in expiration. It continues to contract during the early part of expiration, which opposes some of the lung recoil and results in a slowing of expiration as well as ensuring a smooth transition from inspiration to expiration (Roussos and Macklem, 1982).

### 9.3.1 Physical Properties

The lung is not anatomically connected to the inner chest wall, even though it fills most of the thoracic cavity except for the space occupied by the heart and major blood vessels. No ligaments attach the outer lung surface to the inner chest wall, so if air is allowed access to the gap between the outer lung surface and inner chest wall the lung will recoil away from the inner chest wall and collapse, as illustrated in Figure 9-6. Collapse of the lung creates a condition known as *pneumothorax*.



**FIGURE 9-6 ■**  
Collapse of the lung (pneumothorax) as the chest wall is penetrated.

The outer surface of the lung is covered by a visceral pleura that lies next to the parietal pleura lining the inner chest wall. The gap between the two pleura is called the intrapleural space and is occupied by a small amount of fluid. This intrapleural fluid provides the cohesion that helps link the pleura together. Thus, as the chest wall expands during inspiration the lung is obliged to follow. To preserve the cohesive forces linking the outer lung to the inner chest wall, the intrapleural gap must be kept free of air or excessive fluid. Because the pleura membranes are modestly permeable to gases and water, physiological processes operate to maintain the correct balance (Agostoni, 1972).

The lung is an elastic structure with an anatomical organization that promotes its collapse to a very small volume, much like an inflated balloon. While the elastic properties of the lung are important to bring about expiration, they also oppose lung inflation. As a result, lung inflation depends on contraction of the respiratory muscles. How easily a lung inflates relates to its compliance (the reciprocal of elastance). Therefore, in a lung with a high compliance, a small pressure change would result in a large volume change, and the work performed by the respiratory muscles to inflate the lung would be small.

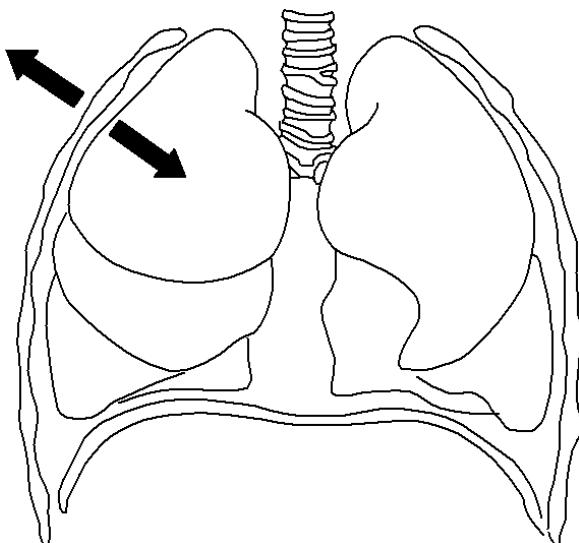
If the volume of the lung (removed from the body) is plotted as a function of its pressure, its derivative gives the compliance. A similar, albeit less practical, experiment could be used to determine the compliance of the chest cage.

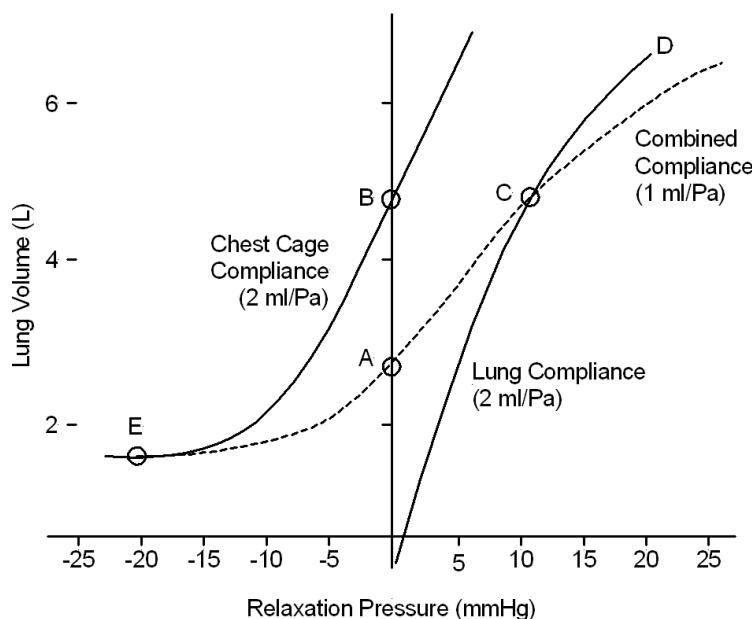
With the lungs placed inside the chest cavity and their respective pleural surfaces held together by cohesive forces, the volume of the lung is higher than its equilibrium volume, whereas the chest cavity volume is less than its equilibrium volume. At equilibrium, the recoil pressure of the lung tending to deflate is opposed by an equal but opposite recoil pressure of the chest wall tending to expand as illustrated in Figure 9-7. These equal but opposite recoil forces are reflected by an intrapleural pressure that is slightly lower than atmospheric pressure. Moreover, with inhalation the intrapleural pressure decreases further, reflecting the force tending to separate the lung and chest wall, but this is prevented by the cohesive forces.

Figure 9-8 shows separate relaxation curves for the lung and the chest cage (both solid), along with the combined lung–chest cage relaxation curve (dotted). The slope of each relaxation curve corresponds to the compliance for the structures. At end expiration (point A), recoil or relaxation pressure for the lung and chest cage alone are equal but

**FIGURE 9-7 ■**

Lung and chest cage in equilibrium.





**FIGURE 9-8** ■ Graph showing the lung volume as a function of the relaxation pressure of the lung, the chest cage, and the combined pressure. The slope of the curves gives the compliance (133.34 Pa per mmHg and 1.36 cmH<sub>2</sub>O per mmHg).

opposite. At this point, lung volume corresponds to functional residual capacity (FRC). With inspiration, the lung is stretched farther and exhibits a greater recoil pressure. At the same time, the chest cage is less compressed, so its negative recoil pressure diminishes as it approaches its equilibrium volume. The chest cage reaches its equilibrium volume (point B), and the lung and lung–chest cage relaxation curves intersect (point C). At this lung volume, all measured relaxation pressure for the lung–chest cage system is from the lung alone. If an even greater air volume is inhaled (point D), both the lung and chest cage are stretched beyond their equilibrium volumes.

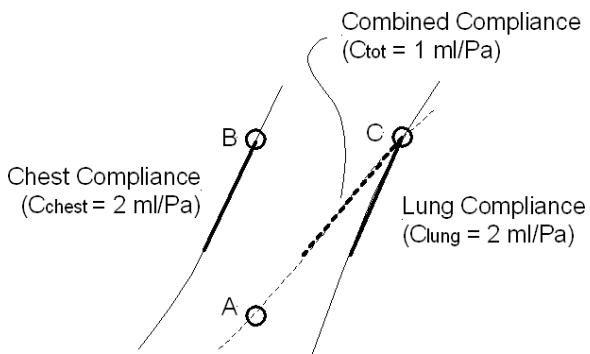
Note that the compliance curve for the combined lung–chest cage becomes more flattened (less compliant) at this point because the lung and chest cage are both tending to recoil toward smaller equilibrium volumes.

If the total lung–chest cage system is returned to resting end expiration (point A) and then more air is expelled, a negative relaxation pressure results for both the chest cage and the combined lung–chest cage (point E). At this point, the chest cage is compressed as more and more air is expelled, with the negative recoil pressure reflecting its tendency to expand toward its equilibrium volume (point B). At the same time, the lung contributes little positive recoil pressure because it is close to its equilibrium volume because it is stretched very little.

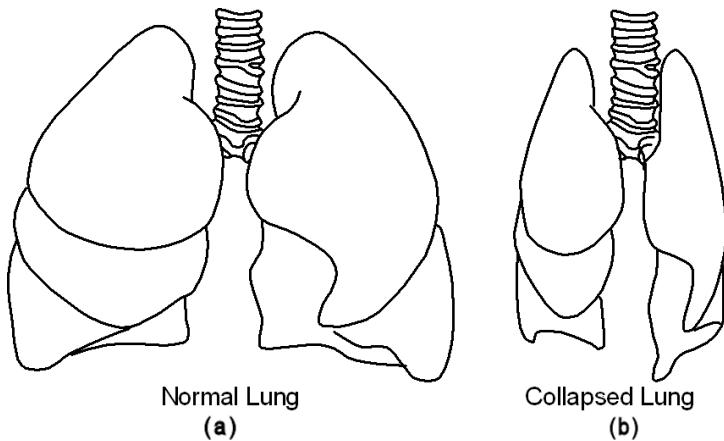
During normal breathing, the tidal volume range is between points A and B. During inspiration, the individual moves up the lung–chest cage compliance curve from point A toward point C and in the opposite direction during expiration. Note that over the normal tidal volume range total pulmonary compliance,  $C_{tot}$ , given by the slope of the curves is less than the compliance of either the lung,  $C_{lung}$ , or chest cage,  $C_{chest}$ , alone, as can be seen in Figure 9-9. This occurs because the lung and chest cage are physically arranged in a series and therefore must be added as reciprocals

$$\frac{1}{C_{tot}} = \frac{1}{C_{lung}} + \frac{1}{C_{chest}} \quad (9.1)$$

**FIGURE 9-9** ■  
Compliance is the slope of the pressure–volume curves.



**FIGURE 9-10** ■  
Diagram showing the lungs.  
(a) Inflated.  
(b) Collapsed.



### 9.3.2 Lung Elasticity

It was thought for a long time that the elastic properties of the lung, illustrated in Figure 9-10, were due to stretching and recoil of individual elastin and collagen fibers. However, recent research has shown that the elastic properties of the lung are related more to the weave than to the stretching of the individual fibers (a bit like knitting).

The fluid film lining alveoli assumes a spherical shape because of the nearly spherical nature of alveoli. In many regards these resemble a soap bubble. In a soap bubble the Young–Laplace equation relates the surface tension,  $\gamma$  (dynes/cm),<sup>1</sup> the pressure,  $P$  (dyne), and the radius,  $a$  (cm), where

$$P = \frac{2\gamma}{a} \quad (9.2)$$

This equation states that the pressure within a soap bubble is directly proportional to surface tension of the air–fluid interface and inversely proportional to the radius of the sphere. If alveolar surface tension is assumed to be constant, say, 50 dynes/cm, then the recoil (collapsing pressure) could be calculated for alveoli of different radii using the Young–Laplace equation. If a number of alveoli are connected, it is obvious that air from the

---

<sup>1</sup>1 dyne = 10  $\mu$ N or the force required to accelerate 1 g by 1 cm/s<sup>2</sup>.

smaller alveoli will pass into the larger ones until only one large alveolus remains. In addition, collapsed alveoli, or those with extremely small radii, would require extremely high opening pressures to inflate. Thus, freshly inspired air would be directed toward large-radius alveoli because the pressure required to expand them is lowest.

This does not occur because, unlike water, the surface tension of the fluid film lining alveoli is not constant but varies in proportion to the alveolar size or exposed area. In fact, the fluid film lining alveoli contains surfactants capable of lowering surface tension far below that of water (70 dynes/cm).

The ability of surfactants to lower surface tension is related to their chemical structure. Because lung surfactants have hydrophobic and hydrophilic groups at different ends of the molecule, they preferentially accumulate at the liquid-air interface. At the interface, the hydrophilic end of the molecule extends into the liquid, and the hydrophobic end projects into the air. Surfactant naturally accumulates at the liquid-air interface and reduces the number of water molecules that would normally occupy it. Their presence disrupts the attracting forces between water molecules so the surface tension is reduced.

During expiration, as surface area is decreased the alveoli deflate and the relative concentration of surfactant molecules per unit area increases so that the surface tension is reduced further. As alveoli inflate, water molecules must be brought to the interface, so surface tension increases as fewer surfactant molecules are present per unit area. Even though the Young-Laplace relationship is still applicable, alveolar surface tension is not constant but decreases as alveolar radius decreases. As a result, pressure in small-radius alveoli is lower than that in large-radius alveoli.

During inspiration, air initially moves into smaller-radius alveoli or from larger to smaller alveoli to ensure uniform filling. In addition, the presence of surfactant reduces the opening and expanding pressures of small-radius alveoli, which enhances alveolar stability and reduces the work of breathing (Fenn and Rahn, 1965).

The elastic properties of the lung relate to both geometric weave of the elastin and collagen fibers and to the surface tension of the fluid film lining the alveoli. While alveolar surface tension is low in alveoli of small radii, it increases as alveoli enlarge. Therefore, surface tension forces contribute to lung elasticity, especially as the lung inflates. Over the normal tidal volume range, the weave of elastic fibers and surface tension contribute about equally to lung elasticity.

### 9.3.3 Frictional Forces

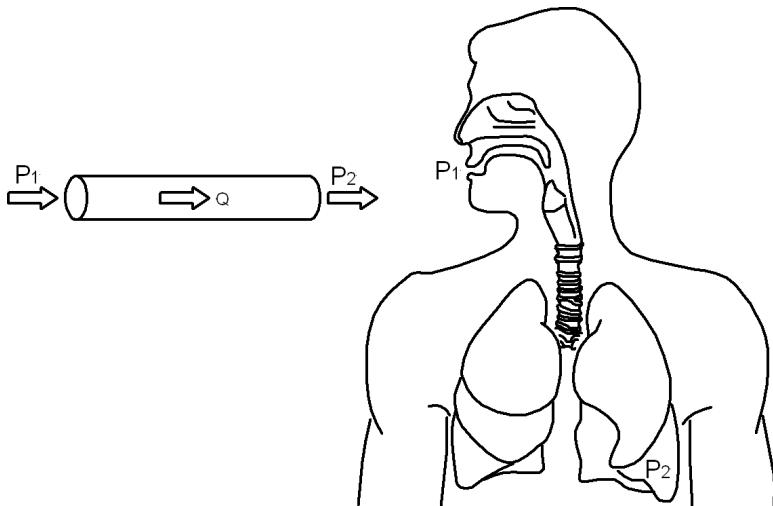
When the respiratory muscles contract, in addition to the elastic recoil of the lung, they must also overcome two types of friction: (1) drag as air passes through the airways; and (2) viscous friction as the lung and chest wall and abdominal organs slide over one another. During normal breathing, airway resistance accounts for 80% of the friction and viscous friction for the remaining 20%.

Viscous friction occurs as the outer surfaces of the lung slide over the inner chest wall and the various lung lobes move over one another during breathing. Even though the adjacent pleura of the lung and chest wall are lubricated with intrapleural fluid, some frictional resistance is present. In addition, as the diaphragm descends with inspiration, it compresses and displaces abdominal contents, and frictional resistance is encountered as abdominal organs are displaced and move over one another.

The magnitude of the frictional resistance encountered by air as it moves between mouth and alveoli depends on the linear velocity of airflow as well as the airflow pattern and

**FIGURE 9-11**

Airflow frictional resistance can be determined using the pneumatic equivalent of Ohm's law.



the physical dimensions and branching pattern of the airways. Two physical relationships can be used to compute frictional resistance to airflow in the lung.: Ohm's law and Poiseuille's equation.

The pneumatic equivalent of Ohm's law states that resistance to airflow,  $R_f$ , is determined by dividing the pressure difference between two points in the airway or driving pressure, equivalent to potential difference, by the airflow rate,  $Q$  (L/min), which is equivalent to the current. This is illustrated in Figure 9-11.

$$R_f = \frac{P_2 - P_1}{Q} \quad (9.3)$$

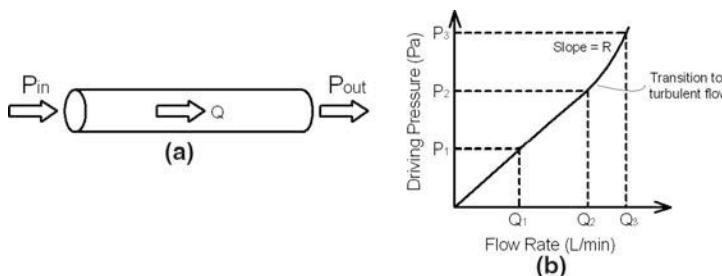
To do this, the pressure difference between the mouth and alveoli needs to be measured at a given flow. While it is difficult to measure alveolar pressure directly, a subject can be placed in a whole-body plethysmograph, which allows alveolar pressure to be measured indirectly.

Poiseuille's equation was derived to calculate airflow resistance through a tube, such as an airway. It takes into account the physical dimensions of the tube (radius and length) and the nature of the fluid moving through it. Poiseuille's equation states that frictional resistance to flow,  $R_f$ , is directly related to viscosity of the fluid,  $\eta$  (Pa.s), and to the length of the tube,  $l$  (m), and indirectly related to the fourth power of tube radius,  $a$  (m).

$$R_f = \frac{8\eta l}{\pi a^4} \quad (9.4)$$

Poiseuille's equation was derived using rigid, perfectly round and smooth, nonbranching tubes with laminar flow. However, lung airways are distensible, compressible, and not perfectly round or smooth. They also branch repeatedly and exhibit changes in radius and length during each breath. Poiseuille's equation also does not compensate for changes in the airflow pattern from laminar to turbulent, where frictional resistance is higher (Fenn and Rahn, 1965).

A rearrangement of Ohm's law reveals that the pressure gradient between two points in a tube is directly proportional to frictional resistance and the velocity of flow. However, if driving pressure is incrementally increased ( $P_1$  to  $P_3$ ) in a tube of fixed radius and length, a driving pressure is reached where the resistance to airflow exhibits a disproportionate increase even though the physical dimensions of the tube are unchanged, illustrated in



**FIGURE 9-12 ■**  
Airflow measurements  
(a) Hardware.  
(b) Relationship between pressure and flow for linear and turbulent regimes.

Figure 9-12. This occurs because the frictional resistance increases when the airflow pattern changes from laminar to turbulent flow at high flow rates.

Laminar or streamlined flow is characterized by concentric cylinders of air flowing at slightly different velocities. Air closest to the wall of the cylinder has the lowest linear velocity, and flow velocity gradually increases toward the center of the airway. When viewed in profile, laminar flow takes the form of a parabola. In contrast, turbulent air flow is more chaotic. With laminar flow, the driving pressure is directly proportional to resistance and flow,  $Q$  (L/min), but with turbulent flow the driving pressure is proportional to the square of the flow rate,  $Q^2$ .

The transition from laminar to turbulent flow is predicted by the Reynolds number,  $R_e$ , which is a function of the fluid density,  $\rho$  ( $\text{kg}/\text{m}^3$ ), the average linear velocity,  $v$  ( $\text{m}/\text{s}$ ), the diameter of the pipe,  $d$  ( $\text{m}$ ), and  $\eta$  ( $\text{Ns}/\text{m}^2$ ), the dynamic viscosity of the fluid.

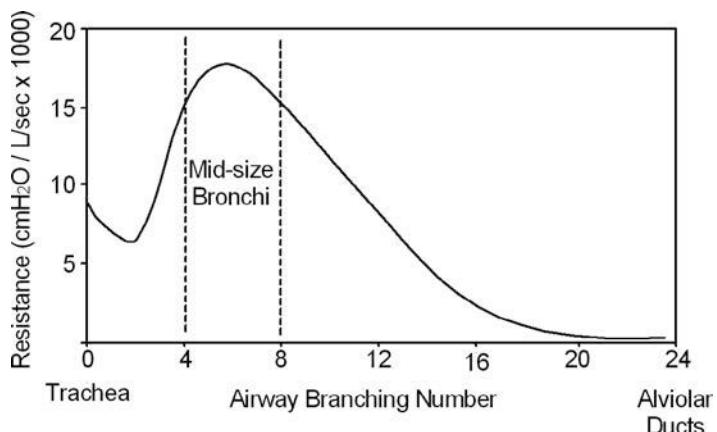
$$R_e = \frac{\rho v d}{\eta} \quad (9.5)$$

If the Reynolds number exceeds 2300, it is highly probable that a turbulent flow pattern is present. The presence of turbulent flow requires a greater driving pressure to generate a given airflow than is necessary with laminar flow.

It is not uncommon to have a mixture of laminar and turbulent flow in the airway. This is referred to as transitional flow, and it is most likely to occur at airway branch points. As predicted by the Reynolds number, turbulent flow is most likely to occur when airflow velocity and tube diameter are large in the trachea and larger airways. Laminar flow is more likely to occur in the smaller airways, where the linear velocity is low because repeated branching of the airways yields a large cross sectional area for airflow.

About 40 to 50% of the total airway resistance is located between the nose and larynx, with the remaining resistance occurring between the larynx and alveolar ducts. Total airway resistance is slightly higher with nasal than with open-mouth breathing. Within the tracheobronchial tree, most resistance to airflow occurs in the medium-sized bronchi between the fourth and eighth order of branching. These midsized bronchi offer more air flow resistance than larger or smaller airways because of the complex relationship among air flow velocity, total cross sectional area, airway length and diameter, and branching frequency. Even though airways become narrower after each branching, which would be expected to increase flow resistance, the flow is divided into two parallel paths, so flow velocity in individual airways decreases. Resistances arranged in parallel are added as reciprocals of their individual resistances, and as a result airways farther from the medium-sized bronchi account for progressively less of the total airway resistance. Increased airway resistance in the bronchioles and smaller airways is often difficult to detect because they represent such a small fraction of total airway resistance (Fenn and Rahn, 1965). These considerations are illustrated in Figure 9-13.

**FIGURE 9-13** ■ Airway resistance distribution in the tracheobronchial tree [Adapted from (Fenn and Rahn 1965).]



### WORKED EXAMPLE

#### Turbulent Flow

If the diameter of each of the eight bronchi after the third branching is 7.5 mm and the tidal volume breathed is 0.6 L inhaled over 0.8 s, will the flow be laminar or turbulent? Using the same breathing parameters, determine whether flow through the trachea with a diameter of 20 mm is laminar or turbulent.

The flow rate in both cases is the same:

$$\begin{aligned} Q &= \frac{0.6}{0.8} \\ &= 0.75 \text{ L/s} (7.5 \times 10^5 \text{ mm}^3/\text{s}) \end{aligned}$$

The total cross sectional area of the eight bronchi, assuming that they are perfectly circular, is

$$\begin{aligned} A &= 8 \frac{\pi d^2}{4} \\ &= 2 \times \pi \times 7.5^2 \\ &= 353 \text{ mm}^2 \end{aligned}$$

Assuming an equal distribution of flow into each bronchus, the air flow velocity will be

$$\begin{aligned} v &= \frac{Q}{A} \\ &= \frac{7.5 \times 10^5}{353} \\ &= 2125 \text{ mm/s} (2.1 \text{ m/s}) \end{aligned}$$

The density of air is 1.2 kg/m<sup>3</sup>, and the dynamic viscosity is  $1.78 \times 10^{-5}$  kg/(m · s); therefore, the Reynolds number is

$$\begin{aligned} R_e &= \frac{\rho v d}{\eta} \\ &= \frac{1.2 \times 2.1 \times 7.5 \times 10^{-3}}{1.78 \times 10^{-5}} \\ &= 1061 \end{aligned}$$

This is smaller than 2300; therefore, the flow will almost certainly be laminar.

The diameter of the trachea is 20 mm, which makes its cross sectional area  $314 \text{ mm}^2$  and the air flow velocity 2.38 m/s for the same breathing rate. In this case the Reynolds number will be

$$\begin{aligned} R_e &= \frac{\rho v d}{\eta} \\ &= \frac{1.2 \times 2.38 \times 20 \times 10^{-3}}{1.78 \times 10^{-5}} \\ &= 3208 \end{aligned}$$

This is larger than 2300; therefore, the flow through the trachea will be turbulent.

---

### 9.3.4 Inertia

When the respiratory muscles contract, they must produce sufficient force to accelerate the lung–chest cage system and move the air in the airway from a standstill to some final velocity. The opposing force of inertia is related to the mass of the object and its rate of acceleration,  $\dot{Q}(\text{L/min}^2)$ . While the lung and chest cage both have considerable mass, their acceleration is small during normal breathing. In contrast, air can be accelerated to a high velocity, but its mass is very small. The effects of inertia are therefore small for both the lung–chest cage structures and air moving through the air passages—typically about 5% of the total forces that oppose respiratory muscle contraction.

## 9.4 | ENERGY REQUIRED FOR BREATHING

Breathing requires that physical work be performed by the respiratory muscles. While the opposing forces of elastance (compliance), frictional resistance, and inertia have been discussed, the amount of work the respiratory muscles must perform to overcome each of these three opposing forces as a portion of the total work performed needs to be considered. As summarized in the Table 9-1, about 60 to 66% of the total work performed by the respiratory muscles is used to overcome the elastic or compliance characteristics of the lung–chest cage, 30 to 35% is used to overcome frictional resistance, and only 2 to 5% of the work is used for inertia. However, this partitioning of the opposing forces is altered by changes in tidal volume or breathing frequency.

**TABLE 9-1** ■ Summary of the Contributions to the Work Done in Breathing

Type of Work	Contributing Components	% of Total
Elastic (compliance)	Lung 50% Chest cage 50%	Surface tension 50 to 80% Tissue 20 to 50% 60 to 66%
Frictional	Viscous 20% Airways 80%	30 to 35%
Inertia	Lung Chest cage Air	2 to 5%

Physical work,  $W$  (J), can be computed by multiplying the force by the distance an object is moved or by multiplying the change in pressure by the change in volume. In general terms this can be written in terms of an integral

$$W = \int_{V_1}^{V_2} P(V) dV \quad (9.6)$$

where  $P$  (Pa) is the pressure, and  $V_1$  and  $V_2$  are the start and end volumes ( $\text{m}^3$ ), respectively.

In the compliance curve for the lung–chest cage system, the change in total volume is plotted as a function of the intrapleural pressure. From the compliance curve, it is possible to estimate the amount of elastic work performed by the respiratory muscles.

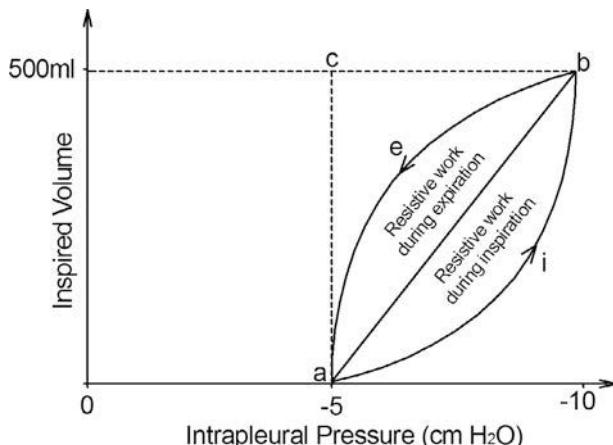
For the inspiration of an average tidal volume of 500 ml, the elastic work,  $W_{el}$  (J), would be proportional to the triangular area a-b-c-a in Figure 9-14. By making some assumptions, the same plot can be used to estimate the frictional work for this tidal volume. During inspiration the intrapleural pressure would initially fall below atmospheric pressure and then gradually return to atmospheric pressure as the lung fills with air at end inspiration. Thus, the friction work,  $W_{insp}$  (J), is proportional to the curved segment a-i-b-a labeled resistive work during inspiration.

In the example shown, the total elastic work performed by the muscles is equal to the area of the triangle a-b-c-a

$$\begin{aligned} W_{el} &= \frac{\Delta P \times \Delta V}{2} \\ &= 500 \times 500 \times 10^{-6}/2 \\ &= 0.125 \text{ J} \end{aligned}$$

From the relaxation curve, it is apparent that most of the work during inspiration is elastic work, with a smaller amount of frictional work. During inspiration, the lung is stretched and sufficient (potential) energy is stored to do the work needed to overcome the frictional resistance encountered during expiration. The elastic work performed during inspiration is recovered and is normally sufficient to overcome the frictional resistance work,  $W_{exp}$  (J), associated with expiration, shown as the curved segment a-b-e-a labeled resistive work during expiration. The negative work,  $W_{neg}$  (J) (done on the inspiratory muscles during expiratory air flow), is proportional to the area a-e-b-c-a.

**FIGURE 9-14** ■  
Work done to  
overcome elastic  
and frictional  
components during  
normal breathing.  
[Adapted from (Fenn  
and Rahn 1965).]



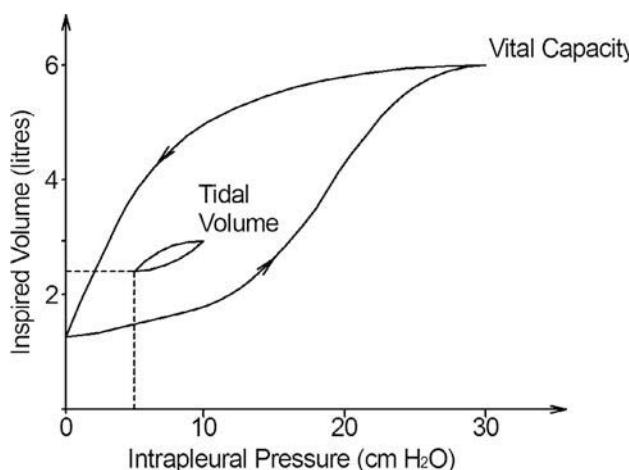
The total work is the sum of the elastic work, the frictional resistive work during inspiration, and the negative work. The frictional resistive work during expiration is provided from the elastic energy stored and thus is not included in the equation.

$$W_{tot} = W_{el} + W_{insp} + W_{neg} \quad (9.7)$$

With an increase in breathing frequency, air moves more rapidly through the airway and results in increased frictional resistance to airflow and tissue movement. If tidal volume remains unchanged but breathing frequency increases, the same elastic work is performed with each breath but frictional work is increased. At higher breathing frequencies, expiratory muscles may also contract to hasten the return of the lung–chest cage to the end expiratory position so that expiration is no longer solely dependent on elastic recoil of the lung. If work is measured over a minute, when tidal volume is unaltered but breathing frequency is increased, both elastic and frictional work increase. With a higher breathing rate, not only is frictional resistance higher for each tidal volume, but also the number of tidal volumes is increased each minute. Thus, more elastic work is also performed.

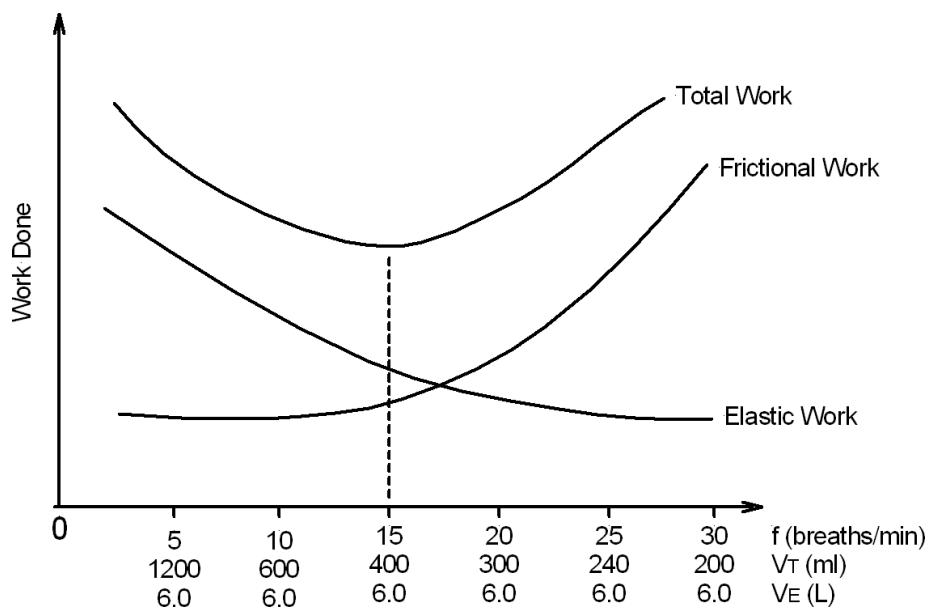
As the tidal volume increases until it reaches the vital capacity of the lungs, elastic work also increases, as shown by the compliance curve of Figure 9-15. For each breath, there may be a small increase in frictional work since more air passes through the airways with the larger volume. However, the increase in elastic work is much larger than the additional frictional work, because pulmonary compliance decreases as total lung capacity is approached. Thus, a greater pressure is required to change lung volume.

Figure 9-16 shows a plot of elastic, frictional, and total work of breathing at different tidal volumes and breathing frequencies. Each point along the x-axis is at the same ventilation rate,  $Q_E$  (L/min), which is accomplished by reducing tidal volume proportionally as breathing frequency,  $f$  (breaths/min), increases. At a high tidal volume and low breathing frequency (left), most of the total work is elastic because the lung is nearly maximally inflated with each tidal volume whereas breathing rate is low. However, when breathing frequency is high and tidal volume is low (right), most of the respiratory muscle work is frictional because of the rapid breathing rate whereas elastic work is small because of the small tidal volume. The total work of breathing is lowest at some optimal point where the sum of elastic and friction work is lowest (Fenn and Rahn, 1965).



**FIGURE 9-15 ■**  
Effect of total volume on elastic work.

**FIGURE 9-16** ■  
The physical work of breathing for different breathing rates and constant ventilation rates  
[Adapted from (Fenn and Rahn 1965).]

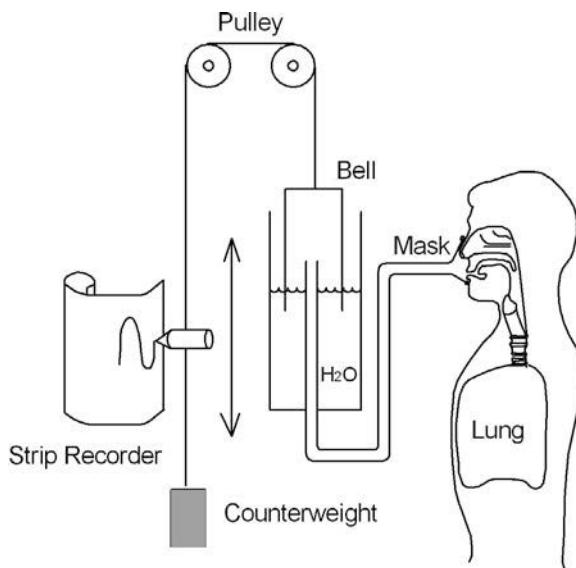


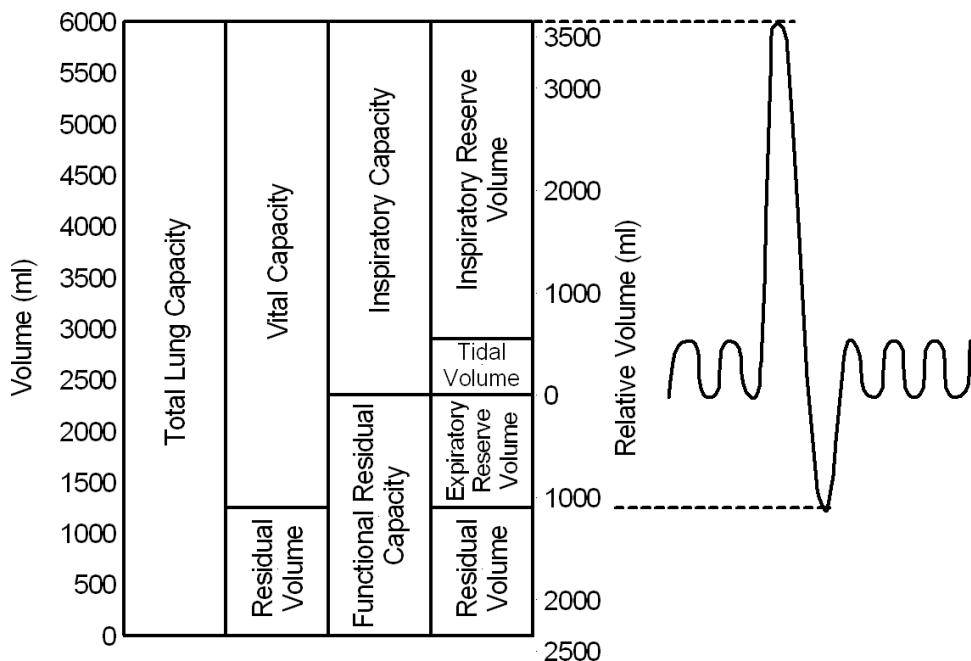
## 9.5 | MEASURING LUNG CHARACTERISTICS

### 9.5.1 Spirometry

Lung volumes are measured using spirometry. A spirometer consists of a bell that is connected to a stylus writing on a rotating drum or a strip recorder, as shown in Figure 9-17. The air-filled bell is inverted over a container of water so that an airtight chamber is formed. The bell is counterbalanced so it moves up and down with respiration with minimal resistance. Volume changes can be recorded on volume- and time-calibrated paper. Input and output check valves may be incorporated into the device to accommodate specific test requirements.

**FIGURE 9-17** ■  
Schematic diagram of a simple spirometer.





**FIGURE 9-18** ■ Volume fractions measured in spirometry [Adapted from (Bronzino 2006).]

### 9.5.1.1 Volume Fraction Definition

There are nine different volume fractions as characterized in spirometry; they are reproduced in Figure 9-18. The residual volume (RV) represents the volume of air left in the lungs after a maximal expiration. The vital capacity (VC) is the maximum volume of air that can be exhaled after a maximal inspiration. VC has three components. The first is the inspiratory reserve volume (IRV), which is the quantity of air that can be inhaled from a normal end inspiratory position. The second is the tidal volume (TV), which is the volume of air inspired and expired with each breath (about 0.5 L during normal breathing). The third is the expiratory reserve volume (ERV), which is the amount of air that can be exhaled from the lungs from a normal end-tidal expiratory position characterized by a relaxed expiratory pause. This is the easiest position to reproduce, and the lung volume in this position is called functional residual capacity ( $FRC = ERV + RV$ ). The total lung capacity (TLC) is the total volume of air in the lungs when they are maximally inflated ( $RV + VC$ ) and is approximately 6 L of air.

### 9.5.1.2 Volume Tests

Several timed respiratory volume tests are used to determine the ability of the respiratory system to move air. These include the forced vital capacity (FVC), forced expiratory volume in  $t$  sec ( $FEV_t$ ), the maximal voluntary ventilation (MVV), and the peak flow (PF). These measurements are obtained using a spirometer without valves or  $CO_2$  absorber or a pneumotachograph and an integrator.

The FVC test is performed by taking the maximum inspiration and forcing all of the inspired air out as rapidly as possible. It uses all the expiratory and accessory muscles. When the strong expiratory accessory muscles are contracted, high airflows at lung volumes near total lung capacity are generated. However, just following peak-expiratory flow (PEF) the airflow velocity decreases linearly with volume no matter how hard the subject tries. This is the effort-independent airflow and is caused by dynamic airway compression.

$\text{FEV}_t$  will be the portion of the FVC that can be expelled within a given period, typically 0.5 or 1 s.

The MVV is the volume of air moved in 1 minute with the subject breathing as deeply and rapidly as possible. The test is generally performed over 20 s to minimize the effects of hyperventilation.

PF is the maximum flow velocity attainable during the FEV maneuver and represents the maximum slope of the expired volume-time curve.

### 9.5.1.3 Oxygen Uptake Test

This is the measurement of oxygen use per unit time, called the  $\text{O}_2$  uptake test. A spirometer with input and output check valves passes the exhaled breath through soda-lime. This is a mixture of calcium hydroxide, sodium hydroxide, and silicates of sodium and calcium and it absorbs  $\text{CO}_2$ . Starting with a spirometer full of oxygen, normal respiration causes the bell to move up and down in a cyclical fashion in the normal way, but as oxygen is absorbed the baseline of the recording rises. By measuring the slope of the baseline on the spirogram, as shown in Figure 9-19, the volume of oxygen consumed per minute can be determined.

To determine the true volume of oxygen consumed, a correction factor,  $F$ , must be applied to compensate for the temperature difference and the pressure. Starting with Boyle's law

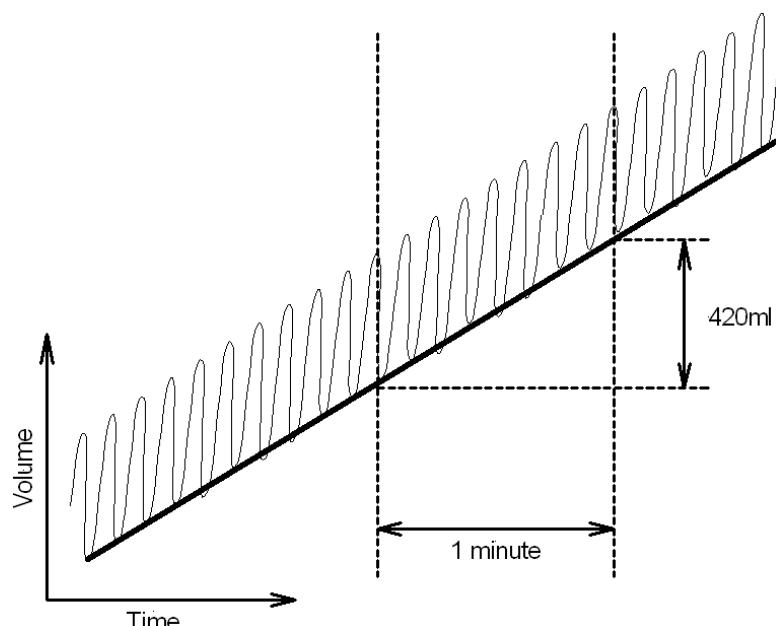
$$\frac{P_1 V_1}{T_1} = \frac{P_2 V_2}{T_2} \quad (9.8)$$

the final volume,  $V_2$ , can be determined as a function of all of the other parameters as

$$\begin{aligned} V_2 &= \frac{P_1 T_2}{P_2 T_1} V_1 \\ &= F V_1 \end{aligned} \quad (9.9)$$

**FIGURE 9-19**

Oxygen consumption measurement.



Remembering that the temperatures  $T_1$  and  $T_2$  are in Kelvin and that the pressures must take into account the partial pressure of water, the correction factor,  $F$ , can be rewritten as

$$F = \frac{P_{at} - P_{H_2O}}{P_{at} - 47} \frac{273 + T_{lung}}{273 + T_{amb}} \quad (9.10)$$

For an atmospheric pressure  $P_{at} = 760$  mmHg,  $P_{H_2O} = 25.2$ , with the lung and the ambient temperature, respectively,  $T_{lung} = 37$  °C and  $T_{amb} = 26$  °C, the correction factor is

$$\begin{aligned} F &= \frac{760 - 25.2}{760 - 47} \times \frac{273 + 37}{273 + 26} \\ &= 1.0685 \end{aligned}$$

For a measured oxygen consumption  $V_1 = 420$  ml/min, the corrected value is

$$\begin{aligned} V_2 &= FV_1 \\ &= 1.0685 \times 420 \\ &= 448.77 \text{ ml/min} \end{aligned}$$

A number of other techniques have been developed to measure oxygen concentration directly. Faraday first demonstrated that oxygen has the unique property among gases of concentrating a magnetic field in the same way that iron does. Oxygen is therefore attracted into a magnetic field and will attempt to displace the other gases in the air mixture. A device to show this consists of a dumbbell filled with nitrogen suspended on a quartz fiber in an asymmetrical magnetic field. The asymmetric distribution of oxygen around the dumbbell results in a torsional force that can be detected using a mirror attached to the fiber.

Another method involves measuring the velocity of sound through the gas, as it is proportional to the molecular weight of the mixture.

The speed of sound,  $c$  (m/s), in an ideal gas can be written in terms of the equations of state ( $PV = nRT$ ) for ideal gases (Randall, 2005).

$$c = \sqrt{\frac{RT\gamma}{M}} = \sqrt{\frac{P\gamma}{\rho_o}} \quad (9.11)$$

where:  $R$  = Universal gas constant (8134.3 J/Kmol)

$T$  = Temperature (K)

$M$  = Molecular weight of the gas (g/mole)

$\gamma$  = Ratio of specific heats (adiabatic exponent)  $C_p/C_v$  for the gas

$P$  = Pressure (1 atm =  $1.013 \times 10^5$  Pa)

$\rho_o$  = Density (kg/m<sup>3</sup>)

The adiabatic exponent  $\gamma$  can be estimated as follows:

- 1.66 for monatomic gases (He, Ne, Ar),
- 1.40 for diatomic gases (H<sub>2</sub>, O<sub>2</sub>, N<sub>2</sub>),
- 1.33 for triatomic and more complex gases (NH<sub>3</sub>, CH<sub>4</sub>, C<sub>7</sub>H<sub>8</sub>),
- 1.286 for very long molecules.

**WORKED EXAMPLE**

The speed of sound in any gas mixture can be calculated by using the molecular weight to determine the density. Air is made up of O<sub>2</sub>, N<sub>2</sub>, and CO<sub>2</sub> in proportions (21:78:1); therefore, the molecular weights are 16 + 16 = 32, 14 + 14 = 28, 12 + 16 + 16 = 44. The molecular mass,  $M$ , of the mixture is therefore

$$M = \frac{21 \times 32}{100} + \frac{78 \times 28}{100} + \frac{1 \times 44}{100} = 29$$

Remembering that at atmospheric pressure and at 0 °C the volume of 1 mole of gas is 22.414 liters the density can be determined

$$\rho_o = \frac{29 \times 10^{-3}}{22.4 \times 10^{-3}} = 1.29 \text{ kg/m}^3$$

And, finally, the speed of sound through the air can be calculated

$$c_{air} = \sqrt{\frac{1.013 \times 10^5 \times 1.4}{1.29}} = 331.6 \text{ m/s}$$

If about 50% of the oxygen has been used and no additional CO<sub>2</sub> has been added, then the proportions become 11:88:1, and the molecular mass changes to

$$M = \frac{11 \times 32}{100} + \frac{88 \times 28}{100} + \frac{1 \times 44}{100} = 28.6$$

This results in a change in density

$$\rho_o = \frac{29.6 \times 10^{-3}}{22.4 \times 10^{-3}} = 1.277 \text{ kg/m}^3$$

which results in a change in the speed of sound of nearly 0.5% to

$$c_{air} = \sqrt{\frac{1.013 \times 10^5 \times 1.4}{1.277}} = 333.25 \text{ m/s}$$

Figure 9-20 shows the theoretical change in the speed of sound for oxygen concentrations varying from 0% up to about 25%. It is interesting to note that the relationship looks very linear.

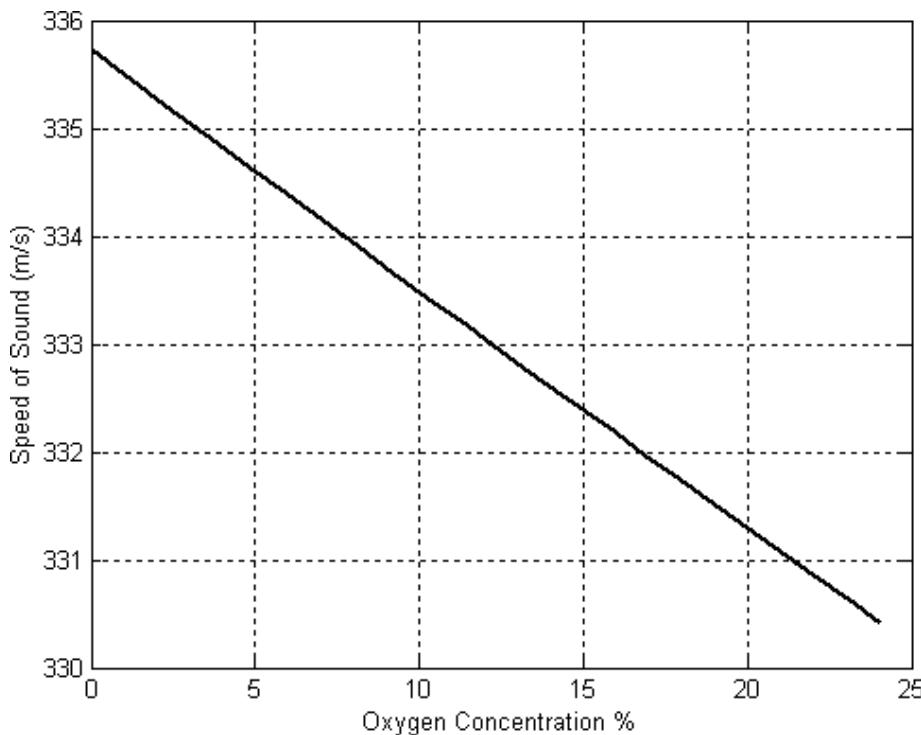
Commercial instruments pass the dried gas into a 1 m long tube and determine the sound velocity at 150 kHz by measuring the phase shift. Sensitivities of 0.004% for oxygen up to 21% concentration and 0.0008% for CO<sub>2</sub> up to 8% concentration are possible.

### 9.5.2 Pneumotachography

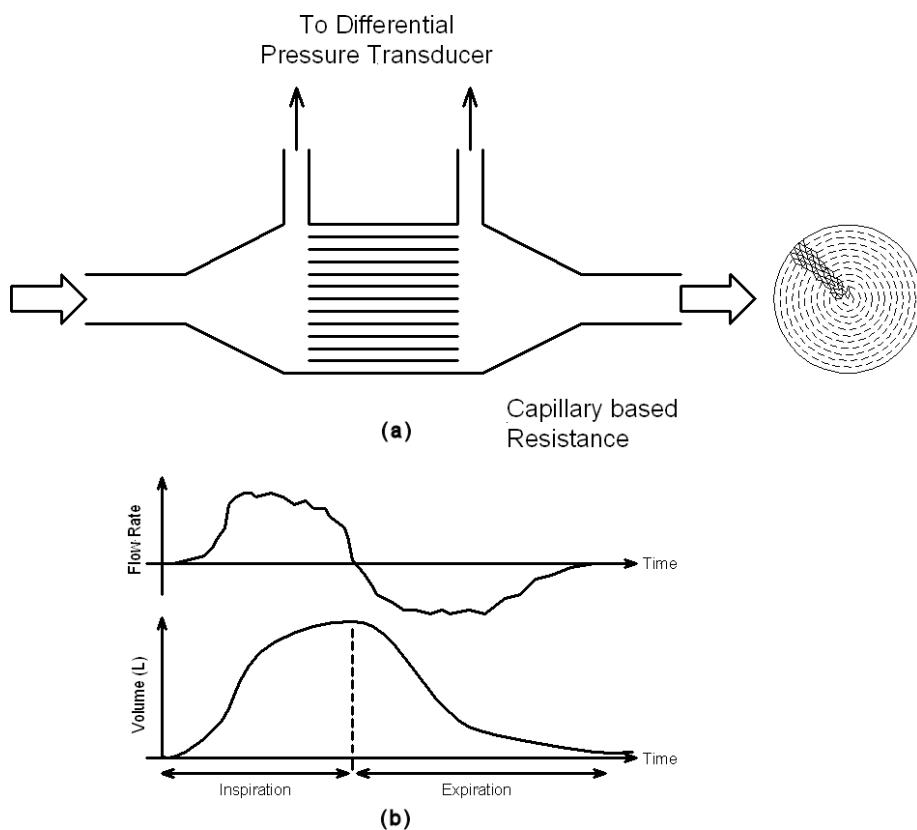
A pneumotachograph is a device for measuring airflow. It consists of a respiratory tube with a small resistance to airflow. The two chambers on either side of the resistance are connected to a differential pressure transducer by thin tubes, as shown in Figure 9-21. During respiration through the pneumotachograph tube, a small pressure difference,  $\Delta P$ , is measured across the resistance. The laminar flow rate is directly proportional to the ratio of the pressure difference,  $\Delta P$ , and resistance,  $R_f$ , according to the fluid version of Ohm's law.

$$Q = \frac{\Delta P}{R_f} \quad (9.12)$$

The resistance element consists either of a wire screen or a series of capillaries that maintains laminar airflow even at the maximum air speed.



**FIGURE 9-20** ■  
Theoretical speed of sound as a function of oxygen content.



**FIGURE 9-21** ■  
Pneumotachograph  
(a) Schematic diagram showing hardware. (b) Typical outputs showing flow rate and volume.

Although tidal volume is probably easier to record with a spirometer, the dynamics of respiration are better displayed using a pneumotachograph, which offers less resistance to the air stream and a much shorter response time. The response is so fast that cardiac impulses are often clearly identifiable on the flow rate versus time record. The volume is obtained by integrating the flow rate.

A typical pneumotachograph offers a resistance of between 5 and 10 mmH<sub>2</sub>O and can measure flow rates of up to 200 L/min. Response times of between 15 and 40 ms are usual.

Other ways of measuring flow using turbines, a heat conduction, or Doppler methods are discussed in Chapter 2.

## 9.6 | MECHANICAL VENTILATION

It is reasonable to classify modern ventilators into two main groups. The first, and largest group, are those used in intensive care to support patients after surgical procedures or to assist patients with acute respiratory disorders. The second includes less complicated machines used at home to help treat patients with chronic respiratory disorders.

### 9.6.1 Early History

*But so that life may in some measure be restored to the animal, you must attempt an opening in the trunk of the trachea and pass into it a tube of rush or reed, and you must blow into this so that the lung may expand and the animal draw breath after a fashion; for at a light breath the lung in this living animal will swell to the size of the cavity of the thorax, and the heart take strength afresh, and exhibit a great variety of motions. Vesalius (1543)*

This is an English translation of a section of the last chapter of Andreas Vesalius famous anatomical treatise, *De humani corporis fabrica*. It is probably the first account of mechanical ventilation and also the first description of the physiological effects resulting from ventilation on a collapsed lung.

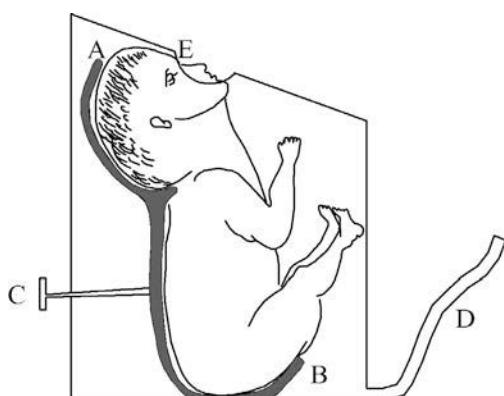
Robert Hooke's rather brutal experiments, described in the proceedings of the Royal Society of October 24, 1667, showed that if the thorax of a dog was opened it was unable to breathe. However, the dog could be kept alive for hours using bellows to inflate the lungs periodically.

John Mayow was probably the first scientist to really understand the mechanics of respiration. In 1670, he showed that air is drawn into the lungs by enlarging the thoracic cavity. He demonstrated the principle by building a model using bellows inside which was inserted a bladder. Expanding the bellows caused air to fill the bladder, and compressing the bellows expelled air from the bladder.

Throughout most of human history, people have sought the ability to restore breath into the bodies of those who have stopped breathing. The devices and theories used in these experiments were built on ideas that came about mostly during the late eighteenth century and were focused on the recovery of the apparently drowned or dead.

Initially, doctors established some basic methods of resuscitation, including warmth, inflation (very similar to modern rescue breathing), fumigation, friction, stimulants, bleeding a vein, and encouraging vomiting. However, most were not effective and thus fell from favor after a few years.

In 1782, the Royal Humane Society of England supported the use of bellows as the best means of inflation for artificial respiration. This method was widely supported



AB. Plaster Mould  
C. Screw for elevating mould  
D. Pipe for exhausting air  
E. Rubber diaphragm surrounding nose and mouth

**FIGURE 9-22 ■**  
Copy of a drawing of the infant resuscitator developed by Dr. Egon Braun.

internationally for over 40 years, until J. J. J Leroy of France challenged its use in an 1829 memoir. Leroy criticized the lack of control in using bellows and suggested a type of bellows that could be regulated for the specific patient's size and weight. As a result of this criticism, the bellows method lost support and went out of use around 1837, by which time many doctors had returned to basic methods of manual inflation and comfort assistance (Keith, 1906).

Many methods of artificial resuscitation were developed during this period, of which most were focused on different medical emergencies that demanded artificial resuscitation, including stillborn infants or chloroform asphyxia (Emerson and Loynes, 1978; Keith 1906).

In 1889, O. W. Doe reported to the Obstetrical Society of Boston on the development of an infant resuscitator box by Egon Braun in Vienna. This early form of artificial respirator obtained a pressure seal by having the child's mouth pressed against a rubber diaphragm opening while the rest of the body was entirely enclosed within the wooden box, as can be seen in Figure 9-22. The operator blew into the pipe to increase the pressure within the box, which forced the chest to compress. Sucking on the pipe reduced the pressure in the box, which in turn caused the chest to expand and draw air into the lungs.

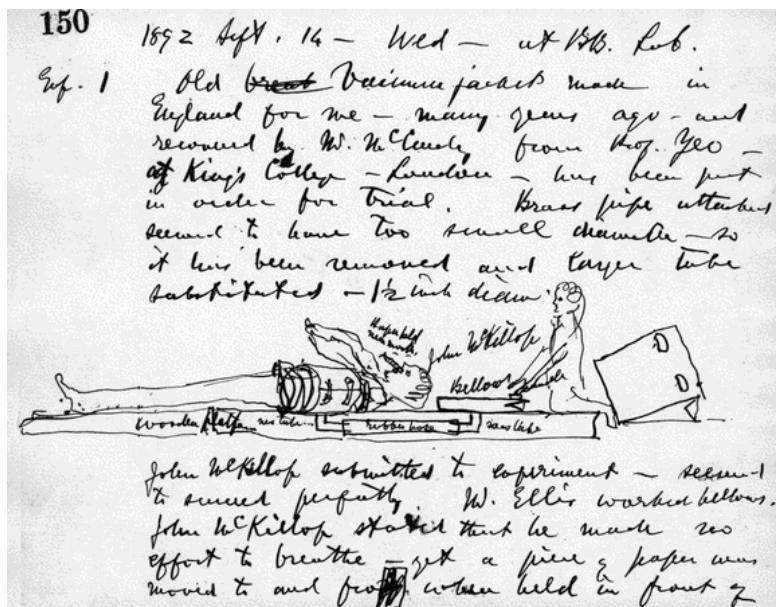
According to the report, the operator would repeat the process 20 to 30 times in a minute. Doe reports that Braun had used the artificial respirator device in 50 cases and was completely successful (Green, 1889).

There were a number of additional early experiments with artificial respiration devices. In 1832, John Dalziel of Scotland developed a box to ventilate a "drowned seaman," and E. J. Woillez designed an artificial respirator called a spirophore in 1876 that was said to have looked and operated much like the Drinker or Emerson iron lung. After the death of his child from respiratory complications in 1881, Alexander Graham Bell designed and built a test-version "vacuum jacket" with hand-operated bellows much like those on the iron lung. His original sketch and notes for the device are shown in Figure 9-23.

## 9.6.2 Polio

Beginning in the second decade of the twentieth century and lasting through the late 1950s, much of the world experienced widespread yearly polio epidemics that reached every part

**FIGURE 9-23 ■**  
 Page from Alexander Graham Bell's diary showing a sketch of his vacuum jacket and notes concerning its operation, (Bell 1892), courtesy of the Library of Congress.



of the population. At first, the disease appeared to exist in poor and overcrowded areas, but before long it was a threat to everyone. U.S. president Franklin Delano Roosevelt contracted polio at a later age than usual and then used his presidential powers to create the National Foundation for Infantile Paralysis.

Poliomyelitis is identified as an acute infectious disease caused by the poliovirus and characterized by fever, motor paralysis, and atrophy of skeletal muscles, often with permanent disability and deformity. It is often accompanied by inflammation of nerve cells in the anterior gray matter in each lateral half of the spinal cord, which can cause paralysis.

The virus is spread mostly through contact with an already infected person—in many cases through the mouth, where it then implants itself in the tissues of the alimentary tract. At this point, the body's natural defense mechanisms usually combat the poliovirus, prevent further spread through developed immunity, and lead to complete recovery from the infection.

For those infected, whose bodies do not manage to eliminate the virus, the outcome is quite different. In these cases the sufferers typically experience neck and back stiffness and muscle weakness as a result of the destruction of the central nervous system, specifically of the motor nerve cells in the gray matter of the spinal cord. The destruction of these cells is usually permanent and induces paralysis (Paul, 1971).

In the United States in the 1940s, there were less than 100 cases of polio reported during the months of January through June. However, between June and September the number of cases regularly increased significantly. As reported by the U.S. Public Health Service, summer 1944 experienced over 1600 reported cases of polio.

Between 1915 and 1945, not a single year had fewer than 1500 cases reported, and in more than half of those years over 5000 cases were reported. Three of the years had over 15,000, with a peak of 27,363 cases in 1916. The number of deaths during this period, attributed to poliomyelitis, ranged from 487 (1938) to 7179 (1916) reported cases. The magnitude of these numbers caused a real fear among the population and prompted a serious effort to combat the disease.



**FIGURE 9-24 ■**  
Iron lung ward of  
Ranchos Los  
Amigos Hospital  
circa 1953.  
(Courtesy of Rancho  
<http://www.rancho.org/>.)

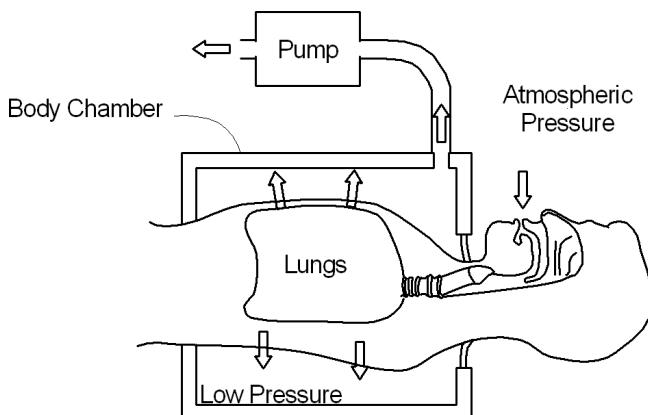
Until the 1920s, medical professionals had a limited variety of treatment options at their disposal for poliomyelitis patients. For example, in 1892 revolutionary public health researcher William Osler concluded that no drugs had the slightest influence upon acute myelitis, and that the child should be put to bed and the affected limb or limbs wrapped in cotton. Many medical caretakers accepted this generally ineffective approach and practiced wrapping or bed rest. Even today, patients suffering from severe poliomyelitis, particularly paralytic poliomyelitis, require intensive assistance especially in breathing (Paul, 1971).

The iron lung, discussed in detail later in this chapter, offered paralytic poliomyelitis patients the assistance they needed to survive. Many people from the generations around the polio epidemics remember images of hospital rooms filled with iron lungs, as seen in Figure 9-24.

### 9.6.3 External Negative-Pressure Ventilators

External negative-pressure ventilators (ENPVs) work by intermittently applying a sub-atmospheric pressure to the chest wall and abdomen, as illustrated in Figure 9-25. This increases transpulmonary pressure and causes atmospheric pressure at the mouth to inflate the lungs. Expiration occurs passively by elastic recoil of the lung and chest wall as pressure within the device rises to atmospheric levels.

**FIGURE 9-25** ■  
Simplified illustration  
of a negative-  
pressure ventilator.



The *iron lung* is the popular name for an ENPV device that consists of an airtight metal cylinder called a plethysmograph that encloses the whole body from the neck down, which is connected to a pump that can lower the internal pressure within the cylinder to below that of the surrounding atmosphere to draw air into the lungs.

Techniques for sustained ventilatory support awaited the widespread availability of a reliable supply of mains electricity, the development of electric motors, and a large number of patients with chronic respiratory failure. This need arose during the polio epidemics of the first half of the twentieth century.

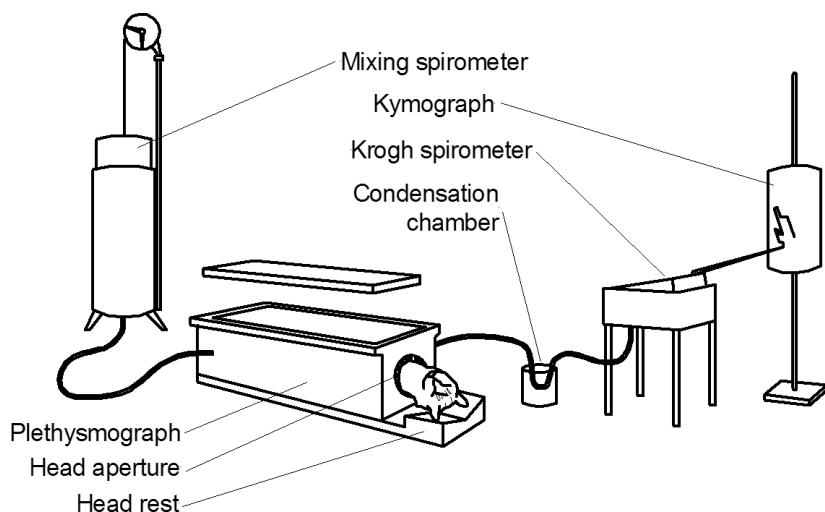
In 1918, South African doctor W. Steuart developed a respirator much like the ones that later made Drinker and Emerson famous. His machine was a sealed wooden box made specifically for treating poliomyelitis that operated by variable-speed, motor-driven bellows. Although his machine was supposedly a great success, the work was never formally reported and therefore became largely forgotten.

In 1926, the Consolidated Gas Company of New York used its Liability Insurance Fund to establish a committee to research resuscitation. As a business program, the executives responsible for the development of this committee intended to improve safety standards for the company's workers.

Among committee members was Harvard University professor Philip Drinker. Drinker suggested that the committee consider supporting research by his brother, the chemical engineer Cecil Drinker, and his colleague, Louis Shaw, who were already working on artificial resuscitation at Harvard University. The committee agreed and sent a check for \$5000 in autumn 1927 to support further investigation (Gorham, 1971).

Drinker and Shaw's research was focused on the discovery of true artificial resuscitation. As part of their research, they placed an anesthetized cat in a sealed box metal box (plethysmograph) with a neck collar, allowing the body to be within a fully pressurized environment. Under these conditions, they were able to record accurate measurements of respiration. Inhalation increased the volume of the cat within the box and made the pressure rise, while expiration produced the opposite effect (Shaw, 1928). A simplified diagram of this apparatus is shown in Figure 9-26.

The men speculated that if the subject could not breathe independently then increasing and decreasing the pressure in the box would induce respiration. To test this hypothesis, Drinker and Shaw injected the cat with curare, a powerful muscle relaxant, to induce respiratory arrest. They then placed the cat into the sealed box and used a hand-operated piston to manually control the pressure. The experiment was successful, and the cat was kept alive until the effects of the curare wore off after a few hours. They had



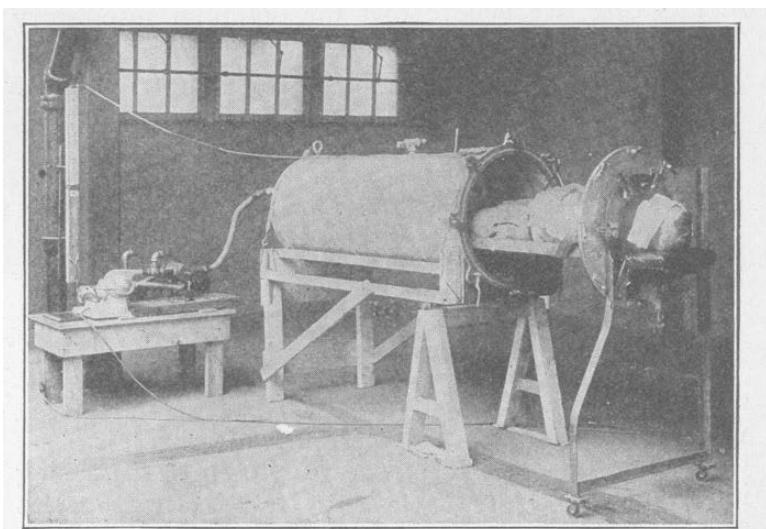
**FIGURE 9-26** ■ Schematic diagram of the Drinker and Shaw respiratory apparatus, tested on a cat. [Adapted from (Shaw 1928).]

convincingly demonstrated that controlled pressure in a sealed environment could induce respiration.

Following this successful experiment, and with an additional \$2000 in funding from Consolidated Gas, Drinker and Shaw extended their research to a larger device that could be tested on human beings. This first unit was constructed by a local tinsmith and used a vacuum cleaner to provide the suction. The patient was slid into the respirator on a garage mechanic's creeper, after which the end plate was secured over the patient's head (Hill, 1995).

#### 9.6.4 The Drinker Respirator

As shown in Figure 9-27, the Drinker respirator was built to accommodate a full range of patients. It could hold a small child or a fully grown man as tall as 2 m and weighing



**FIGURE 9-27** ■ Early prototype of the Drinker-McKhann respirator, (Drinker and McKhann 1929). Courtesy of JAMA, with permission.

Fig. 1.—The mechanical respirator, showing patient ready to be pushed into the tank. The pumps and manometer for controlling the pressure are shown in the background, to the left.

100 kg. The machine satisfied all of the technical goals outlined already and was still a very functional and accessible respiratory care device.

As shown in Figure 9-27, patients lie flat on their back with their head lying on a stand outside the lid of the tank and a rubber collar providing a seal around the neck. This rubber collar is designed to offer the seal necessary to maintain a pressurized environment but is still be comfortable for the patient. In this arrangement the doctors only have to slide the bed out to examine the patient.

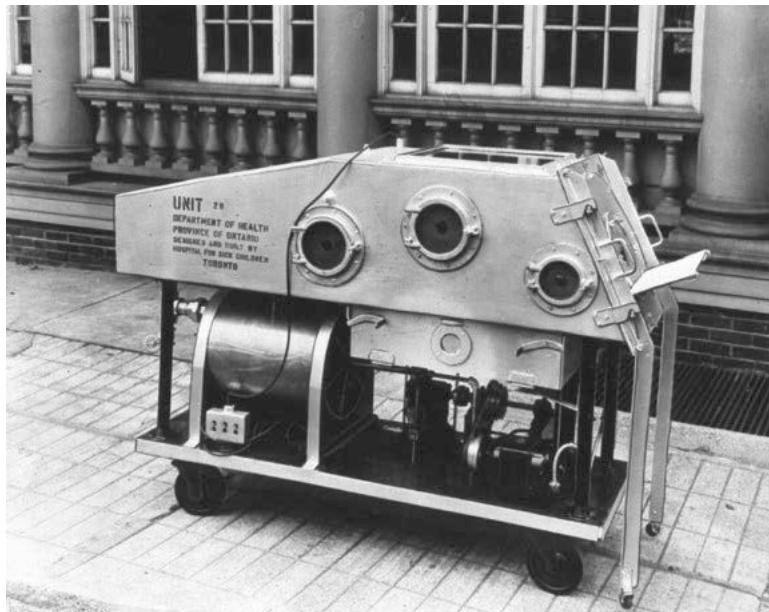
Consolidated Gas purchased a fully operable respirator and donated it to Bellevue Hospital in New York City, and within a few months the first clinical trial was under way. An 8-year-old child, comatose from lack of oxygen, was revived by the device. Subsequent to that, most of the early patients of the iron lung were polio sufferers with chest paralysis (Bellis, 2008; Drinker and McKhann, 1929; Gorham, 1971).

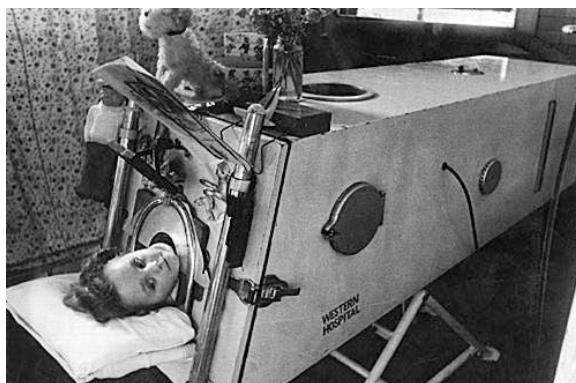
Drinker and McKhann aimed to offer all patients suffering from paralytic anterior poliomyelitis the opportunity to recover normal respiration with the assistance of artificial respiration for several hours, several days, or however long might be necessary. Their study had found that the existing manual methods of artificial resuscitation were ineffective in providing the necessary oxygen interchange and could not be used for extended periods of time. Additionally, other artificial resuscitators, including the pulmator, were too forceful and damaged other organs.

The main requirements proposed for the respirator design included long and steady function, adaptability to many ages and sizes, the ability to regulate the rate and depth of respiration, and the ability to provide proper artificial respiration without harming the patient (Drinker and McKhann, 1986).

It was dubbed the iron lung by an unknown American journalist, and after numerous well-publicized improvements and modifications, the device went into production by Warren E. Collins Company (Boston, MA). By 1931, 70 Drinker respirators of the kind shown in Figure 9-28 were in use across the United States (Hill, 1995).

**FIGURE 9-28 ■**  
Photograph of a Drinker type respirator manufactured in the workshops of the Hospital for Sick Children in 1937. (Courtesy of Hospital Archives, the Hospital for Sick Children, Toronto, with permission.)





**FIGURE 9-29 ■**  
Photograph of the Both portable respirator. (Courtesy of Western Fever Hospital, Fulham.)

In a historical review, biomedical engineer Philip A. Drinker, son of the inventor, identified the Drinker and Shaw iron lung as an early example of biomedical engineering long before the field was even conceptualized. He suggested that the success of the iron lung could be attributed to the availability of electricity and the immediate need for treatment of polio patients as well as the involvement of an engineer at all stages of the development of the device.

The 1937 polio outbreak was particularly virulent, with the influx of patients exceeding the available respirators. As patients often needed to spend 2 years or more in a respirator, hospital workshops had no alternative but to build their own devices. Figure 9-28 shows one of the respirators built at the Toronto Hospital for Sick Children in 1937 to cater for this influx (Uleryk, 2010).

Further afield, Drinker respirators were marketed in the United Kingdom by Siebe, Gorman and Company Ltd who had for many years lead the development of deep-sea diving gear.

### 9.6.5 The Both Respirator

In 1937, the London County Council in conjunction with the South Australia health authority commissioned Edward Booth, a medical apparatus manufacturer at Adelaide University, to design a cheaper alternative to the Drinker respirator. The Both portable respirator, shown in Figure 9-29, consisted of a plywood cabinet and a separate cylindrical motor-driven bellows, both on wheels for easy mobility (Hicks, 2003). Because they were made of wood, they were much lighter and less expensive than the alternatives and thus became the respirator of choice in hospitals in Australia and throughout the British Empire (Meacham, 2004).

Lord Nuffield, founder of the Morris Motor Car Company, had 800 of these devices manufactured at one of his car factories. He donated them to hospitals around the country but was bitterly attacked for his donation by the medical press for imposing the experimental Both respirator on the profession (Hill, 1995).

### 9.6.6 Homemade Iron Lungs

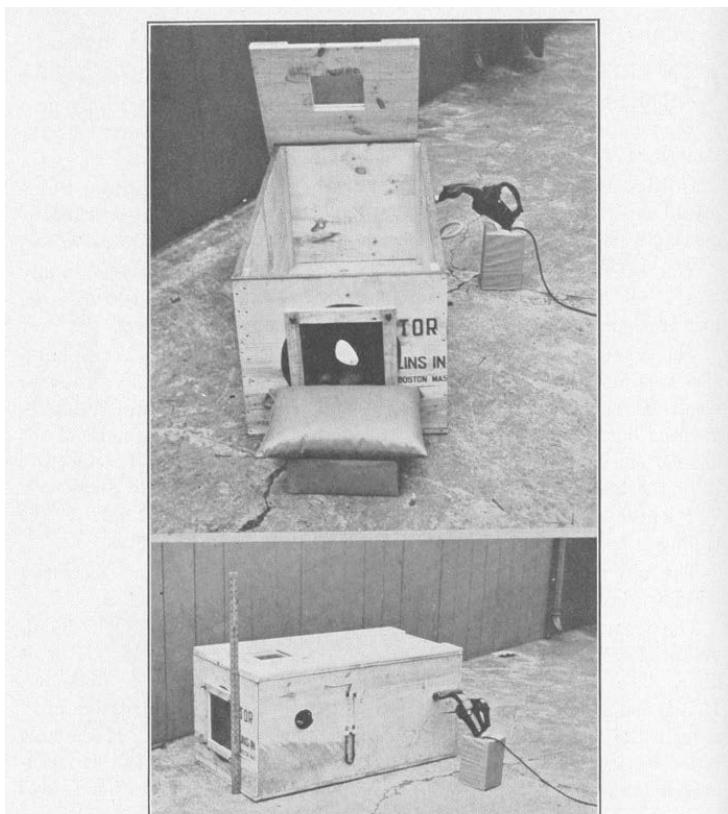
In September 1937, when their only Drinker respirator was in full-time use and more paralyzed children were admitted to Toronto's Hospital for Sick Children, chief engineer Harry Balmforth and his carpenter William Hall assembled an emergency "wooden lung" at the request of hospital superintendent Joseph Bower. It was completed in 7 hours from

pine boards, three hinges from a trunk, some metal rings, a rubber sheet, an air hose, and a vacuum pump. A few weeks later, with money provided by the Ontario government, six steel replicas of their wooden respirator were manufactured in the hospital workshop at a cost of less than \$500 each (Anon., 1937).

At about the same time, Maxwell Reynolds, a prominent Marquette citizen, and hospital engineer Lowell Reynolds also designed a wooden lung. This saved the lives of numerous children throughout the Upper Peninsula of Michigan and elsewhere. Apparently plans for building the wooden lung were written up in medical journals and used throughout the country. No one patented the design, and no profit was made by anyone involved in the project.

During the height of the polio epidemic years, Philip Drinker and Edgar Roy also provided instructions for building an emergency respirator for cases of life-threatening paralysis. Although this emergency solution was intended only for use until a production version arrived, it was functional enough to save lives. In contrast to the original, manufactured iron lung, Drinker and Roy employed common household and conveniently available hardware-store materials in their emergency respirator. The materials included a car inner tube for a rubber collar, a common vacuum cleaner to provide pressure, a 6-inch square piece of double-thick glass, a piece of sole leather to serve as the valve, a glass U-tube with colored water to show pressure, and several pieces of spruce. The detailed construction plans, of which one page is shown in Figure 9-30, clearly indicated that it

**FIGURE 9-30 ■**  
Photographs of a  
homemade plywood  
lung.' (Drinker and  
Roy 1938).  
Copyright Elsevier,  
reproduced with  
permission.



Figs. 1 and 2.—Wooden respirator, 43 in. × 22 in. × 21 in., with vacuum cleaner attached at rear right and hand-operated valve.

was an emergency respirator that could only accommodate small children and was not meant to replace the standard-size iron lung. Nonetheless, they fully supported its use in emergency situations (Drinker and Roy, 1938).

### WORKED EXAMPLE

---

#### Pressure, Volume, and Flow

- Determine the reduction in pressure required to produce the required tidal volume,  $TV = 0.25$  liters ( $250 \text{ cm}^3$ ), in a 5-year-old child in the plywood lung shown in Figure 9-30.

The combined lung and chest compliance,  $C_{tot} = 1 \text{ cm}^3/\text{Pa}$ ; therefore, the difference in pressure required is

$$\begin{aligned}\Delta P &= \frac{TC}{C_{tot}} \\ &= \frac{250}{1} \\ &= 250 \text{ Pa}\end{aligned}$$

- What volume of air must be removed from the plywood lung (assuming no leaks) to achieve this change in pressure?

If the mass of the child,  $M_{child} = 19 \text{ kg}$ , of which 1 kg is the mass of his head, and his average density,  $\rho = 0.98 \text{ g/cm}^3$ , his volume can easily be determined.

$$\begin{aligned}V_{child} &= \frac{M_{child} - 1}{\rho} \\ &= \frac{(19 - 1) \times 10^3}{0.98} \\ &= 18.4 \times 10^3 \text{ cm}^3\end{aligned}$$

The internal volume of the box determined in  $\text{in}^3$  and converted to  $\text{cm}^3$

$$\begin{aligned}V_{box} &= L.B.H \\ &= 43 \times 22 \times 21 \times 2.54^3 \\ &= 325.5 \times 10^3 \text{ cm}^3\end{aligned}$$

The volume of air in the box is just the difference between the internal volume of the box and the volume of the child

$$V_{air} = V_{box} - V_{child} = 307.1 \times 10^3 \text{ cm}^3$$

Boyle's law can be used to obtain the change in volume to achieve the required reduction in pressure. Assuming that the temperature remains constant

$$P_1 V_1 = P_2 V_2$$

As air is pumped out of the chamber, the pressure in the chamber decreases slightly as the volume increases to equal the sum of the volume in the chamber and that which has been pumped out.

Additionally, because the lung volume expands by 250 cm<sup>3</sup>,  $V_{air}$  decreases proportionally. Therefore,

$$P_{air}V_{air} = (P_{air} - 250)(V_{air} - 250 + \Delta V)$$

Changing the subject gives

$$\begin{aligned}\Delta V &= \frac{P_{air}V_{air}}{P_{air} - 250} - V_{air} + 250 \\ &= \frac{101338.4 \times 307.1 \times 10^3}{101338.4 - 250} - 307.1 \times 10^3 + 250 \\ &= 1009.5 \text{ cm}^3\end{aligned}$$

Note that the volume of air that must be removed from the box is a factor of four greater than the actual change in volume of the lungs.

3. If the average breathing rate of a 5-year-old child is  $f_{breath} = 23/60 \text{ Hz}$ , what is the minimum flow rate required from the vacuum cleaner?

Assuming that the breathing cycle is 50:50 inspiration:expiration, the time available for inspiration is,  $t_{insp} = 0.5 \times 60/23 = 1.3 \text{ s}$ .

The flow rate,  $Q$  (cm<sup>3</sup>/s), is

$$\begin{aligned}Q &= \frac{\Delta V}{t_{insp}} \\ &= \frac{1009.5}{1.3} \\ &= 777 \text{ cm}^3/\text{s}\end{aligned}$$

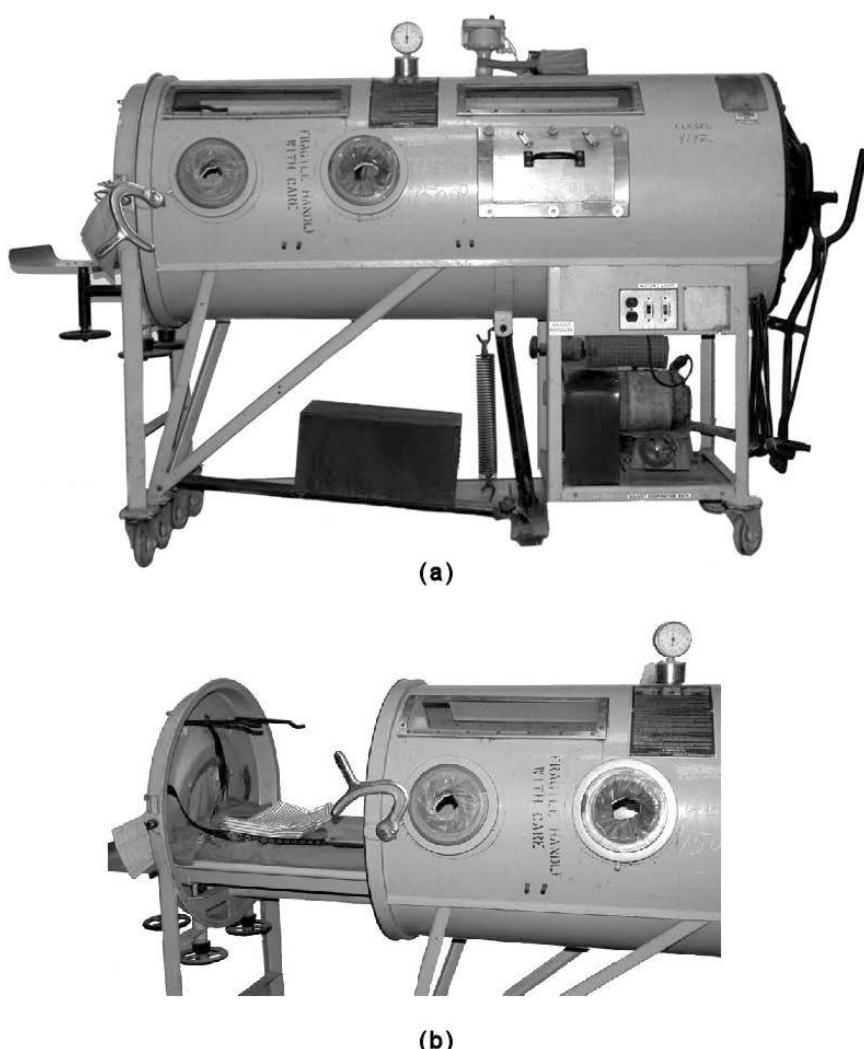
A modern vacuum cleaner can reduce the pressure by 20 kPa and provides a flow rate of between 20 and 50 lit/s. Assuming that technology has not advanced dramatically since 1938, the portable device shown would have been more than capable of providing the pressure and flow required by the plywood lung. In fact, it is likely that a deliberate leak would have been introduced to reduce the possibility of embolism in the child's lungs.

---

### 9.6.7 The Emerson Respirator

John Emerson, a Harvard engineer, simplified the Drinker respirator making it easier to manufacture and thus less expensive. His company J. H. Emerson Co. (Cambridge, MA) manufactured hundreds of the respirators shown in Figure 9-31, which were widely available in the 1940s and 1950s.

As can be seen from Figure 9-31, the machine is a large cylindrical metal drum into which a patient is admitted by opening the left end and rolling out a flat metal bed. The patient's head extends through the rubber/plastic collar and reclines on the headrest. After the patient has been installed, the bed is rolled back into the drum and clamped into place. A motor and pump mechanism underneath the body of the iron lung increases and decreases the internal air pressure in a cyclical manner, which forces air into and out of the lungs. Four windows on the top and six rubber-lined openings on the sides provide access to the patient. Whenever patients needed bathing or other medical care, a plastic dome was installed over their head that took over breathing automatically.



**FIGURE 9-31 ■**  
Emerson respirator manufactured in 1957. (a) Photograph of complete unit closed.  
(b) Photograph of unit partially open.  
(Courtesy of the Kansas Museum of History, reproduced with permission.)

According to a pamphlet issued by J. H. Emerson Co. in April 1956, “In the slow progress against poliomyelitis the iron lung has become a symbol of victory. Respirators have brought the breath of life to thousands, and of those who owe their lives to this temporary mechanical aid, the great majority are now completely free again from reliance on it.”

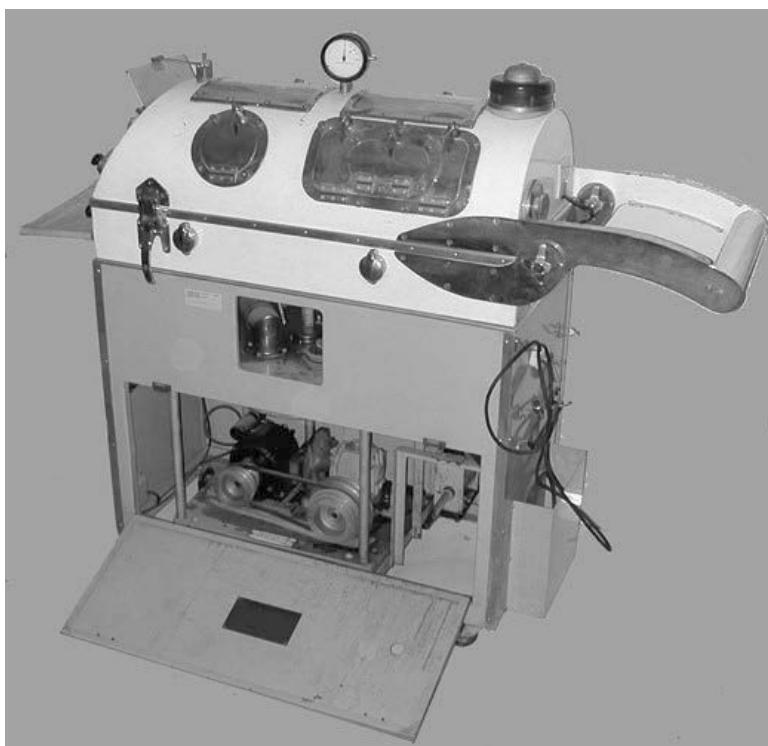
This pamphlet was issued in 1955, just 1 year after Jonas Salk developed the first effective polio vaccine. The rate of polio infection had dropped dramatically by 1957, the first year the vaccine was widely available.

### 9.6.8 The Alligator Cabinet Respirator

Meanwhile, in the United Kingdom, G. T. Smith-Clarke was commissioned in 1952 by the Birmingham Regional Hospital Board to update the Both design. He did this but also suggested that access could be improved if the whole cabinet could be split with the lid hinged at the foot end, as shown in the junior version in Figure 9-32. This new design, known as the Smith-Clarke “alligator” cabinet respirator, was manufactured and sold by Cape Engineering of Warwick. It made its debut in November 1954.

**FIGURE 9-32 ■**

The Smith-Clarke junior cabinet respirator. (Courtesy of the Sheffield Museum of Anaesthesia, reproduced with permission.)



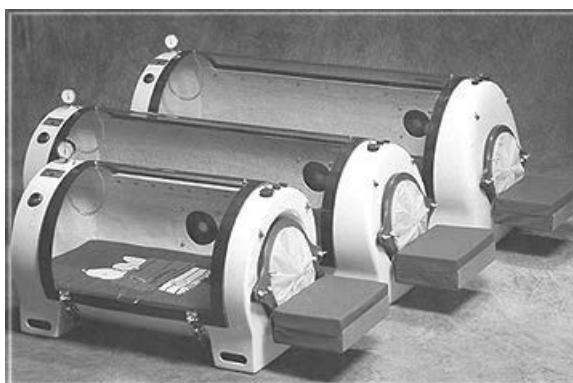
### 9.6.9 Portable Respirators

Tank-type respirators that operated using negative pressure, such as the Emerson iron lung or Drinker respirator, were the mainstay of ventilatory support during the polio epidemics in the 1950s. These were reliable but bulky (3 meters long) and heavy (300 kg), which meant that patients were restricted to remaining in hospital for the duration of their illness. More portable fiberglass tank ventilators became available for home use. These included the Portalung, shown in Figure 9-33, but these still weighed approximately 50 kg and required two people to move them.

Subsequently, less bulky, even more portable negative-pressure ventilators were developed. The most commonly used today are the poncho wrap (or jacket) ventilator, which

**FIGURE 9-33 ■**

Three different sizes of Portalung. (Courtesy of Portalung Inc <http://www.portalung.com/>.)





**FIGURE 9-34 ■**  
John Prestwich  
using a cuirass  
respirator.<sup>2</sup>  
(Courtesy of Maggie  
Prestwich,  
<http://www.johnprestwich.btinternet.co.uk/>  
with permission.)

consists of an impermeable nylon jacket suspended by a rigid chest piece that fits over the chest and abdomen, and the cuirass (or tortoise shell) ventilator, which consists of a rigid plastic or metal dome over the chest and abdomen, as seen in Figure 9-34. These chest and wrap ventilators are lightweight themselves, but both must be connected to negative-pressure generators, which weigh 15 to 30 kg.

In 50 years, the ventilators that kept John Prestwich alive have taken more than 420 million breaths on his behalf. Should one have failed and not been repaired or replaced for only 5 minutes during that time, John would have died. This speaks volumes for the quality and reliability that is required from life-support hardware such as this.

Until fairly recently, paralyzed polio patients and quadriplegics relied almost entirely on human support for every activity. However, the development of breath and sound controlled interfaces has increased their independence considerably. In Prestwich's case a series of whistles, similar to those used by shepherds to communicate with their dogs, are picked up by a microphone and relayed to a computer to control peripheral devices. He could control the TV, video, and hi-fi, open and close curtains, and control lighting and heating as well as unlock the front door. The interface allowed him to make and receive phone calls, while state-of-the-art software facilitated the use of his personal computer (PC) for word processing, emailing, and playing games (Hill, 1996).

### 9.6.10 Other Uses for Negative-Pressure Ventilation

Though the iron lung is remembered most notably from the days of the polio epidemic, it continued to receive occasional use after the development of the Salk and Sabin vaccines.

From the 1960s onward, most respiratory care patients were treated with new techniques that used improved endotracheal or tracheostomy tubes, but some survivors of

<sup>2</sup>Prestwich was paralyzed by polio at the age of 17 and spent more than 50 years using ventilators until his death in 2006. His indefatigable spirit was an inspiration to all who knew him.

earlier polio epidemics, as well as other intensive medical cases, still used of the iron lung. Lifecare Services, Inc., reported that as of January 1985 approximately 300 iron lungs were still being used in the United States (Drinker and McKhann, 1986).

In the last few decades, the United States has been largely free of polio and has been able to manage the rare occurrences with existing machines. As a result, J. H. Emerson Co. ceased production of its respirator in 1970, and Resironics Colorado discontinued repair, service, and parts on iron lungs as of March 2004.

According to a report from September 2004, there are an estimated 40 survivors of polio still living in the iron lung (Nelson, 2004). On November 1, 2009, June Middleton died in a Melbourne nursing home. She was 83 and had been confined to an iron lung for 60 years after being struck down by polio.

Despite wide-scale vaccination efforts by the World Health Organization, recent outbreaks of poliomyelitis in Africa and South America have led health workers back to negative-pressure ventilators as a possible life-saving technology for emergency cases of paralytic poliomyelitis. In addition, several studies have reported benefits of intermittent negative-pressure ventilation in patients with chronic respiratory failure due to chest wall deformity, neuromuscular diseases, and central hypoventilation.

However, in recent years, negative-pressure ventilation has been used infrequently for the management of patients with acute respiratory failure. In a review of the literature on noninvasive ventilation, Hillberg and Johnson (1997) note that the role of negative-pressure ventilation in the management of acute respiratory failure is unclear. In some studies of the use of the body ventilator or poncho wrap for patients with acute respiratory failure and chronic obstructive pulmonary disease (COPD), neuromuscular disease, or chest wall deformity, some benefit from these devices has been found. However another large randomized controlled trial found that negative-pressure ventilation had no benefit (Shapiro, Ernst et al., 1992).

Negative-pressure ventilation has not been widely used because of poor acceptance by patients, inadequate effectiveness for many patients, the awkward size of the devices, and the development of upper airway obstruction in some. Patients with stable or slowly progressive neuromuscular diseases, central hypoventilation, or chest wall deformities are the best candidates for noninvasive negative-pressure ventilation.

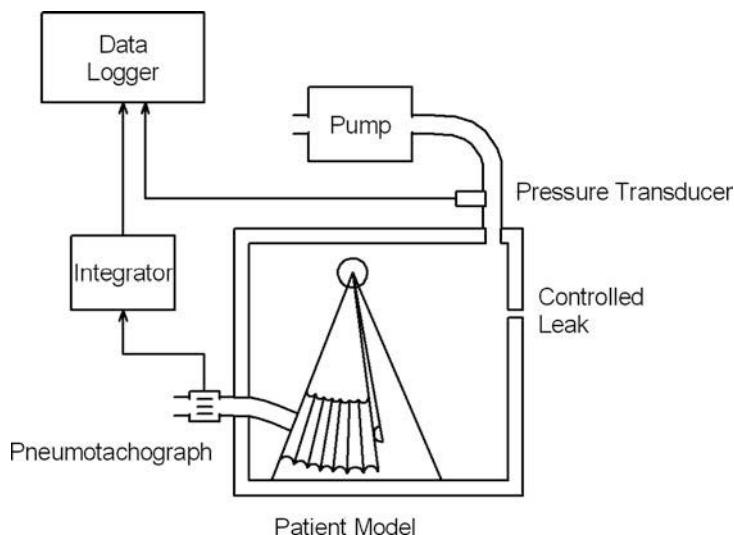
Noninvasive ventilation is not effective in diseases that affect the upper airways, such as amyotrophic lateral sclerosis. Additionally, patients with rapidly progressive neuromuscular processes like Guillain-Barre syndrome are also poor candidates. In general, noninvasive ventilation should not be used in patients who are unable to cooperate or who have impaired consciousness, problems with retained secretions, or hemodynamic instability.

## 9.7 | THE PHYSICS OF EXTERNAL NEGATIVE-PRESSURE VENTILATION

External ventilation systems consist of two main components: the chamber and the pump mechanism. The function of the chamber is to enclose either the whole body or the chest cavity effectively so that negative pressure will result in an expansion of the chest and draw air into the lungs. The pump must be capable of providing the correct negative pressure level in a cyclical fashion to drive the forced ventilation function. Various parameters of the pumps should be controllable to suit the differing requirements of a range of patients.

**TABLE 9-2** ■ Specifications of Five Negative-Pressure Pumps

Variable	NEV-100	Negavent	Maxivent	Newmarket	CCP-001
Breaths/min	4 to 60	1 to 50	5.6 to 25	6 to 30	10 to 30
Inspire: Expire ratio	1:29 to 1:0.6	1:99 to 1:0.2	1:1.2 fixed	1:1.5 to 1:0.7	1:1.5 to 1:0.7
Max negative pressure	-94 cmH <sub>2</sub> O	-89 cmH <sub>2</sub> O	-74 cmH <sub>2</sub> O	-49 cmH <sub>2</sub> O	-38 cmH <sub>2</sub> O
Max positive pressure	30 cmH <sub>2</sub> O	—	70 cmH <sub>2</sub> O	51 cmH <sub>2</sub> O	41 cmH <sub>2</sub> O
Max flow	845 L/min	944 L/min	630 L/min	780 L/min	660 L/min
Weight	14.5 kg	14.5 kg	16.5 kg	34.8 kg	14.7 kg



**FIGURE 9-35** ■ Experimental setup to measure negative-pressure pump characteristics. [Adapted from (Smith, King et al., 1995).]

In an interesting paper, the performances of five different pumps were compared using a mechanical lung simulator and pressure and flow measurement sensors (Smith, King et al., 1995). The five pumps were the NEV-100, Negavent respirator DA-1, Thompson Maxivent, the CCP-001, and Newmarket device, and their specifications are listed in Table 9-2.

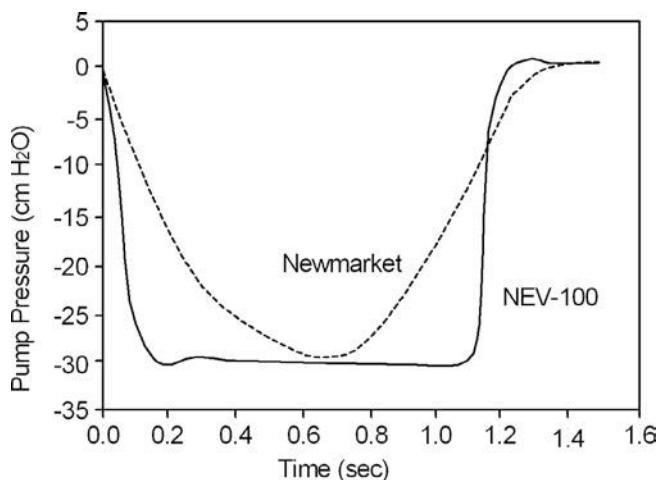
In the experimental setup shown in Figure 9-35, a sealed chamber with an adjustable leak houses the lung simulator. A pressure sensor connected to the chamber monitors the internal pressure. The output of the lung simulator (the mouth) was open to the atmosphere and was instrumented using a pneumotachograph to measure air flow. The negative-pressure pumps in the comparison were fitted to the chamber in turn, and their characteristics were logged and later analyzed.

Analysis showed that the selected inspire:expire ratio was within 1% for all of the pumps but that the variation in peak negative pressure was larger, with the NEV-100 being the most accurate at -29.8 to -31.2 cmH<sub>2</sub>O for a 30 cmH<sub>2</sub>O set point and the Maxivent the least accurate at -29.7 to -32.8 cmH<sub>2</sub>O.

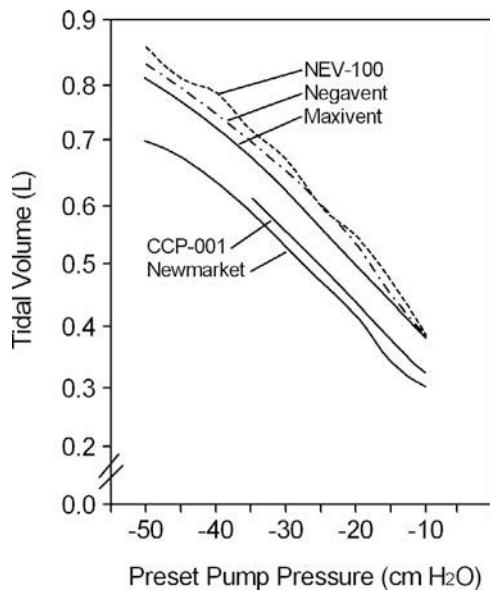
Most interesting, however, was the variation in the pressure during the pump cycle. The NEV-100 and Negavent produced square waves over the inspiratory time of 1.33 s, whereas the CCP-01 and Newmarket pumps approximated a half-sine wave, as shown in Figure 9-36.

Because the total inspired volume is proportional to the time integral of the negative pressure if the inspiratory time is reasonably short, the pumps that produced half-sine waveforms would generate a smaller tidal volume. This effect is illustrated in Figure 9-37.

**FIGURE 9-36 ■**  
Pressure waveforms measured within the patient model for 15 breaths/min, input:output ratio 1:2, and a preset pressure of  $-30\text{ cmH}_2\text{O}$ . [Adapted from (Smith, King et al., 1995).]

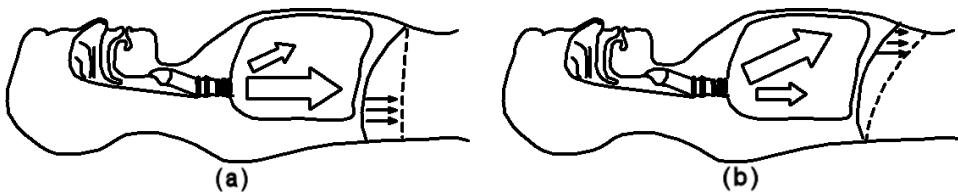


**FIGURE 9-37 ■**  
Tidal volume produced by the five pumps as a function of negative-pressure set point. [Adapted from (Smith, King et al., 1995).]



Compensation for leaks also varied by a large margin across the pump types. The Maxivent had no compensatory mechanism, so the drop in tidal volume was directly proportional to the size of the leak. The Newmarket pump and CCP-001 devices have a servo-controlled rotary valve to adjust the feed pressure. The time constant of this control loop was about 25 ms. In contrast, the Negavent and NEV-100 compensate for leaks by increasing the motor speed, which, though slower, proved to be more accurate.

Although it may appear that negative-pressure respirators incorporate the same principles as natural respiration, there are a number of significant differences. One of the main challenges has been to design a chamber that creates negative pressure around the thoracic walls. The iron lung solved this problem by encasing the entire body and sealing only around the neck. However, this leads to negative pressure applied to the chest and abdominal walls, which creates venous pooling in the abdomen and reduced cardiac output.



**FIGURE 9-38** ■ Airflow distribution in the lungs (a) Natural breathing. (b) Breathing assisted by negative-pressure ventilation.



**FIGURE 9-39** ■ New generation of cuirass respirators employ a biphasic approach to breathing. (a) Medical hardware. (b) Device installed on a human patient. (Courtesy of United Hayek Medical <http://www.unitedhayek.com/>.)

Attempts to reduce the size of the chamber and seal around the chest, as used in cuirass designs, have not been particularly successful because of sealing difficulties.

In addition, negative-pressure ventilators make the patient less accessible for observation, they are generally noisy, and in some cases synchronization with the patient's own attempts to breathe is difficult (Bronzino, 2006).

One final issue with negative-pressure ventilation is that the air flow distribution within the lungs is unnatural, as can be seen in Figure 9-38.

The latest cuirass respirators, of which one example is shown in Figure 9-39, have gone some way in addressing these problems by employing a biphasic approach that employs negative pressure for inspiration and positive pressure for expiration. This capability eliminates the dependency of passive recoil of the patient's chest and can therefore increase the frequency of breathing while simultaneously facilitating the removal of fluid and mucus from the lungs (Darwin, 2009).

## 9.8 | POSITIVE-PRESSURE VENTILATORS

### 9.8.1 Historical Background

Positive-pressure ventilation has evolved primarily from the use of bellows for resuscitation and in the development of technology to sustain divers underwater. In regard to the latter, it all started in 1771 when John Smeaton invented the air pump from which a hose could be connected to a submerged diving barrel, allowing air to be pumped to the diver. A year later, Sieur Freminet invented a rebreathing device that recycled the exhaled air

from within the barrel. Unfortunately, it didn't work very well, and the inventor died from lack of oxygen after being in his own device for 20 minutes.

Just over 50 years later, in 1825, William James designed another self-contained breather comprising a cylindrical iron hoop attached to a copper helmet. It provided enough air for a 7-minute dive. This was followed in 1876 by a closed-circuit oxygen rebreather invented by Henry Fleuss. This was first used in the repair of an iron door of a flooded ship's chamber and later in a 10 m deep dive. Unfortunately, oxygen under pressure is toxic, and Fleuss died.

Christian Lambson designed a more successful system in 1939 as part of the military Self Contained Underwater Breathing Apparatus (SCUBA) program. However, divers were still injured or killed frequently from oxygen toxicity. It wasn't until 3 years later when Emile Gagnan and Jacques Cousteau invented a demand regulator that a safe and effective method of providing fresh air when a diver breathed became available. A year later they began marketing the Aqua-Lung.

Research undertaken during WWII to deliver oxygen to fighter pilots operating at high altitude also played a part in the design of the modern positive-pressure ventilator.

Intensive use of positive-pressure ventilation gained momentum during polio epidemics in Scandinavia and the United States in the early 1950s. In Copenhagen, patients with respiratory paralysis were supported by manually forcing a rich oxygen mixture through a tracheostomy and had reduced mortality rates. However, this activity required the effort of 1400 medical students recruited from universities. The reduction in mortality rates from 80 to 25% lead to the development and adoption of positive-pressure machines (Byrd, Kosseifi et al., 2010).

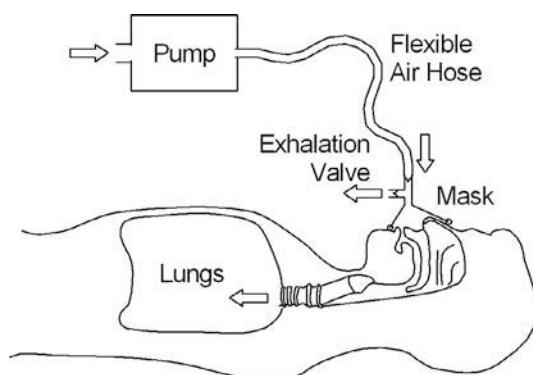
### 9.8.2 The Need for Positive-Pressure Ventilation

Positive-pressure ventilators generate inspiratory airflow by applying a positive pressure (greater than atmospheric pressure) to the airways. During inspiration, the inspiratory flow delivery system creates a positive pressure in the tubes connected to the patient's airway, while the expiration system closes the outlet valve to ensure that air flows into the patient's lungs. When the ventilator switches to expiration, the inspiratory flow is stopped and the expiration system opens the outlet valve, allowing the patient's exhaled breath to flow to the atmosphere, as shown in Figure 9-40.

The use of a positive-pressure gradient in creating the flow allows treatment of patients with high lung resistance and low compliance. As a result, positive-pressure ventilators have been very successful in treating a number of different breathing disorders and have all but displaced the use of negative-pressure ventilation from almost all respiratory

**FIGURE 9-40 ■**

Simplified schematic showing operation of a positive-pressure ventilator.



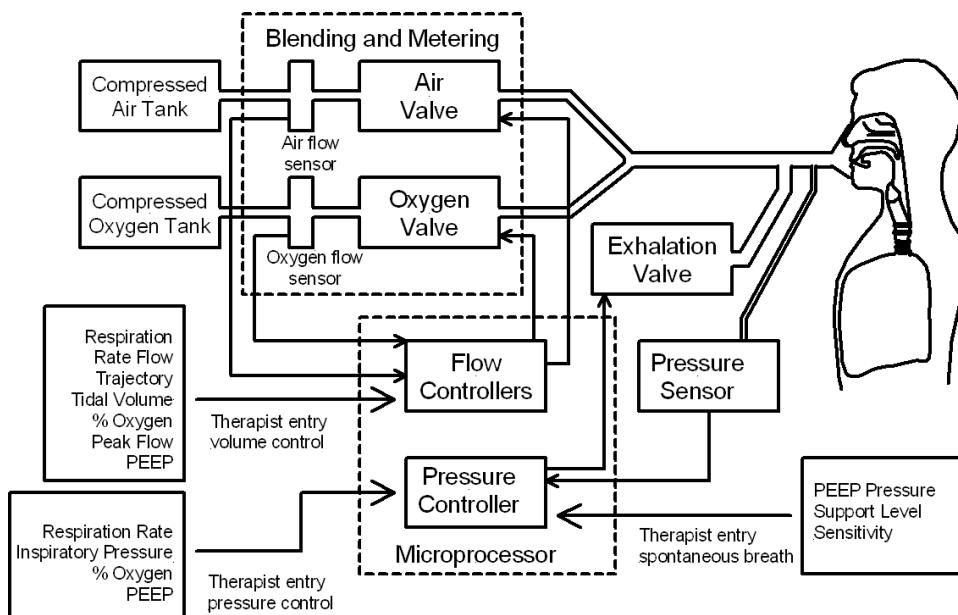
applications (Bronzino, 2006). The following are some of the more common indications for positive ventilatory support:

- Bradypnea or apnea with respiratory arrest.
- Acute lung injury.
- Tachypnea (respiration rate > 30 breaths per minute).
- Vital capacity less than 15 ml/kg.
- Inability to maintain arterial O<sub>2</sub> saturation.
- Respiratory muscle fatigue.
- Low blood pressure.
- Neuromuscular disease.

In some ventilation systems, compressed air and oxygen are stored in high-pressure tanks attached to the ventilator inlet, whereas in others an air compressor replaces the compressed air tank. Controlled mixing enriches the inspiratory airflow to the required level. In Figure 9-41 the air and oxygen valves are controlled by feedback from their respective flow sensors as programmed. During inspiration, a microprocessor controls each of the valves to deliver the required oxygen and air as well as closing the expiration valve. When expiration starts, the microprocessor opens the expiration valve by the correct amount to maintain the required positive end expiratory pressure (PEEP) as measured by an airway pressure sensor.

### 9.8.3 Ventilation Modes

Modern ventilators are classified (and named) by their method of cycling from the inspiratory phase to the expiratory phase. A number of different modes of ventilation have been devised to suit patient conditions. At one extreme, a patient may need a ventilation system to completely take over the respiration task. In this case, the device operates in mandatory mode and delivers mandatory breaths. Patients may be able to initiate breathing but may



**FIGURE 9-41** ■ Block diagram of a positive-pressure ventilator. [Adapted from (Bronzino 2006).]

need oxygen-enriched air or a slightly elevated airway pressure. When the ventilator assists a patient who is capable of breathing, the ventilator delivers spontaneous breaths and operates in spontaneous mode.

#### 9.8.4 Controlled Mandatory Ventilation

Controlled mandatory ventilation (CMV) may be divided into two distinct approaches: volume controlled and pressure controlled. Volume-controlled ventilation delivers a specific tidal volume to the patient during the inspiratory phase. Pressure-controlled ventilation raises the airway pressure to an adjustable set level during the inspiratory period.

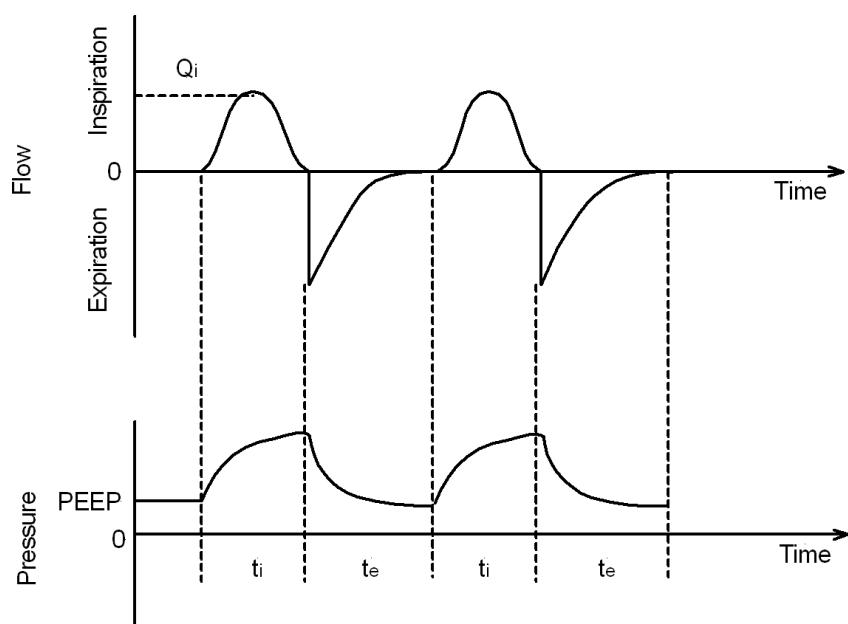
#### 9.8.5 Volume-Controlled Mandatory Ventilation

In this mode, the ventilator delivers a specific tidal volume to the patient during the inspiratory phase. However, at the end of the expiration phase, as shown in Figure 9-42, the airway pressure may not end at atmospheric pressure. This PEEP is sometimes maintained to keep alveoli from collapsing during expiration. The increased lung volume increases the surface area available for oxygen diffusion and reduces the volume of poorly oxygenated blood returning to the left atrium. PEEP therapy is also effective in improving lung compliance with the partially inflated lung, requiring less volume and energy to return to full inflation than a completely deflated lung.

The clinician specifies the following parameters on the ventilator:

- Respiration rate (breaths/min)
- Flow waveform
- Tidal volume
- Delivered oxygen concentration
- Peak flow
- PEEP

**FIGURE 9-42** ■  
Inspiratory flow for volume-controlled ventilation. [Adapted from (Bronzino 2006).]



The microprocessor then calculates the required inspiratory flow trajectory. In the example shown in Figure 9-42, which uses a half-sinewave flow waveform, the therapist has selected a tidal volume,  $V_t$  (L), and a respiration rate of  $n$  (breaths/min).

The total desired respiratory flow,  $Q_d(t)$ , for a single breath can be determined from the peak flow rate,  $Q_i$  (L/s).

$$Q_d(t) = \begin{cases} Q_i \sin \frac{\pi t}{t_i}, & 0 \leq t \leq t_i \\ 0, & t_i \leq t \leq t_e \end{cases} \quad (9.13)$$

where  $t_i$  is the duration of inspiration, calculated from the tidal volume and the peak flow rate.

$$\begin{aligned} V_t &= \int_0^{t_i} Q_i \sin \frac{\pi t}{t_i} dt \\ &= -\frac{Q_i t_i}{\pi} \cos \frac{\pi t}{t_i} \Big|_0^{t_i} \\ &= \frac{2Q_i t_i}{\pi} \end{aligned}$$

Therefore, the duration of inspiration is

$$t_i = \frac{\pi V_t}{2Q_i} \quad (9.14)$$

The duration of expiration in seconds is determined from the breathing rate and the inspiration time

$$t_e = \frac{60}{n} - t_i \quad (9.15)$$

### WORKED EXAMPLE

A ventilator has been programmed for the following:

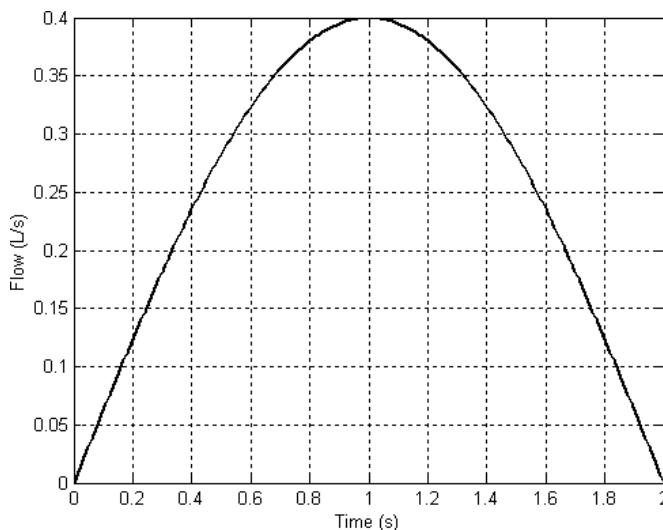
- Respiration rate: 12 breaths/min
- Flow waveform: half sine
- Tidal volume: 0.5 L
- Peak flow: 0.4 L/s

The duration of inspiration is

$$\begin{aligned} t_i &= \frac{\pi V_t}{2Q_i} \\ &= \frac{\pi \times 0.5}{2 \times 0.4} \\ &= 1.96 \text{ s} \end{aligned}$$

The duration of expiration is

$$\begin{aligned} t_e &= \frac{60}{n} - t_i \\ &= \frac{60}{12} - 1.96 \\ &= 3.03 \text{ s} \end{aligned}$$



The ratio of the inspiratory-to-expiratory period is often used to adjust the respiration rate. This is represented by the I:E ratio and is usually expressed in terms normalized to the inspiratory period. This makes the ratio  $1:R$ , where  $R = t_e/t_i$ .

A number of control strategies are used to control the oxygen and air delivery valves. One common type is the proportional plus integral (PI) controller, as discussed in Chapter 4, which can quickly adjust the amount of oxygen in the enriched breath gas, usually within a single breath (Bronzino, 2006).

It is often desirable to maintain a slightly positive pressure in a patient's lungs rather than allowing them to deflate completely during expiration. In this case, the controller closes the expiration valve when the airway pressure reaches PEEP. In a microprocessor-based system, the pressure is measured using a transducer and is monitored while the expiration valve is open. The response time must be fast because no new air is being provided, so any pressure overshoot will remain until the following inspiration phase.

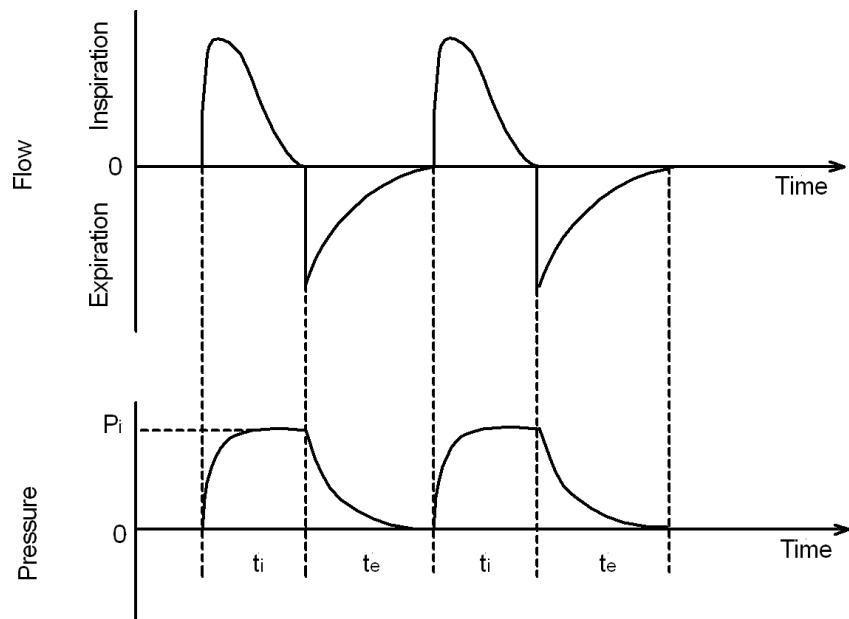
### 9.8.6 Pressure-Controlled Mandatory Ventilation

In this mode, the respirator raises and maintains the airway pressure at the desired level independent of patient compliance and resistance. The pressure level,  $P_i$ , shown in Figure 9-43, is set by the therapist. It should be noted that even though the pressure remains constant the actual flow rate will be different for each patient.

This controller uses the following therapist-selected parameters to compute the desired inspiratory pressure trajectory:

- Inspiratory pressure
- Respiration rate
- I:E ratio
- Delivered oxygen concentration

**FIGURE 9-43** ■  
Inspiratory flow for pressure-controlled ventilation. [Adapted from (Bronzino 2006).]



- Peak flow
- PEEP

The trajectory then serves as an input to the controller, which, as in the previous case, can be implemented using a PI algorithm (Bronzino, 2006).

### 9.8.7 Spontaneous Ventilation

In spontaneous ventilation, the ventilator supplies air on demand only during the inspiration cycle. In pressure support ventilation, the controller tries to maintain an airway pressure higher than PEEP. The common continuous positive airway pressure (CPAP) is a subset of this mode in which pressure support is maintained at the PEEP level.

### 9.8.8 Continuous Positive Airway Pressure

In this mode the ventilator maintains a positive pressure at the airway as the patient attempts to inspire. Figure 9-44 illustrates a typical pressure waveform during CPAP pressure delivery. The therapist sets the inspiration sensitivity level lower than PEEP, so that when the patient attempts to inhale the pressure drops below the threshold sensitivity and the ventilator responds by supplying the gas mixture to raise the pressure back to the PEEP level. Typically, the PEEP level and the sensitivity are selected so that the patient is required to exert some effort to breathe independently. A common level is  $-2 \text{ cm H}_2\text{O}$ . If the level is too high, weak patients may be unable to trigger a breath, and if it is too low the machine may auto-cycle, causing overventilation.

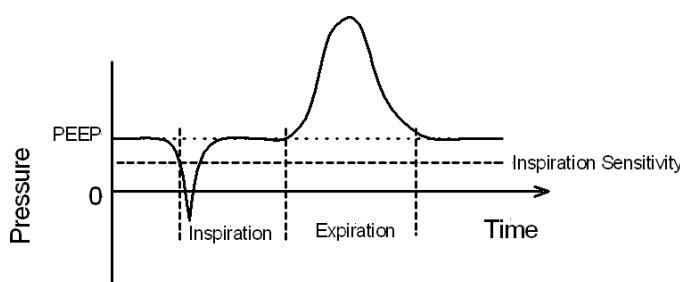
As with the mandatory modes, when the patient exhales, the ventilator shuts off the flow of gas, and the expiration valve is opened to vent exhaled gas into the atmosphere.

Figure 9-41 shows a small pipe that leads from the patient airway to a pressure transducer within the ventilator. The desired airway pressure is calculated by the system, from the programmed inputs, which include PEEP pressure, support level, and sensitivity.

The multiple-loop controller within the ventilator is then used to deliver a CPAP breath by comparing the measured airway pressure with the desired airway pressure and then calculating the total inspiratory flow level required to raise the airway pressure to that level. This calculated flow trajectory may differ from breath to breath, and therefore a reasonably fast control system must be implemented. As with the previous cases a conventional PI controller can be used.

### 9.8.9 Portable Ventilators

Advantages of using mechanical ventilation at home rather than in a hospital include decreased exposure to infections, increased mobility, improved nutrition, resumption of



**FIGURE 9-44** ■  
Airway pressure during CPAP spontaneous breath delivery. [Adapted from (Bronzino 2006).]

more normal interactions and routines of daily living, and lower health-care costs. Most portable ventilators are simple to operate and normally do not have the sophisticated controls necessary for patients in intensive care units.

All portable ventilator systems include the appropriate controls for setting operating modes and alarms; some systems also include special breathing circuits, oxygen accumulators, and heated humidifiers or heat-moisture exchangers (HMEs). Power can be supplied through mains, an external battery, or an internal battery.

Portable ventilators deliver room air or oxygen-enriched gas to the inspiratory limb of the breathing circuit, where it can be humidified by a heated humidifier or an HME before delivery to the patient. Typically, these ventilators drive air into the breathing circuit with a motor-driven pump. For oxygen enrichment, most portable ventilators use an accumulator, which collects oxygen and mixes it with air before it is drawn into the ventilator; alternatively, oxygen can be delivered directly into the breathing circuit from a separate source, such as an oxygen tank.

Nasal ventilation delivered through a face mask, as shown in Figure 9-45, is sometimes used, usually for patients who require ventilation only at night. The expiration valve, which occludes the expiration limb during inhalation to divert gas into the patient's lungs and opens during expiration to release the exhaled breath, is located close to the patient connection. This eliminates the need for an expiration hose and simplifies the breathing circuit.

Portable ventilators offer several ventilation modes as discussed in the previous section. They also use various methods of cycling from inspiration into expiration. In volume-cycled ventilation, the ventilator cycles into expiration when a preset tidal volume (the volume of a single breath) has been delivered. In time-cycled ventilation, the ventilator cycles into expiration after a preset time. A variant of volume- or time-cycled ventilation is volume- or time-cycled pressure-limited ventilation. A mechanical pressure-relief valve that vents and hence limits excess pressure is used on the inspiratory side. When the inspiratory pressure reaches the set level, excess gas is vented, and the inspiratory pressure remains at the set level until the end of inspiration.

**FIGURE 9-45 ■**  
Nasal ventilator mask. (Courtesy of Vitality Medical <http://www.vitalitymedical.com/>.)



Portable ventilators monitor airway pressure and have adjustable low- and high-pressure alarm limits. Airway pressures are measured at the patient connection of the breathing circuit; measuring at the patient connection produces measurements more reliable than those obtained at other points in the breathing circuit. The high-pressure alarm can warn of peak inspiratory pressure increases caused by decreases in patient compliance, breathing circuit occlusion, or increases in airway resistance (e.g., resulting from a buildup of secretions in the airway). The low-pressure or low minute-volume alarm can warn of a breathing-circuit disconnection, a leak, or a failure to deliver a breath. Delivery of an inappropriately high tidal volume or occlusion of the breathing circuit's expiratory limb can result in overdistention of alveoli, which can damage them.

Switchover to battery operation (either internal or external) is usually automatic and is signaled by an alarm. A large-capacity external battery (e.g., 12 VDC) is used to supply power during portable operation or extended alternating current (AC) power loss. An internal battery allows temporary backup (usually up to 1 hour) in the event of external battery depletion and, if the external battery is disconnected, can be used briefly as a power source while the ventilator user moves to another area, such as the toilet.

### 9.8.10 Sleep Apnea

A common application for CPAP ventilators is to provide relief for severe cases of sleep apnea. This effective treatment provides pressure to the person's airway using a portable ventilator. The airflow is delivered through a mask that fits on the face and covers the nose or the nose and mouth, as shown in Figure 9-46. The positive air pressure acts as a "splint"



**FIGURE 9-46** ■  
Photograph of a face mask and CPAP ventilator.  
(Courtesy of Philips Respironics  
<http://www.healthcare.philips.com/main/homehealth/respironics.wpd.>)

to keep the airway open during sleep, allowing breathing to become more regular. Snoring stops, restful sleep is restored, and risk factors associated with untreated sleep apnea are greatly reduced.

## 9.9 REFERENCES

- Agostoni, E. (1972). "Mechanics of the Pleural Space." *Physiological review* 52: 57–128.
- Anon. (1937). "Medicine: Polio and Lungs." *Time Magazine*, September 13.
- Bell, A. (1892). "The Alexander Graham Bell Family Papers." *Library of Congress*.
- Bellis, M. (2008). "History of the Iron Lung—Respirator." Retrieved October, 2008, from [http://inventors.about.com/od/istartinventions/a/iron\\_lung.htm](http://inventors.about.com/od/istartinventions/a/iron_lung.htm).
- Bronzino, J. (Ed.). (2006). *Medical Devices and Systems*. Boca Raton, FL: CRC Press.
- Byrd, R., S. Kosseifi, et al. (2010). "Mechanical Ventilation." *eMedicine*. Retrieved from <http://emedicine.medscape.com/article/304068-print>
- Darwin, M. (2009). "CPR and the Breath of Death." Retrieved June 2010 from <http://www.depressedmetabolism.com/2009/07/06/cpr-and-the-breath-of-death/>
- Drinker, P. and C. McKhann. (1929). "The Use of a New Apparatus for the Prolonged Administration of Artificial Respiration: I. A Fatal Case of Poliomyelitis." *JAMA* 92(20): 1658–1660.
- Drinker, P. and C. McKhann. (1986). "The Iron Lung: First Practical Means of Respiratory Support." *JAMA* 255(11): 1478.
- Drinker, P. and E. Roy. (1938). "The Construction of an Emergency Respirator for use in Treating Respiratory Failure in Infantile Paralysis." *Journal of Pediatrics* 13(1): 71–74.
- Emerson, J. and J. Loynes. (1978). *The Evolution of Iron Lungs: Respirators of the Body-Encasing Type*. Cambridge, MA: JH Emerson Co.
- Fenn, W. and H. Rahn (Eds.). (1965). *Handbook of Physiology, Section 3: Respiration*. Washington, DC: American Physiological Society.
- Gorham, J. (1971). "A Medical Triumph: The Iron Lung." *Respiratory Therapy* 9(1): 71–73.
- Green, C. (1889). "Reports of Societies." *Boston Medical and Surgical Journal* 120(1): 9.
- Hicks, M. (2003). "The 'Iron Lung' in Australia." *The HaMMer—Health and Medicine Museums Newsletter*. 24(1)
- Hill, R. (1995). "A Being Breathing Thoughtful Breath: The History of the British Iron Lung, 1832—1995." Retrieved October 2008 from <http://richardhill.co.uk/ironlung/>
- Hill, R. (1996). "Each Saving Breath: The Story of Maggie and John Prestwich." Retrieved June 2010 from <http://www.johnprestwich.btinternet.co.uk/each-saving-breath.htm>
- Hillberg, R. and D. Johnson. (1997). "Current Concepts: Noninvasive Ventilation." *New England Journal of Medicine* 337(24): 1746–1752.
- Keith, A. (1906). "Three Huntarian Lectures on the Mechanism Underlying the Various Methods of Artificial Respiration." *Lancet*, March 13, 746–749, 825–828.
- Marburg. (1541). *Mundinus Anatomica*. London: Wellcome Institute Library.
- Murray, J. and J. Nadel. (Eds.). (2008). *Textbook of Respiratory Medicine*, 3d ed. Philadelphia: W. B. Saunders.
- Nelson, R. (2004). "On Borrowed Time." *AARP Bulletin*, September, p. 20.
- Paul, J. (1971). *A History of Poliomyelitis*. New Haven, CT: Yale University Press.
- Randall, R. (2005). *An Introduction to Acoustics*. Mineola, NY: Dover Publications.
- Roussos, C. and P. Macklem. (1982). "The Respiratory Muscles." *New England Journal of Medicine* 307: 786–797.
- Shapiro, S., P. Ernst, et al. (1992). "Effect of Negative-Pressure Ventilation in Severe Chronic Obstructive Pulmonary Disease." *Lancet* 1992(340): 1425–1429.
- Shaw, L. (1928). "Cutaneous Respiration of the Cat." *American Journal of Physiology* 85: 158–167.

- Meacham, S. (2004). Memories of Polio survivors and those who were disabled by It. Sydney Morning Herald. Sydney. December 07.
- Smith, I., M. King, et al. (1995). "Choosing a Negative Pressure Ventilation Pump: Are There Any Important Differences." *European Respiratory Journal* 1995(8): 1792–1795.
- Uleryk, E. (2010). Respirator Information—Toronto Hospital for Sick Children, Personal communication.
- Vesalius, A. (1543). "De Humanis Corporis Fabrica." Retrieved June, 2009, from <http://archive.nlm.nih.gov/proj/ttp/flash/vesalius/vesalius.html>.
- Widdicombe, J. (2008). "The Lung." Retrieved October 2008 from <http://www.answers.com/topic/lung>
- Wilson, L. (1960). "The Transformation of Ancient Concepts of Respiration in the Seventeenth Century." *Isis* 51(2).



# Active and Passive Prosthetic Limbs

## Chapter Outline

10.1	Introduction .....	524
10.2	Structure of the Arm.....	529
10.3	Kinematic Model of the Arm.....	531
10.4	Structure of the Leg .....	532
10.5	Kinematic Model of the Leg .....	534
10.6	Kinematics of Limb Movement .....	536
10.7	Sensing .....	538
10.8	Passive Prosthetics.....	538
10.9	Active Prosthetics .....	547
10.10	Prosthesis Suspension .....	582
10.11	References .....	584



## 10.1 INTRODUCTION

The ability to interact with the environment is one of the main characteristics of any living organism. As human beings our interaction surpasses that of any other animals, as we have developed a pair of extremely dexterous manipulators—our hands. These, coupled with an ability to reason and to plan, give us the ability to modify our environment to suit our requirements. Inevitably, without our hands and arms, and to a lesser extent our legs and feet, this ability is curtailed. History is replete with examples of ingenious mechanical replacements for lost limbs that are able, to a limited extent, to accommodate lost capability. However, to date, no prosthesis comes anywhere near to reproducing the dexterity of the human hand, and though specialist devices are capable of replacing legs and feet for certain activities (like running fast) none provides an all-around capability to match the original.

### 10.1.1 A Brief History of Prosthetics

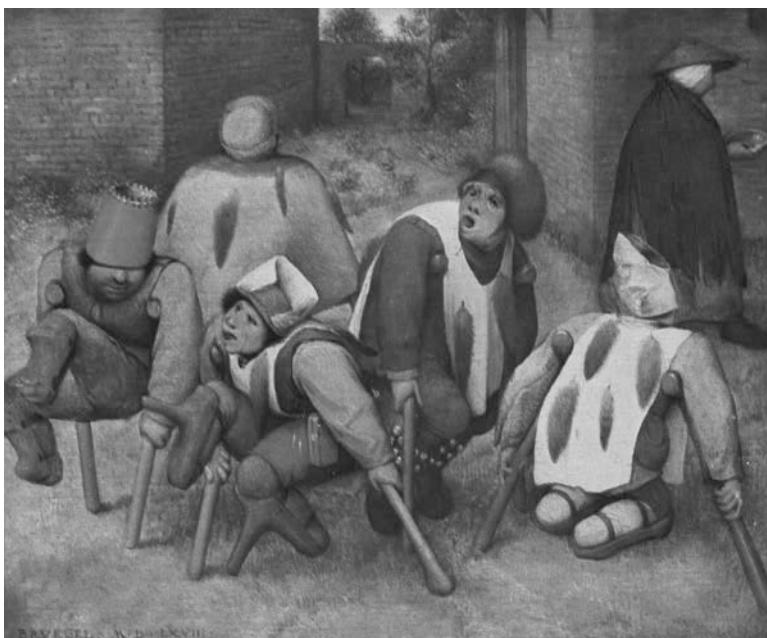
It is almost certain that had one of our cave-dwelling ancestors survived a crushed or broken leg, some form of crutch would have been fashioned to aid with locomotion. In more recent times, more sophisticated prosthetics with both cosmetic and utilitarian functions have been developed. An early example of a prosthesis unearthed by archeologists recently is a big toe made of wood and leather attached to the almost 3000-year-old mummified remains of an Egyptian noblewoman, shown in Figure 10-1.

It is believed that most Egyptian prostheses were produced for a sense of wholeness, but analysis showed that the big toe shown in Figure 10-1 was functional and would have aided with the balance and locomotion of the wearer.

In 424 BC, Herodotus wrote of a Persian seer who had been condemned to death but escaped by amputating his own foot and making a wooden replacement to walk about 50 km to the next town. In Capua, Italy, archeologists unearthed an artificial leg dating from about 300 BC that was made of bronze and iron, with a wooden core, apparently for a below-the-knee amputee (Norton 2007).

**FIGURE 10-1 ■**  
Prosthetic toe made from wood and leather attached to a mummy between 950 and 710 BC.  
(Courtesy of the Museum of Egyptian Antiquities in Cairo.)





**FIGURE 10-2 ■**  
“The Beggars,”  
painted by Peter  
Bruegel in 1568.  
(Courtesy of the  
Louvre Museum.)

Wars have always played a big role in depriving people of their limbs, and particularly since the invention of the sword and the battle-axe the requirement for replacement arms and hands has been well documented. An example described by Pliny the Elder is that of the Roman general Marcus Sergius, who lost his right hand during the Second Punic War (218 to 210 BC). He recovered from the injury and had an iron replacement fashioned to hold his shield so that he was able to return to war (Healy, 1991; Norton, 2007).

Very little changed for 1500 years, and simple wooden peg legs and hooks remained the best to offer, as seen in Peter Bruegel’s 1568 painting shown in Figure 10-2.

Most prostheses of the time were cosmetic and were made to hide deformities or injuries sustained in battle. A knight would be fitted with a prosthesis that was designed to hold a shield or for a leg to appear in the stirrup, with little attention to general functionality. These were often made by the same blacksmith who made the rest of the armor.

The earliest surviving hands of this kind are possibly those made for the well-known Franconian knight Götz von Berlich (circa 1480–1562). Two are preserved in Schloss Jagsthausen and a third in Schloss Grüningen bei Riedlingen. They were made after June 22, 1504, when the 24-year-old Götz lost his right hand to a cannonball at the siege of Landshuf, but probably before 1512. In the case of the hand shown in Figure 10-3, both the thumb and the paired fingers are capable of being independently locked in several positions by means of an elaborate system of ratchets and spring-operated pawls. The push button that can be seen protruding from the back of the hand unlocked the mechanism. According to a verse composed by Count Franz Poccii in 1861, this “iron hand” allowed Götz to securely grip both his lance and his sword and stood him in good stead during many years of soldiering.

When not in battle, knights and nobles replaced these metal hands with cosmetic wooden hands similar to the one shown in Figure 10-4.

As technology improved, artisans began to include more functionality in the artificial limbs they produced, and around 1512 an Italian surgeon traveling in Asia recorded

**FIGURE 10-3 ■**  
Prosthetic hand of  
the type made  
famous by the knight  
Götz von Berlich.  
(Courtesy of Schloss  
Jagsthausen.)



**FIGURE 10-4 ■**  
Modern copy of a  
cosmetic wooden  
hand with limited  
movement.  
(Courtesy of  
Mantiques Modern,  
with permission.)



observations of a bilateral upper-arm amputee who was able to remove his hat, open his purse, and even sign his name.

French army barber and surgeon Ambroise Paré is considered to be the father of modern amputation surgery and prosthetic design. In 1529 he introduced modern amputation procedures to the medical community and a few years later started making prostheses for upper- and lower-extremity amputees. He also invented a kneeling peg leg and foot prosthesis with a fixed position, adjustable harness, knee lock control, and other engineering features that are still in use today. His work showed the first true understanding of how a prosthesis should function, as illustrated in the extract from his collected works shown in Figure 10-5 (Norton, 2007).



**FIGURE 10-5 ■**  
 Illustration showing  
 the mechanism for  
 an artificial hand  
 from Ambroise  
 Paré's *Les Oeuvres*  
 (Collected Works)  
 published in 1575.  
 (Courtesy of National  
 Library of Medicine.)

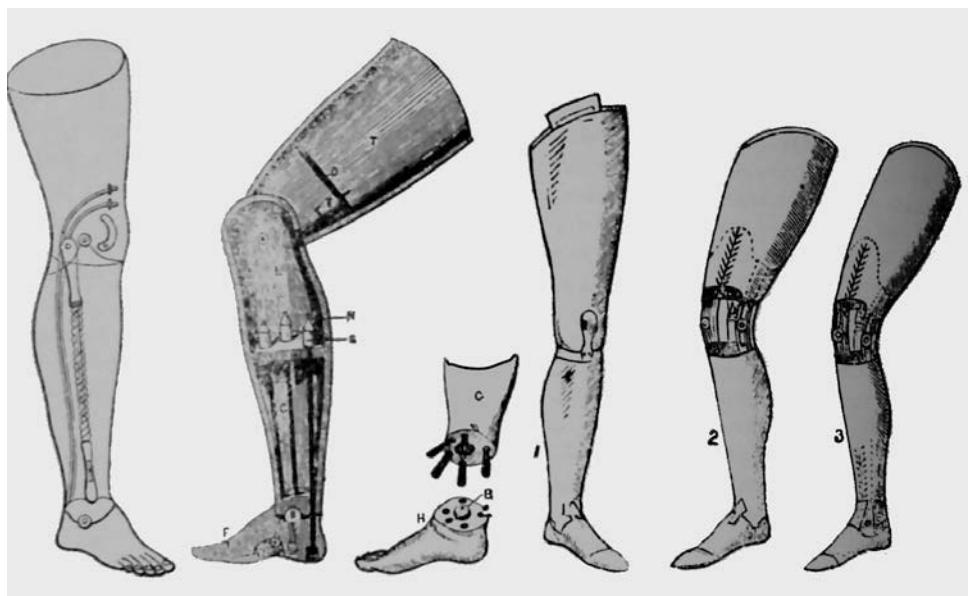
More than 100 years later, in 1690 Dutch surgeon Pieter Verduyn developed the first nonlocking below-the-knee prosthesis and a leather cuff for improved attachment to the body. Many of the features of this design, like Paré's, are still common in modern prostheses.

Such refinements were not available to the common man or to pirates who sometimes relied on hooks and peg legs, as these could be made from materials available onboard ship. However, a trained doctor to perform the amputation or to clean up the stump would have been rare, and instead the ship's cook typically performed any required surgery, generally with a poor success rate.

In 1800 James Potts designed the prosthesis shown in Figure 10-6. It consisted of a wooden shank and socket, a steel knee joint, and an articulated foot that was controlled by catgut tendons from the knee to the ankle. It became known as the Anglesey Leg, after the Marquis of Anglesey, who had it developed after losing his leg in the Battle of Waterloo. A full 40 years later, the leg was taken to the United States by William Selpho where it became known as the "Selpho Leg." This design was improved a few years later, in 1846, when Benjamin Palmer added an anterior spring, smooth appearance, and concealed tendons to simulate natural-looking movement. The design remained in common use until WWI.

With improvements in hygiene, the numbers of survivors of amputations increased, and advances were made in prosthetic design, such as joint technology and suction-based

**FIGURE 10-6 ■**  
 Bigg and Grossmith limbs on display at the Great International Exhibition at St. Thomas's Hospital in 1862. (Courtesy of Best Foot Forward, with permission.)



attachment methods. One good example that occurred in 1812 was the development of a prosthetic arm that could be controlled by the opposite shoulder with connecting straps. By the 1840s, with the introduction of gaseous anesthetic, doctors could perform more meticulous amputation surgeries, allowing them to operate on the limb stump in such a way as to prepare it for interfacing to a prosthesis.

During the American Civil War (1861 to 1865), the number of amputations rose astronomically, forcing the United States to enter the field of prosthetics. James Hanger, one of the first amputees of the civil war, developed the “Hanger Limb” carved from barrel staves. In his prosthesis, the catgut tendons commonly used at that time were replaced with rubber bumpers to control plantar and dorsiflection. He and other inventors such as Selpho, Palmer, and Marks helped advance the prosthetics field with other refinements in the mechanisms and materials of the devices.

A good example of the range of artificial legs available at this time can be seen in Figure 10-6, which shows the merchandise produced by Bigg and Grossmith on display at the Great International Exhibition of 1862 (Phillips, 1990).

Prosthetic legs of this type were very heavy as they were made primarily from wood, steel, and leather, so in 1868 Gustav Hermann suggested the use of aluminium instead of steel to make artificial limbs lighter and more functional. However, the lighter device did not become a reality until 1912 when Marcel Desoutter, a famous English aviator, lost his leg in an airplane accident and made the first aluminium prosthesis with the help of his brother Charles, an engineer.

It is interesting to note that WWI did not engender much technical advancement in the prosthetics field for some reason. However, the surgeon general of the army at the time did at least realize the importance of technology for development of prostheses, and this eventually led to the formation of the American Orthotic & Prosthetic Association (AOPA).

Following WWII, veterans were dissatisfied with the lack of technology in their devices, which had hardly changed in a century, and they demanded improvement. This

forced the U.S. government to react, and it entered into a deal with some military companies to improve prosthetic function. This agreement paved the way to the development and production of modern prostheses, and aluminium and plastic composites and other new materials led to the development of lighter and stronger devices. However it was still some time before the basic articulated hook was replaced by something more advanced.

The problem with prosthetics research has always been the small size of the market. Notwithstanding increased use of improvised explosive devices (IEDs) and antipersonnel mines in war zones, and of course the proliferation of higher-speed motor cars in both the First and Third World, there are still insufficient amputees for industry to recover investment in expensive research and development (R&D).

In the conflicts in Afghanistan and Iraq up until February 2009, 862 U.S. troops became amputees, of whom 186 had lost arms. In the entire United States, there are fewer than 100,000 arm amputees. Fortunately, public opinion has influenced the U.S. government, and in 2007 the Defence Advanced Research Projects Agency (DARPA) began the Revolutionising Prosthetics program, the cost of which is now close to \$100 million.

The first phase of the project headed by DEKA Research and Development Corp was a 2-year program to make an advanced prosthetic arm using the world's best existing technology. The 2009 program is driven by Johns Hopkins University's Applied Physics Laboratory, and its mandate is to create a prosthesis that would be "mind controlled" and restore the patient's ability to feel heat, cold, pressure, and even surface texture (Kuniholm, 2009).

Though these projects have been moderately successful, as reported later in this chapter, their outputs are still experimental, and at the time of writing the amputee public at large has yet to benefit significantly from the research.

## 10.2 | STRUCTURE OF THE ARM

---

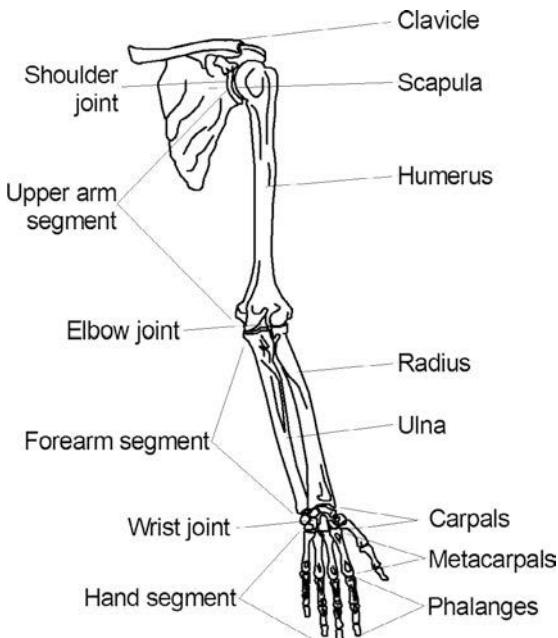
Arm mechanisms consist of up to three articulated joints and the structures between them: the wrist, elbow, and shoulder separated by the forearm and the upper arm segments. The hand and trunk are each considered to be single segments, as shown in Figure 10-7.

### 10.2.1 Wrist

The wrist is one of the most complex joints in the body, combining high mobility with the ability to carry heavy loads. It provides two degrees of freedom (DoFs): flexion–extension and abduction–adduction. Flexion of the wrist allows the palm to rotate toward the forearm up to a maximum of about  $90^\circ$ , whereas extension is rotation in the opposite direction up to about  $80^\circ$ . Abduction is wrist rotation toward the radius and is limited to about  $15^\circ$ , whereas adduction is rotation toward the ulna and can reach between  $20^\circ$  and  $30^\circ$  in a normal human arm.

A prosthetic wrist unit provides a means of orientating the terminal device in space. This can be positioned manually or by cable operation and, once positioned, is held in place by friction or with a mechanical lock. Wrist flexion is also sometimes provided as it improves an amputee's performance of midline activities such as shaving or manipulating buttons.

**FIGURE 10-7 ■**  
Anatomy of the arm showing joints and segments.



### 10.2.2 Elbow

The elbow joint can also be considered to provide two degrees of freedom. These are flexion–extension and pronation–supination. The range of the former is between  $0^\circ$  (full extension) and  $145^\circ$  (full active flexion). The latter is a rotation about the long axis of the forearm from about  $80^\circ$  down (maximum pronation) to about  $85^\circ$  up (maximum supination).

Elbow units are chosen based on the level of amputation. For a transhumeral prosthesis, an internal locking elbow joint is the most common. They provide about  $135^\circ$  of flexion and can be locked into a number of different positions. Elbow spring lifts are often used to counterbalance the weight of the forearm. Between 80 and 100 mm of forearm is required to fit an internal locking elbow, and if insufficient length is available then forearm extensions are required to ensure that the prosthesis can reach the body midline or mouth.

### 10.2.3 Shoulder

The shoulder and the shoulder girdle is another extremely complex joint group in the human body. The range of motion afforded by a combination of the ball-and-socket joint of the humerus with the scapula and rotation and displacement of the shoulder girdle covers 65% of a sphere. It is therefore difficult to restore function if amputation at the shoulder or forequarter level is required. This is due to a combination of the weight of the prosthesis and the diminished overall function that results when combining multiple prosthetic joints. This issue in conjunction with the increased energy expenditure results in many amputees using only a lightweight prosthesis for cosmetic purposes.

## 10.3 KINEMATIC MODEL OF THE ARM

Human limbs can be modeled as a chain of rigid segments, where each segment corresponds to one of the body segments with joints in between, as discussed earlier in this chapter. For the real or prosthetic limb to control the terminal device, the relative angles of all of the segments in between it and a reference point in the torso must be known. Kinematic models generally make some simplifications regarding the number of degrees of freedom by ignoring linear displacements and bone flexibility.

A detailed model of the arm can include seven DoFs if motion of the shoulder blade is included. These are three degrees of freedom at the shoulder, followed by 2 degrees of freedom at each the elbow and the wrist as illustrated in Figure 10-8.

In the Denavit–Hartenberg (D–H) model, the base coordinate system ( $X_0, Y_0, Z_0$ ) is located in the body halfway between the shoulders, then three frames corresponding to the three DoFs are located at the center of the shoulder joint:

- Circumduction ( $X_1, Y_1, Z_1$ )
- Adductio–abduction ( $X_2, Y_2, Z_2$ )
- Flexion–extension ( $X_3, Y_3, Z_3$ )

The elbow joint consists of a pair of frames corresponding to the following:

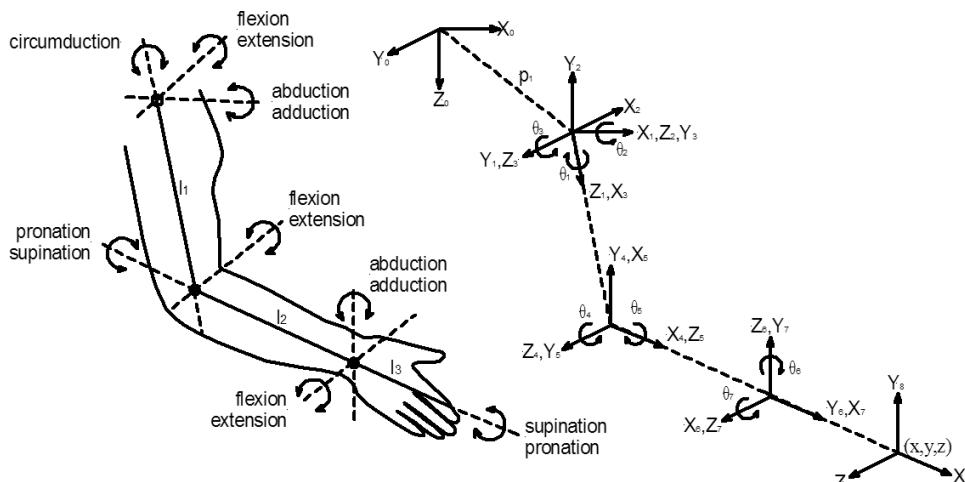
- Flexion–extension ( $X_4, Y_4, Z_4$ )
- Supinatio–pronation ( $X_5, Y_5, Z_5$ )

The movement in the wrist is described by two more frames:

- Adduction–abduction ( $X_6, Y_6, Z_6$ )
- Flexion–extension ( $X_7, Y_7, Z_7$ )

Finally, the end effector frame is ( $X_8, Y_8, Z_8$ ).

The rotation of a frame with respect to the previous one in the kinematic chain corresponds to a physiological DoF. Multiple DoF joints, where several frames have the same origin, can be considered to be part of a kinematic chain in which some of the links have zero length.



**FIGURE 10-8 ■**  
Denavit–Hartenberg  
model of the arm.

**TABLE 10-1** ■ Denavit–Hartenberg Parameters for the Human Arm

Joint	$\beta_i$	No	$\alpha_i$	$a_i$	$d_i$	$\theta_i$
Base	0°	0–1	0	$a_0$	$d_0$	0
Shoulder	−90° medial rotation to +90° lateral rotation	1–2	−90°	0	0	$\beta_1 + 90^\circ$
Shoulder	−180° abduction to +50° adduction	2–3	+90°	0	0	$\beta_2 + 90^\circ$
Shoulder	−180° flexion to +80° extension	3–4	0	$l_1$	0	$\beta_3 + 90^\circ$
Elbow	−10° flexion to +145° extension	4–5	+90°	0	0	$\beta_4 + 90^\circ$
Elbow	−90° pronation to +90° supination	5–6	+90°	0	$l_2$	$\beta_5 + 90^\circ$
Wrist	−90° flexion to +90° extension	6–7	+90°	0	0	$\beta_6 + 90^\circ$
Wrist	−15° abduction to +40° adduction	7–8	0	$l_3$	0	$\beta_7$

Source: Pons, J. (Ed.), *Wearable Robots—Biomechatronic Exoskeletons*, Chichester, UK: John Wiley & Sons, 2008.

The D–H parameters for the human arm are defined in Table 10-1 where the angle  $\theta_i$  around axis  $Z_i$  corresponds to the variable around the  $i$ -th DoF of the model. The range of motion depends on the physiological limits of that joint,  $\beta_i$  (deg), as listed Table 10-1. The parameters  $a_i$  and  $d_i$  are the body segment lengths that remain constant for each human being, but they vary from individual to individual (generally as a function of their height and sex).

Using the D–H convention, the general form of the transformation matrix between two consecutive coordinate systems is described by

$$T_{i-1}^i = \begin{bmatrix} \cos \theta_i & -\cos \alpha_i \sin \theta_i & \sin \alpha_i \sin \theta_i & a_i \cos \theta_i \\ \sin \theta_i & \cos \alpha_i \cos \theta_i & -\sin \alpha_i \cos \theta_i & a_i \sin \theta_i \\ 0 & \sin \alpha_i & \cos \alpha_i & d_i \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (10.1)$$

The position and orientation of the terminal device can be determined by developing a single transformation matrix as a combination of the transformations of each successive joint.

$$T_0^8 = T_0^1 T_1^2 \dots T_7^8 \quad (10.2)$$

In the case of most prosthetic arms, only a subset of the complete transformation would be applied depending on the position of the amputation.

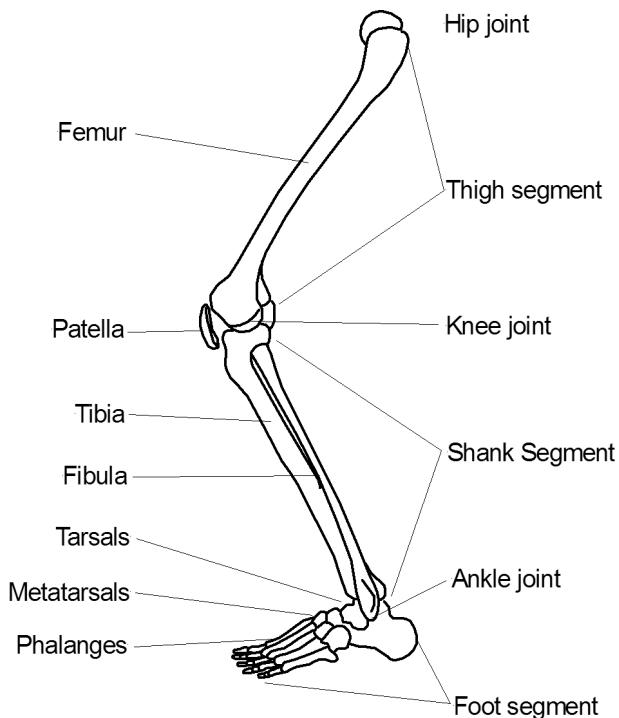
From a dynamic perspective, the movement of each joint can be related to the external forces transmitted to the joint, the inertia of the body in motion, and the torque at each joint actuator.

## 10.4 | STRUCTURE OF THE LEG

As shown in Figure 10-9, the leg consists of three segments: (1) the thigh; (2) shank and foot; and (3) three joints—hip, knee, and ankle.

### 10.4.1 The Hip Joint

The hip joint consists of the cup-shaped acetabulum on the pelvis and the spherical head of the femur. It functions as a spherical joint allowing three degrees of freedom: flexion–extension, abduction–adduction, and medial–lateral rotation. The first allows rotation in



**FIGURE 10-9 ■**  
Anatomy of the leg  
showing joints and  
segments.

the lateral plane with flexion being the motion that brings the thigh forward and upward, up to about  $120^\circ$ , and extension being the motion that moves it backward. In abduction, the leg is moved away from the midline of the body up to about  $40^\circ$ , whereas in adduction the limit is between  $30^\circ$  and  $35^\circ$ . Medial and lateral rotations are those around the long axis of the femur, with the limit of medial rotation being between  $15^\circ$  and  $30^\circ$  and that of lateral rotation being up to  $60^\circ$  (Pons, 2008).

### 10.4.2 The Knee Joint

The knee joint consists of two parts: the femero-patellar joint; and the femero-tibial joint. It is well restrained by powerful muscles and ligaments to limit motions that may damage it and by others that control walking and standing. The knee is capable of flexion–extension as well as lateral and medial rotation. In flexion, the shank approaches the thigh, with extension being the reverse, while femur and the tibia remain in the same plane. The maximum flexion angle is  $120^\circ$  normally, although in extension when the leg is straightened for load bearing the maximum extension ranges from  $0^\circ$  to  $10^\circ$ . During the final stages of extension, some medial rotation brings the knee to the locked position for maximum stability. Lateral rotation is needed during the early stages of flexion to unlock the knee.

### 10.4.3 The Ankle Joint and the Foot

The ankle and the foot together contain 26 bones connected by 33 joints. The ankle comprises two joints—the talocrural and the talocalcaneal—that from a biomechanical perspective are considered to be a single hinge-type joint. Dorsiflection brings the foot up toward the anterior surface of the leg with a maximum angle of  $20^\circ$ . Plantarflexion is the opposite and can exceed  $40^\circ$ .

## 10.5 | KINEMATIC MODEL OF THE LEG

Legs can be considered to have six degrees of freedom as illustrated in Figure 10-10: three at the hip joint; one at the knee; and two at the ankle. Using the D-H convention, the base frame is situated on the pelvis between the hips at  $X_0, Y_0, Z_0$ . The hip frames correspond to the following:

- Circumduction ( $X_1, Y_1, Z_1$ )
- Adduction–abduction ( $X_2, Y_2, Z_2$ )
- Flexion–extension ( $X_3, Y_3, Z_3$ )

The knee joint frame corresponds to the following:

- Flexion–extension ( $X_4, Y_4, Z_4$ )

The movement in the ankle is described by two more frames:

- Dorsiflexion–plantarflexion ( $X_5, Y_5, Z_5$ )
- Inversion–eversion ( $X_6, Y_6, Z_6$ )

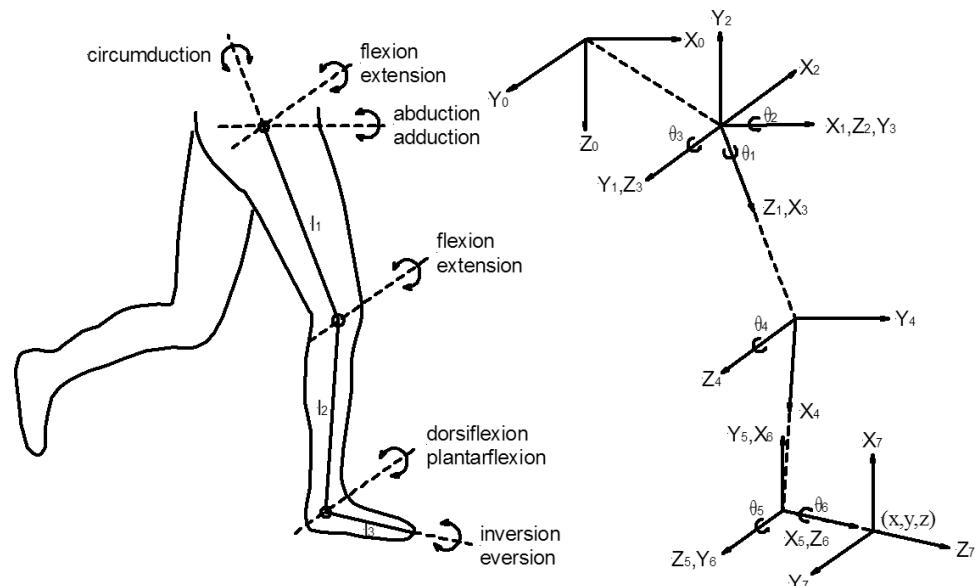
Finally, the end effector frame, at the tip of the big toe, is ( $X_7, Y_7, Z_7$ ).

As with the arm, the D-H parameters for the human leg are defined in Table 10-2, where the angle  $\theta_i$  around axis  $Z_i$  corresponds to the variable around the  $i$ -th DoF of the model and the range of motion depends on the physiological limits of that joint,  $\beta_i$  (deg).

### 10.5.1 Walking

Bipedal walking is an extremely complex movement involving the unstable process of preventing a fall by repeatedly placing one foot in front of the other. Human beings have evolved over millions of years to minimize their energy expenditure while performing this function. These gait efficiency determinants include three pelvic movements (rotation,

**FIGURE 10-10** ■  
Denavit–Hartenberg  
model of the leg.



**TABLE 10-2** ■ Denavit–Hartenberg Parameters for the Human Leg

Joint	$\beta_i$	No	$\alpha_i$	$a_i$	$d_i$	$\theta_i$
Base	0°	0–1	0	$a_0$	$d_0$	0
Hip	−50° medial rotation to +40° lateral rotation	1–2	−90°	0	0	$\beta_1 + 90^\circ$
Hip	−20° abduction to +45° adduction	2–3	+90°	0	0	$\beta_2 + 90^\circ$
Hip	−30° extension to +120° flexion	3–4	0	$l_1$	0	$\beta_3$
Knee	0° extension to +150° flexion	4–5	0	$l_2$	0	$\beta_4 + 90^\circ$
Ankle	−40° plantarflexion to +20° dorsiflexion	5–6	+90°	0		$\beta_5 + 90^\circ$
Ankle	−35° inversion to +20° eversion	6–7	0	0	$l_3$	$\beta_6$

Source: Pons, J. (Ed.), *Wearable Robots—Biomechatronic Exoskeletons*, Chichester, UK: John Wiley & Sons, 2008.

tilt, and lateral displacement), knee flexion, and knee–ankle interaction. Together, they minimize the amplitude of any vertical sinusoidal oscillations in the body center of mass. Additionally, the various leg muscles work only during certain phases of the locomotion cycle, thus reducing energy use. Other energy-saving adaptations include the storage and release of elastic energy in the passive elements of the leg (ligaments and tendons). Finally, it has also been shown that people select a specific walking speed (the product of the step rate, known as cadence, and stride length) related to the natural period of the leg pendulum to minimize energy consumption per unit distance traveled (Pons, 2008).

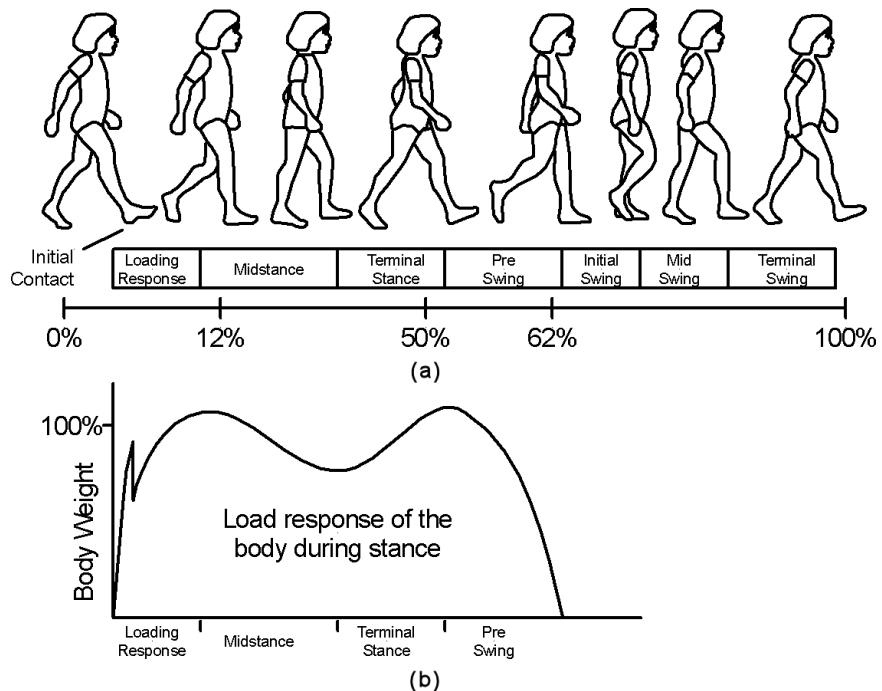
It is not possible for an amputee fitted with a transfemoral (above-the-knee) prosthesis to achieve the energy efficiency of an uninjured human being, with the best current prostheses using about 50% more energy during normal walking. This increased effort leads to a limitation of the activity or even an avoidance of some activities by amputees. In addition, deviations from normal motion during different activities produces alternative strains on the human body that often lead to other injuries (Burke, Roman et al., 1978). A great deal of research is being undertaken to improve energy efficiency while simultaneously maintaining the balance between the two halves of the body and providing comfort for long-term use. Some modern prosthetics incorporate intelligent structures using sensors on both the sound leg and the prosthesis to provide information on positioning and loading, and these are addressed later in this chapter. However, there is still a place for lower-cost, well-designed, lightweight passive prostheses.

Different classes of prostheses are used for transtibial (below-the-knee) amputations and transfemoral amputations. People with the former are usually able to regain fairly normal movement and, in some cases like South African athlete Oscar Pistorius, are able to outperform their uninjured colleagues. Transfemoral amputees require a prosthetic knee, lower leg, and foot and seldom regain normal movement because of the complex movement associated with the knee. However, that does not necessarily result in curtailed activity. Douglas Bader, for example, became a fighter pilot during WWII with more than 22 confirmed German aircraft shot down even though he was a bilateral amputee.

## 10.5.2 Normal Walking Dynamics

To understand the problems associated with the design of lower limb prostheses, it is important to understand normal walking dynamics as described by the gait cycle. Figure 10-11 shows that the gait cycle can be divided into two main periods: (1) stance, the period when the foot is in contact with the ground; and (2) swing, the time when the foot is in the air (Dehghani, 2010; Perry, 1992).

**FIGURE 10-11** ■ Illustration of the gait cycle. (a) Definition of gait cycle stages. (b) Normalised load response on the active leg. [Adapted from (Perry 1992; Dehghani 2010).]



**Initial swing:** The limb is advanced by hip flexion, with increased knee flexion and a slight lifting of the foot. The objective of this phase is to provide foot clearance to avoid tripping or stubbing. The gracilis and sartorius muscles control the three-dimensional (3-D) path of the limb and induce knee flexion as they act at the hip. The popliteus and the short head of the biceps provide direct knee flexion up to 40°, and the pretibiala muscles perform toe lift.

**Mid-swing:** Knee flexion reaches 60° at an angular velocity of about 350°/s. Hip activity is reduced with the advancement of the limb being a continuation of the action of the initial swing. The popliteus and the short head of the biceps continue to provide direct knee flexion, and at the ankle there is continuing dorsi-flexion.

**Terminal swing:** This is the final preparation in the transition from swing to stance. Muscle activity prepares the advancing limb for load acceptance. The hamstring muscles start to decelerate. Some muscles also produce a slight hyperextension of the knee, whereas others maintain a slight degree of flexion with the final knee angle being about 5° of flexion in preparation for initial contact.

**Initial contact and loading response:** At initial contact, the knee flexes from 5° to a maximum of 18° to absorb shock. The quadriceps functions eccentrically to restrain the degree of flexion to allow for stable weight bearing.

## 10.6 | KINEMATICS OF LIMB MOVEMENT

### 10.6.1 Center of Mass and Moment of Inertia of a Limb Segment

Each segment of a limb can be described by its total mass, the center of mass, and the moment of inertia around that point. The mass can be determined by dividing the segment

into  $N$  slices of mass  $m_i$  displaced from the end by a distance  $x_i$ . The total mass,  $M$  (kg), is the sum of the masses of the individual segments

$$M = \sum_{i=1}^N m_i \quad (10.3)$$

with the center of mass at a distance  $x$  (m) from the end of the segment.

$$x = \frac{1}{M} \sum_{i=1}^N m_i x_i \quad (10.4)$$

The moment of inertia,  $I$  (kg · m<sup>2</sup>), about the endpoint is

$$I = \sum_{i=1}^N m_i x_i^2 \quad (10.5)$$

The moment of inertia is generally specified around the center of mass of the segment (where it is at its lowest), and the parallel axis theorem is used to determine its value there:

$$I_0 = I - Mx^2 \quad (10.6)$$

where  $I_0$  is the moment of inertia around the center of mass.

Obviously this formula can also be used to determine the moment of inertia around one of the ends of the segment if it has been provided at the center of mass.

The radius of gyration,  $\rho_0$  (m), of the limb segment is defined as the distance from the center of mass for each of two point masses,  $M/2$  (kg), such that

$$I_0 = M\rho_0^2 \quad (10.7)$$

Note that both the total mass and the center of mass remain unchanged in this case.

### WORKED EXAMPLE

---

#### Moment of Inertia

A modern prosthetic leg has a mass of 2.5 kg, with its center of mass 200 mm from the knee joint. The radius of gyration is 141 mm. Calculate the moment of inertia about the knee joint.

The moment of inertia around the center of mass is determined from the given mass and the radius of gyration

$$\begin{aligned} I_0 &= M\rho_0^2 \\ &= 2.5 \times (141 \times 10^{-3})^2 \\ &= 0.0497 \text{ kg} \cdot \text{m}^2 \end{aligned}$$

The parallel axis theorem is then used to determine the moment of inertia around the knee joint:

$$\begin{aligned} I &= I_0 + Mx^2 \\ &= 0.0497 + 2.5 \times (200 \times 10^{-3})^2 \\ &= 0.1497 \text{ kg} \cdot \text{m}^2 \end{aligned}$$


---

### 10.6.2 Angular Acceleration

If a torque,  $\tau$  (Nm), is applied at one end of the segment, then the angular acceleration,  $\alpha$  (rad/s<sup>2</sup>), around that axis is

$$\alpha = \frac{\tau}{I} \quad (10.8)$$

### 10.6.3 Center of Mass and Moment of Inertia of a Complete Limb

Now consider the complete limb made up of three segments with masses  $M_1$ ,  $M_2$ , and  $M_3$ , respectively, and centers of mass at coordinates  $(x_1, y_1, z_1)$ ,  $(x_2, y_2, z_2)$ , and  $(x_3, y_3, z_3)$ . The mass for the complete system is

$$M_0 = M_1 + M_2 + M_3 \quad (10.9)$$

at coordinates

$$\begin{aligned} x_0 &= \frac{M_1 x_1 + M_2 x_2 + M_3 x_3}{M_0} \\ y_0 &= \frac{M_1 y_1 + M_2 y_2 + M_3 y_3}{M_0} \\ z_0 &= \frac{M_1 z_1 + M_2 z_2 + M_3 z_3}{M_0} \end{aligned} \quad (10.10)$$

The time history of the center of mass can be used to determine balance issues. However, if driving torques or limb angular accelerations are required, the moment of inertia must be determined and functions as the center of mass and the relative limb angle with respect to the point of application.

## 10.7 | SENSING

The two main sensing mechanisms relevant to locomotion are proprioception and mechanoreception. Proprioception is a sense of perception of movement and the positions of the limbs and the body gained primarily from the sensory nerve terminals in muscles and tendons. Mechanoreceptors respond to mechanical distortion or applied pressure (Dehghani, 2010).

Proprioception and mechanoreception are used by some passive prostheses, but in most active prostheses feedback sensors include accelerometers, rate gyros, as well as angle and force sensors. These provide inputs to onboard microprocessors that control the position of the prosthesis by actuating drive motors.

## 10.8 | PASSIVE PROSTHETICS

Passive lower and upper body prostheses, sometimes called conventional or body-powered prostheses, are generally driven by gross body movements. These can be coupled to the limb through a harness in the case of an arm or by momentum transfer in the case of a prosthetic leg.

A prosthesis needs to be comfortable to wear, easy to put on and remove, lightweight, and robust. It must function well mechanically and need little maintenance. Its selection

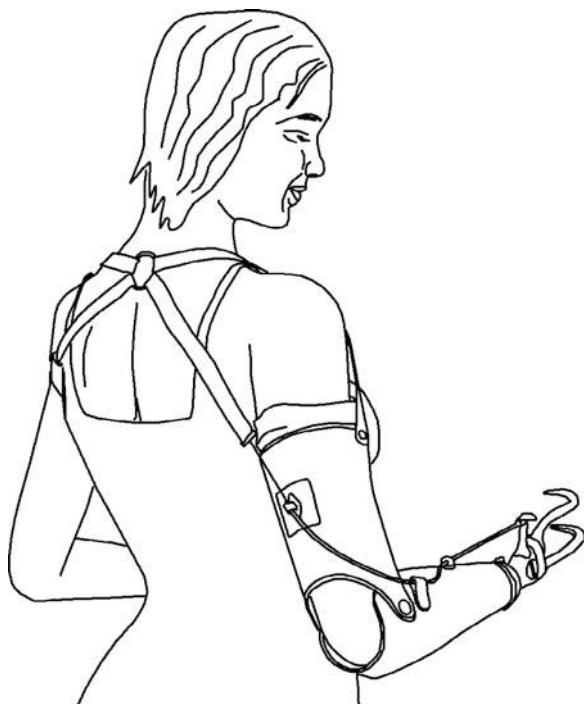
depends on a number of diverse considerations (Kelley, Pangilinan et al., 2009):

- Amputation level.
- Contour of the residual limb.
- Expected function of the prosthesis.
- Cognitive function of the patient.
- Vocation of the patient (desk job or manual labor).
- Patient interests (e.g., hobbies).
- Cosmetic importance of the prosthesis.
- Financial resources of the patient.

### 10.8.1 Actuation and Control of Upper Limb Prostheses

For upper limb prostheses, gross motions of the shoulder or the upper arm and sometimes of the chest are generally captured by a harness connected to a terminal device (TD) such as a hook or hand, as illustrated in Figure 10-12. For above-the-elbow amputees, double- or even triple-cable harnesses may be required with one to control the hook, a second to restore elbow flexion, and a third to lock it in place.

For patients to be able to control a body-powered prosthesis, they must be capable of one or more of the following gross body movements: glenohumeral flexion, scapular abduction or adduction, shoulder depression and elevation, or chest expansion. Additionally, they need sufficient muscle power and range of motion to provide effective actuation of the TD.



**FIGURE 10-12 ■**  
Shoulder harness controlled articulated hook.

An extension of this basic process is cineplasty, in which a loop of muscle in the chest or arm is isolated and covered with skin. Contraction of the muscle operates the attached prosthetic actuator. This process has the potential to provide excellent sensory feedback and control for the prosthesis user. The technique was originally developed by German surgeon Ernst Saurbruch prior to WWI and was improved with the introduction of the biceps tunnel cineplasty by M. Lebsche, one of Saurbruch's students.

The procedure was performed on many amputees during and after WWII. It was used to control the opening and closing of a prosthetic hand with the force applied to closing the fingers being proportional to the muscle force. Because sensory feedback is inherent in the process, this technique can be extremely effective in replacing lost functionality.

A disadvantage of the original process was that the amount of power available from the cineplasty was limited and sometimes could not provide enough force to actuate the prosthesis. However, this is no longer an issue with the application of servo assistance, which can amplify the muscle force in a similar fashion to power steering in cars.

One of the most successful prosthetic devices is the body-powered split hook, which has remained largely unchanged since it was patented in 1912. Upgrades have mostly been to the materials used, with silicon and plastic replacing leather for the socket, carbon fiber instead of wood or fiberglass for the frame, titanium instead of steel for the hook, and Spectra instead of steel cable for the control line (Kuniholm, 2009).

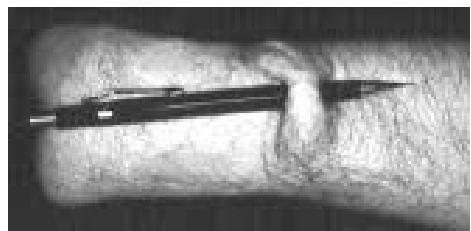
The major advantages of these mechanisms are their simple and robust design, which can survive in dusty or wet environments, their moderately low cost, and their increased control due to proprioception. The latter gives users feedback regarding the position of the terminal device so they know whether the hand or hook is open or closed by the pressure fed back to the shoulder area through the harness.

The main disadvantage is the restrictive and generally uncomfortable harness. It must be tight to capture the movement of the shoulder and therefore restricts arm motion to directly in front of the user between the waist and the mouth. Many patients also dislike the look of the hook or of any latex hand facsimile that may cover it.

A typical split hook such as the one shown in Figure 10-13 has canted fingers for grasping and is often textured or covered with rubber to improve the grip. Some even include a notch to hold a cigarette. It is opened using the body harness and closes automatically by means of an elastic or spring component. D. W. Dorrance designed the first commercially produced split hooks and started to market them in 1912. As with many other prosthesis makers, he was a user who was unhappy with existing devices and so proceeded to invent his own.

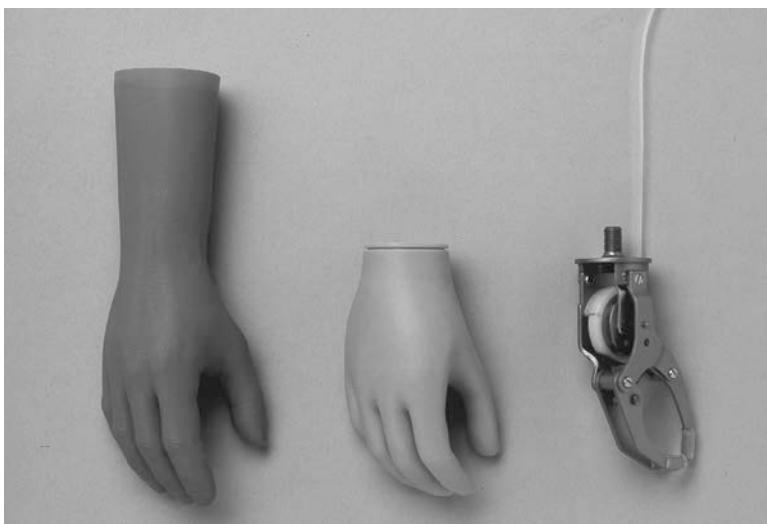
Split hooks are particularly well suited for activities that demand extreme precision for grasping small objects such as screws or nails. They are a reliable, robust, and resilient tool for manual labor or activities that are performed in a wet or corrosive environment, as they are made from inert materials like stainless steel or titanium.

Obviously, hooks are completely utilitarian and so can be somewhat intimidating; therefore amputees often resort to less effective but more cosmetic devices such as the





**FIGURE 10-13 ■**  
Otto Bock split hook. (Courtesy of Otto Bock, with permission.)



**FIGURE 10-14 ■**  
Mechanical hand showing the finger-control mechanism.  
(Courtesy of Otto Bock, with permission.)

Otto Bock hand shown in Figure 10-14. As with the split hook, the fingers of this prosthesis are controlled by a cable actuation driven by a harness.

Hybrid prostheses fall between passive and active types insofar as they use a cable-operated elbow and an electrically controlled terminal device. The advantage of this format is that the control strap can be less restrictive as it controls only elbow flexion (Muzumdar, 2004).

### 10.8.2 Walking Dynamics Using a Passive Prosthesis

A transfibial amputation has a major effect on joint motion and muscle activity during the different stages of the gait cycle. Prosthetics attempt to restore as much synergy as possible using advanced materials and mechanics, but there are still some difficult problems (Pons, 2008):

**Initial swing:** Gracilis and Sartorius muscle function is partially affected by the amputation insofar as length and effectiveness are compromised. Amputees accommodate for this using increased hip-flexor muscle activity to lift prosthetic feet off the ground and initiate the swing phase. In addition, amputees need to rotate the pelvis and lower trunk while applying load on the toes of the prosthetic. Because there is no muscle activity at the level of the knee or foot to initiate toe off, ground clearance is compromised. The loss of the popliteus and the short head of the biceps and the pretibial muscles eliminate active knee flexion and toe lift.

**Mid-swing:** Limitation of knee flexion becomes more critical and results in the prosthesis being lifted by using the hip (hip hiking), vaulting, or circumduction to compensate for reduced ground clearance. These gait deviations result in asymmetrical loading of the body, which puts strain on the lower back and requires more energy. Some of these limitations are due to knee design that often requires toe load to initiate flexion.

**Terminal swing:** Prosthetic knees all suffer from friction, which reduces angular acceleration of the knee during the advancement of the lower leg in the swing phase. This reduces the pendulum motion, and users must overcome this by means of additional muscle activity and abnormal motion patterns to complete the swing. Typically, users kick the knee into extension by contracting the gluteus muscles. This is abnormal and reduces the efficiency of the gait. The reaction also shifts the center of mass backward, counter to the walking direction, which affects balance and walking effectiveness. This passive swing extension also limits the possibility of stumble recovery because if the knee does not reach full extension in time it cannot accept the load of the next step and users generally fall.

**Initial contact and loading response:** This phase is responsible for stance stability and shock absorption. A passive prosthesis cannot produce the knee-flexion motion for shock absorption because the knee needs to be fully extended or even hyperextended to prevent it from buckling into flexion. Because the amputee kicks the prosthesis into extension, the initiation of flexion after initial contact is slower and more difficult with the result that there is a larger vertical displacement of the center of mass and therefore lower efficiency.

### 10.8.3 Knee Prosthetics

Knee kinematics is complex and difficult to quantify exactly due to variations between subjects and because the joint does not operate as a simple hinge. However, following motion studies undertaken by Walker (Walker, Kurosawa et al., 1985), it is possible to describe the instantaneous position of the hinge axis in the anterior-posterior axis,  $z_{dis}$  (mm), and the proximal-distal translation,  $y_{dis}$  (mm), in terms of the flexion angle,  $\theta_f$  (deg):

$$\begin{aligned} y_{dis} &= -0.05125\theta_f + 0.000308\theta_f^2 \\ z_{dis} &= -0.0602\theta_f + 0.0000178\theta_f^2 \end{aligned} \quad (10.11)$$

A guide or cam can be used to produce this motion curve, but these present drawbacks in terms of high manufacturing cost and resistance and pinching between the parts. An alternative is the crossed four-bar linkage described in Chapter 3, which has the advantages of simplicity, robustness, and ease of design.



**FIGURE 10-15** ■  
Range of Otto Bock  
passive knee  
prostheses.  
(a) Mechanical.  
(b) Fluid controlled.  
(Courtesy of Otto  
Bock, with  
permission.)

The displacement of a four-bar linkage can be described in a closed mathematical form depending only on the lengths and locations of the four bars. The point where the linkages cross defines the instantaneous hinge axis. It is therefore possible to develop a linkage that follows the path described by equation (10.11) with a high degree of accuracy (Forner-Cordero, Pons et al., 2008).

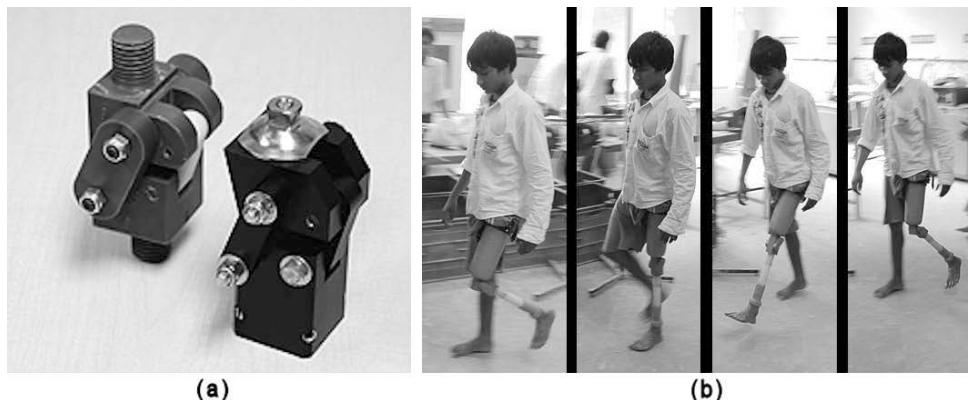
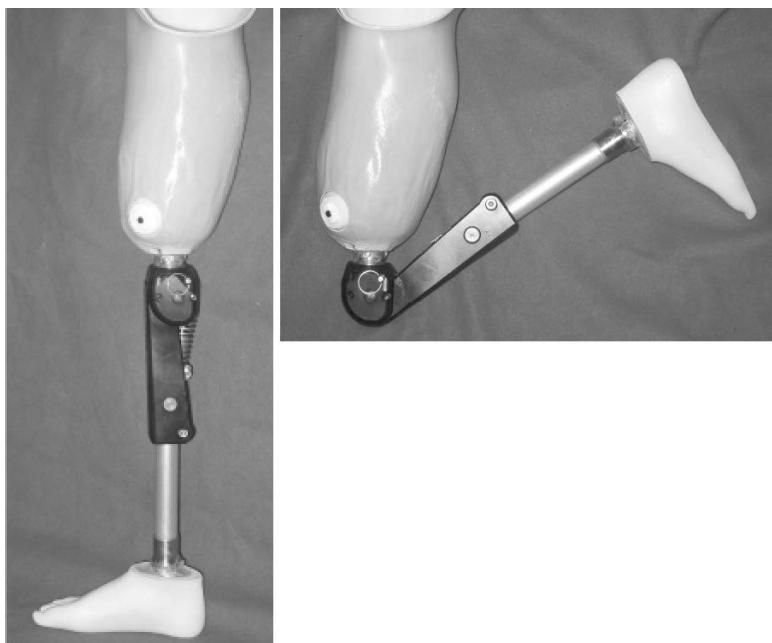
A number of manufacturers have developed passive knee prosthetics to minimize the issues described in the previous section. The first designs used constant-friction and friction-brake mechanisms developed as a result of research conducted after WWI. They include the Otto Bock 3R22 and 3R15 designs shown in Figure 10-15. They were superior to the old peg leg but were not effective for working on uneven surfaces or at a range of speeds (Torrealba, Fernandez-Lopez et al., 2008).

In the 1950s, Hans Mauch introduced the fluid-controlled prosthesis, and many prosthetic knees were produced with Mauch® cylinders. The introduction of fluid control, which accommodated variable torques during the gait cycle, improved both swing control and stability during the stance phase. Typical examples are the Gaitmaster from Össur and the Otto Bock 3R80 and 3R92, which are equipped with innovative, load-dependent brake mechanisms. The brake responds to a heavy heel load and stabilizes the prosthesis and maintains a high level of security during the stance phase. Adjustable stance flexion acts as a shock absorber and reduces the amount of stress on the body. A forefoot load automatically turns off the brake, which facilitates a graceful transition to the swing phase. The friction brake knee joint reduces the requirement for unnatural movements like hip hike. The 3R92 has a progressive pneumatic swing phase controller.

Another lightweight knee, shown in Figure 10-16, is the Aulie 802. It is made from nylon and stainless steel for use in harsh environments and is apparently the only knee prosthetic that can be used in the water. It includes an adjustable hydraulic control that is both light and simple. The adjustment clamp controls fluid flow by deforming the walls of the cylinder, while a stainless steel spring provides extension assist.

Most of these general-purpose knees come with standard fittings and are therefore interchangeable. They weigh between 600 g and 1 kg, can generally carry loads of 100 kg and more, and cost thousands of dollars. At the other extreme are the JaipurKnee and the LEGS M1 knees, both of which were developed by university students in the United States for the Third World.

**FIGURE 10-16 ■**  
The Aulie 802 is a typical passive prosthetic knee joint shown here extended and flexed. (Courtesy of Aulie, with permission.)



**FIGURE 10-17 ■** Low-cost knee prostheses (a) Comparison between the JaipurKnee and the LEGS M1. (b) JaipurKnee in action. (Courtesy of LeTourneau University and Melanie Worley, Re:Motion Designs, with permission.)

The JaipurKnee shown in Figure 10-17, was developed by biomedical device design and evaluation students at Stanford University based on their examination of the mechanics of high-end titanium knee joints costing between \$10,000 and \$100,000. They also surveyed a range of materials to find one that would be suitable to build a cheap prosthetic. The JaipurKnee is a polycentric knee joint made from oil-filled nylon polymer that is self-lubricating and can be manufactured for under US\$20 (Greig, 2009).

Prior to this, starting in 2004 students at LeTourneau University in Texas were developing the LeTourneau Engineering Global Solutions (LEGS) M1 knee. It is a block-shaped, four-bar, polycentric joint made from Delrin, a low-cost polymer used to make journal bearings and stainless steel bolts. The M1 costs about \$15 to make and is being locally manufactured at some of the 20 prosthetics clinics in developing nations (Stanfield, 2010).

### 10.8.4 Foot Prosthetics

Prosthetic feet must be capable of ankle rollover, must function as shock absorbers, and must return as much energy to the system as possible at the start of the swing phase. As shown in Figure 10-18 their design has evolved considerably since the introduction of the first Flex-Foot® in 1984.

As a result of changing attitudes among amputees, some of these prostheses are designed to be worn uncovered because they are both beautiful and extremely efficient. The two people shown in Figure 10-19 who have done much to promote this attitude are athlete, model, and activist Aimee Mullins and Oscar Pistorius, who was banned from the Beijing games because of a supposed unfair advantage over able-bodied athletes.

Should the amputee choose to cover the prosthesis, Dorset Orthopaedic and other companies can provide handcrafted silicone cosmeses based either on an existing limb or on a suitable alternative. These “skins” can be extremely lifelike, as shown in Figure 10-20, but alternatively can be patterned to suit the mood of the user.

As with the knee, there is a strong demand for prosthetic feet that are affordable by the world’s poor. This is a particular problem in regions such as Angola, Vietnam, and Iraq, among others, where millions of antipersonnel mines still take their toll.

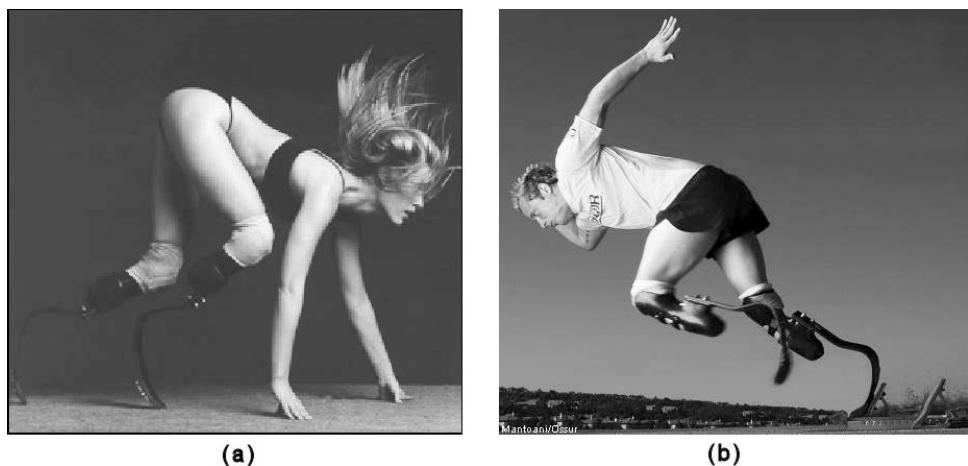
According to a survey conducted at the S.M.S Medical College in Jaipur, India, most existing passive foot prostheses are designed for First-World environments and are unsuitable for barefoot use. As a result of the survey, the following guidelines were drafted:

- The foot should be wearable with or without a shoe; it should therefore look like a foot.

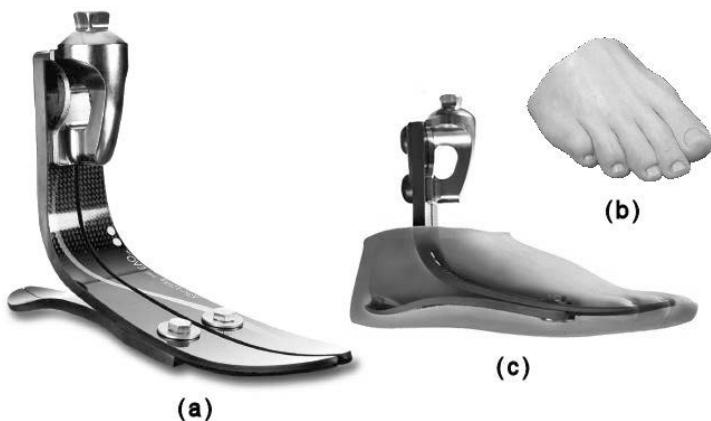


**FIGURE 10-18 ■**  
Evolution of Össur  
artificial feet for  
different  
applications.  
(Courtesy of Össur,  
with permission.)

**FIGURE 10-19 ■**  
**Famous amputees in action.** (a) Aimee Mullins. (b) Oscar Pistorius. (Courtesy of Howard Schatz and Össur, with permission.)



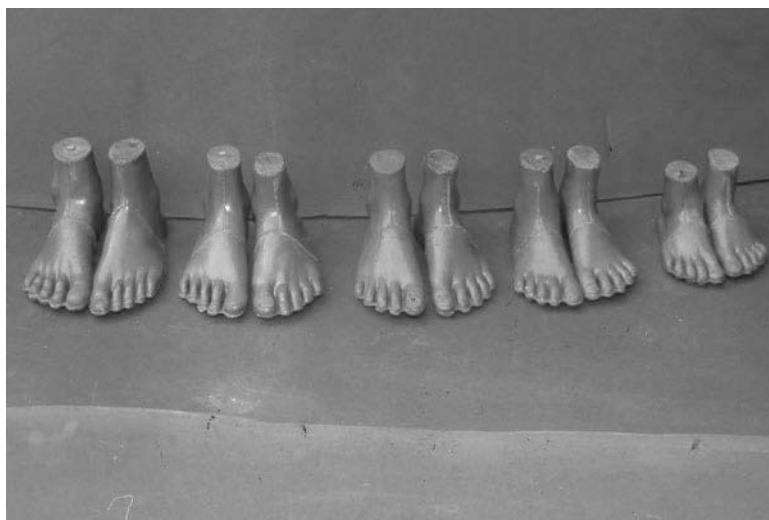
**FIGURE 10-20 ■**  
**Foot prosthesis**  
 (a) Hardware.  
 (b) Detail of silicone cosmesis.  
 (c) Transparent view of fitted cosmesis.  
 (Courtesy of Dorset Orthopaedic.)



- The exterior should be waterproof and durable.
- It should allow sufficient dorsiflection to permit the amputee to squat.
- It should support sufficient transverse rotation on the leg to facilitate walking and sitting cross-legged.
- It should have a good range of inversion and eversion for comfortable walking on uneven ground.
- It should be inexpensive.
- It should be made from materials that are readily available in the country of manufacture.

The Jaipur foot is a rubber-based prosthetic for people with below-the-knee amputations who meet those criteria. It was developed under the guidance of P. K. Sethi by Masterji Ram Chander in 1969 for victims of landmine explosions.

The foot, shown in Figure 10-21, was designed to be quick to fit, robust, and low cost. It is fitted for free by Bhagwan Mahavir Viklang Sahyata Samiti (BMVSS), a non-governmental organization (NGO) founded by Devendra Raj Mehta. Although it costs approximately US\$40, Dow India supports the NGO with free polyurethane, which is used in its production.



**FIGURE 10-21 ■**  
Examples of the Jaipur foot.  
(Courtesy of BMVSS, with permission.)

More than 20,000 Jaipur feet are fitted from clinics in 21 countries annually, with total fittings since its debut exceeding 350,000.

## 10.9 | ACTIVE PROSTHETICS

An active prosthesis is one that uses an additional energy supply for motion and control over and above that provided by muscle power from the user. These range from the most complex, 20 DoF upper limb prosthetics currently under development to relatively unsophisticated, single DoF knee joints and everything in between.

There are of course pros and cons of selecting an active prosthesis above a passive or cosmetic one, as detailed in Table 10-3 (Kelley, Pangilinan et al., 2009).

The process of operating our arms and hands to perform even very complex functions has been so well learned that it has become completely automatic and can often be done while the brain is engaged in other activities. Consider, for example, drinking from a cup,

**TABLE 10-3 ■** Pros and Cons Relating to the Selection of Upper Limb Prosthesis Types

Type	Pros	Cons
Cosmetic	Lightest Best cosmesis Least harnessing	High cost if custom made Least function
Body powered	Moderate cost Reasonably light Durable Good sensory feedback Wide range of actuators for different activities	Most body movement needed to operate Most harnessing Least satisfactory appearance Increased energy expenditure
Battery powered	Moderate or no harnessing Least body movement needed to operate Moderate cosmesis More function-proximal areas Stronger grasp in some cases	Heaviest Most expensive Most maintenance Limited sensory feedback Extended therapy time for training

something that babies aged from 18 months can do perfectly. The brain instructs your arm from the shoulder to the fingertips to perform the function, and it occurs in a continuous smooth manner without your even being aware of the process.

In contrast, a person with a conventional active prosthesis must rethink the whole process with each action consciously and painstakingly thought out and executed as follows:

- Shoulder—forward
- Elbow—bend
- Forearm—rotate the thumb up
- Hand—open around the cup
- Hand—close slowly
- Elbow—lift without spilling
- Head—move forward to the cup
- Drink

Each joint movement must be executed in sequence, and then, of course, the process must be reversed to put the drink back down. Additionally, because there is no proprioception, the person must look at the cup and the arm continuously to provide the correct visual feedback to drive the process. Compared with our natural ability, operating a prosthesis can be time-consuming and awkward. This is only one of the many reasons amputees often refuse to use such devices.

Others problems include the fact that they are generally uncomfortable and heavy. For an active prosthesis to operate effectively it must operate from a stable frame. This increased intimacy usually results in an interface with a large contact area and high contact pressure to prevent rotational instability. This larger skin contact increases issues of heat dissipation. In addition, patients with residual limbs often rely on those for their sense of touch, so if they are enclosed an important tactile surface is lost (Muzumdar, 2004).

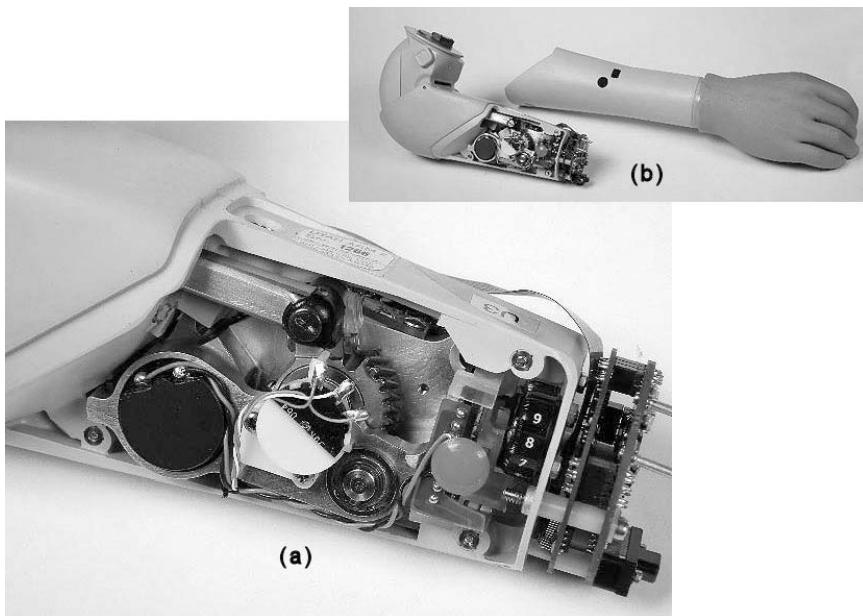
### 10.9.1 Arm Mechanisms

Depending on the site of the amputation or the position of any residual limb, a prosthesis may need a number of degrees of freedom to replace wrist rotation and flexion, elbow flexion, and two axes of shoulder rotation. Issues with the hand itself are considered in the following section.

A number of different methods of performing the actuation have been developed, including electric motor and gearbox combinations, linear electrical actuators, shape memory alloy (SMA), hydraulics, and pneumatics. The most common are the motor and gearbox types, of which the Utah Arm shown in Figure 10-22 is a good example.

The original Utah Arm was developed in 1981 by Steve Jacobsen from the University of Utah Center for Engineering Design (Sears, Jacobsen et al., 1989). Motion Control had been established by faculty members to commercialize medical technology developed by the center, and it released the Utah Arm 2 with upgraded electronics and improved nickel metal hydride (NiMH) batteries in 1997. It provides full proportional control of both the elbow and any attached terminal device, with the terminal device being the default channel in multiplexed operation.

Finally, in 2004 microprocessor technology was incorporated into the design of the Utah Arm 3. This device continues to provide proportional control to the wrist and elbow



**FIGURE 10-22 ■**  
Electrically powered Utah Arm. (a) Open section showing internal mechanics and electronics. (b) Forearm and hand attachment. (Courtesy of Motion Control.)

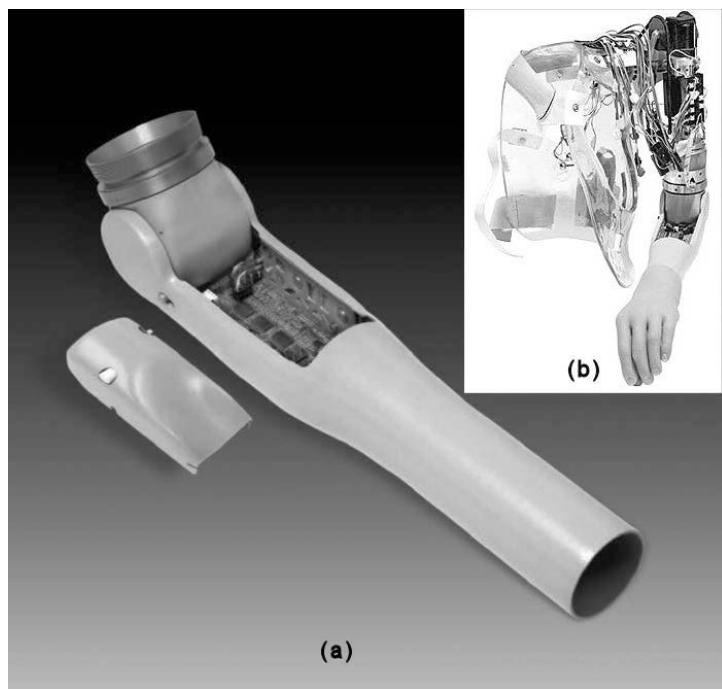
but is now capable of controlling two separate actuation motions simultaneously. It includes a computer interface that allows the user to fine-tune performance.

Technical specifications of the device are as follows (Motion Control, 2009):

- Excursion range: 135°
- Excursion time with myoelectric hand TD: 1.20 seconds
- Active lift: 1 kg in the terminal device and using a fully charged battery
- Load limit: 22.7 kg, with elbow locked at 90° flexion, 15.9 kg, when forearm extension installed
- Humeral rotation: unlimited
- Wrist rotation: quick-disconnect wrist: 360° in each direction
- Weight: 913 grams
- Heat tolerances: operating temperatures: 0° to 44 °C
- Storage temperatures: -18° to 60 °C
- Current: maximum: 4.0 amps quiescent: 10 mA
- Battery specifications: rechargeable NiMH (1100 mA hours capacity)
- Voltage: dual-supply, ± 6 volts direct current (DC)
- Charge time: 2.5 hours maximum
- Forearm length: (from rearmost point of the forearm to the end of the wrist)
  - Standard: 27.3 cm
  - With extension: 32.4 cm
  - Minimum: 24.8 cm

Liberating Technologies developed a high-performance prosthetic elbow called the Boston Digital Arm System, shown in Figure 10-23, that was first marketed in 2001. It

**FIGURE 10-23 ■**  
**Liberating**  
 technologies Boston  
 Digital Arm. (a) Arm  
 detail. (b) Fitted with  
 a shoulder  
 mechanism and  
 hand cosmesis.  
 (Courtesy of  
 Liberating  
 Technologies, Inc.,  
 with permission.)



provides a torque of 14 Nm and a flexion time of 1.1 s and weighs only 888 g. The powerful embedded Texas Instruments digital signal processor (DSP) can control up to four other prosthetic devices in addition to the elbow, making it the hub for a complete upper limb prosthetic device (Liberating Technologies, 2009).

One of the most advanced prosthetic arms in the world today is the “Luke” arm developed by Dean Karmen’s DEKA Research with funding from DARPA. The arm is modular, allowing a person with any level of amputation to use some or all of the modules. The hand contains separate electronics, as does the forearm. The elbow is powered by electronics in the upper arm, and the shoulder is unique among prosthetics as it allows users to reach above their head. The complete arm, shown in Figure 10-24, has 18 degrees of freedom (compared with 22 degrees for a natural arm) and weighs only 3.6 kg including

**FIGURE 10-24 ■**  
 Complete DEKA  
 arm. (a) Arm  
 including shoulder  
 mechanism.  
 (b) Detail of hand  
 with pinch grip.  
 (Courtesy of DEKA  
 Research.)





**FIGURE 10-25** ■  
DEKA arm being used by a transhumeral amputee. (Courtesy of DEKA Research.)

batteries. At present it can be controlled using pressure switches or myoelectric signals obtained from electrodes connected to the amputee (Adee, 2008). An example of the capability of the DEKA arm on a transhumeral amputee is shown in Figure 10-25.

Real specifications for the arm are hard to come by, but apparently the hand has motor control fine enough for test subjects to pick up chocolate-covered coffee beans individually, to unlock a door, or to use a power drill. These are made possible by some of the six preconfigured grip settings including chuck grip, key grip, and power grip.

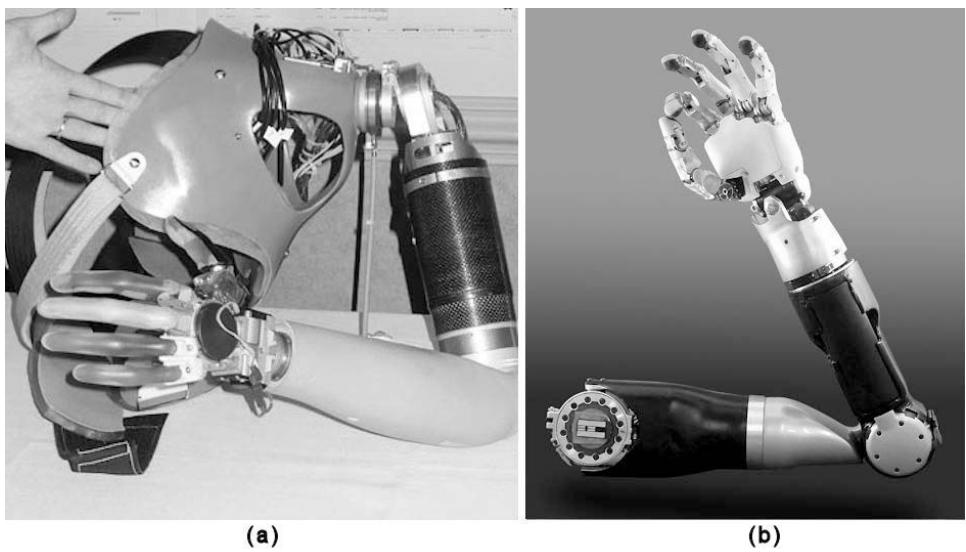
Whereas most prosthetics manufacturers' devices conform to standard interfaces to allow users to mix and match components, DEKA Research does not and uses proprietary electrical and mechanical interfaces. It is not alone in this decision, and the next generation of Otto Bock prostheses, which also use technology from the DARPA funded program, will use a proprietary and encrypted digital communication standard called the Axon bus (Kuniholm, 2009).

Depending on the degree of amputation, today's state-of-the-art prosthetic arms can cost patients \$100,000 or more. The goal with the DEKA arm is to keep as close to that as possible.

Another important player in the DARPA-supported prosthetic arm business is the Johns Hopkins University Advanced Physics Laboratory (APL). Its mandate was to oversee the development of a fully integrated prosthetic arm that could be controlled naturally and to provide sensory feedback. The latest version of their modular prosthetic limb (MPL), shown in Figure 10-26, provides 22 degrees of freedom and is controlled with signals from the reinnervated pectoral of the amputee. Work is now under way to move the control site to the brain, and the APL team is collaborating with the University of Pittsburgh and the California Institute of Technology, which both have some experience in brain-computer interfaces, and with the University of Chicago for its expertise in sensory perception.

In contrast to DEKA and Otto Bock, APL has decided to open up the framework of its project. First, it is making the virtual environment open source. This integration environment is a training simulation that allows signal processing and control techniques to be evaluated by having an amputee control a virtual reality arm. Second, the team also

**FIGURE 10-26 ■**  
**Advanced Physics Laboratory arm developed to provide sensory feedback.**  
 (a) Complete arm with harness housing sensory electrode array.  
 (b) Complete arm and hand. (Courtesy of Rehabilitation Institute of Chicago and Johns Hopkins University Applied Physics Laboratory, with permission.)



plans to publish an open-control communication architecture as well as the mechanical interfaces for each of the physical components (Kuniholm, 2009). This decision will ensure that the small players, including manufacturers of specialist terminal devices like Touch Bionics (makers of iHand), will continue to fill niche markets in the prosthetics business.

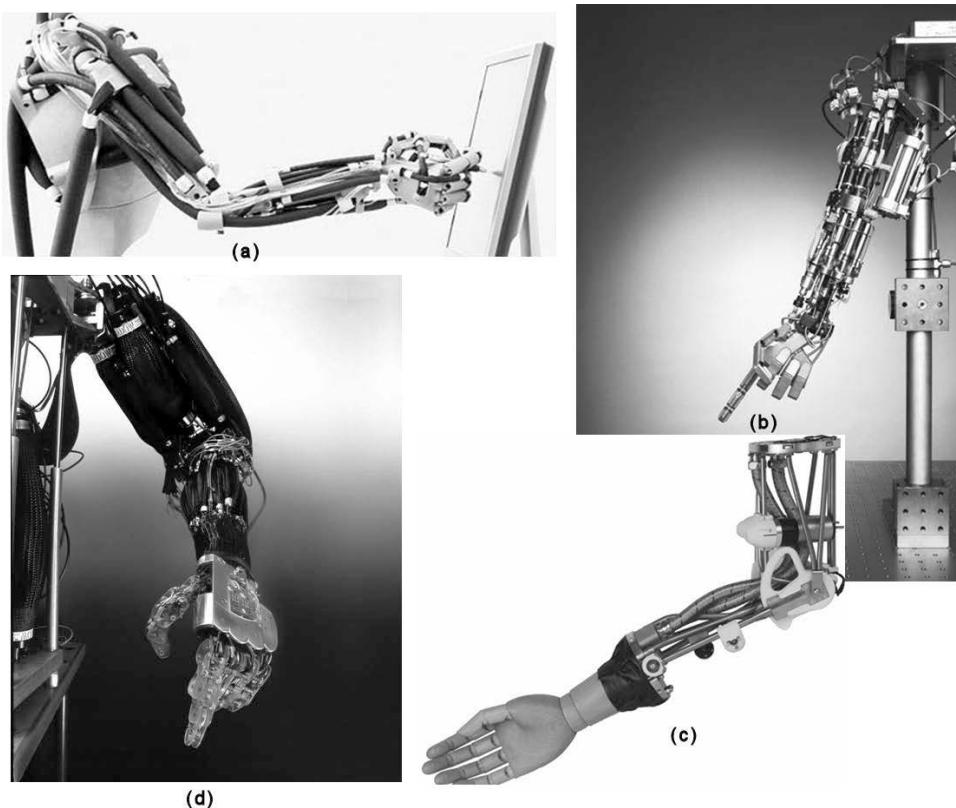
Problems with conventional electric motor-driven prosthetics are that they are not usually compliant, the power-to-weight ratio of electric motors is not particularly high, and gears suffer from backlash. For this reason a number of developments using pneumatics and even the catalysis of hydrogen peroxide are under way, as shown in Figure 10-27.

Pneumatic artificial muscles (PAMs) or air muscles used by the Shadow Robotics prosthesis are a good choice for the actuation of prosthetic arms as they provide smooth jerk-free motion because there is no stiction. Additionally, they can produce the force required but still offer compliance and will yield if an obstacle is encountered. Finally, they are quiet and light and have a relatively good power-to-weight ratio.

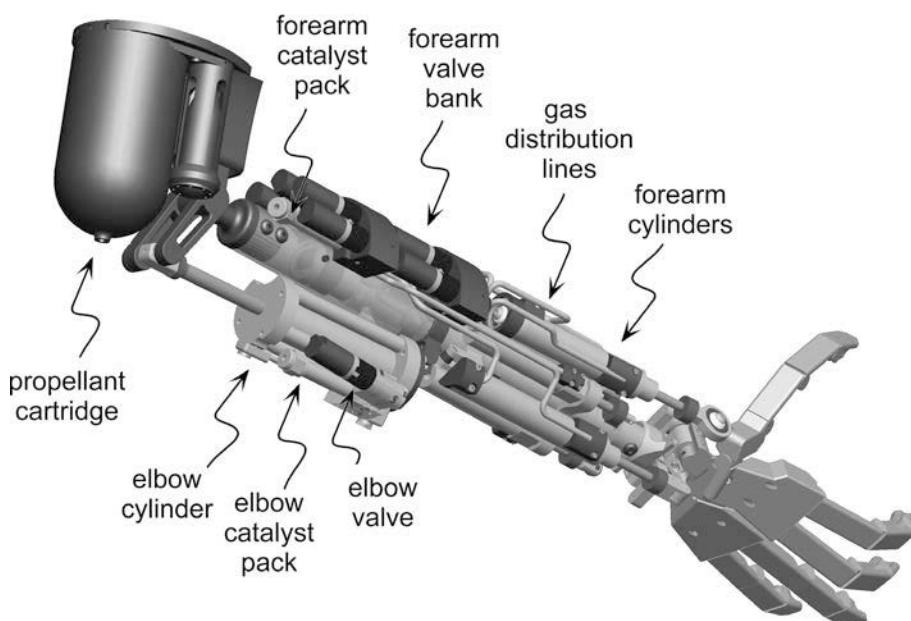
The Vanderbilt arm, shown in Figure 10-28, was developed by Michael Goldfarb's team at the Center for Intelligent Mechatronics at Vanderbilt University to investigate a power source with a superior power-to-weight ratio to that provided by batteries. It uses a catalytic converter that burns hydrogen peroxide to produce high-temperature steam, which is then passed through a bank of precision valves to drive spring-loaded cylinders that actuate the joints through pulleys. The small canister of  $H_2O_2$  that fits easily within the upper arm socket provides sufficient propellant to power the system for 18 hours of normal use.

Thermal management was an issue, and though the catalyst pack produces a lot of heat it has been covered by a thin insulating layer to ensure that it remains cool enough to touch. Exhaust steam is released through a porous cover on which it condenses. It generates about as much water as a person would sweat from a normal arm on a warm day.

The prototype, which is close to normal in function and power to a human arm, provides about 10 times the power of most other prosthetic arms. It can lift about 10 kg, three to four times more than current commercial arms, and can do so three to four times faster even though it has not yet been optimized for these capabilities.

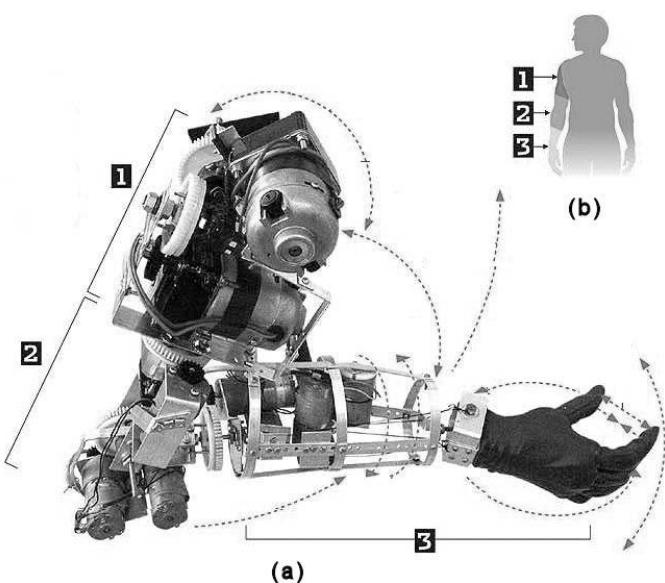


**FIGURE 10-27** ■ Other prosthetic arms. (a) Airicís pneumatic arm. (b) Vanderbilt rocket-powered arm. (c) Fraunhofer ISELLA arm. (d) Shadow Robotics PAM arm.



**FIGURE 10-28** ■ Vanderbilt University H<sub>2</sub>O<sub>2</sub>-powered prosthetic arm. (Courtesy of the Center for Intelligent Mechatronics, Vanderbilt University.)

**FIGURE 10-29** ■  
 Low-cost prosthetic arm developed by Simón Guerrero Castillo of the Instituto Politécnico Nacional in Mexico. (a) Photograph of the prosthesis hardware. (b) Relationship between hardware and a human arm. (Castillo 2007).



Not all prosthetic arms are so costly, as has been demonstrated by Simón Guerrero Castillo, a student from the Instituto Politécnico Nacional of Mexico. He developed a myoelectric-controlled prosthesis at a total cost of 18,000 pesos that mimics the natural motions of an arm including wrist flexion, forearm rotation, and movement of the elbow and the shoulder. The device, shown in Figure 10-29, does not look like a conventional arm, but it works.

### 10.9.2 Hand Mechanisms

Two classes of hands are available: (1) anthropomorphic, with many degrees of freedom and mostly prototypes or in limited production for use as research tools; and (2) terminal devices (TDs), which are much lighter and generally simpler devices with fewer degrees of freedom for use as prosthetics. Specifications of some of the research hands under development, or in limited production, are listed in Table 10-4. Hands of various complexities are available from a number of manufacturers, including Motion Control Inc., Otto Bock, RSL Stepper, Liberating Technologies Inc., and Touch Bionics. The more primitive hands contain a small geared motor that drives the fingers to produce pinch forces of up to 10 kg at the fingertips, which allows them to grip most objects securely even without the compliance of a natural hand. However, without the appropriate feedback such hands tend to crush delicate objects, so most modern active TDs include force sensors or motor drive current monitoring with some form of feedback to the user.

The earliest devices had digital control (they were either on or off), but the introduction of proportional control, in which the rate of closure is proportional to the actuation signal amplitude, has given these TDs more precision. Actuation signals could be provided by myoelectrics, force sensors, or linear potentiometers.

Most TDs come with quick-release attachments so that the correct hand for the job can be fitted easily and quickly. In the past these included cosmetic hands with limited movement, power grippers, and hooks, as shown in Figure 10-30. However, in the past decade or so, hands with individually articulated fingers and opposable thumbs have become available.

TABLE 10-4 ■ Some Anthropomorphic Research Hands

BH8-series Barrett hand Barrett technology, inc. Robot Hand identification		Robonaut Hand NASA Johnson Space Center USA		DLR Hand II DLR-German Aerospace Center Germany		Ultraglight Hand Research center of Karlsruhe Germany		Gifu Hand Gifu University Japan		Shadow Hand Robot Company Ltd United-Kingdom		UB Hand III Bologna University Italy	
Year	Status	Year	Status	Year	Status	Year	Status	Year	Status	Year	Status	Year	Status
1997	Commercial use	1998	Research prototype	1999	Research prototype	2000	Research prototype	2000	Research prototype	2001	Research prototype	2004	Research prototype
<b>Structures and materials</b>	3 fingers (opposable), Electrical revolute brushless motors, Spur and worm gear transmissions, Fair contact surface smoothness	3 fingers and one opposable thumb, Electrical revolute brushless motors, Tendons routed through pulleys	4 fingers and one opposable thumb, Electrical revolute brushless motors, Flex-shaft + lead screw transmissions, Kevlar body armor coated with Teflon surface	4 fingers and one opposable thumb, Electrical revolute brushless motors, Harmonic drives/ gears transmissions, Silicone finger surface	4 fingers and one opposable thumb, Electrical revolute brushless motors, Direct drive transmissions, Silicone-rubber glove	4 fingers and one opposable thumb, Flexible fluidic actuators, Worm gear transmissions, Thin polycarbonate fingernails	4 fingers and one opposable thumb, Built-in DC Maxon servomotors, Worm gear transmissions, Silicone-rubber glove	4 fingers and one opposable thumb, Air Muscles, Tendons, Layer of soft polyurethane flesh, Thin polycarbonate fingernails	4 fingers and one opposable thumb, Built-in DC Maxon servomotors, Worm gear transmissions, Silicone-rubber glove	4 fingers and one opposable thumb, Air Muscles, Tendons, Layer of soft polyurethane flesh, Thin polycarbonate fingernails	4 fingers and one opposable thumb, Air Muscles, Tendons, Continuous compliant pulps	4 fingers and one opposable thumb, DC brushed motor, Pulling tendons, Continuous compliant pulps	
<b>Sensory-motor skills</b>	4 controlled DOF Motor position sensors (optical incremental encoders) Strain-gauges based joint torque sensors	16 controlled DOF Motor position sensors based joint position sensors 3-axis fingertip force sensors	14 controlled DOF Motor position sensors (2 wrist + 12 fingers) Joint position sensors Tendon tension sensors 3-axis fingertip force sensors	13 controlled DOF Motor position sensors (3 wrist + 10 fingers) Joint position sensors (potentiometers) Strain-gauged based sensors Tactile force sensors (FSR Force Sensing Resistor technology)	13 controlled DOF Motor position sensors (3 wrist + 10 fingers) Joint position bending sensors Pressure sensors in finger links 6-axis fingertip force sensors	16 controlled DOF Motor position sensors Joint position bending sensors 6-axis fingertip force sensors	16 controlled DOF Motor position sensors Joint position bending sensors Pressure sensors in finger links Distributed resistive tactile sensors	16 controlled DOF Motor position sensors 1 palm) Muscle pressure sensors	16 controlled DOF Motor position sensors Joint position bending sensors 6-axis fingertip force sensors	16 controlled DOF Motor position sensors Joint position bending sensors Pressure sensors in finger links Distributed resistive tactile sensors	16 controlled DOF Motor position sensors Joint position bending sensors 1 palm) Muscle pressure sensors	16 controlled DOF Motor position sensors Joint position bending sensors 6-axis fingertip force sensors	
<b>Feedback control</b>	No sensory feedback Patented clutch mechanism distributes grasp forces Patented reconfigurable spreading fingers	Position feedback	No sensory feedback	Position feedback (impedance control)	No sensory feedback	Position feedback	No sensory feedback	Position feedback (low level joint position and joint stiffness control)	Position feedback	Position feedback	Position feedback (low level joint position and joint stiffness control)	Incomplete sensory feedback	Incomplete sensory feedback
<b>Functional capabilities</b>	Power grasp and precision grasp	Power grasp and precision grasp	Power grasp, precision grasp and human-like fine manipulation	Power grasp, precision grasp and human-like fine manipulation	Power grasp and precision grasp	Power grasp and precision grasp	Power grasp and precision grasp	Power grasp and precision grasp	Power grasp and precision grasp	Power grasp and precision grasp	Power grasp and precision grasp	Power grasp, precision grasp and human like fine manipulation	Power grasp, precision grasp and human like fine manipulation
<b>Autonomy</b>	Non autonomous	Non autonomous	Remotely operated via visual feedback	Pre-planning; performs previously stored trajectories and hand poses	Pre-planning; performs patterns of movements, position and torque of every joints	Non autonomous	Non autonomous	Non autonomous	Non autonomous	Non autonomous	Non autonomous	Semiautonomous	Semiautonomous

**FIGURE 10-30 ■**

Terminal devices.  
 (a) Photograph of the Vari-Plus Speed Hand. (b) Electric Greifer. (Courtesy of Otto Bock, with permission.)



### 10.9.2.1 Ultralight hand

The Ultralight hand was developed at the Forschungszentrum Karlsruhe (FZK) using flexible fluidic actuators. These consist of a flexible chamber between two members configured so that inflation of the chamber using air or a liquid results in an increasing separation between the members or angle if they are hinged. Unlike PAMs, these actuators provide a reasonably linear relationship between pressure and expansion force. They are low cost and can be cascaded to produce very complex movements while remaining lightweight.

The hand shown in Figure 10-31 consists of 18 miniature flexible fluidic actuators integrated into the fingers and wrist. The fingers contain actuators, flex, and tactile sensors, while the metacarpals each house a microcontroller, microvalves, and a small pump and a power source. Joint extension is passive using elastomeric spring elements.

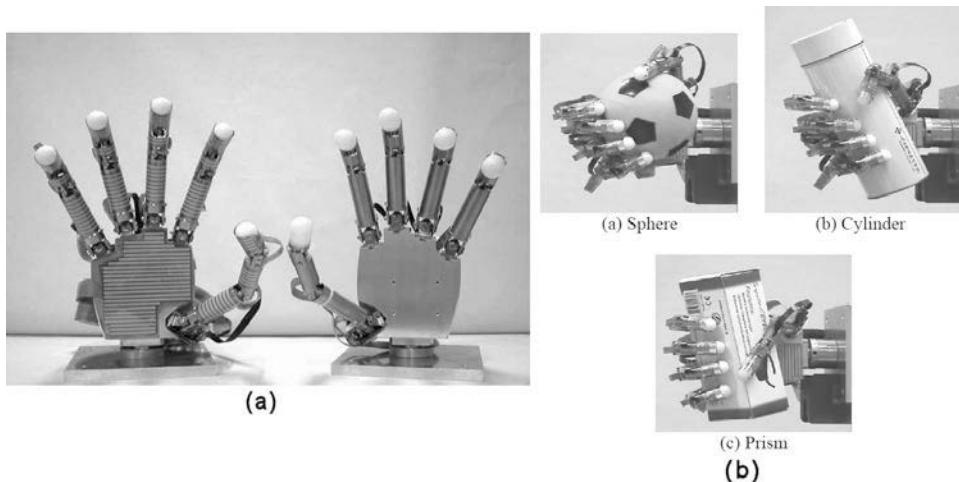
### 10.9.2.2 GIFU Hand

Developed at the Virtual System Laboratory and the Faculty of Engineering of GIFU University, this hand weighs 1.4 kg and consists of an opposable thumb and four fingers supported by a palm plate and wrist, as shown in Figure 10-32. The thumb has four joints with four degrees of freedom, while each of the fingers has four joints and three degrees of freedom, making 16 DOF in total. All the DC motor actuators of the joints are mounted within the hand providing a fingertip force of up to 2.7 N. Each of the driven joints includes

**FIGURE 10-31 ■**

Ultralight hand by  
 Forschungszentrum  
 Karlsruhe. (Schulz,  
 Pylatiuk et al., 2001)





**FIGURE 10-32 ■**  
The Gifu hand.  
(a) With and without tactile sensors.  
(b) Grasping various objects. (Mouri, Kawasaki et al., 2002).

a motor with a 16-count magnetic shaft encoder and a gearbox to increase output torque. The third and fourth joint of each of the fingers is linked by a four-bar linkage that results in an almost linear relationship between those joint angles.

A total of 16 individual power amplifiers and power device (PD) controllers ensure fast but well-damped response for all the joints. Measurements made of the first joint of the thumb show a step response time of less than 0.1 s to a 15° angle command and a 10.4 Hz bandwidth for the velocity transfer function. The minimum bandwidth of any of the joints exceeds 7.4 Hz, making the hand faster than its human counterpart, which has a bandwidth of 5.5 Hz.

To aid with control, the current to each of the motors is monitored, as this is a good indication of the output torque. For the second joint of the thumb, the relationship between torque,  $\tau$ (Nm), and current,  $I$ (A), is

$$\tau = 0.79I - 0.0098 \quad (10.12)$$

The torque required to overcome static friction is therefore 0.0098 Nm.

A six-axis force sensor in each fingertip provides force feedback to aid with grasping control. As shown in Figure 10-32, a 624-point distributed tactile sensor is attachable to the palm and finger surfaces.

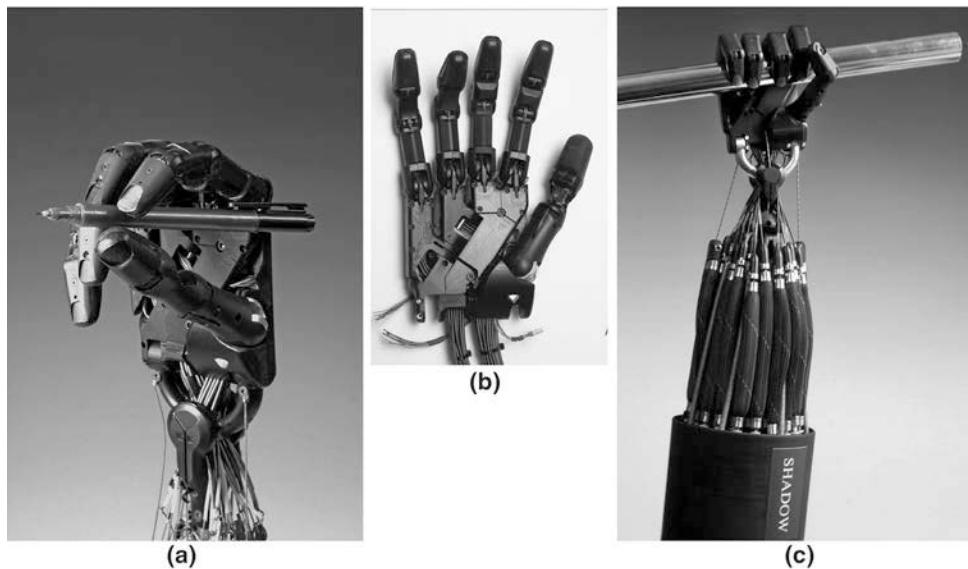
### 10.9.2.3 Shadow Hand

The shadow hand is probably the robot hand closest to the human hand currently available. It provides 24 movements driven by an integrated bank of 40 PAMs, allowing a direct mapping from a human hand to the robot, as seen in Figure 10-33. Because the muscles are compliant, the hand can be used to manipulate soft or fragile objects.

It has integrated sensing and position control, allowing precise control from off-board computers, and it can be fitted with touch sensing on the fingertips, offering sensitivity sufficient to detect a single small coin.

The thumb has five joints and five degrees of freedom, while each finger has four joints and three degrees of freedom, as shown in Figure 10-34. Depending on the joint, actuation can be by a pair of antagonistic PAMs or a single PAM with return spring. Hall effect sensors measure the angles of the joints to a resolution of 0.2° digitized to 12 bits

**FIGURE 10-33 ■**  
 PAM actuated shadow hand.  
 (a) Pencil grip.  
 (b) Open. (c) Power grip. (Courtesy of Shadow Robot Company, with permission.)

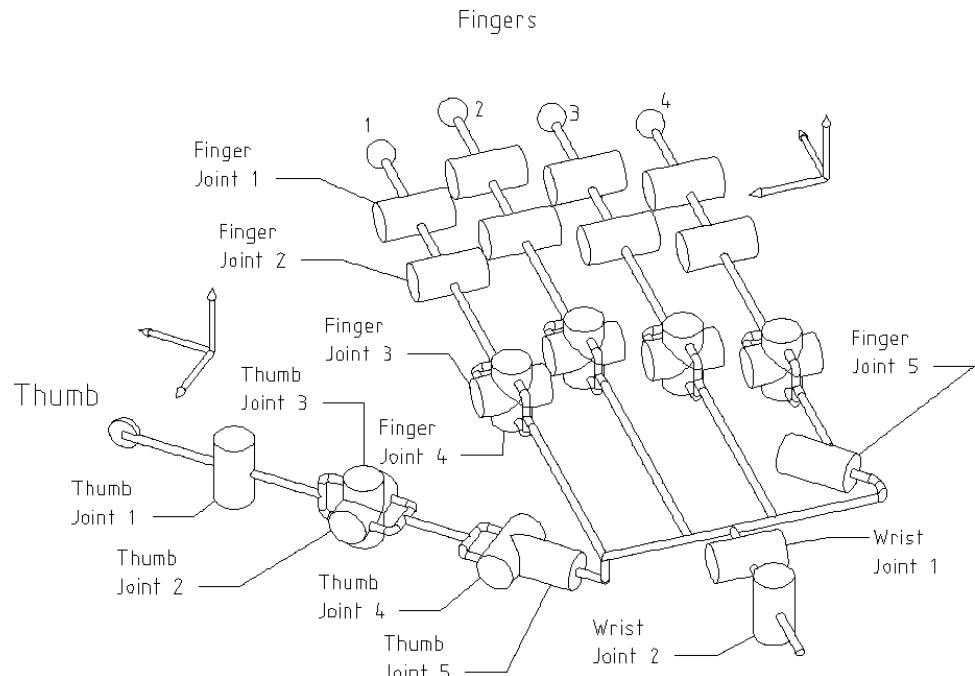


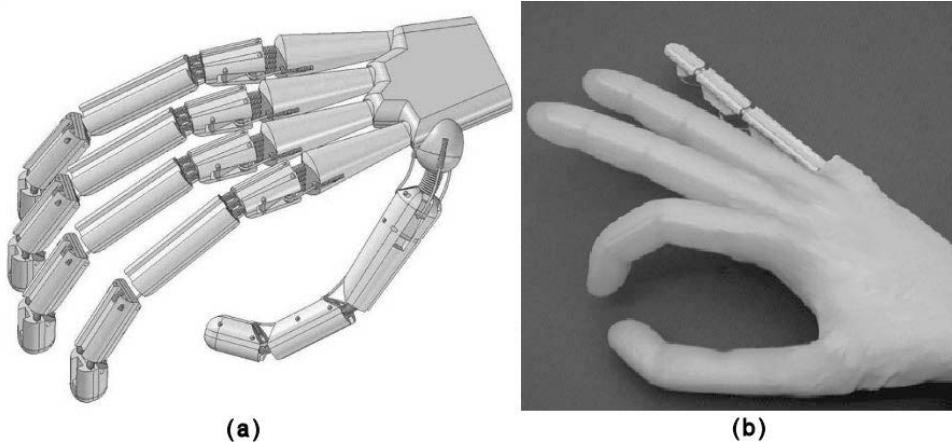
and sampled at 180 Hz. The pressure in each muscle (zero- to four-bar) measured by a solid-state pressure sensor is also digitized to 12 bits and transmitted on the CANBUS.

#### 10.9.2.4 University of Bologna Hand III

The University of Bologna (UB) hand shown in Figure 10-35 uses deformable (compliant) elements as joint hinges to provide a structure that is able to support distributed tactile sensors and a continuous compliant outer layer.

**FIGURE 10-34 ■**  
 Kinematic construction of the shadow hand.  
 (Courtesy of Shadow Robot Company, with permission.)





**FIGURE 10-35** ■  
UB hand. (a) CAD model.  
(b) Assembled prototype. (Courtesy of University of Bologna.)

Finger actuation is based on “tendons” and has been designed to be independent of any particular form of actuation, so it can be used with motors and pulleys or PAMs.

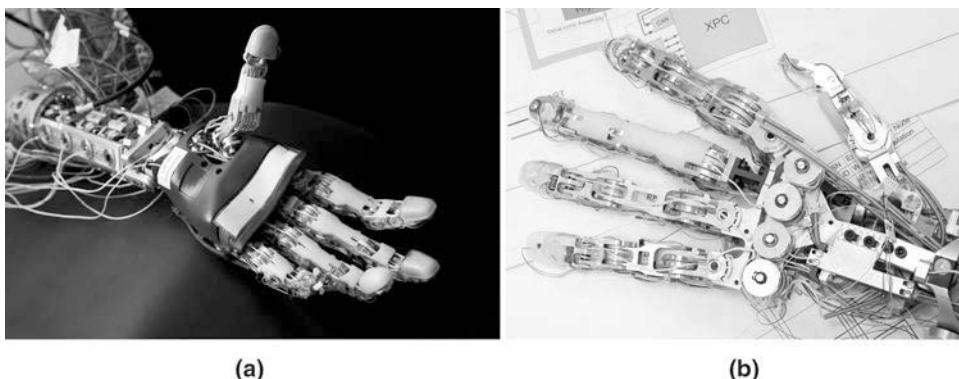
#### 10.9.2.5 APL Hand

Advances in the design of hand prosthetics, either as independent devices or as part of an arm, have been dramatic since the introduction of the DARPA Revolutionising Prosthetics Programme. A good example of this is the second-generation prototype (Proto-2) of the device developed at APL. The hand alone, shown in Figure 10-36, is driven by 15 electric motors and is capable of unprecedented mechanical dexterity.

Sensors in the hand are capable of providing pressure, texture, and even temperature to the user using both conventional interfaces and more recently through reinnervated pectoral muscles, as discussed later in this chapter.

#### 10.9.2.6 i-Limb Hand

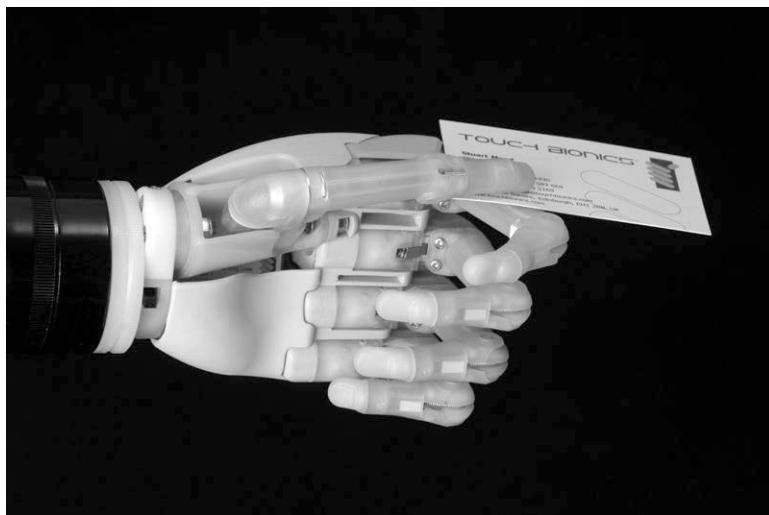
One of the most advanced but practical prosthetics available today is the i-Limb Hand by Touch Bionics, shown in Figure 10-37. It is actuated through a simple two-electrode myoelectric interface that controls five individual fingers with the thumb being positioned manually. The grasp of the hand is like that of a human hand, with the articulating fingers able to close tightly around objects. Built-in stall detection tells each individual finger when



**FIGURE 10-36** ■  
APL Proto-2 hand.  
(a) Hand and forearm. (b) Close up of hand mechanism. (Courtesy of Johns Hopkins University Advanced Physics Laboratory, with permission.)

**FIGURE 10-37 ■**

The i-Limb hand by Touch Bionics.  
(Courtesy of [www.touchbionics.com](http://www.touchbionics.com), with permission.)



it has sufficient grip on an object and, therefore, when to stop powering. Individual fingers lock into position until the patient triggers an open signal through a simple muscle flex.

The inclusion of a thumb that can, like the human thumb, be rotated into different positions enables important grip configurations, many of which have not been available to amputees before, such the following:

**Key grip:** The thumb closes down onto the side of the index finger. This grip is used to hold items such as a plate or a business card. The addition of wrist rotation enables the patient to turn a key in a lock in a “human” fashion.

**Power grip:** All five fingers and the thumb close together to create a full-wrap grip. This grip can be used to hold a can of drink while opening the ring-pull or for carrying large objects such as briefcases.

**Precision grip:** The index finger and thumb meet (or index finger, middle finger, and thumb meet) to pick up small objects and to hold objects when performing finer control tasks.

**Index point:** The thumb and fingers close, but the index finger remains extended. Patients have found this grip very useful for operating computer keyboards, telephone dial pads, ATM cash machines, and a host of other everyday activities.

The power grip and index point positions are shown in Figure 10-38.

The i-Limb Hand is anatomically correct both when resting and in motion, so in conjunction with an excellent cosmesis it is hardly distinguishable from the real thing, as seen in Figure 10-39. This innovation has been very much appreciated by patients, many of whom simply wish to blend back into society without others noticing their amputation.

People interested in technology are often keen to show off their prosthesis. However, because of the need to provide a grip surface and to protect the hand from dust and water, Touch Bionics has developed the i-Limb Skin. This is a thin layer of semitransparent material that has been computer modeled to wrap accurately to every contour of the hand.



(a)



(b)

**FIGURE 10-38 ■**  
The i-Limb hand in action. (a) Index point configuration. (b) Power grip configuration. (Courtesy of [www.touchbionics.com](http://www.touchbionics.com), with permission.)



**FIGURE 10-39 ■**  
The i-Limb hand cosmesis makes it difficult to see which is the real hand and which is the prosthesis. (Courtesy of [www.touchbionics.com](http://www.touchbionics.com), with permission.)

Touch Bionics was founded in 1963 at the Princess Margaret Rose Hospital in Edinburgh. Original research was conducted with children with limbs damaged by Thalidomide. By the 1990s, the partial hand system started to receive international recognition, and Touch EMAS (Edinburgh Modular Arm System) was then established. The name was soon changed to Touch Bionics (Kurmala, 2007).

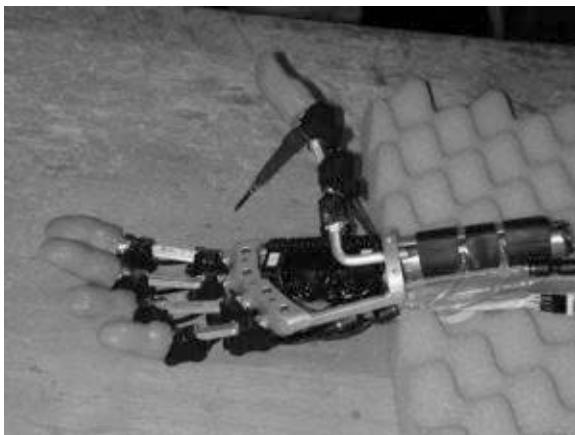
#### 10.9.2.7 Fluidhand

The prototype Fluidhand from Karlsruhe, shown in Figure 10-40, being tested at the Orthopedic University Hospital in Heidelberg, uses hydraulics to drive individual fingers and the thumb. The new hand can close around objects, even those with irregular surfaces. A large contact surface and soft, passive form elements greatly reduce the gripping power required to hold objects. The hand also feels softer, more elastic, and more natural than conventional hard prosthetic devices.

The flexible drives are located directly in the movable finger joints and operate on the biological principle of the spider leg. To flex the joints, elastic chambers are pumped up by miniature hydraulics. In this way, index finger, middle finger, and thumb can be moved

**FIGURE 10-40 ■**

Heidelberg University Fluidhand is driven by hydraulics. (Courtesy of Orthopedic University Hospital in Heidelberg.)



independently. The prosthetic hand also provides feedback, enabling the amputee to sense the strength of the grip.

So far only one patient in Heidelberg has tested both the i-Limb hand and the Fluidhand. “This experience is very important for us,” says Simon Steffen, director of the Department of Upper Extremities at the Orthopedic University Hospital in Heidelberg. The two new models were the best of those tested, with a slight advantage for Fluidhand because of its better finish, programmed grip configurations, power feedback, and more easily adjustable controls. However, this prosthetic device is not yet in production, so the comparison is not really fair.

#### 10.9.2.8 Hands Using Shape Memory Alloy Actuators

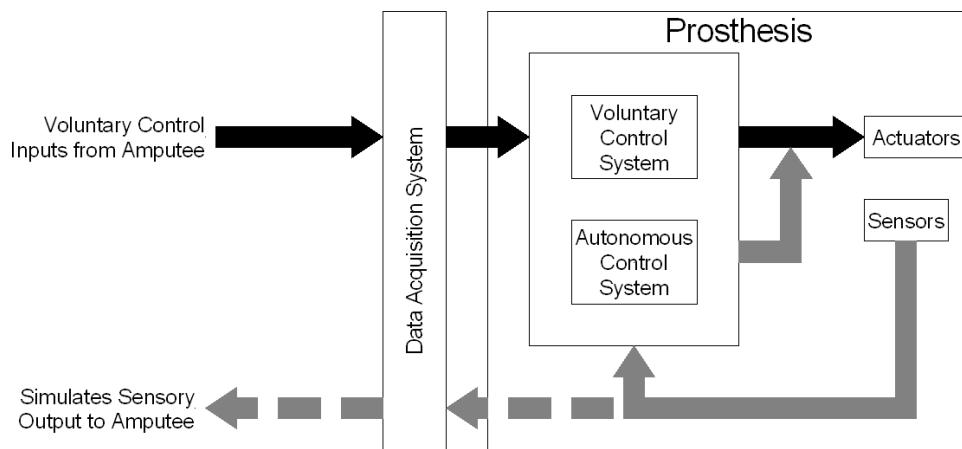
A number of research groups have developed prosthetic hands that use SMA actuation because it is possible to incorporate all of the finger actuators within the palm or the wrist. An example of this is the development of a hand that provides 12 degrees of freedom for four actuated fingers at the Dublin Institute of Technology (O’Toole and McGrath, 2007). Similar research is being conducted at the University of Patras in Greece (Andrianesis and Tzes, 2008) and the University of Victoria in British Columbia (Bundhoo and Park, 2005).

#### 10.9.3 Hand Research and Applications

The anthropomorphic hands described reproduce most of the movements of the human hand and generally provide comparable force output and sensitivity. This means that they can pick up or handle small to medium-sized objects and perform precision tasks, so robots using them can have the versatility of human beings.

Examples of research applications for the shadow hand include situated learning at the University of Bielefeld and grasping at Carnegie Mellon, while NASA’s Robonaut group bought one to inspire their engineers. Other universities are using the hands as a component in their neurological and rehabilitation projects as part of humanoid robots and in many more applications.

Systems using the technology will allow an operator to work in inaccessible areas, isolated by chemical, biohazard, or radiation. Remote control using visual and haptic feedback has the potential of putting a specialist anywhere the hand is.



**FIGURE 10-41** ■ Block diagram showing control structure of an active prosthetic limb.

Finally, these hands are ideal tools for investigations into rehabilitation and assistive technology where safety, flexibility, and compliance are essential.

#### 10.9.4 Control of Prosthetic Arms and Hands

Control of a prosthetic limb can be described by the block diagram shown in Figure 10-41. The input indicated in black allows the user control over basic motions required of the limb. The feedback paths indicated in gray allow the prosthesis to autonomously fine-tune its function based on feedback from local sensors. This functionality is more subtle than voluntary control because it usually goes unnoticed when controlling a natural limb. For instance, when we pick up a glass of water, we do not think about how hard we should grasp. Instead, our nervous system automatically takes care of that for us so that we can hold the glass without breaking or dropping it. The dashed gray arrows in Figure 10-41 indicate a path that allows users to be made aware of the sensory feedback provided by the hand.

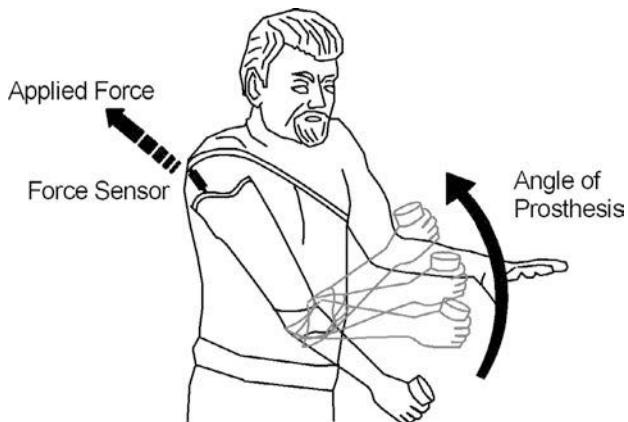
Voluntary control inputs from the body can include outputs from devices that measure gross body motions captured by a harness or cineplasty. These devices include switches and potentiometers, among others. An alternative source of inputs, which is becoming increasingly popular, is myoelectric signals. Simulated sensory outputs can be provided by force (haptic) feedback or through vibrotactile transducers.

##### 10.9.4.1 Microswitches, Force Sensors, and Linear Potentiometers

Harnesses such as those described earlier in this chapter can be designed to operate switches, to extend linear potentiometers, or to apply a force to a strain gauge. The outputs of these transducers are typically read into microcontrollers and then are used to control the servo systems, which actuate the prosthesis. In Figure 10-42, a servo pro force sensor from Motion Control is used to provide commands to control the position of an elbow joint.

Relatively simple transducers are often still used to control state-of-the-art prostheses because they provide a reliable interface that requires little training to operate. For example, the Luke arm can be controlled by the toes using joystick-like controllers embedded into special shoes.

**FIGURE 10-42 ■**  
 Direct relationship of force to elbow positioning can be achieved using a force sensor.  
 [Adapted from (Lake and Miguelez 2003).]



#### 10.9.4.2 Servo Assisted Cineplasty

An extension of this process is the cineplasty discussed earlier. The availability of external power sources and servo assistance has enabled smaller cineplasties to be performed. This creates the possibility of building multiple cineplasties in a group of muscles to control more degrees of freedom. Attaching the externalized muscle to a prosthetic component through a controller that supports extended physiological proprioception (EPP) allows the position, speed, and force of the controlling muscle to be directly correlated to the position, speed, and force of the prosthesis. The physiological sensory feedback inherent in the skin and muscle of the cineplasty informs the amputee of the state of the prosthesis in a subconscious and natural manner.

#### 10.9.4.3 Myoelectric Signals

A myoelectric signal can be defined as the electrical activity produced by a contracting muscle. The characteristics of this signal as measured by surface electrodes are a function of the depth of the muscle, its size, the strength of contraction, and the nature of the overlying tissue. The electrode type, positioning, and orientation also have an effect.

Muscles are made up from fibers arranged in bundles called fascicles. Each fiber is a long cylindrical cell with multiple oval nuclei arranged just beneath the cell membrane (sarcolemma). This sarcolemma exhibits dark and light bands that give it a striated appearance. Each muscle fiber contains a large number of rod-like myofibrils that are the contractile elements in skeletal muscle.

The dark band of the myofibril is composed primarily of the protein myosin, whereas the light band that overlaps it is made from another protein called actin. The contraction process involves chemical interactions between the two fiber types in which the thin actin filaments slide over the thick myosin filaments, increasing the overlap. This results in a shortening of the overall muscle length to between 60 and 70% of its original length. Because muscles cannot actively lengthen, they work in antagonistic pairs.

Several muscle fibers, called a motor unit, are innervated from a single axon whose cell body lies within the spinal cord. The junction between the axon and the muscle fiber is called the end plate, of which there is generally one per fiber. In muscles for fine control there are between 1 and 10 fibers per motor unit, whereas for gross movement there can be more than 1000.

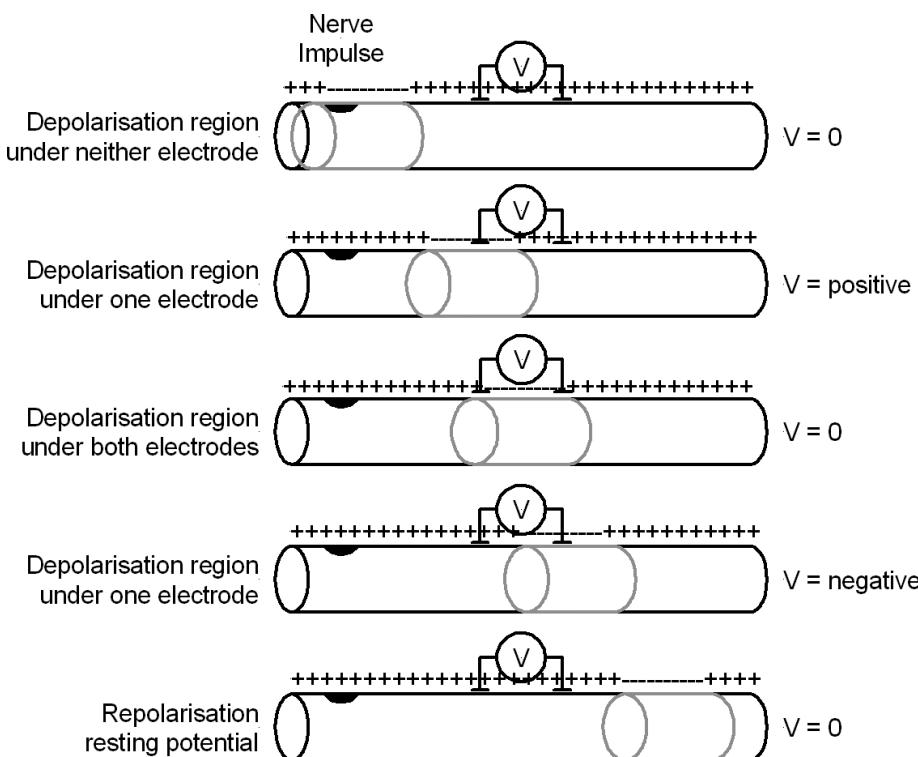
When the nerve stimulus is strong enough (about  $-70$  mV), the characteristics of the muscle cell membrane are altered, and the region around the end plate becomes depolarized. This depolarization spreads up and down the fiber. Sodium ions pour into the cell, and the membrane potential reverses polarity from  $-90$  mV to  $+30$  mV in about  $0.5$  ms. This process is self-regulating, and as the interior of the cell becomes more positive the gradient opposes further influx of  $\text{Na}^+$  ions. As the sodium entry declines, voltage regulated potassium gates open, potassium flows out of the cell, and the cell membrane drops back toward the resting potential.

The arrival of the nerve impulse at the end plate also causes an increase in cellular calcium that initiates the contraction process.

If a pair of electrodes is placed on the outer surface of a single muscle fiber in line with the long axis of the cell, a potential difference will be registered as the depolarization region propagates past the electrodes. As shown in Figure 10-43 and Figure 10-44, this impulse first goes positive and then negative before returning to zero.

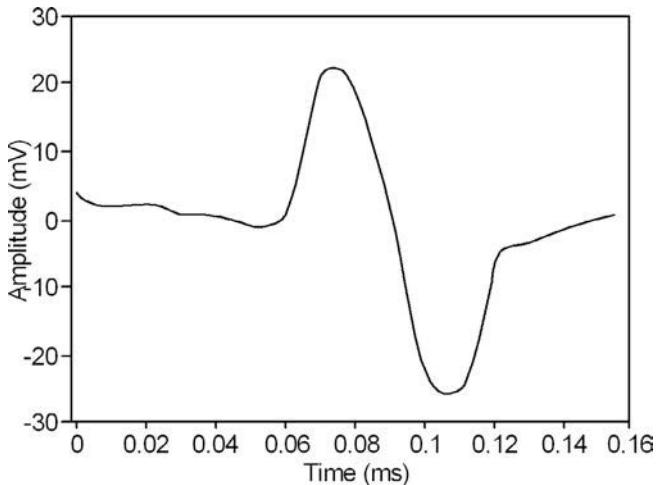
As soon as a number of muscle fibers are involved, even if they are innervated by a single axon, because the end plates are not aligned the spatial phase of the transient depolarization signals will be different and the signals picked up by an electrode pair in proximity will be more complex. For a very light contraction when only a few fibers are involved, it may still be possible to pick out the individual impulses, as shown in Figure 10-45. In this example, three fibers at different depths, shown by the different amplitudes and with different firing rates, have been recorded using surface electrodes.

As the contraction level increases, more fibers from different nerves are triggered, and the firing rate of each speeds up, making the signal more noise-like as shown in



**FIGURE 10-43 ■**  
Propagating action potential along a muscle fiber and the associated voltage across the exterior electrodes.

**FIGURE 10-44** ■  
Typical motor unit action potential (MUAP).



**FIGURE 10-45** ■  
Measured weak myoelectric signal from three fibers at different depths for a very light contraction.  
[Adapted from (Muzumdar 2004).]

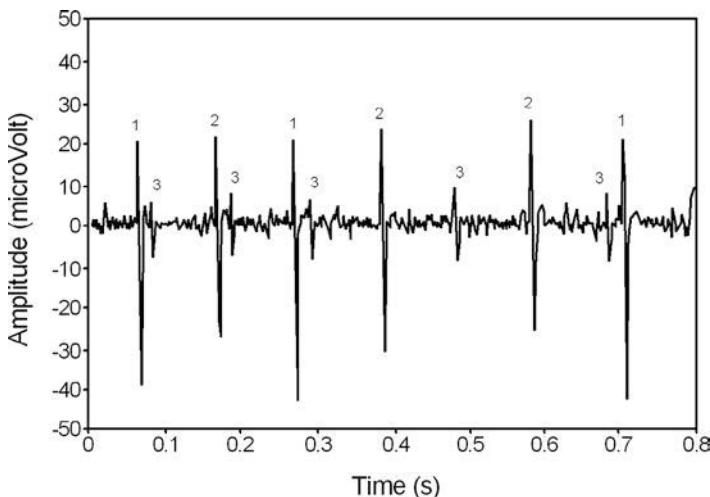


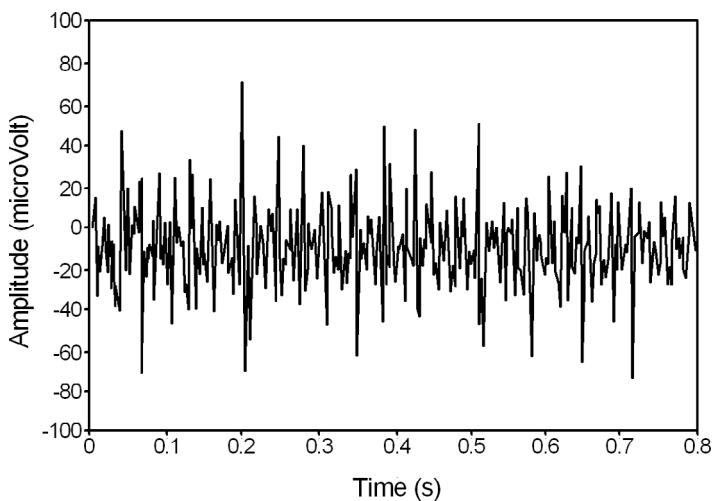
Figure 10-46. This noise has all of the characteristics of shot noise and can therefore be described by the Poisson distribution.

The tension of a muscle contraction is determined by the number of muscle fibers firing and their rate of firing. Fibers become active only when needed, with the small motor units deep within the muscle being activated first to allow for fine motor control. As more force is required the firing rate increases, and finally, at high levels of contraction, larger motor units near the surface are recruited and begin firing.

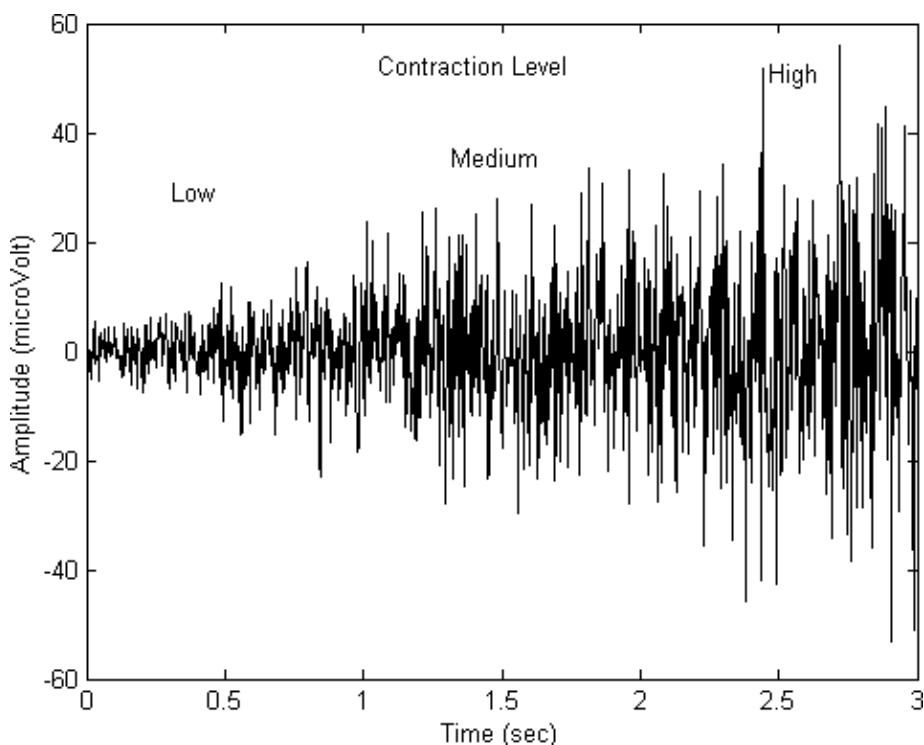
As the firing rate increases and as more fibers close to the surface are activated, the amplitude of the myoelectric signal picked up by a pair of surface electrodes increases dramatically. As shown in Figure 10-47, the signal shows very little structure compared to the previous examples and can be modeled as a white noise source with zero mean.

#### 10.9.4.4 Myoelectric Control

At the end of WWII a few myoelectric control systems were being used by amputees to perform repetitive tasks in the workplace. The prosthesis was hard wired to a large console housing vacuum-tube amplifiers and signal processing electronics. By the early 1960s a



**FIGURE 10-46** ■  
Measured strong (high-level contraction) myoelectric signal for a heavy contraction.  
[Adapted from (Muzumdar 2004).]

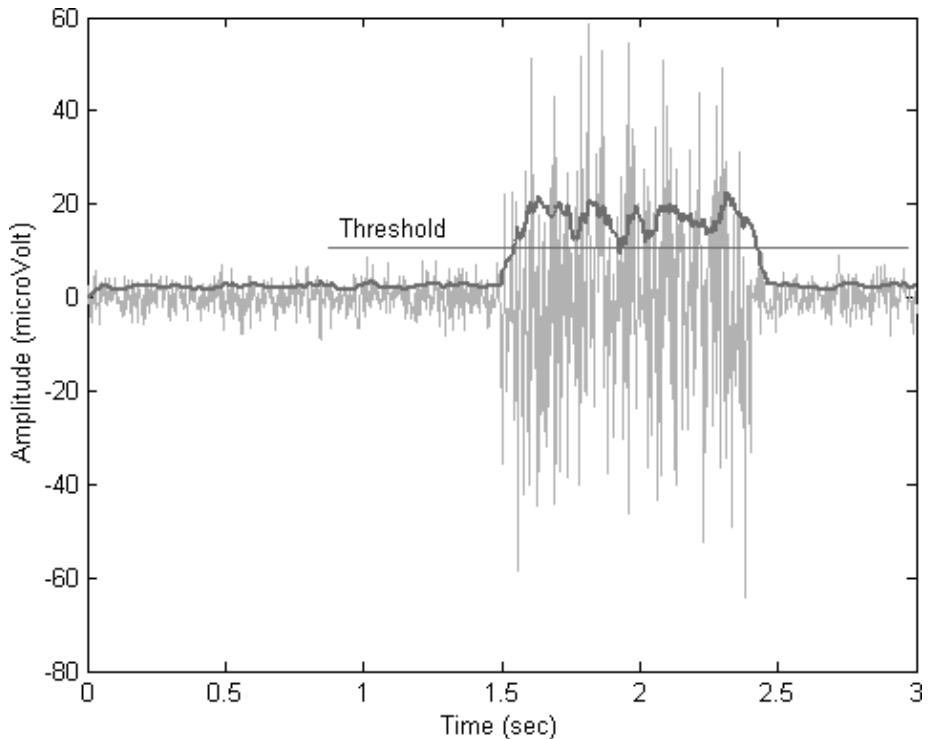


**FIGURE 10-47** ■  
Variation in myoelectric signal level with contraction level as measured by surface electrodes.

Russian scientist, A. Kобринский, had built a completely self-contained battery-operated system that was later licensed for use in Canada. Within a decade systems were starting to address cosmetic issues and contained rechargeable batteries.

The most basic control strategy is based on the amplitude of a single myoelectric signal envelope, as shown in Figure 10-48. The process of obtaining the envelope was discussed in detail earlier in the book and involves rectifying the signal and then passing it through a low-pass filter.

**FIGURE 10-48** ■ Control based on the threshold of the myoelectric signal envelope.



The control of the prosthesis is one in which some action is taken if the envelope exceeds a set threshold. This is known as a one site system and supports two binary states (on-off).

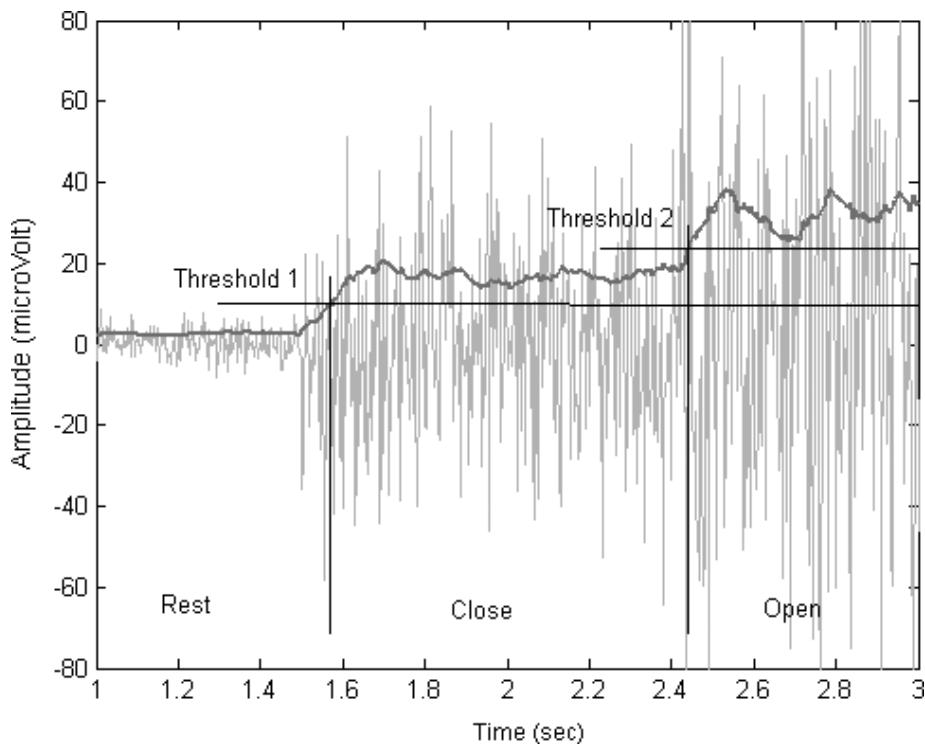
Binary control can be extended by including a number of different threshold levels, which result in different functions. For example, a two-threshold system can encode for three states, as illustrated in Figure 10-49.

To switch through to the open state without activating the closed state requires a small delay in the control strategy. This does delay the response of the prosthesis, but it also reduces the occurrence of operator error.

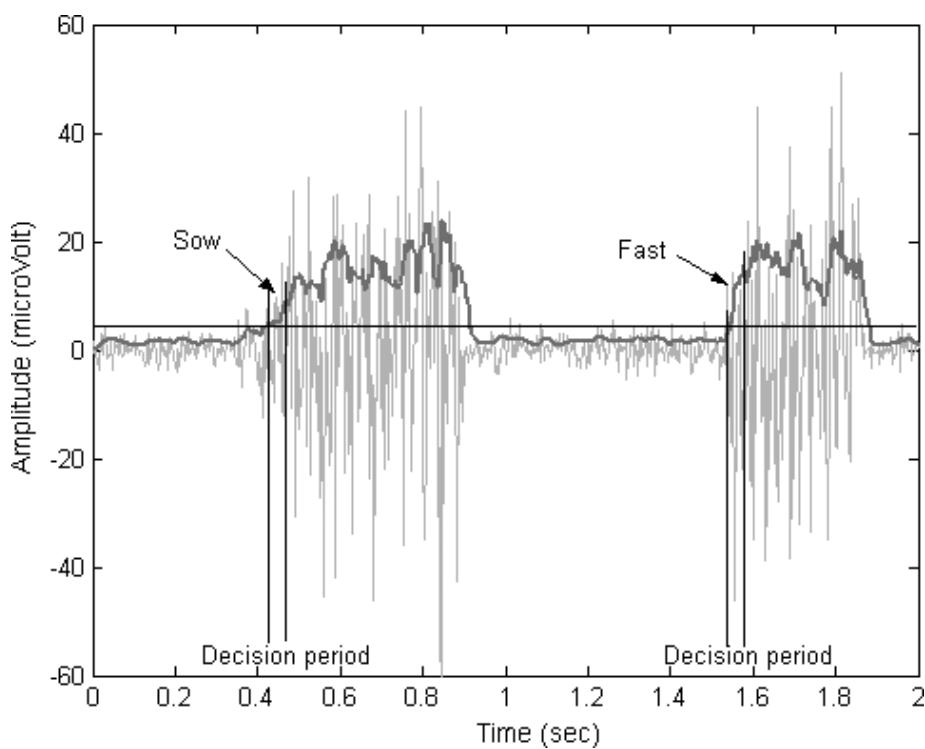
It is obviously possible to go beyond threshold control and to use the actual changes in amplitude of the myoelectric signal envelopes to generate a proportional response in the prosthesis. One implementation of this process is rate coding, in which the rate of change of the envelope voltage is used to select the open or close option of a terminal device, as shown in Figure 10-50. The subsequent signal level can then be used to control the rate of closing or opening. To be effective, the time constant of the low-pass filter used by the envelope detector must be faster than the rate of contraction.

In a two-site system, two pairs of electrodes are used, each with its own signal processor and threshold setting. This supports four binary states, of which the 00 state is generally inactive (rest state), so even with three active states it can be used for much more intricate control.

Binary control can be extended to rate control or proportional control as before. The Utah Arm, shown in Figure 10-51, is a good example of a modern prosthesis controlled by two myoelectric channels. Fast co-contraction of both the myoelectric channels switches control to the elbow, whose speed is then controlled by the difference signal between

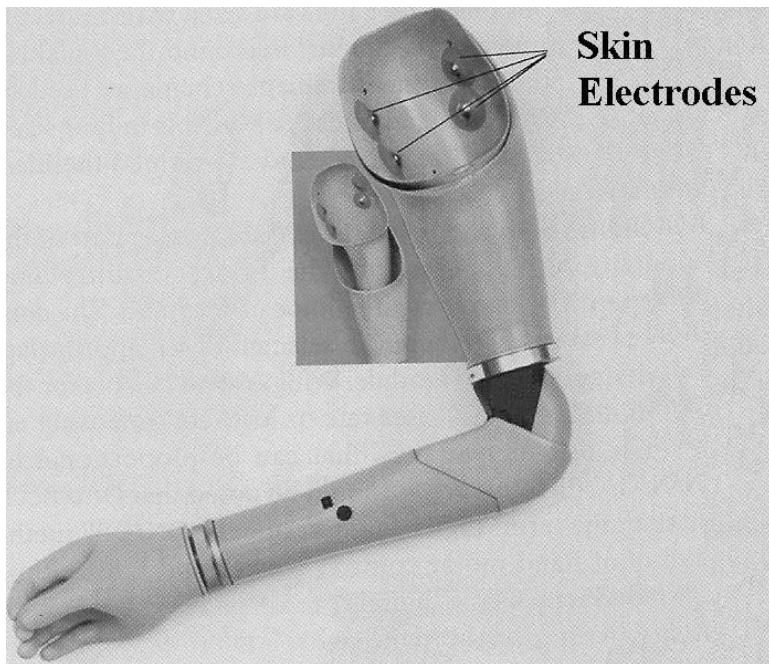


**FIGURE 10-49** ■  
Level coding using  
the myoelectric  
envelope.



**FIGURE 10-50** ■  
Rate coding of a  
myoelectric signal.

**FIGURE 10-51** ■  
The Utah Arm  
showing the  
positions of the  
surface electrodes.  
(Courtesy of Motion  
Control.)



the biceps and triceps channels. On relaxation, the elbow locks and control reverts to the terminal device.

An example of using myoelectric control to flex the elbow and then to extend the elbow before flexing it again is shown in Figure 10-52. This is achieved by maintaining a fairly constant contraction on the biceps and varying the contraction of the triceps.

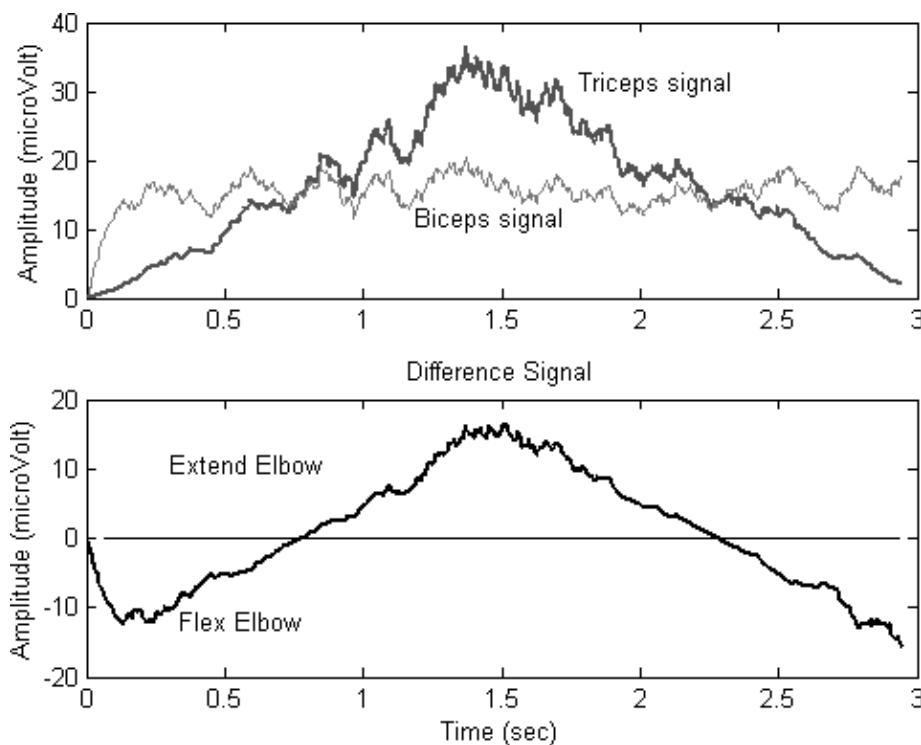
As microcontroller and DSP chips have become less power hungry, new control schemes based on complex pattern recognition have become possible. Research has shown that the contribution of a given muscle within a group varies according to the intended limb action. The sum of the contributions of all of the muscles within the group can be used to identify actions.

Initially pattern-classification-based control systems were limited to simple-to-calculate time-domain signal statistics such as variance, zero crossings, and waveform length to represent the myoelectric signal of interest. Now, more computationally complex feature sets are being investigated, including autocorrelation coefficients, spectral measures, time-series model parameters and time-frequency coefficients based on wavelet and wavelet packet transforms, and higher-order spectral analysis (Hernandez-Arieta, Dermitzakis et al., 2008; Parker, Englehart et al., 2004).

Problems arise with high-level limb amputees as more degrees of freedom must be controlled with fewer available muscle sites. In addition, the remaining muscle sites are not physiologically appropriate as they bear no natural relationship with the lost degrees of freedom. No sophisticated pattern recognition can overcome this problem, and the amputee needs to develop a contrived suite of contractions to control the prosthesis.

#### 10.9.4.5 Targeted Muscle Reinnervation

After amputation, the residual nerves often retain the capability of transmitting and receiving messages but just do not go anywhere. Targeted muscle reinnervation (TMR) is a



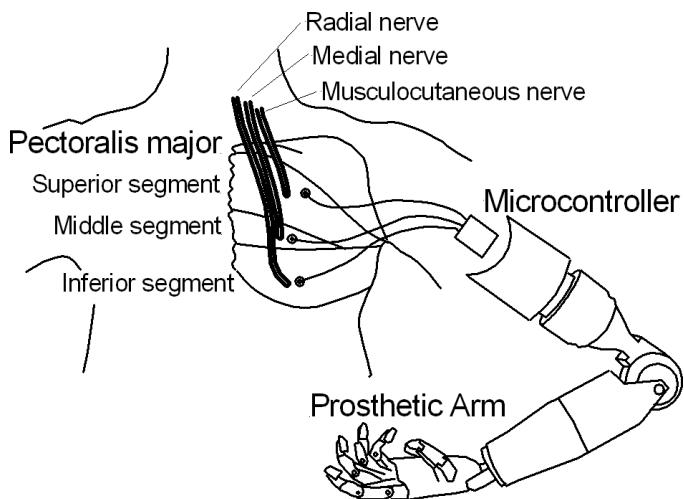
**FIGURE 10-52 ■**  
Proportional control  
of the Utah Arm.

surgical procedure that transfers residual nerves from an amputated limb onto alternative muscle groups that are not biomechanically functional since they are no longer attached to the missing arm. Normally, the nerves travel from the upper spinal cord across the shoulder, down into the armpit, and into the arm. In TMR they are pulled away from the armpit and passed under the clavicle to connect to the pectoral muscles, as shown in Figure 10-53. The reinnervated muscle then serves as a biological amplifier of the amputated nerve motor commands, which are then intuitively coupled to the intended action. The muscle provides the physiologically appropriate surface myoelectric control signals that are related to functions in the lost arm. When the patient thinks about moving the arm and signals travel down nerves that were formerly connected to the native arm but are now connected to the chest, the chest muscles contract in response and the myoelectric signals are sensed by electrodes on the chest, which then control the prosthesis.

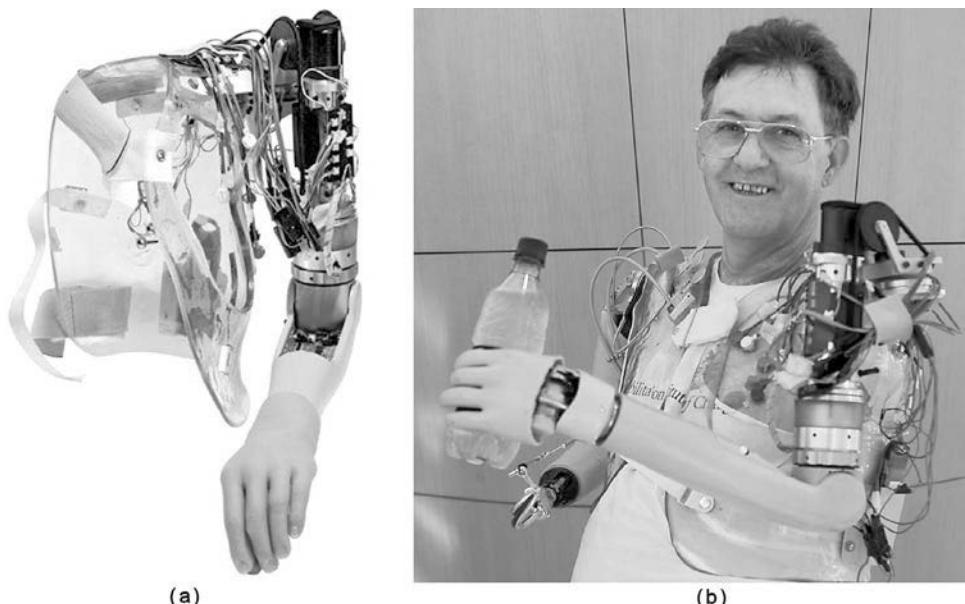
The first person to receive TMR was a 54-year-old male, Jesse Sullivan, who had suffered severe electrical burns working as a high-power lineman. The damage was so severe that it had required that both his arms be amputated at the shoulder. When Sullivan received the TMR surgery, it was to reduce sensitivity from the skin grafts at his amputation sites that were causing him pain and also to give him more control of his prosthetic arm. In a striking discovery, Todd Kuiken's team at the Rehabilitation Institute of Chicago (RIC) found that the reinnervation procedure allowed him to regain sensation in his transferred nerves. When patients are touched on the patch of skin covering the transferred nerves, they feel as if their lost hand is being touched.

Within a few months of the surgery, surface myoelectric signals could be recorded from the reinnervated pectoral segments. In Sullivan's case, TMR allows the musculocutaneous nerve transfer to control elbow flexion, the radial nerve transfer to control elbow extension,

**FIGURE 10-53** ■  
Schematic diagram  
of TMR technique.

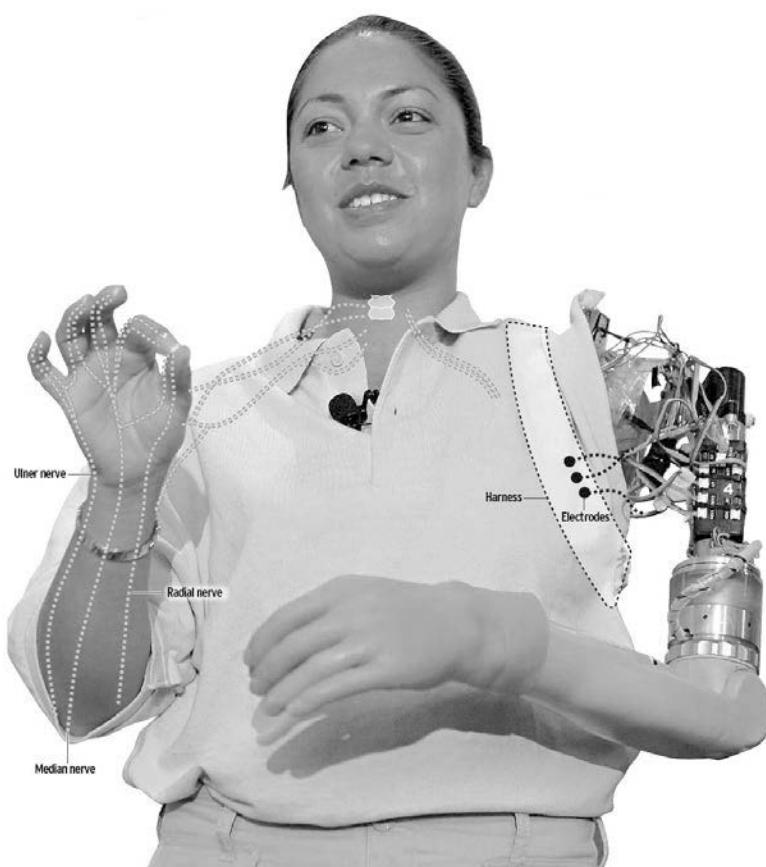


**FIGURE 10-54** ■  
RIC prosthetic arm.  
(a) Hardware detail.  
(b) Arm fitted to  
Jesse Sullivan  
(Courtesy of  
Rehabilitation  
Institute of Chicago,  
with permission.)



the median nerve flexor region to control hand closing, and the median nerve thumb abductor region to control hand opening. Fitted with a sophisticated artificial arm that interfaced with the pectoral muscle group shown in Figure 10-54, Sullivan was able to operate his elbow, wrist, and terminal device simultaneously with greater ease and speed than he could previously (MacIsaac and Englehart, 2006).

In 1995 Dr Kuiken's bionic arm project at RIC was publicly unveiled when Sullivan began using a six-motor version of the arm that provided more dexterity than the original prosthesis. It was also the year that Claudia Mitchell became the first female to receive the TMR surgery (Kuiken, Miller et al., 2007). A total of 3 months after the surgery, she started to feel the reinnervation taking effect. When she thought about squeezing her nonexistent hand she could tell that her chest muscles were working, and when the patch of skin on



**FIGURE 10-55** ■  
Claudia Mitchell with her prosthetic arm.  
(Courtesy of Rehabilitation Institute of Chicago, with permission.)

the left side of her chest was touched it felt as if her hand were being touched. A total of 6 months after the procedure, she was fitted with a prosthesis, as shown in Figure 10-55, and began physical therapy.

Doctors at RIC are now able to map sensitive spots on the chest to specific parts of the missing fingers and hands and have discovered that the reinnervation process allows the patients to feel heat and cold as well as pain (Adee, 2007).

When RIC presented her as the first bionic woman in 2006, she received a great deal of media attention, but for her the focus has always been on the research. “Whether it’s an engineer or a senator, the more people realize what we are doing and the need that exists, the better,” she says (Young, 2007).

DARPA wants amputees to have all this capability, especially those who have lost limbs while serving in the military. In 2006 it launched the Revolutionizing Prosthetics 2007 and 2009 initiatives. To date it has given a combined \$50 million in grant funding to the Johns Hopkins University APL and DEKA Research and Development Corp. to bring the “complexities of biology into the world of engineering” (Young, 2007).

#### 10.9.4.6 Injectable Myoelectric Sensors

Driven by the requirement to improve the quality of signals coming from the residual muscles, Richard Wier has been working on an injectable myoelectric sensor (IMES). The IMES are encapsulated cylinders about 2 mm in diameter and 12 mm long that, when

injected into the muscle of the residual limb, send signals wirelessly to a surface receiver to control the prosthesis. Because they can be placed within individual muscle bundles, the signals obtained offer much improved discrimination (Weir, Troyk et al., 2009).

#### 10.9.4.7 Feedback

One of the problems with using myoelectric signals in isolation is that there is no proprioception or tactile feedback. This requires that the amputee look at the prosthesis to gauge position and force. One alternative is to use auditory feedback (Smith, 1990) or a vibrotactile or electrotactile element (known as a tacter) secured against the user's skin to convey some information about the angle of a joint or the force applied by the terminal device. A good example is the pressure sensor on the Luke hand that generates a signal proportional to the grip strength. This signal controls a tacter that vibrates slightly when the grip is light, and as the user's grip tightens the frequency of the vibration increases. This enables a user to pick up and drink out of a flimsy paper cup without crushing it or to firmly hold a heavy cordless drill without dropping it. System dynamics should be considered when stimulating skin sensors. A response time of a few hundred milliseconds is typical from the application of the stimulus to the user registering the effect. Additionally, the skin becomes desensitized to continuously applied stimuli after a few minutes.

For grasping, it is impractical to provide feedback from each finger, so prosthetic hands like the Gufu III include velocity control of each joint until contact, as described by

$$E_i = K_{Vi}(\dot{\theta}_i - \dot{\theta}_{di}) \quad (10.13)$$

where  $E_i$  is the motor input,  $K_{Vi}$  is the velocity feedback gain, and  $\dot{\theta}_{di}$  is the required rotation rate of the  $i$ -th joint.

After contact, the grasping force on each link controls the rotation rate of the adjacent joint,

$$\dot{\theta}_{di} = -K_{Pi}(F_i - F_{di}) - K_{Ii} \int (F_i - F_{di}) dt \quad (10.14)$$

where  $K_{Pi}$  is the proportional force feedback gain,  $K_{Ii}$  is the integral force feedback gain, and  $F_{di}$  is the required force of the  $i$ -th link.

This control strategy ensures that the object is grasped quickly and that the grasping force is distributed reasonably uniformly across the links of each of the fingers.

A recent development is to use reinnervation of the pectoral segment to provide sensory information directly to the nerves that used to carry the sensory signals from the hands. However, though this technique can provide some tactile feedback it is not yet sufficiently advanced to provide the required proprioception for reliable control of a prosthesis.

#### 10.9.5 Leg Mechanisms

A prosthesis that combines intelligence and motorized actuation has the ability to regenerate the correct gait kinematics. This process involves programming the knee to execute normal gait dynamics during all phases of the cycle. It must also be capable of meeting other activities for which passive prosthetics are ill suited. These include climbing and descending stairs or simply sitting down or standing up (Pons, 2008). Advances in microprocessor speed, available memory, and low-power operation in conjunction with developments in artificial intelligence are now making it more feasible to produce

intelligent transfemoral prostheses (Dehghani, 2010). How this is achieved is described in the following section.

**Initial swing:** An active knee prosthesis contains a sensor that monitors knee flexion and a motor that can drive the knee at up to  $300^{\circ}/\text{s}$ . This flexion removes the foot from the ground and reduces the requirement for hip hiking, vaulting, and circumduction. Foot lift also improves safety on uneven terrain by reducing the likelihood of snagging.

**Mid-swing:** As the user continues to flex the hip to execute leg advancement, the active prosthesis increases knee flexion up to  $60^{\circ}$  to improve ground clearance still further. Following a short pause in mid-swing, the knee starts to extend at a speed comparable with the initial swing flexion. This transition into swing extension is required to advance the lower leg. During this phase the motor replaces the function of the quadriceps and hamstring to accelerate and decelerate the lower leg to produce smooth motion. The characteristics of the forward pendulum motion maintain the correct relationship between the leg and the center of mass of the body, which improves energy efficiency. In addition, users become confident that the knee will reach full extension, so they become less aware of its function. In addition, an active swing phase goes some way in overcoming the resistance of obstacles in the path of the foot and thus reduces the likelihood of tripping.

**Terminal swing:** The active prosthetic replaces the deceleration function of a nonamputee's hamstring, ensuring that the extension is properly controlled. The knee is positioned with a flexion of about  $5^{\circ}$  that is correct for both shock absorption and stability.

**Initial contact and loading response:** As the foot touches the ground, flexion angles are allowed to increase under software control to between  $5^{\circ}$  and  $15^{\circ}$  to absorb the shock. This spring-like action contributes to the user's feeling of stability without requiring the forceful extension of the prosthetic knee joint.

**Standing up from a chair:** The active prosthesis applies the same amount of force as does the good knee to maintain symmetry and to reduce strain on the sound limb.

**Climbing stairs:** With a passive prosthesis the user can only lift the mechanism to the level of the good leg. However, with an active prosthesis, the knee identifies that the user is climbing stairs, and it drives the knee with the same force and speed as the sound leg to move the user up to the next step.

### 10.9.5.1 Control Strategies

Control strategies for intelligent passive leg prostheses involve some form of state machine that can adjust of the damping of the knee joint during the different gait phases for different walking speeds. Early work, starting in the 1970s, was based on "echoing" the actions of the sound leg to control the prosthetic one. Myoelectric-based control came into vogue in the 1980s, and, of course, fuzzy logic has been tried.

Ideally, however, leg control should be based only on parameters that can be measured on and around the prosthetic. The C-Leg detects angle, ankle force, and torque, which it uses to calculate the required damping for flexion and extension during the swing phase and in addition offers damping control during stance. The system is not user adaptive, and a trained clinician programs the various damping levels until the user is comfortable. The

knee cannot adapt to changes in the terrain or changes in gait if the user were to carry a backpack, for example. More recently, systems have been developed that can adapt the knee state to changes in the environment. The Rheo-Knee from Össur was developed as a result of research conducted by Herr a few years ago (Herr and Wilkenveld, 2003).

### 10.9.5.2 Knee Prosthetics

The two main concerns in prosthetic leg design are a healthy, natural walk and safety. Amputees actually fall quite often because they cannot feel their foot and do not know what the knee joint is doing because effective proprioception is limited. In the past, safety concerns were not compatible with a natural movement, which is fundamental to diminish the negative effects of uneven posture and the resulting back pain. However, this has changed with the introduction of intelligent passive prostheses in the 1970s and active knee and leg prosthetics only a decade later. Examples of these intelligent knee prostheses are shown in Figure 10-56.

An example of an intelligent passive prosthesis is the C-Leg® by Otto Bock. First introduced in 1997, it struck a balance between these two needs. It has had numerous revisions, and new models have been created, with the most recent being the C-Leg compact, shown in Figure 10-57.

Strain gauges in the tube adaptor and a knee-angle sensor are used to determine the ankle moment above the foot adapter as well as the angle and angular velocity of

**FIGURE 10-56 ■**

Examples of knee prostheses.

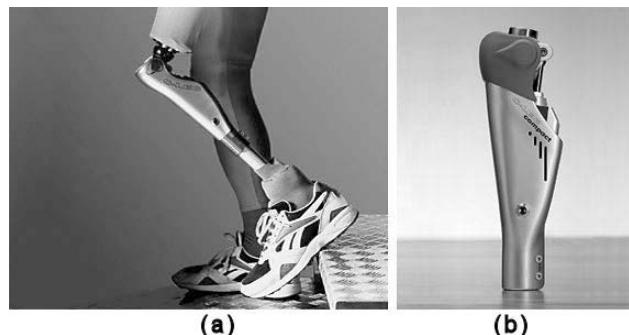
- (a) Intelligent passive prostheses.
- (b) Active knee prosthetic. (Courtesy of Endolite, Otto Bock, and Össur, with permission.)



**FIGURE 10-57 ■**

The C-Leg range by Otto Bock is a state-of-the-art actively controlled knee prosthesis.

- (a) Fitted to a human patient.
- (b) Prosthesis hardware. (Courtesy of Otto Bock, with permission.)



the knee joint every 20 ms. The microprocessor uses this information to determine the characteristics of the gait, including stride length and frequency, from which it can predict the current phase of the gait cycle and then adjust itself in real time. Whether sitting in a chair or walking on an uneven surface, such as a slope or stairs, the mechatronic hydraulic stance phase safety system is active, stabilizing the joint from heel impact right up to the point when it switches precisely to the hydraulically controlled swing phase.

Power comes from a lithium ion battery that, depending on activity, runs for between 40 and 45 hours between charges. It can be recharged from the mains or using the 12 V supply from a car lighter socket.

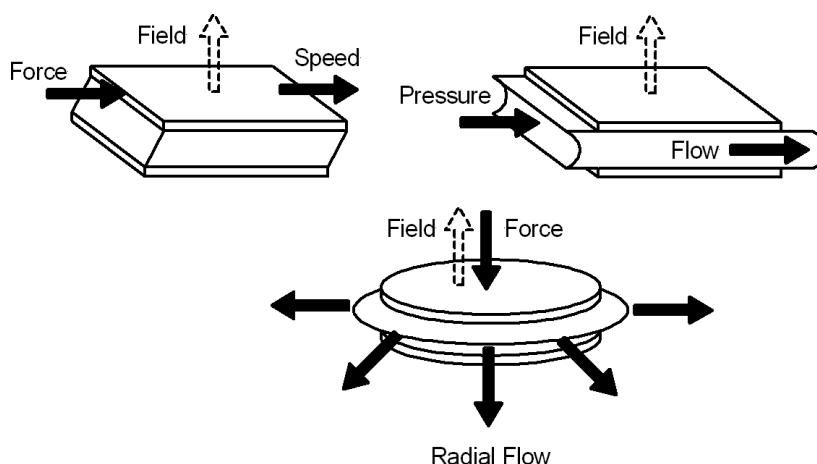
A wireless remote control is provided with the leg that allows the user to change the mode of the prosthesis to a user defined setting. This can be one that holds the knee at a specific flexion to accommodate long periods of standing, or it could be a highly dynamic mode to accommodate a rigorous exercise regime.

Compared with the \$20 price tag of the Jaipur knee, the C-Leg, with a cost of about \$70,000, is definitely not aimed at providing improved locomotion for Third-World amputees.

Electrorheological (ER) and magnetorheological (MR) fluid actuators are considered to be semiactive because they cannot provide motive power but dissipate it only in an active manner. They are field responsive fluids whose rheological properties (specifically viscosity) are altered dramatically when subjected to an external electric or magnetic field. A MR fluid contains small iron particles about  $1\text{ }\mu\text{m}$  in diameter suspended in an oil. When a magnetic field is applied the particles form chains that stiffen the fluid. Varying the magnetic field by adjusting the current through the coils of an electromagnet adjusts the stiffness level.

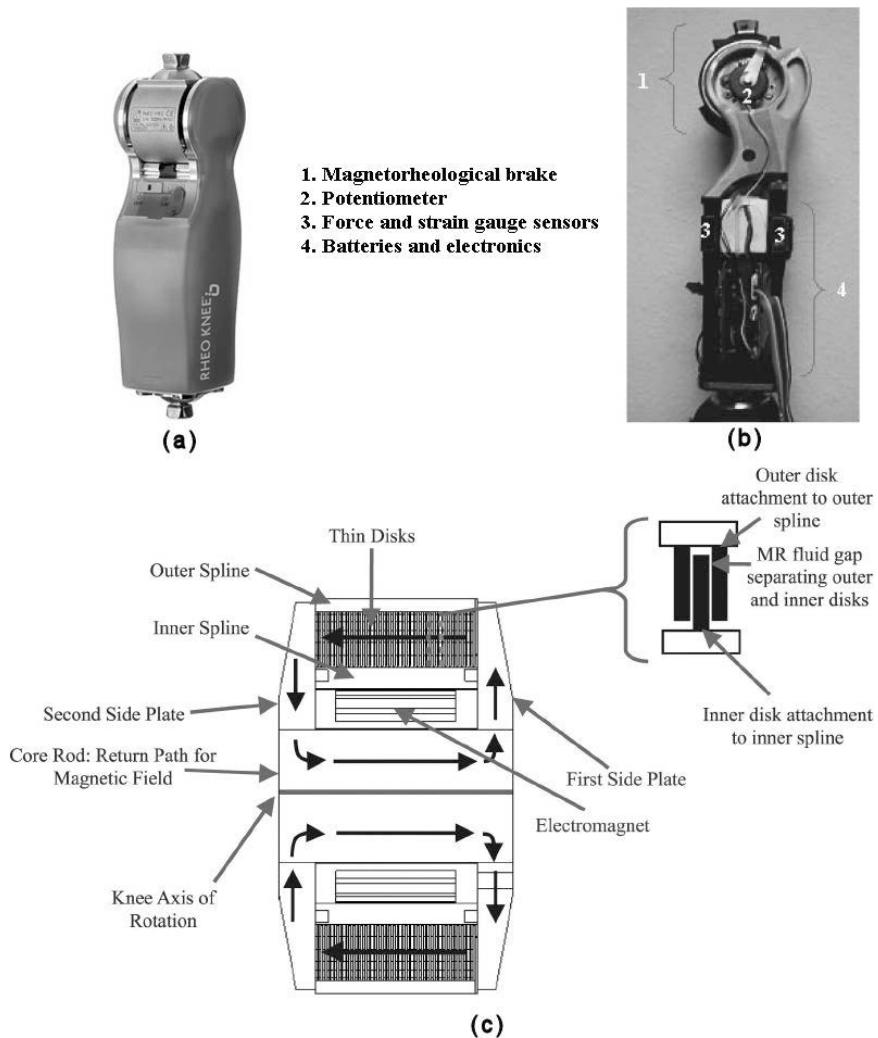
One of three different operating modes—shear mode, flow mode, and squeeze mode—for the actuators can be exploited, as shown in Figure 10-58.

The Rheo knee by Össur, shown in Figure 10-59, is a state-of-the-art actively controlled prosthetic based on a braking system that uses a MR fluid in the shear mode as the primary torque coupling strategy. In this case, the fluid is contained between closely spaced disks that are free to rotate relative to each other. As the electromagnet is progressively energized, the fluid becomes more and more viscous, which increases the stiffness of the knee joint. Sensing consists of a potentiometer to measure the flexion angle and the differentiated signal to provide angular rate. Two aft and two fore strain gauges measure the force applied



**FIGURE 10-58 ■**  
Possible operating modes for ER and MR fluids. [Adapted from (Moreno, Bueno et al., 2008).]

**FIGURE 10-59 ■**  
 Construction of the MR Rheo knee from Össur. (a) Prosthesis hardware.  
 (b) Identification of individual components  
 (c) Schematic showing how the rheological mechanism operates. (Herr and Wilkenveld 2003.)

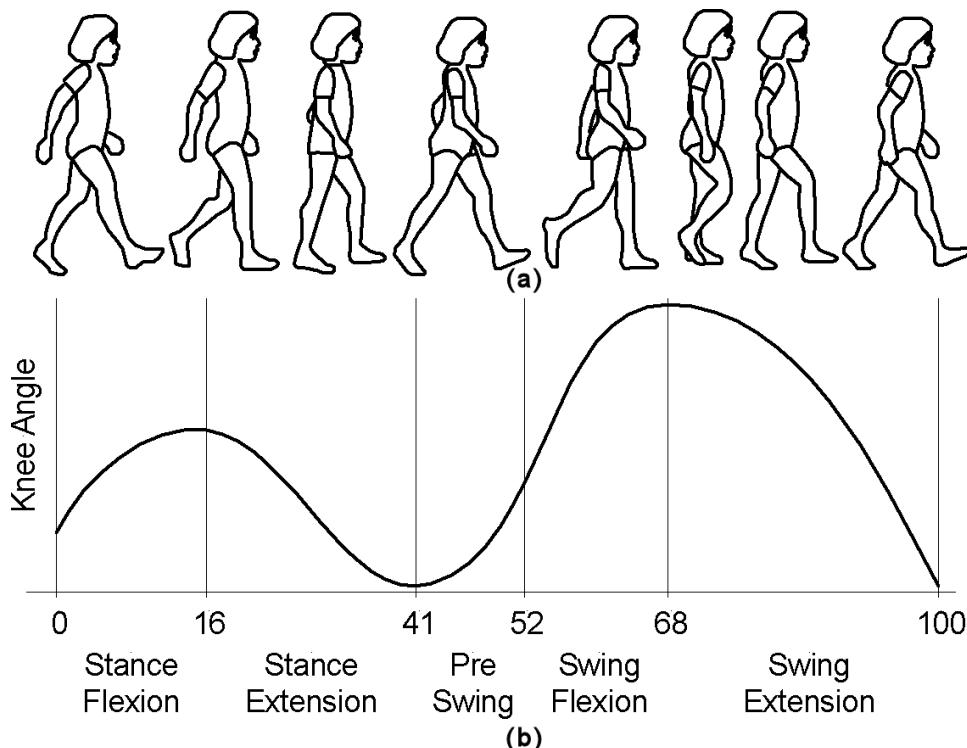


to the knee from the ground in the direction of the longitudinal axis when the signals are added. Subtracting the strain gauge signals provides a measure of the knee torque.

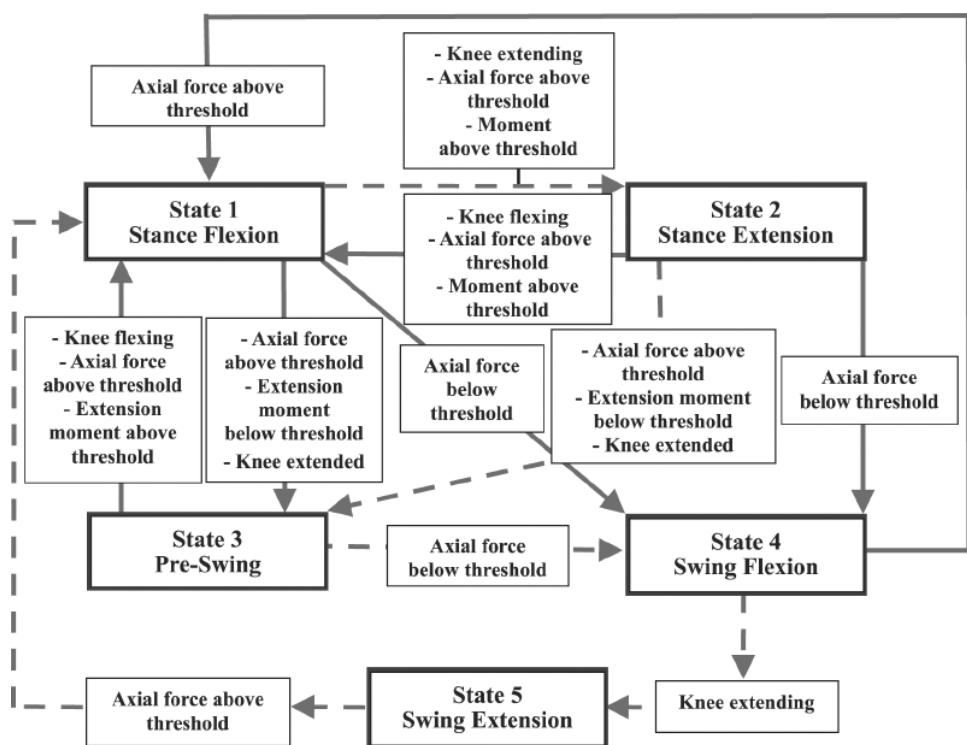
Five critical states were identified, and a state machine was developed to determine the requirements for transition between the states, as shown in Figure 10-60 and Figure 10-61.

**State 1: Stance flexion (0–16%):** The prosthetic knee applies a high level of damping to inhibit the knee from buckling under the user's weight. Initially this starts out at the maximum to ensure that buckling does not occur, but it is progressively reduced to a value proportional to the peak axial force during the stance period.

**State 2: Stance extension (16–41%):** High damping is applied to damp knee extension to stop the knee hitting its extension end stop. Flexion and extension damping is dependent on the size of the amputee. Initially this starts out low to ensure that the knee extends completely, but it is progressively increased to a value proportional to the peak axial force as in State 1.



**FIGURE 10-60** ■ Normal walking.  
(a) Graphic showing leg positions.  
(b) Changes in knee angle with the different gait phases.



**FIGURE 10-61** ■ State transition diagram for an actively controlled knee prosthesis.  
(Herr and Wilkenveld 2003.)

**State 3: Preswing (41–52%):** The current in the electromagnet is reduced to zero so that the only damping is due to the viscous oil between the plates.

**State 4: Swing flexion (52–68%):** Damping starts out at the minimum and increases progressively during the amputees first few strides until it reaches a maximum allowable flexion of 70° during this phase. This makes the gait look natural.

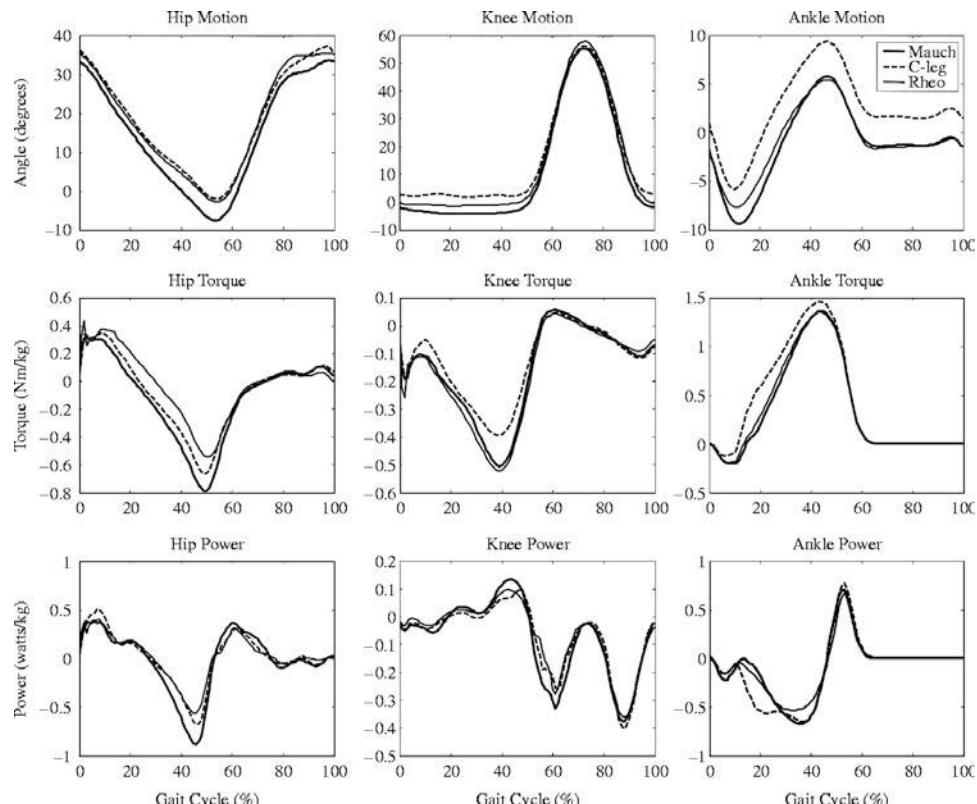
**State 5: Swing extension (68–100%):** Damping also starts out at the minimum and gradually increases but remains proportional to the damping setting for State 4.

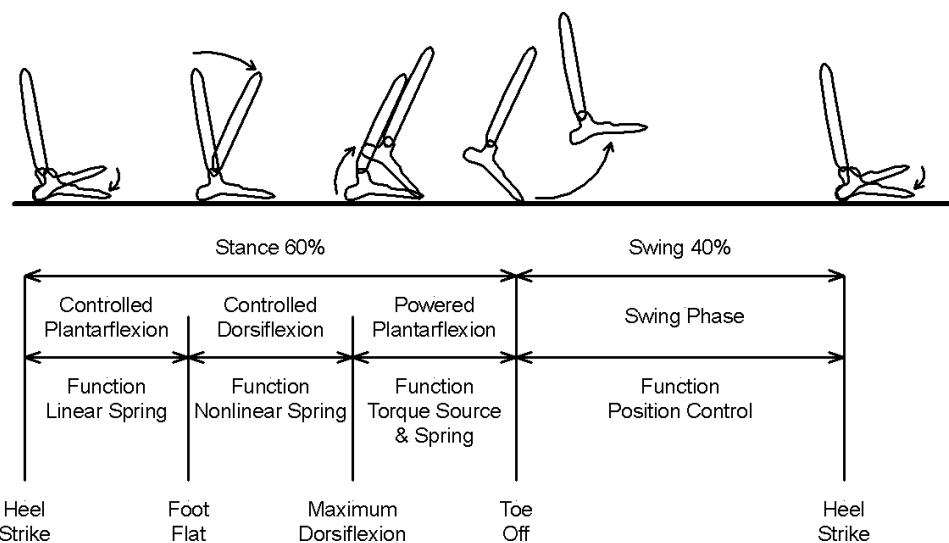
A number of researchers have performed detailed comparisons between similar prosthetic knees. An interesting clinical comparison between variable-damping and mechanically passive knee prostheses examined the C-Leg from Otto Bock, the Rheo knee from Össur, and the Mauch leg. It found that oxygen consumption by amputees using the Rheo knee was 3% lower than for those using the C-Leg, and peak hip torques were lower as well (Johansson, Sherrill et al., 2005). A comparison among angle, torque, and power for the three prostheses is shown in Figure 10-62.

### 10.9.6 Ankle–Foot Mechanisms

Transtibial amputees fitted with passive prostheses generally experience problems such as nonsymmetrical gait patterns and slower walking speeds as well as higher gait metabolic speeds. These are typically 20 to 30% higher than those of able-bodied individuals (Dehghani, 2010). The human ankle performs more positive mechanical work than

**FIGURE 10-62 ■**  
Performance comparison of a number of intelligent passive knee prostheses.  
(Johansson, Sherrill et al., 2005.)





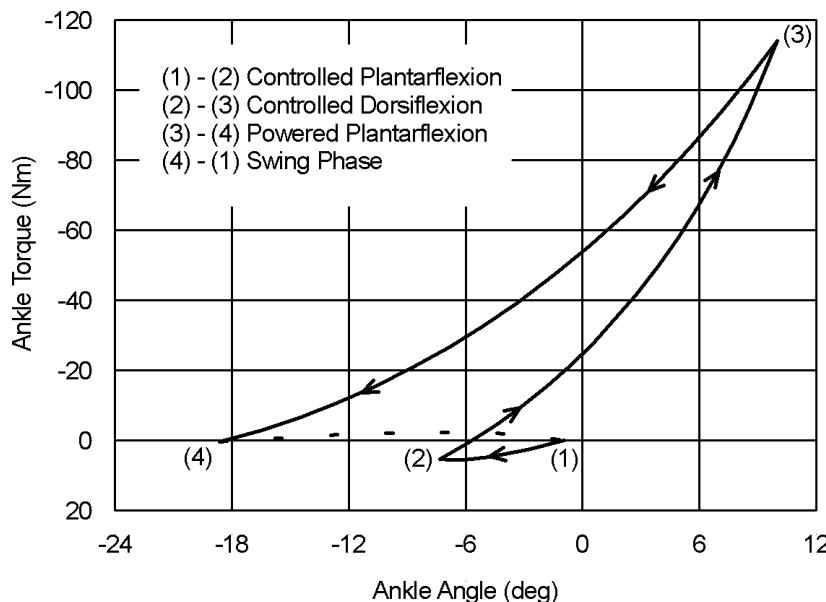
**FIGURE 10-63** ■  
Ankle–foot  
relationship during  
normal walking.  
[Adapted from (Au,  
Weber et al., 2007).]

negative, but passive prostheses do not yet have the ability to reach this state (Au, Weber et al., 2007).

An ankle–foot prosthesis should mimic the normal ankle–foot relationship, shown in Figure 10-63.

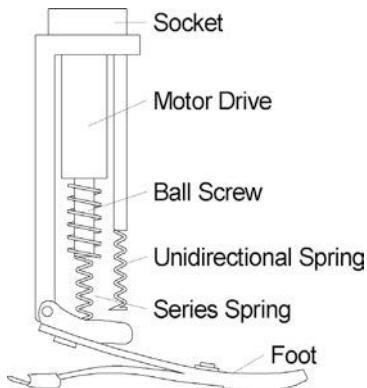
This can be used to provide a model for the angle, torque, power, and stiffness requirements for the prosthesis, as shown in Figure 10-64, for the torque component.

A powered ankle–foot prosthesis based on these torque–angle requirements can be constructed using a powerful linear ball-screw drive, a series elastic actuator, a number of springs, and a carbon composite leaf spring prosthetic foot of the kind shown in Figure 10-18. A schematic of the device is shown in Figure 10-65.

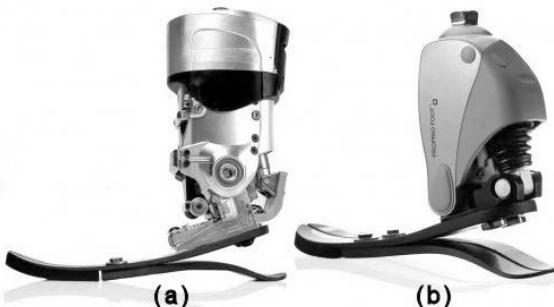


**FIGURE 10-64** ■  
Torque–angle plot  
for a 75 kg person.  
[Adapted from (Au,  
Weber et al., 2007).]

**FIGURE 10-65** ■  
Mechanical design  
of a powered  
ankle-foot  
prosthesis. [Adapted  
from (Au, Weber  
et al., 2007).]



**FIGURE 10-66** ■  
Examples of active  
prosthetic feet.  
(a) PowerFoot from  
iWalk Inc. (b) Proprio  
Foot from Össur.  
(Courtesy of iWalk  
and Össur, with  
permission.)



Examples of the state-of-the-art in active foot prosthetics are the PowerFoot® and the Proprio Foot™ devices, shown in Figure 10-66. The PowerFoot is the latest lightweight prosthesis developed by iWalk Inc. It integrates three microprocessors and 10 environmental sensors to evaluate and adjust ankle position, stiffness, and damping in real time. The Proprio Foot is a motor-powered foot from Össur with an integrated motion sensing system that detects and adapts to terrain changes in real time. The active ankle motion identifies slopes and stairs after the first step and instructs the ankle to flex appropriately and allows wearers to more easily sit down or rise from a chair.

In even the most advanced passive foot prosthetics, some energy is lost during each stride so less energy is provided at push-off than was absorbed during contact. However, in these active prostheses, power assist provides additional spring at push-off to reduce energy expenditure by the amputee.

## 10.10 | PROSTHESIS SUSPENSION

One of the major problems with any prosthesis is the fact that it cannot easily be attached to the articulated support structure (skeleton) to which the original limb was attached. This requires that the full weight of the device and any applied loads be supported by soft tissue.

Often the contact between the soft tissue on the limb and the prosthetic attachment causes problems such as blisters, cysts, edema, and other skin irritations. Another common problem is that the weight load on a prosthetic limb can be painful and can restrict blood

flow limiting use or the time the patient can be mobile. To a lesser extent, the straps and belts that hold the prosthetic limb on can be uncomfortable, and, of course, they are a hassle.

### 10.10.1 Conventional Suspension Methods

Various conventional suspension methods with their associated advantages and disadvantages are described in Table 10-5.

DEKA engineers have investigated this problem and come up with a good solution. In their design, the inside of the socket is coated with a mosaic of thin air bladders that can be individually filled with air to offer padding and rigidity necessary to support the prosthesis during normal activity. When the arm is not in use the system deflates to release pressure on the soft tissue. An added capability is its ability to alternately fill and empty to offer a massage effect.

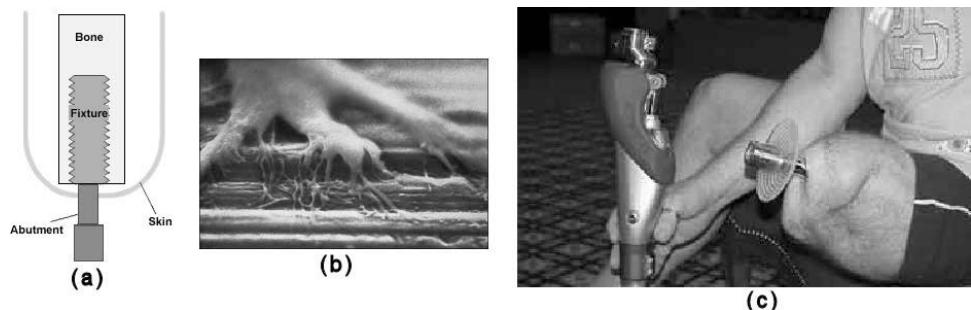
### 10.10.2 Osseointegration

Osseointegration is the permanent incorporation of a nonbiological component into bone. It has been used to facilitate the attachment of bone anchored hearing apparatus (BAHA) devices, as discussed in Chapter 6, and it can be used to attach artificial limbs. A titanium

**TABLE 10-5 ■ Suspension Methods for Prostheses**

Suspension	Indications	Advantages	Disadvantages
Harness	Figure-8 Transradial Transhumeral Light to normal activities	Simple, durable, adjustable	Axillary pressure produces discomfort
	Shoulder saddle and chest strap Transradial Transhumeral Heavy lifting	Greater lifting ability, more comfortable than figure-8 harness	Reduced control compared with figure-8 harness, difficult to adjust in women because straps cross breasts
Self-suspending	Muenster Northwestern Supracondylar	Wrist disarticulation Elbow disarticulation Short transradial Myoelectric transradial	Ease of use  Limited lifting capacity compared with harness systems, compromised cosmesis, reduced elbow flexion
Suction	Suction socket with air valve Transhumeral with good soft tissue cover	Secure suspension, elimination of suspension straps	Requires stable residual volume, harder to put on than other suspension systems
	Gel sleeve with locking pin Transradial Transhumeral Compromised limbs with scarring or impaired skin integrity	Accommodate limb volume change with socks, reduced skin shear	Greater cleaning and hygiene requirements, can be uncomfortable in hot climates

*Source:* Kelley, B., P. Pangilinan et al., Emedicine. Retrieved October 2009 from <http://emedicine.medscape.com/article/317234-print>, 2009, with permission.



**FIGURE 10-67** ■ Osseointegration. (a) Schematic. (b) Scanning electron microscopy (SEM) of bone fusing onto titanium. (c) Photo of attachment for artificial knee. (Courtesy of Össur, with permission.)

stud, machined to have threads much like a screw at one end, is inserted into the cavity of one of the long bones in the arm or leg at the site of the amputation as illustrated in Figure 10-67. Living bone cells migrate into the threads and attach themselves firmly to the titanium so that after about 6 months the insert is firmly anchored.

A second titanium component called an abutment is attached to the implant, and it protrudes through the skin at the end of the amputee's stump, as shown in Figure 10-64. The limb can be attached to this abutment using a wrench or a quick-release coupling. The main complication is that the abutment permanently pierces the skin, which increases the possibility of infection (Mooney, 2001). However, this is a small price to pay for the advantages of this method suspension, as listed:

- No feeling of increased weight.
- No lack of positive prosthesis control.
- No increase in energy consumption to operate the prosthesis.
- No need to bear weight on other parts of the body.
- No fitting problems due to weight gain or loss.
- No socket induced skin irritation.
- No problems attaching or removing the prosthesis.

Because the procedure is still experimental, it is generally performed only on younger unilateral amputees who cannot be successfully fitted by conventional means. Additionally, candidates must have a sound bone in the residual limb, not have systemic diseases such as diabetes or peripheral vascular disease, and not be smokers.

## 10.11 REFERENCES

- Adee, S. (2007). "Artificial Arm Researchers Restore Feeling of Missing Limb." *IEEE Spectrum* 45(12).
- Adee, S. (2008). "Dean Karmen's 'Luke Arm' Prosthesis Readies for Clinical Trials." *IEEE Spectrum: Biomedical/Bionics*, February.
- Andrianesis, K. and A. Tzes. (2008). "Design of an Anthropomorphic Prosthetic Hand Driven by Shape Memory Alloy Actuators." *2nd IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics*, 2008.

- Au, S., J. Weber, and H. Herr. (2007). Biomedical Design of a Powered Ankle-Foot Prosthesis. *10th International Conference on Rehabilitation Robotics, ICORR2007*.
- Bundhoo, V. and E. J. Park. (2005). Design of an Artificial Muscle Actuated Finger towards Biomimetic Prosthetic Hands. *12th International Conference on Advanced Robotics, 2005 Proceedings*.
- Burke, M., V. Roman, and V. Wright. (1978). "Bone Changes in Lower Limb Amputees." *Annals of Rheumatic Diseases* 37(3): 252–254.
- Castillo, S. (2007). "Prosthetic Arm Developed by a Mexican Engineer." Retrieved October 2009 from <http://www.unboundedmedicine.com/2007/01/24/prosthetic-arm-developed-by-a-mexican-engineer/>
- Dehghani, A. (2010). "Intelligent Prosthesis—A Biomechatronics Approach." In *Mechatronics in Action*, D. Bradley and D. Russell (Eds.). London: Springer.
- Forner-Cordero, A., J. Pons, E. Turowska and A. Schiele. (2008). "Kinematics and Dynamics of Wearable Robots." In *Wearable Robots—Biomechatronic Exoskeletons*, J. Pons (Ed.). Chichester, UK: Wiley.
- Greig, D. (2009). "A \$20 Prosthetic Knee to Bring Relief to Disadvantaged Amputees." Retrieved June 2010 from <http://www.gizmag.com/a-20-prosthetic-knee-to-bring-relief-to-disadvantaged-amputees/11514/>
- Healy, J. (1991). *Natural History: A Selection, by Pliny the Elder*. London: Penguin Classics.
- Hernandez-Arieta, A., K. Dermitzakis, D. Damian, M. Lungarella and R. Pfeifer. (2008). "Sensory-Motor Coupling in Rehabilitation Robotics." In *Handbook of Service Robotics*, Y. Takahashi (Ed.). Vienna, 21–36.
- Herr, H. and A. Wilkenveld. (2003). "User-Adaptive Control of a Magnetorheological Prosthetic Knee." *Industrial Robot: An International Journal* 30(1): 42–55.
- Johansson, J., T. Sherrill, P. Riley, P. Bonato and H. Herr. (2005). "A Clinical Comparison of Variable-Damping and Mechanically Passive Prosthetic Knee Devices." *American Journal of Physical Medicine and Rehabilitation* 84(8): 563–575.
- Kelley, B., P. Pangilinan, R. Mipro, G. Rodriguez and V. Bodeau. (2009). "Upper Limb Prosthetics." *Emedicine*. Retrieved October 2009 from <http://emedicine.medscape.com/article/317234-print>
- Kuiken, T., L. Miller, R. Lipschutz, B. Lock and K. Stubblefield. (2007). "Targeted Reinnervation for Enhanced Prosthetic Function in a Woman with a Proximal Amputation: A Case Study." *Lancet* 369(9559): 371–380.
- Kuniholm, J. (2009). "Open Arms." *IEEE Spectrum*. March, 37–41.
- Kurmala, A. (2007). "Transforming Lives: The i-Limb Prosthetics System." Retrieved September 2009.
- Lake, C. and J. Miguelez. (2003). "Comparative Analysis of Microprocessors in Upper Limb Prosthetics." *Journal of Prosthetics and Orthotics* 15(2): 48–65.
- Liberating Technologies. (2009). "Boston Digital Arm." Retrieved October 2009 from <http://www.liberatingtech.com/>
- MacIsaac, D. and K. Englehart. (2006) "The Science Fiction's Artificial Men." *CrossTalk—The Journal of Defense Software Engineering*, October.
- Mooney, D. (2001). "Osseointegratio—New Hope for Future Amputees?" *Amputee Coalition Communicator* 2(3).
- Moreno, J., L. Bueno, and J. Pons. (2008). "Wearable Robot Technologies." In *Wearable Robots: Biomechatronic Exoskeletons*, J. Pons (Ed.). Chichester, UK: Wiley.
- Motion Control. (2009). "Utah Arm." Retrieved October 2009 from <http://www.utaharm.com/>
- Mouri, T., H. Kawasaki, K. Yoshikawa, J. Takai and S. Ito. (2002). Anthropomorphic Robot Hand: Gifu Hand III. *International Conference on Control, Automation and Systems (ICCAS2002)*. Jeonbuk, Korea: 1288–1293.
- Muzumdar, A. (Ed.). (2004). *Powered Upper Limb Prostheses*. Berlin: Springer.
- Norton, K. (2007). A Brief History of Prosthetics. *in Motion*. 17.

- O'Toole, K. and M. McGrath. (2007). "Mechanical Design and Theoretical Analysis of a Four Fingered Prosthetic Hand Incorporating Embedded SMA Bundle Actuators." *World Academy of Science, Engineering and Technology* 31: 142–149.
- Parker, P., K. Englehart, et al. (2004). "The Control of Upper Limb Prostheses." In *Electromyography, Physiology, Engineering and Non Invasive Applications*, R. Merletti and P. Parker (Eds.). New York: Wiley-IEEE Press.
- Perry, J. (1992). *Gait Analysis: Normal and Pathological Function*. Englewood Cliffs, NJ: Slack Inc.
- Phillips, G. (1990). *Best Foot Forward*. Cambridge: Granta Editions.
- Pons, J. (Ed.). (2008). *Wearable Robots—Biomechatronic Exoskeletons*. Chichester, UK: John Wiley & Sons.
- Schulz, S., C. Pylatiuk, and G. Brethauer. (2001). "A New Lightweight Anthropomorphic Hand." Paper presented at the International Conference on Robotics and Automation (ICRA2001), Seoul, Korea.
- Sears, H., S. Jacobsen, et al. (1989). *Experience with the Utah Arm, Hand and Terminal Device, Comprehensive Management of the Upper-Limb Amputee*. New York: Springer-Verlag.
- Shadow Robot Company. (2008). Retrieved July 2008 from <http://www.shadowrobot.com/>
- Smith, R. (1990). "Audibiofeedback for Tasks." In *Rehabilitation Engineering*, R. Smith and H. Leslie (Eds.). Boca Raton, FL: CRC Press, 29–78.
- Stanfield, M. (2010). "R&D Ethics: The Case of the \$20 Knee." Retrieved June 2010 from [http://www.oandp.com/articles/2010-02\\_03.asp](http://www.oandp.com/articles/2010-02_03.asp)
- Torrealba, R., G. Fernandez-Lopez, et al. (2008). "Towards the Development of Knee Prostheses: Review of Current Researches." *Kybernetes* 37(9–10): 1561–1576.
- Walker, P., H. Kurosawa, et al. (1985). "External Knee Joint Design Based on Normal Motion." *Journal of Rehabilitation Research and Development* 22(1): 9–22.
- Weir, R. F., P. R. Troyk, et al. (2009). "Implantable Myoelectric Sensors (IMESs) for Intramuscular Electromyogram Recording." *IEEE Transactions on Biomedical Engineering*, 56(1): 159.
- Young, E. (2007). "The Future Is Now." *inMotion*, 17(7).

# Index

Page numbers followed by *f* and *t* indicate figures and tables, respectively.

- A**
- AbioCor Implantable Replacement Heart, 415
- AbioMed, 412
- AbioCor, 415
  - Impella, 442–443
- Absolute encoders, rotary optical, 34, 35–36, 35*f*, 36*f*
- Absolute pressure, defined, 56
- Accelerometers, 50–55. *See also* Sensors and transducers
- amplitude response, 53*f*
  - characteristics, 51
  - construction, 50*f*
  - manufacturers of, 55*t*
  - piezoelectric, 53–54, 53*f*, 54*f*
  - theory, 50–53
- AC motors
- induction motors, 134–137
  - efficiency, 136–137
  - speed, 135, 135*f*
  - torque, 135–136, 136*f*
- Actuation/stimulus, in biomechatronic systems, 3
- Actuators
- bellows, 129, 130*f*
- Actuators, 91–157
- design, 312–314
  - electromechanical, 91–137. *See also* Electromechanical actuators
  - hydraulic, 137–139. *See also* Hydraulic actuators
  - mechanical amplification, 145–154. *See also* Mechanical amplification, actuators
  - overview of, 91
  - pneumatic, 139–142. *See also* Pneumatic actuators
  - prosthetic hand actuation, 154–157, 155*f*, 156*f*. *See also* Prosthetic hand actuation
  - shape memory alloys, 142–145. *See also* Shape memory alloys (SMA)
- ADCs. *See* Analog-to-digital converters (ADCs)
- Advanced combination encoder (ACE) strategies, 326
- Advanced Physics Laboratory (APL), 551, 552
- Age-related macular degeneration (AMD), 339
- Airflow
- frictional resistance, 482
  - measurements, 482–483
- Air pressure, measurement of, 59–60
- Aliasing, signal, 245–247, 246*f*–248*f*
- Alkaline battery, 17, 18. *See also* Batteries
- Alligator cabinet respirator, 505–506
- Aluminium prosthesis, 528
- Alveolar ducts, 475–476
- Amplitude shift keying (ASK), 272
- Analog amplifiers, 295–297
- Analog filters. *See also* Digital filters
- band-pass filters, 230, 230*f*
  - band reject filters, 231–232, 232*f*
  - high-pass filters, 228–230, 229*f*
  - implementation, 232–234, 233*f*
  - low-pass filter, 227–228, 227*f*–229*f*
  - notch filters, 230–231, 231*f*
- Analog signal processing. *See also* Digital signal processing (DSP)
- analog circuits, 234–240
  - current-to-voltage converter, 234, 234*f*
  - differentiator, 235–237, 235*f*
  - envelope detection, 237–238, 237*f*
  - integrator, 235–237, 236*f*
  - myoelectric signal processing, 211*t*, 238–240, 239*f*, 240*f*
  - summing amplifier, 234–235, 234*f*
- Analog filters. *See* Analog filters
- frequency content of signal, 224–225, 225*f*, 226*f*
- Analog switches, 12, 13*f*. *See also* Switches
- Analog-to-digital converters (ADCs), 241, 298
- DACs and, 243–245, 243*f*, 244*f*, 245*f*
- Anglesey Leg, 527
- Angular acceleration, 538
- Ankle–foot prosthesis, 580–582
- Ankle joint and foot, bones in, 533
- Anthropomorphic research hands, 554, 555, 562
- Aortic semilunar valve, 399
- APL hand, 559
- Aqueous humor, 336
- Aristotle, 396
- Arm
- Denavit–Hartenberg (D–H) model, 531–532
  - mechanisms of, 548–554
  - motion, energy use for, 20
  - structure of, 529, 530
  - elbow, 530
  - shoulder, 530
  - wrist, 529
- Arrow/LionHeart, 425–427
- Arteries, 397
- Artificial hearts, 408–409
- history, 409
  - AbioMed/AbioCor, 415
  - Jarvik-7 artificial heart, 410–412
  - Liotta-Cooley heart, 409–410
  - Syncardia Systems/CardioWest, 412–414
  - implantation of, 417
- Artificial intelligence, 267, 574–575
- Artificial silicon retina (ASR)
- microchip, 375
- ASK. *See* Amplitude shift keying (ASK)
- Aspirin, 286
- Atrial systole, 399
- Atrioventricular (AV) valves, 398
- Attack time, 300
- Auditory brain stem implant (ABI), 328
- electrodes, 329
  - stimulus mapping, 330

Auditory brain stem response, 288  
 Auditory substitution, 351–352  
 Aulie 802, 543, 544  
 Australian Vision Prosthesis Group, 385–386  
 Automatic gain control (AGC), 321  
 Averager  
   digital filters, 256–258, 256*f*, 257*f*  
 Axial pumps, 455, 456

**B**

Back EMF, 98  
 Back telemetry, 272, 273*f*  
 Band-pass configuration  
   filter implementation, 233, 233*f*  
 Band-pass filters, 230, 230*f*  
   bank, 320–321  
 Band reject filters, 231–232, 232*f*  
   *vs.* notch filter, 231–232, 232*f*  
 Batteries, 16–19, 18*f*. *See also* Power supplies  
   advantages, 17  
   alkaline, 17, 18  
   disadvantages, 18  
   lead-acid, 17  
   lithium ion (Li-Ion), 17, 18  
   lithium polymer (Li-Po), 17  
   mercury, 17  
   nickel cadmium (NiCd), 17, 18  
   nickel metal hydride (NiMH), 17, 18  
   rechargeable, 17–18, 18*f*  
   silver oxide, 17  
   zinc air, 17  
   zinc carbon, 17  
   zinc chloride, 17  
 Baxter, 412  
 Bayesian (belief) networks, 267  
 Bayesian nets, 267  
 Bayesian statistics, 267  
 Bearings, 466  
 Bell, Alexander Graham, 495, 496  
 Bellows actuators, 129, 130*f*  
 Belt drives, 149, 152–153  
   power transmitted by, 152  
 Bernoulli's equation  
   flowmeters, operation of, 61–62  
 "Big blue," 414  
 Bimetallic thermostats, 68, 68*f*. *See also* Temperature sensors  
 Binaural sensory aid (BSA), 352  
 Binaural Sonic Aid (Sonicguide), 345  
 Bioacoustic signals, 208  
 Biochemical signals, 208  
 Biochemical system, 4  
 Bioelectric signals, 207, 208–210,  
   209*f*, 210*f*  
 Bioimpedance signals, 208

Biological feedback mechanisms, 160, 160*f*  
 Biomagnetic signals, 208  
 Biomechanical signals, 208  
 Biomechatronic applications  
   brushless DC (BLDC) motors, 117  
   linear actuators, 129–130, 130*f*  
   of shape memory alloys, 145  
   solenoids, 99  
   stepper motors, 124  
   voice coil actuators, 99  
 Biomechatronic feedback mechanisms, 160–162  
   limit switches, 161, 162*f*  
   proportional and higher-order controllers, 161–162  
 Biomechatronics  
   overview, 1–2  
 Biomechatronic systems, 2–4, 2*f*  
   feedback elements in, 4  
   future of, 6–7  
   human subject in, 2–3  
   recording and display in, 3  
   signal processing elements in, 3  
   stimulus/actuation in, 3  
   transducers and sensors in, 3  
 Biomedical pressure sensors, 56  
 Biomedical signals  
   bioacoustic signals, 208  
   biochemical signals, 208  
   bioelectric signals, 207, 208–210,  
   209*f*, 210*f*  
   bioimpedance signals, 208  
   biomagnetic signals, 208  
   biomechanical signals, 208  
   bio-optical signals, 208  
   source, characterized by, 210, 211*t*  
   type, characterized by, 210–211  
     periodic signals, 210  
     quasi-static signals, 210  
     quasi-transient signals, 210–211  
     repetitive signals, 210  
     static signals, 210  
     stochastic signals, 211  
     transient signals, 210–211  
 Bionic Vision Australia (BVA), 385  
 Bio-optical signals, 208  
 Biosensors, 77  
 Bipedal walking, 534–535  
 Bipolar coils in stepper motors, 118  
 Bjork, Viking, 404  
 BLDC. *See* Brushless DC (BLDC) motors  
 Blindness, causes of, 339  
 Blood oxygen concentration,  
   measurement, 82–83  
 Blood pressure (BP), 401  
   long-term regulation, 402  
   measurement of, 56–59  
   short-term regulation, 401  
 Blood vessels, types of, 397  
 Body-powered prosthesis, 539  
 Body-surface biopotential electrode, 84,  
   85. *See also* Electrodes  
 Bone anchored hearing apparatus (BAHA), 300, 583  
 Bone conduction devices, 300–302  
 Boston Digital Arm System, 549, 550  
 Boston retinal implant project, 379–381  
 Both portable respirator, 501  
 Braille, 361  
 Bränemark, Per Ingvar, 300  
 Braun, Egon, 495  
 Breathing  
   energy required for, 485–488  
   energy use for, 20  
 Brindley, Giles, 335  
 Brushless DC (BLDC) motors, 113–117,  
   113*f*–115*f*, 116*t*  
   biomechatronic applications, 117  
   selecting, parameters for, 115–116  
   torque–speed characteristics of,  
   115, 115*f*  
   types, 117, 117*f*  
   *vs.* brushed DC motors, 116*f*  
   *vs.* brushless DC motors, 116*f*  
 BTR (bridge to recovery) event, 420, 442

**C**

Cameras, input from, 353–355  
 Cams, 148–149, 148*f*  
 Capacitance, 57  
 Capacitive displacement sensors, 34. *See also* Sensors and transducers  
 Capacitive tactile sensors, 75–76, 75*f*. *See also* Tactile sensors  
 Capillaries, 397  
 Carbon microcoils (CMC) tactile sensors, 76–77, 76*f*, 77*f*. *See also* Tactile sensors  
 Cardiac output (CO), 401  
 Cardiovascular system, 5, 396, 397. *See also* Heart  
 Catalytic sensors, 78, 78*f*. *See also* Chemical sensors  
 CCD. *See* Charge-coupled device (CCD) array  
 Centrifugal pumps, 406, 455, 456  
   energy, 456–457  
   power, 456–457  
   pressure and head, 457–459  
   specific speed, 459  
   torque, 456–457  
 Characteristic impedance, 278–279  
 Charge amplifiers, 221–222, 222*f*

Charge-coupled device (CCD) array, 41  
 CHEMFET, 78, 79*f*  
 Chemical sensors, 77–80. *See also*  
     Sensors and transducers  
     electrochemical, 78  
     enzyme and catalytic, 77–78, 78*f*  
     microbalance odor, 80, 81*f*  
     oscillating, 79–80  
     resistive, 78–79  
 Cisplatin, 286  
 Clark, Graeme, 314  
 C-5 laser cane, 350  
 Class-A amplifiers, 295  
 Class-B amplifiers, 295  
 Class-D amplifiers, 296  
 Classification trees, 268, 268*f*  
 C-Leg®, 576, 577  
 Closed-loop system, 159,  
     162, 163  
 Clustering, 269  
 CMC. *See* Carbon microcoils (CMC)  
     tactile sensors  
 Cochlea, 283  
 Cochlear implants, 314  
     historical background, 314–315  
     installation of the electrode, 320  
         safe current density, 320  
     signal processing and cochlear  
         stimulation, 320  
         band-pass filter bank, 320–321  
         continuous interleaved sampling  
             (CIS), 321–325  
     spectral maxima strategies, 326  
     strategies to enhance vocal pitch,  
         326–327  
     working of, 315–319  
 Common-mode noise, 215–216, 215*f*  
 Comparator, 241, 241*f*  
 Completely-in-the-canal (CIC) hearing  
     aids, 290  
 Compressed analog (CA) principle, 321  
 Compression ratio, 300  
 Conductive hearing loss, 285  
 Cones, 336, 337  
 Congestive heart failure (CHF), 402  
 Continuous interleaved sampling  
     (CIS), 321  
     compression function, 323–324  
     design parameters, 323  
     envelope detection, 324–325  
     filter spacing, 325  
     stimulation rate, 323  
 Continuous positive airway pressure  
     (CPAP), 517  
 Continuous signals, 208  
 Controlled mandatory ventilation  
     (CMV), 514

Controllers, 188–201  
     implementation of, 201–205  
     gains, selection of, 201–202,  
         202*f*, 202*t*  
     hardware, 202–205, 203*f*, 204*f*  
     integral. *See* Integral controller  
     PID, 200–201, 200*f*, 201*f*  
     proportional. *See* Proportional  
         controllers  
 Conventional prosthesis, 539  
 Cooley, Denton, 409  
 Copper aluminum nickel (CuAlNi), 142  
 Copper zinc aluminum (CuZnAl), 142  
 CorAide, 439  
 Coriolis effect, in rate gyros, 46–47,  
     47*f*, 48*f*  
 Cornea, 335  
 Coronary heart disease (CHD), 402  
 Correlation coefficient, 265, 449  
 Cousteau, Jacques, 512  
 C-pulse counterpulsation, 444, 445  
 Crossover distortion, 295  
 Current-to-voltage converter, 234, 234*f*

**D**

DACs. *See* Digital-to-analog converters  
     (DACs)  
 Dalziel, John, 495  
 D'Armata, Salvino, 334  
 Data mining  
     signal processing, 267  
 Dead time control, 115  
 Decision trees, 268, 268*f*  
 DEKA arm, 550, 551  
 Dennis, Clarence, 404  
 DeVries, William, 410  
 Diaphragm pumps, 463–465  
 Diastole, 401  
 Difference equations  
     digital filters, 248–249, 250*f*  
 Differential amplifiers, 218–219,  
     219*f*, 220*f*  
 Differential mode noise, 215–216  
 Differential pressure, defined, 56  
 Differential pressure flowmeter, 61–63.  
     *See also* Flow measurement  
     Bernoulli's equation in, 61–62  
     nozzle method, 63, 63*f*  
     orifice plate, 62–63, 63*f*  
     Venturi tube, 63, 63*f*  
 Differentiator  
     analog circuits, 235–237, 235*f*  
     digital filters, 254, 254*f*  
 Digital amplifiers, 297  
     digital feedback reduction (DFR), 299  
     digital hearing aids as signal  
         generators, 299–300

digital noise reduction (DNR), 299  
 digital speech enhancement (DSE), 299  
 directional microphones and DSP, 299  
 frequency shifting, 299  
 gain processing, 298–299  
 Digital feedback reduction (DFR), 299  
 Digital filters, 248–258. *See also* Analog  
     filters  
     averager, 256–258, 256*f*, 257*f*  
     difference equations, 248–249, 250*f*  
     differentiator, 254, 254*f*  
     envelope detection, 259–260, 261*f*  
     finite impulse response (FIR), 250–251,  
         251*f*  
     impulse response, 259, 259*f*  
     infinite impulse response (IIR), 250  
     integrator, 255–256, 255*f*  
     spectral estimation, 260–264, 261*f*  
     step response, 259, 260*f*  
     time-domain response, 258–259, 258*f*,  
         259*f*, 260*f*  
     tracking filter, 251–254, 253*f*  
     transfer functions, 248–249, 250*f*  
 Digital hearing aids as signal generators,  
     299–300  
 Digital noise reduction (DNR), 299  
 Digital optical encoders, 34–38, 34*f*, 35*f*.  
     *See also* Encoders  
     absolute. *See* Absolute encoders  
     incremental, 36–38, 37*f*, 38*f*  
 Digital signal processing (DSP), 292. *See*  
     also Analog signal processing  
 ADCs and DACs, 243–245, 243*f*,  
         244*f*, 245*f*  
 comparator, 241, 241*f*  
 digital filters, 248–258. *See also* Digital  
     filters  
     signal acquisition, 241–243, 241*f*, 242*f*  
     signal aliasing, 245–247, 246*f*–248*f*  
     electrocardiogram, 246–247,  
         247*f*, 248*f*  
 Digital speech enhancement (DSE), 299  
 Digital-to-analog converters (DACs), 242  
     ADCs and, 243–245, 243*f*, 244*f*, 245*f*  
 Dilator muscle, 335  
 Direct Acoustic Cochlear Stimulation  
     (DACS) device, 288, 312  
     actuator design, 312–314  
 Direct current motors, 99–113, 100*f*  
     brushless, 113–117. *See also* Brushless  
         DC (BLDC)  
     model, 170–171  
     multiple-coil direct current motor,  
         100–103, 101*f*, 103*f*  
     powering, 110–113, 111*f*, 112*f*  
     real motor characteristics, 103–107, 104*f*  
     selecting, 108–110

- Direct current motors (*cont.*)  
 single-coil DC motor, 100, 101f  
 types, 107–108, 107f, 108f
- Directional microphones and DSP, 299
- Discrete signals, 208
- Discriminant analysis, 268
- Display, in biomechatronic systems, 3
- Distributed artificial heart (DAH), 409
- Doe, O.W., 495
- Doherty Eye Institute, 382–384
- Dorrance, D.W., 540
- Dorsal cochlear nucleus (DCN), 328
- Double-acting cylinder  
 hydraulic actuators, 138–139, 139f
- Double pole double throw (DPDT)  
 switch, 10, 10f
- Doyle, John, 314
- DPDT. *See* Double pole double throw (DPDT) switch
- Drinker, Cecil, 498
- Drinker, Philip A., 498–499, 501, 502
- Drinker respirator, 499–501
- Drivers  
 stepper motors, 118–120, 119f, 120t
- Dry cell devices, 17
- Dry electrode, 86–87. *See also*  
 Electrodes
- Dynamic characteristics of stepper motors, 122, 122f
- Dynamic deflection mode, 26–27, 26f
- Dynamic pumps. *See* Turbo pumps
- E**
- Ear, working of, 281  
 hearing statistics, 283–285  
 inner ear, 283  
 middle ear, 281–283  
 outer ear, 281
- EDA. *See* Electronic design automation (EDA)
- Elbow, 530
- Electret microphones, 294
- Electrical elements, in system models, 168–169. *See also* System models
- Electrically erasable programmable read-only memory (EEPROM), 297
- Electrical model, 169–171, 169f, 170t. *See also* System models
- Electric motors  
 prosthetic hand actuation based on, 155, 156f
- Electrocardiogram  
 signal aliasing, 246–247, 247f, 248f
- Electrochemical sensors, 78, 79f. *See also*  
 Chemical sensors
- Electrocochleography, 288
- Electrocutaneous stimulation. *See*  
 Electrotactile stimulation
- Electrode, installation of, 320  
 safe current density, 320
- Electrodes, 83–88, 83t, 329  
 body-surface biopotential, 84, 85f  
 dry, 86–87  
 for internal use, 87, 87f, 88f  
 for long term use, 85–87, 86f  
 metal plate, 85, 85f  
 needle, 87, 88f  
 nonpolarized, 84  
 polarizable, 84  
 silver–silver chloride, 84
- Electromagnetic (EM) wave  
 radio frequency telemetry and, 272
- Electromagnetic hearing devices, 307  
 semi-implantable middle ear  
 electromagnetic hearing device (SIMEHD), 311  
 Soundtec direct system, 309–311  
 vibrant soundbridge device, 307–309
- Electromechanical actuators, 91–137  
 AC motors, 134–137. *See also* AC motors  
 BLDC motors, 113–117, 113f–115f, 116t. *See also* Brushless DC (BLDC) motors  
 direct current motors, 99–113. *See also*  
 Direct current motors  
 linear actuators, 124–130. *See also*  
 Linear actuators  
 servo motors, 130–134. *See also* Servo motors  
 solenoids, 94–96, 94f, 95f, 96f  
 stepper motors, 117–124. *See also*  
 Stepper motors  
 voice coil, 96–99, 97f, 97t, 98f, 99f
- Electronic design automation (EDA), 234
- Electroretinogram (ERG), 377
- Electrorheological (ER) fluid actuators, 577
- Electrotactile and vibrotactile transducers, 356  
 electrotactile displays, 365–369  
 skin, 356–358  
 thermotactile displays, 369  
 vibrotactile devices construction, 363–365  
 displays, 358–363
- Electrotactile display, 365–369
- Electrotactile stimulation, 356
- EM. *See* Electromagnetic (EM) wave
- Emerson, John, 504
- Emerson respirator, 504–505
- Encoders  
 digital optical. *See* Digital optical encoders  
 incremental. *See* Incremental encoders  
 linear/angular displacement, 32–33
- Energy, 456–457
- Energy scavenging, 19–22. *See also*  
 Power supplies  
 external devices, 19–22  
 internal devices, 22  
 piezoelectric/rotary generators in, 21–22
- Engineering, in heart assist devices  
 estimation and control of blood flow, 450–452  
 fluid dynamics in centrifugal and axial LVADs, 448–449  
 fluid dynamics in pulsatile LVADs, 446–448  
 transcutaneous energy transfer, 452–454
- Envelope detection  
 analog circuits, 237–238, 237f  
 digital filters, 259–260, 261f
- Envoy system, 306–307
- Enzyme sensors, 77. *See also* Chemical sensors
- Epiretinal implants, 381  
 Australian Vision Prosthesis Group, 385–386
- Doherty Eye Institute, 382–384
- EPI-RET project and IIP technologies, 384–385
- Harvard Medical School–MIT Collaboration, 381–382
- EPI-RET project and IIP technologies, 384–385
- Error, least squared, 265
- Error amplifier  
 radio control servos, 132
- Estimation and control of blood flow, 450–452
- Ethicon Inc., 423
- EvaHeart, 439
- Evanescence wave spectroscopy, 81–82
- Exercitatio Anatomica de Motu Cordis et Sanguinis in Animalibus, 396
- Expert systems, 267
- Expiratory reserve volume (ERV), 489
- External negative-pressure ventilators (ENPVs), 497–499  
 physics of, 508–511
- Extracorporeal ventricular assist devices, 420–421
- Eyeglasses, 339
- EyePlusPlus, 367

**F**

- Fast Fourier transform (FFT), 322  
 Feedback, 159–205  
     elements, in biomechatronic systems, 4  
     mechanisms  
         biological, 160, 160f  
         biomechatronic. *See* Biomechatronic feedback mechanisms  
 FET. *See* Field-effect transistors (FET)  
 Fiber optic gyro (FOG) rate sensor, 48–49, 48f  
 Fiber optic pressure sensor, 59, 59f  
 Field-effect transistors (FET), 78  
 Field-programmable gate array (FPGA), 156  
 Finite impulse response (FIR), synthesizing  
     digital filters, 250–251, 251f  
 FIR. *See* Finite impulse response (FIR), synthesizing  
 Fleuss, Henry, 512  
 Flex-Foot®, 545  
 Flexing plate gas sensor, 79, 79f  
 Floating mass transducer (FMT), 307  
 Flow measurement, sensors for, 61–67  
     differential pressure flowmeter. *See* Differential pressure flowmeter  
     impeller flowmeters, 67, 67f  
     magnetic flowmeters, 65, 65f  
     target flowmeters, 65–66  
     temperature flowmeters, 64  
     turbine flowmeters, 66–67, 66f  
     ultrasound flowmeters, 64–65  
 Fluid dynamics  
     in centrifugal and axial LVADs, 448–449  
     in pulsatile LVADs, 446–448  
 Fluid flow elements, in system models, 171–174, 172f, 173f, 174t. *See also* System models  
 Fluidhand, 561–562  
 FOG. *See* Fiber optic gyro (FOG) rate sensor  
 Foot prosthetics, 545–547  
 Forced expiratory volume in t sec (FEV<sub>t</sub>), 489, 490  
 Forced vital capacity (FVC), 489  
 Forehead recognition sensory system (FSRS), 366  
 Fourier transform, digital filters, 261–264, 263f, 264f  
 Fovea, 336  
 FPGA. *See* Field-programmable gate array (FPGA)  
 Freedom driver, 414  
 Freminet, Sieur, 511–512

Frequency-modulated continuous wave (FMCW) techniques, 344

- Frequency shifting, 299  
 Frictional forces, 481–485  
 Functional magnetic resonance imaging (fMRI), 355  
 Functional residual capacity (FRC), 489  
 Futaba, 133, 133f  
 Futurists, 6–7

**G**

- Gagnan, Emile, 512  
 Gain processing, 298–299  
 Gaitmaster, 543  
 Galen, 396  
 Galvanic isolation, 273–274  
 Gauge factor ( $\gamma$ ), on strain gauge, 24  
 Gauge pressure, defined, 56  
 Gear pumps, 460–461  
 Gears, 149–153  
     planetary gearheads, 150, 151–152, 151f, 152f, 152t  
     ring, 150, 150f  
     spur gearheads, 150, 151–152, 152t  
     sun, 150, 150f  
 Generation 1 LVADs, 421  
     Arrow/LionHeart, 425–427  
     Thoratec/Heartmate, 424  
     WorldHeart/Novacor, 421–423  
     WorldHeart/Novacor II, 427–428  
 Generation 2 VADs, 433  
     Thoratec/Heartmate II, 433–435  
 Generation 3 VADs, 435  
     VentraCo/VentrAssist, 435–437  
         Mohawk Innovative Technology/MiTiHeart, 437–439  
 Generation 4 VADs, 439  
     AbioMed/Impella, 442–443  
     HeartWare LVAD, 442  
     Jarvik-2000, 440–441  
     Micromed/DeBakey LVAD, 441–442  
 Genetic algorithms, 267  
 German Aerospace Centre (DLR), 155  
 Giant magnetoresistance (GMR), 31  
 Gibbon, John, 403  
 GIFU hand, 556–557  
 GMR. *See* Giant magnetoresistance (GMR)  
 Goldfarb, Michael, 552  
 Götz von Berlich, 525, 526  
 GPS-based systems, 370–371  
 Graphical model, 359  
 Gray code disc  
     encoder pattern for, 35–36, 36f  
 Guerrero, Simón, 554  
 GuideCane, 348–349

**H**

- Half-cell potential, defined, 84  
 Hall effect, magnetic field strength with, 31  
 Hall switches, 11. *See also* Switches  
 Hand mechanisms, 554–562  
     APL hand, 559  
     Fluidhand, 561–562  
     GIFU hand, 556–557  
     hands using shape memory alloy  
         actuators, 562  
         i-Limb Hand, 559–561  
         shadow hand, 557–558  
         Ultralight hand, 556  
         University of Bologna (UB) Hand III, 558–559  
 Hand research and applications, 562–563  
 Hands using shape memory alloy  
     actuators, 562  
 Hanger, James, 528  
 Hans Mauch, 543  
 Harasaki, Hiroaki, 412  
 Harbin Institute of Technology ( HIT ), 155  
 Harvard Medical School-MIT Collaboration, 381–382  
 Harvey, William, 396  
 Hearing aids, 277, 289  
     analog amplifiers, 295–297  
     auditory brain stem implant (ABI), 328  
         electrodes, 329  
         stimulus mapping, 330  
     bone conduction devices, 300–302  
     cochlear implants, 314  
         historical background, 314–315  
         installation of the electrode, 320  
         signal processing and cochlear stimulation, 320–325  
         spectral maxima strategies, 326  
         strategies to enhance vocal pitch, 326–327  
         working of, 315–319  
 digital amplifiers, 297  
     digital feedback reduction (DFR), 299  
     digital hearing aids as signal generators, 299–300  
     digital noise reduction (DNR), 299  
     digital speech enhancement (DSE), 299  
     directional microphones and DSP, 299  
     frequency shifting, 299  
     gain processing, 298–299  
 Direct Acoustic Cochlear Stimulation (DACS), 312  
     actuator design, 312–314  
 ear, working of, 281

- Hearing aids (*cont.*)  
 hearing statistics, 283–285  
 inner ear, 283  
 middle ear, 281–283  
 outer ear, 281  
 hearing loss, 285  
 causes, 285–286  
 diagnosis, 286–288  
 treatment, 288  
 history, 289–292  
 microphones, 292–294  
 middle ear implants, 302  
 electromagnetic hearing devices, 307–311  
 issues with implantable middle ear devices, 311–312  
 piezoelectric devices, 303–307  
 power consumption, 300  
 signal compression, 300  
 sound, 278  
 characteristic impedance and sound pressure, 278–279  
 sound intensity, 279
- Heart  
 artificial hearts, 408–409  
 history, 409  
 implantation of, 417  
 disease, 402  
 engineering, in heart assist devices  
 estimation and control of blood flow, 450–452  
 fluid dynamics in centrifugal and axial LVADs, 448–449  
 fluid dynamics in pulsatile LVADs, 446–448  
 transcutaneous energy transfer, 452–454  
 heart–lung machines  
 history, 403–404  
 modern machines, 404–407  
 as pump, 397–398  
 biomechatronic perspective, 402–403  
 cardiac output, 401  
 components of, 416  
 cycle, 399–400  
 heart disease, 402  
 heart valves, 398–399  
 pressure regulation, 401–402  
 pump types, 455  
 axial pump characteristics, 456–459  
 bearings, 466  
 centrifugal pump characteristics, 456–459  
 reciprocating pump characteristics, 462–465  
 rotary pump characteristics, 460–462
- structure of, 398  
 ventricular assist devices (VADs), 417–418  
 CorAide, 439  
 EvaHeart, 439  
 extracorporeal ventricular assist devices, 420–421  
 generation 1 LVADs, 421–428  
 generation 2 VADs, 433–435  
 generation 3 VADs, 435–439  
 generation 4 VADs, 439–443  
 history, 419–420  
 ideal replacement heart, 443  
 intracorporeal left ventricular assist devices, 421  
 pulsatile pump technology, 429–433  
 pump types, 443–444  
 Terumo Heart, Inc., 439
- Heart–lung machines  
 history, 403–404  
 modern machines, 404–407
- HeartWare LVAD, 442
- Heat–moisture exchangers (HMEs), 518
- Heimlich, Henry, 409
- Hermann, Gustav, 528
- Higher-order controllers, 161–162
- High-pass filters, 228–230, 229f
- Hilbert transform, 325
- Hip joint, 532–533
- HIT. *See* Harbin Institute of Technology (HIT)
- Hitec, 132f, 132t
- Homemade iron lungs, 501–504
- House, William, 314
- Human activity, energy use for, 19–20, 19t
- Human subject, in biomechatronic systems, 2–3
- Hybrid stepper motors, 117, 118
- Hydraulic actuators, 137–139  
 components of, 137, 137f  
 double-acting cylinder, 138–139, 139f  
 overview of, 137  
 single-acting cylinder, 138, 139f  
 valves, 137–138, 138f
- Hydraulic capacitance, defined, 172, 172f
- Hyperplane, 269
- I**
- Ideal mechanical replacement heart, 443
- IIR. *See* Infinite impulse response (IIR)
- I-Limb Hand, 559–561
- Illinois Institute of Technology Research Group, 389–391
- Image intensifiers, 340
- Impella Recover system, 442
- Impeller flowmeters, 67, 67f. *See also* Flow measurement
- Implant systems, 270–271, 270f  
 back telemetry, 272, 273f  
 inductive telemetry, 271–272, 272f  
 optical telemetry, 271  
 radio frequency telemetry, 272
- Impulse response  
 digital filters, 259, 259f
- Incremental encoders. *See also* Encoders  
 magnetic, 33, 33f  
 measuring rate and angular rate, 43, 43f  
 rotary optical, 36–38, 37f, 38f
- Induction motors, 134–137  
 efficiency, 136–137  
 speed, 135, 135f  
 torque, 135–136, 136f
- Inductive (brushless) resolvers, 31
- Inductive displacement sensors, 28–31.  
*See also* Sensors and transducers  
 inductosyns and resolvers, 30–31  
 LVDT, 28–30, 29f, 30f
- Inductive telemetry, 271–272, 272f
- Inductosyns, 31
- Infinite impulse response (IIR)  
 digital filters, 250
- Infrared light-emitting diode (IR LED), 11
- Initial swing  
 leg mechanisms, 575  
 walking dynamics, 536, 542
- Injectable myoelectric sensors (IMES), 573–574
- In-line meters, 67, 67f
- Inner ear, 283
- Inspiratory reserve volume (IRV), 489
- Instrumentation amplifiers, 219–221, 220f
- Integral controller, 198, 198f, 199f. *See also* Controllers  
 proportional plus, 198–200, 199f
- Integrator  
 analog circuits, 235–237, 236f  
 digital filters, 255–256, 255f
- Interaural amplitude difference, 345
- Intercontinental ballistic missiles (ICBM), 46
- Interferometry, optical, 38–39, 39f
- In-the-canal (ITC) hearing aids, 290
- In-the-ear (ITE) hearing aids, 290
- Intracorporeal left Ventricular assist devices, 421
- Inverting amplifiers, 217–218, 217f
- Iris, 335
- IR LED. *See* Infrared light-emitting diode (IR LED)
- Iron lung, 498  
 homemade, 501–504
- Isolation amplifiers, 272–274, 274f

- Isolation barrier, 270–274  
 implant systems, 270–272. *See also*  
     Implant systems  
 isolation amplifiers, 272–274, 274*f*  
 in personal computers, 273  
 Isovolumetric contraction, 400  
 Isovolumetric relaxation, 400  
 IWALK Inc, 582
- J**
- J. H. Emerson Co., 504, 508  
 Jacket ventilator, 506–507  
 Jaipur foot, 546, 547  
 JaipurKnee, 543, 544  
 James, William, 512  
 Jarvik-2000, 440–441  
 Jarvik-7 Artificial heart, 410–412
- K**
- KASPA system, 352  
 Kessler, Thomas, 410  
 Kirchhoff's laws, 169  
 K-means, 269  
 Knee joint, 533  
 Kneepoint, 300  
 Knee prostheses, low-cost, 544  
 Knee prosthetics, 542–544,  
     576–580  
 Kobrinski, A., 567  
 Kolf, Willem, 409, 410  
 K-Sonar, 344  
 Kuiken, Todd, 571, 572
- L**
- Lambson, Christian, 512  
 Laplace transform, in system response,  
     175–178, 177*t*  
     process of using, 175, 176*f*  
     rules for handling, 175, 176*t*  
 Laser-based systems, 350  
 Lateral geniculate nucleus (LGN), 338  
 Lead-acid battery, 17. *See also* Batteries  
 Lead zirconate titanate (PZT), 21–22, 21*t*  
 Least squared error, 265  
 LED. *See* Light-emitting diode (LED)  
 Leg  
     kinematic model of, 534  
     normal walking dynamics, 535–536  
     walking, 534–535  
     structure of, 532  
         ankle and the foot, 533  
         hip joint, 532–533  
         knee joint, 533  
 Leg mechanisms, 574  
     control strategies, 575–576  
     knee prosthetics, 576–580  
 Leonardo da Vinci, 472  
 Leroy, J. J. J., 495  
 LeTourneau Engineering Global Solutions  
     (LEGS) M1 knees, 543, 544  
 Levers, 147–148, 147*f*  
 Liberating Technologies Inc., 549, 554  
 Light-emitting diode (LED)  
     back telemetry and, 272  
 Limb movement, kinematics of, 536  
     angular acceleration, 538  
     center of mass and moment of inertia  
         of complete limb, 538  
         of segment of limb, 536–537  
 Limit switches, 10–11, 11*f*, 12. *See also*  
     Switches  
     feedback mechanisms, 161, 162*f*  
 Linear actuators, 124–130, 125*f*  
     bellows actuators, 129, 130*f*  
     biomechatronic applications,  
         129–130, 130*f*  
     piezoelectric actuators, 126–129, 127*f*,  
         127*t*, 128*f*  
     worm gear drive, 125, 125*f*  
 Linearity, defined, 28  
 Linearization, of thermistors  
     resistance-mode, 70, 71*f*  
     voltage-mode, 70, 71*f*  
 Linear Motor Driven LVADs and TAHs,  
     430–433  
 Linear power supplies, 13–14, 14*f*, 15*t*.  
     *See also* Power supplies  
 Linear regulators, 14, 15*t*  
 Linear time invariant (LTI), 233–234  
 Linear variable differential transformer  
     (LVDT), 28–30, 29*f*, 30*f*, 98  
 Linear voice coil actuators, 96–98, 97*f*,  
     97*t*, 98*f*  
     moving coil, 96, 97*f*  
     moving magnet types, 96, 97*f*  
 Linkages, 145–147, 145*f*–147*f*  
     usage of, 145  
 LionHeart, 425–427  
 Liotta, Domingo, 409  
 Liotta-Cooley heart, 409–410  
 Lithium ion (Li-Ion) battery, 17, 18. *See*  
     also Batteries  
 Lithium polymer (Li-Po) battery, 17. *See*  
     also Batteries  
 LM10 linear magnetic encoder, 33–34, 33*f*  
 LM35Z temperature sensor, 72  
 Lobe pumps, 461  
 Lorentz force law, 91–92  
 Loud sounds, 286  
 Low-pass filter, 227–228, 227*f*–229*f*  
 LTI. *See* Linear time invariant (LTI)  
 Luke arm, 550  
 Lung characteristics, measuring, 488  
     pneumotachography, 492–494  
     spirometry, 488–492  
 Lung elasticity, 480–481  
 LVDT. *See* Linear variable differential  
     transformer (LVDT)
- M**
- Machine learning (ML), 3, 267–270  
     clustering, 269  
     decision/classification trees, 268, 268*f*  
     supervised learning, 267  
     support vector machines (SVMs),  
         269–270, 269*f*  
     unsupervised learning, 267  
 Macula, 336  
 Magnetic displacement sensors, 31–34,  
     32*f*. *See also* Sensors and  
     transducers  
 Magnetic flowmeters, 65, 65*f*. *See also*  
     Flow measurement  
 Magnetomotive force (*mmf*), 94  
 Magnetoresistance (MR), magnetic field  
     strength with, 31  
 Magnetorheological (MR) fluid  
     actuators, 577  
 Magnets  
     stepper motors and  
         magnets, 123  
 Malpighi, 472  
 Masking, 350  
 Maximal voluntary ventilation (MVV),  
     489, 490  
 Maxon micro-drive servo motor,  
     134, 134*f*  
 Maxon RE 16 motor  
     Simulink model of, 192*f*, 193*f*  
     specifications, 191*t*  
 Mayow, John, 494  
 McKhann, C., 500  
 Mean arterial blood pressure  
     (MABP), 450  
 Mechanical amplification, actuators,  
     145–154  
     cams, 148–149, 148*f*  
     gears and belt drives, 149–153,  
         149*f*–153*f*, 152*t*  
     levers, 147–148, 147*f*  
     linkage, 145–147, 145*f*–147*f*  
     translation screw devices, 153–154,  
         153*f*  
 Mechanical elements, system models,  
     164–166, 166*t*. *See also* System  
     models  
 Mechanical model, 166–167, 166*f*, 167*f*.  
     *See also* System models  
 Mechanical ventilation, 494  
     alligator cabinet respirator, 505–506  
     Both respirator, 501

- Mechanical ventilation (*cont.*)  
Drinker respirator, 499–501  
early history, 494–495  
Emerson respirator, 504–505  
external negative-pressure ventilators (ENPVs), 497–499  
homemade iron lungs, 501–504  
negative-pressure ventilation, uses for, 507–508  
polio, 495–497  
portable respirators, 506–507  
Mechanoreception, 538  
Mechatronic engineering, 1  
Med-El, 307  
Medquest, 412  
MEIHD. *See* Middle ear implantable hearing devices (MEIHD)  
MEMS. *See* Microelectromechanical systems (MEMS)  
Mercury battery, 17. *See also* Batteries  
Merkel cells, 356  
Metal plate electrode, 85, 85*f*. *See also* Electrodes  
Microbalance odor sensors, 80, 81*f*. *See also* Chemical sensors  
Microelectromechanical systems (MEMS)  
rate gyros, 46–48, 47*f*  
strain gauge pressure transducer, 57, 58, 58*f*  
Micromed/DeBakey LVAD, 441–442, 452  
MicroMo Electronics, 153  
Microphones, 60–61, 292–294  
Microphotodiodes, 377  
Middle ear, 281–283  
Middle ear implantable hearing devices (MEIHD), 99, 166, 166*f*, 288  
Middle ear implants, 302  
electromagnetic hearing devices, 307  
semi-implantable middle ear  
electromagnetic hearing device (SIMEHD), 311  
Soundtec direct system, 309–311  
vibrant soundbridge device, 307–309  
issues with implantable middle ear devices, 311–312  
piezoelectric devices, 303  
envoy system, 306–307  
middle ear transducer (MET), 305–306  
Rion Device E-type (RDE), 304  
totally integrated cochlear amplifier (TICA), 304  
Middle ear transducer (MET), 305–306  
Mid-swing  
leg mechanisms, 575  
walking dynamics, 536, 542  
MiTiHeart, 437–439  
Mixed-flow pumps, 455  
ML. *See* Machine learning (ML)  
MLX90609 rate gyro, 49, 50*f*  
Mohawk Innovative Technology, 437–439  
Moment of inertia, 537, 538  
Motion Control Inc., 554  
Moving coil  
linear voice coil actuators, 96, 97*f*  
Moving magnet types  
linear voice coil actuators, 96, 97*f*  
Mowat sensor and derivatives, 344  
MR. *See* Magnetoresistance (MR)  
Multiphotodiode array (MPDA), 378  
SEM image of, 379  
Multiple-coil direct current motor, 100–103, 101*f*, 103*f*  
Multiple regression, 266  
Musculoskeletal system, 5  
Myoelectric control systems, 566–570  
Myoelectric signal processing, 211*t*, 238–240, 239*f*, 240*f*  
Myoelectric signals, 564–566, 567
- N**  
Nafis, Ibn, 396  
blood circulation model of, 396  
Nasal ventilation, 518  
NavBelt, 345–348  
Needle electrode, 87, 88*f*. *See also* Electrodes  
Negative-pressure ventilation, uses for, 507–508  
Negative temperature coefficient (NTC) thermistors, 69–70, 70*f*  
Negavent, 509, 510  
Neodymium-iron-boron (NdFeB) magnet, 309  
Nervous system, 5  
Neural interface, 374  
Neural networks, 267  
Neural stimulator, 373–374  
Neurovisual prostheses. *See* Visual neuroprostheses  
NEV-100, 509, 510  
New England Journal of Medicine, 435  
Nickel cadmium (NiCd) battery, 17, 18. *See also* Batteries  
Nickel metal hydride (NiMH) batteries, 17, 18, 423. *See also* Batteries  
Nickel titanium (NiTi), 142  
Night vision binoculars, 340  
Night-vision cameras. *See* Image intensifiers  
Noise, signal acquisition, 212–216  
Noninverting amplifiers, 218, 218*f*  
Notch filters, 230–231, 231*f*  
vs. band reject filters, 231–232, 232*f*
- Nottingham obstacle detector (NOD), 344  
Novacor, 412  
NTC. *See* Negative temperature coefficient (NTC) thermistors  
Nuffield, Lord, 501
- O**  
Occipital tactile-visual (LOtv) region, 355  
Op amp specifications, 223–224, 223*t*  
Open-loop system, 159, 162, 163  
Optacon (OPtical TActile CONverter), 356, 361  
Optical character recognition (OCR), 362  
Optical chemical sensors, 80–83. *See also* Sensors and transducers  
blood oxygen concentration, measurement of, 82–83  
evanescent wave spectroscopy, 81–82  
methods, 81  
molecules, measuring other, 83  
optical fiber sensors, 82, 82*f*  
SPR, 82  
Optical displacement sensors, 34–38. *See also* Sensors and transducers  
digital optical encoders. *See* Digital optical encoders  
Optical fiber sensors, 82, 82*f*. *See also* Optical chemical sensors  
Optical prosthetics, 339–340  
Optical switches, 11. *See also* Switches  
Optical telemetry, 271  
Optic nerve stimulation, 387–388  
Optobionics, 375–378  
Organ of Corti, 283  
Orifice plate, differential pressure flowmeter, 62–63, 63*f*  
Oscillating chemical sensors, 79–80, 79*f*, 80*t*. *See also* Chemical sensors  
Osler, William, 497  
Osseointegration, 583–584  
Otto Bock, 554  
Otto Bock 3R15, 543  
Otto Bock 3R22, 543, 576  
Outer ear, 281  
Oximetry, 82–83  
Oxygen uptake test, 490–492
- P**  
Paddle-wheel meters. *See* Impeller flowmeters  
Palmer, Benjamin, 527  
PAM. *See* Pneumatic artificial muscles (PAMs)  
Papillary sphincter muscle, 335  
Paré, Ambroise, 526  
Partial fraction expansion method, 178–179

- Partially implantable hearing aid (PIHA), 304
- Pathsounder, 344
- Pattern recognition, 264, 385, 570
- PCI. *See* Peripheral component interconnect (PCI)
- PCs. *See* Personal computers (PCs)
- PDF. *See* Probability density function (PDF)
- Peak flow(PF), 489, 490
- Peak torque
- BLDC motor selection and, 115–116
- Periodic signals, 210
- Peripheral component interconnect (PCI), 156
- Peristaltic pumps, 405, 406, 462
- Permanent magnet (PM) stepper motors, 117, 118
- Personal computers (PCs)
- isolation barrier, 273
- Personal protective equipment (PPE), 285
- Phosphene, 371
- Photoreceptors, 336
- Physiological systems, 4–5, 4*f*
- biochemical system, 4
  - cardiovascular system, 5
  - musculoskeletal system, 5
  - nervous system, 5
  - respiratory system, 5
- PID. *See* Proportional–integral–derivative (PID) controllers
- Piezoelectric accelerometers, 53–54, 53*f*, 54*f*. *See also* Accelerometers
- Piezoelectric actuators, 126–129, 127*f*, 127*t*, 128*f*
- Piezoelectric devices, 303
- envoy system, 306–307
  - middle ear transducer (MET), 305–306
  - Rion Device E-type (RDE), 304
  - totally integrated cochlear amplifier (TICA), 304
- Piezoelectric effect, 126
- Piezoelectric generators, 21–22
- Piezoelectric (PZT) materials, 341
- properties of, 127*t*
- Piezoelectric tactile sensors, 76. *See also* Tactile sensors
- Planetary gearheads, 150, 151–152, 151*f*, 152*t*
- vs. spur gearheads, 152*t*
- PM. *See* Permanent magnet (PM) stepper motors
- Pneumatic actuators, 139–142
- components of, 139, 139*f*
  - pneumatic muscles, 140–142, 140*f*, 141*t*, 142*f*, 142*t*
- Pneumatic artificial muscles (PAMs), 140–142, 140*f*, 141*t*, 142*f*, 142*t*, 238, 552
- advantage of, 140
  - commercially available, specifications of, 142*t*
  - force-length relationship, 140, 141*f*
  - force-to-weight ratios, 142, 142*t*
- Pneumotachography, 492–494
- Pneumothorax, 477
- Poiseuille's equation, 482
- Poisson distribution
- shot noise, 214, 214*f*
- Poisson's ratio ( $\nu$ ), 23
- Polarization, defined, 84
- Poles, defined, 181
- Polio, 495–497
- Polymer Technology Inc., 429
- Polyvinylidene fluoride (PVDF), 21–22, 21*t*
- Poncho wrap ventilator, 506–507
- Portable respirators, 506–507
- Portable ventilators, 517–519
- Portalung, 506
- Portescap 26DAM series, 125
- Position-sensitive detector (PSD)
- measuring range with, 39–40, 40*f*
- Position sensor
- radio control servos, 132
- Positive displacement pump, 455
- Positive end expiratory pressure (PEEP), 514
- Positive-pressure ventilators
- continuous positive airway pressure, 517
  - controlled mandatory ventilation (CMV), 514
  - historical background, 511–512
  - modes, 513–514
  - need for, 512–513
  - portable ventilators, 517–519
  - pressure-controlled mandatory ventilation, 516–517
  - sleep apnea, 519–520
  - spontaneous ventilation, 517
  - volume-controlled mandatory ventilation, 514–516
- Positive temperature coefficient (PTC) thermistors, 69
- Potentiometers, 27–28, 27*f*, 28*f*. *See also* Resistive displacement sensors
- Potts, James, 527
- Power, 456–457
- Power consumption, 300
- PowerFoot®, 582
- Power spectral density (PSD), 211
- Power supplies, 13–22
- batteries. *See* Batteries
  - energy scavenging. *See* Energy scavenging
  - linear, 13–14, 14*f*, 15*t*
  - switch-mode, 15–16, 15*f*, 16*t*
- Presbycusis, 285
- Pressure
- absolute, 56
  - defined, 56
  - differential, 56
  - gauge, 56
  - measurement, 56–60
  - air pressure, 59–60
  - blood, 56–59
  - fiber optic pressure sensor, 59, 59*f*
  - MEMS strain gauge pressure transducer, 57, 58, 58*f*
  - SM5822 pressure sensor, 60, 60*f*
  - sphygmomanometer, 57
  - strain gauges for, 57
  - units for, 56, 57*t*
  - sensing technologies, properties of, 60*t*
- Pressure-controlled mandatory ventilation, 516–517
- Pressure regulation, 401–402
- long-term, 402
  - short-term, 401
- Pretectum, 338
- Primary visual cortex, 339
- Probability density function (PDF), 211
- Proportional controllers, 161–162, 188–198, 190*f*. *See also* Controllers
- classic configuration of, 189*f*
  - disadvantages of, 198
  - for motor angular position, 197*f*
  - of motor speed, 195*f*
  - motor speed and position control, 189–198, 190*f*, 191*f*, 192*f*, 194*f*–195*f*, 196*f*–197*f*
  - Maxon RE 16 motor. *See* Maxon RE 16 motor
  - plus integral, 198–200, 199*f*
  - with saturation, 189*f*
  - Simulink model of, 195*f*
- Proportional–integral–derivative (PID) controllers, 200–201, 200*f*, 201*f*. *See also* Controllers
- Proprioception, 538
- Prosthesis
- ankle-foot, 580–582
  - hybrid, 541
- Prosthetic arm, 161–162, 162*f*
- Prosthetic arms and hands, control of, 563
- feedback, 574
  - force sensors, 563

- Prosthetic arms and hands (*cont.*)  
 injectable myoelectric sensors (IMES), 573–574  
 linear potentiometers, 563  
 microswitches, 563  
 myoelectric signals, 564–566, 567  
 myoelectric control, 566–570  
 servo assisted cineplasty, 564  
 targeted muscle reinnervation (TMR), 570–573
- Prosthetic hand actuation, 154–157  
 based on electric motors, 155, 156f  
 challenges, 155  
 pneumatic artificial muscles, 156–157, 156f  
 using shape memory alloys, 155, 155f
- Prosthetic limbs, 523  
 active prosthetics, 547  
   ankle–foot mechanisms, 580–582  
   arm mechanisms, 548–554  
   hand mechanisms, 554–562  
   hand research and applications, 562–563  
   leg mechanisms, 574–580  
   prosthetic arms and hands, control of, 563–574  
 arm  
   kinematic model of, 531–532  
   structure of, 529–530  
 brief history, 524–529  
 leg  
   kinematic model of, 534–536  
   structure of, 532–533  
 limb movement, kinematics of, 536  
   angular acceleration, 538  
   complete limb, center of mass and moment of inertia of, 538  
 limb segment, center of mass and moment of inertia of, 536–537
- passive prosthetics, 538  
   foot prosthetics, 545–547  
   knee prosthetics, 542–544  
   upper limb prostheses, actuation and control of, 539–541  
   walking dynamics, 541–542
- prosthesis suspension, 582  
   osseointegration, 583–584  
   suspension methods, 583
- sensing, 538
- PSD. *See* Position-sensitive detector (PSD); Power spectral density (PSD)
- PTC. *See* Positive temperature coefficient (PTC) thermistors
- Pulmonary semilunar valve, 399
- Pulsatile pumps, 417
- Pulsatile pump technology, 429
- Linear Motor Driven LVADs and TAHs, 430–433
- Roller-Screw LVAD, 429–430
- Pulsatile rotary pumps, 444
- Pulsatility index (PI), 435
- Pulse-width modulated (PWM) output, 296
- Pulse-width-to-voltage converter radio control servos, 131
- Pump cycle, 399  
   Stage 1 atrial systole, 399  
   Stage 2 isovolumetric contraction, 400  
   Stage 5 isovolumetric relaxation, 400  
   Stage 3 rapid ejection, 400  
   Stage 6 rapid ventricular filling, 400  
   Stage 4 reduced ejection, 400  
   Stage 7 reduced ventricular filling, 400
- Pump types, 443, 455  
   bearings, 466  
   centrifugal and axial pump characteristics, 456  
   energy, 456–457  
   power, 456–457  
   pressure and head, 457–459  
   specific speed, 459  
   torque, 456–457  
   pulsatile rotary pumps, 444  
   reciprocating pump characteristics, 462  
   diaphragm pumps, 463–465  
   reciprocating piston pumps, 462–463  
   rotary pump characteristics, 460  
   gear pumps, 460–461  
   Lobe pumps, 461  
   peristaltic pumps, 462  
   shape memory alloy driven VAD, 444  
   Sunshine Heart/C-Pulse, 444
- Push-button switches, 10, 10f, 11f. *See also* Switches
- PVDF. *See* Polyvinylidene fluoride (PVDF)
- PZT. *See* Lead zirconate titanate (PZT)
- Q**
- Quartz ( $\text{SiO}_2$ )  
   piezoelectric effect and, 126
- Quasi-static signals, 210
- Quasi-transient signals, 210–211
- Quinine, 286
- R**
- Radio control servos, 131–133, 131f, 132f, 132t, 133f  
   error amplifier, 132  
   Futaba, 133, 133f  
   Hitec, 132f, 132t  
   position sensor, 132
- pulse-width-to-voltage converter, 131
- Radio frequency identification (RFID) tags, 271
- Radio frequency telemetry, 272
- Radius of gyration, 537
- Randomized Evaluation of Mechanical Assistance for the Treatment of Congestive Heart Failure (REMATCH), 417
- Ranging sensors, 38–43. *See also* Sensors and transducers  
   interferometry, 38–39, 39f  
   time-of-flight ranging, 41–43, 42f  
   triangulation, 39–41, 40f, 41f, 42f
- Rapid ejection, 400
- Rapid ventricular filling, 400
- Rate gyros, measuring rate and angular rate, 44–50, 50f  
   Coriolis effect, 46–47, 47f, 48f  
   FOG rate sensor, 48–49, 48f  
   manufacturers of, 49f  
   MEMS, 46–48, 47f  
   MLX90609, 49, 50f  
   precession, 45–46, 45f, 46f  
   restraining precession of, 45, 46f  
   in single gimbal, 44f  
   spinning mass, 46
- RDC. *See* Resolver-to-digital converter (RDC)
- Reaction curve method, 201–202, 202f, 202t
- Rechargeable batteries, 17–18, 18f. *See also* Batteries
- Reciprocating piston pumps, 462–463
- Reciprocating pump characteristics, 462  
   diaphragm pumps, 463–465  
   reciprocating piston pumps, 462–463
- Recording, in biomechatronic systems, 3
- Reduced ejection, 400
- Reduced ventricular filling, 400
- Reed switches, 11. *See also* Switches
- Regression analysis, 264–267, 266f  
   multiple, 266
- Regulators  
   linear, 14, 15t  
   switching, 15  
   three-terminal, 14
- Rehabilitation Institute of Chicago (RIC) prosthetic arm, 571–573
- Relays, switches, 12, 12f. *See also* Switches
- Repetitive signals, 210
- Repetitive transcranial magnetic stimulation (rTMS), 355
- Residual gap  
   in solenoids, 96
- Residual volume (RV), 489

- Resistance-mode linearization, of thermistors, 70, 71*f*
- Resistance temperature detector (RTD), 68–69, 69*f*. *See also* Temperature sensors
- Resistive chemical sensors, 78–79. *See also* Chemical sensors
- Resistive displacement sensors, 22–28. *See also* Sensors and transducers
- potentiometers, 27–28, 27*f*, 28*f*
  - strain gauges, 22–24
  - Wheatstone bridge, measuring with, 24–27, 25*f*
- Resistive tactile sensors, 74–75, 75*f*. *See also* Tactile sensors
- Resolvers, 30–31
- Resolver-to-digital converter (RDC), 31
- Respiratory aids, 471
- breathing, energy required for, 485–488
  - construction, 473–476
  - external negative-pressure ventilation, physics of, 508–511
  - lung characteristics, measuring, 488
  - pneumotachography, 492–494
  - spirometry, 488–492
  - mechanical ventilation, 494
  - alligator cabinet respirator, 505–506
  - Both respirator, 501
  - Drinker respirator, 499–501
  - early history, 494–495
  - Emerson respirator, 504–505
  - external negative-pressure ventilators (ENPVs), 497–499
  - homemade iron lungs, 501–504
  - negative-pressure ventilation, uses for, 507–508
  - polio, 495–497
  - portable respirators, 506–507
- mechanics of respiration, 476
- frictional forces, 481–485
  - inertia, 485
  - lung elasticity, 480–481
  - physical properties, 477–479
- positive-pressure ventilators
- continuous positive airway pressure, 517
  - controlled mandatory ventilation (CMV), 514
  - historical background, 511–512
  - need for, 512–513
  - portable ventilators, 517–519
  - pressure-controlled mandatory ventilation, 516–517
  - sleep apnea, 519–520
  - spontaneous ventilation, 517
  - ventilation modes, 513–514
- volume-controlled mandatory ventilation, 514–516
- Respiratory system, 5
- Retina, 335, 336
- Retinal ganglion cells (RGCs), 337
- Retinal pigmented epithelium (RPE), 336
- Retinitis pigmentosa (RP), 339
- Reynolds number, 483
- RFID. *See* Radio frequency identification (RFID) tags
- Rheo knee, 576, 577–580
- Ring gear, 150, 150*f*
- Rinne tuning fork test, 287
- Rion Device E-type (RDE), 304
- RMS. *See* Root mean square (RMS)
- torque
  - Robert, Dowling, 417
  - Rods, 336, 337
  - Roller pumps. *See* Peristaltic pumps
  - Roller-Screw LVAD, 429–430
  - Roosevelt, Franklin Delano, 496
  - Root locus method, in system stability, 184–187, 185*f*
  - MATLAB version of, 187*f*
  - for quadratic system, 186*f*
  - for simple controller, 184*f*
  - Root mean square (RMS) output noise, 15–16
  - Root mean square (RMS) torque
  - BLDC motor selection and, 116
  - Root mean square (RMS) value, 278
  - Rotameters. *See* Variable-area meters
  - Rotary generators, 21
  - Rotary pump characteristics, 460
  - gear pumps, 460–461
  - Lobe pumps, 461
  - peristaltic pumps, 462
  - Rotary pumps, 417
  - Rotary switches, 11. *See also* Switches
  - Roto-dynamic pumps. *See* Turbo pumps
  - Rotor, defined, 31
  - Roy, Edgar, 502
  - RSL Stepper, 554
  - RTD. *See* Resistance temperature detector (RTD)
  - Running, energy use for, 20
- S**
- Sallen Key filter topology, 233, 233*f*
- Sanyo Denki 103H8222-0941 stepper motor
- specifications of, 124, 124*f*
- Sanyo Denki 103-4902-0650 stepper motor
- specifications of, 123, 123*f*
- Saunders, Rod, 315
- Saurbruch, Ernst, 540
- SAW. *See* Surface acousticwaves (SAW)
- chemical sensors
- Schmitt trigger, 241
- Schottky noise. *See* Shot noise
- Seebeck, Thomas, 72
- Seebeck effect, 72
- Selpho Leg, 527
- Semiconductor sensors, 70, 72. *See also* Temperature sensors
- Semi-implantable middle ear electromagnetic hearing device (SIMEHD), 311
- Semilunar valves, 398
- SensComp 40LT16, 341
- SensComp 600 series, 341
- Sensing mechanisms, 538
- Sensors and transducers, 9–88
- accelerometers. *See* Accelerometers
  - in biomechatronic systems, 3
  - capacitive displacement sensors, 34
  - chemical sensors. *See* Chemical sensors
  - for flow measurement. *See* Flow
  - measurement
  - inductive displacement sensors. *See* Inductive displacement sensors
  - magnetic displacement sensors, 31–34, 32*f*
  - for measuring rate and angular rate, 43–50
  - incremental encoders, 43, 43*f*
  - rate gyros. *See* Rate gyros
  - tachogenerators, 43–44
  - optical chemical sensors. *See* Optical chemical sensors
  - optical displacement sensors. *See* Optical displacement sensors
  - for pressure measurement, 56–60
  - ranging sensors. *See* Ranging sensors
  - resistive displacement sensors. *See* Resistive displacement sensors
  - for sound pressure, 60–61
  - tactile sensors. *See* Tactile sensors
  - temperature sensors. *See* Temperature sensors
  - tilt sensors, 55–56, 55*f*
- Sensory cortex, 358, 359
- Sensory homunculus model, 359
- Sensory substitution, 350–351
- auditory substitution, 351–352
  - definition of, 350
  - electrotactile and vibrotactile transducers, 356
  - input from cameras, 353–355
  - input from sonar, 352–353
- Servo assisted cineplasty, 564
- Servo mechanism, 130

- Servo motors, 130–134  
 overview of, 130–131  
 professional, 134, 134/  
 radio control servos, 131–133, 131/  
 132f, 132t, 133f. *See also* Radio control servos
- Sethi, P.K., 546
- Sewell, William Jr., 419
- Shadow Hand, 156–157, 156f, 557–558
- Shape memory alloy driven VAD, 444
- Shape memory alloys (SMA), 142–145  
 biomechatronic applications, 145  
 principle of operation, 142–144,  
 143f, 144f  
 properties, 142  
 prosthetic hand actuation using,  
 155, 155f
- Sharp Electronics GP2Y0A21YF, 41, 41f
- Shaw, Louis, 498–499
- Shot noise, 214–215, 214f  
 power spectrum for, 215
- Shoulder, 530
- Signal acquisition, 211–224  
 amplifiers, 216–222, 216f  
 charge, 221–222, 222f  
 differential, 218–219, 219f, 220f  
 instrumentation, 219–221, 220f  
 inverting, 217–218, 217f  
 negative feedback, 216–217, 216f  
 noninverting, 218, 218f
- digital signal processing, 241–243,  
 241f, 242f
- noise, 212–216  
 common-mode noise, 215–216, 215f  
 differential mode noise, 215–216  
 shot noise, 214–215, 214f  
 thermal noise, 212–213, 213f
- op amp specifications, 223–224, 223t  
 practical considerations, 222–223, 223f
- Signal aliasing, 245–247, 246f–248f
- Signal compression, 300
- Signal processing  
 analog. *See* Analog signal processing  
 biomedical signals. *See* Biomedical signals  
 data mining, 267  
 digital. *See* Digital signal processing (DSP)  
 elements in biomechatronic systems, 3  
 isolation barrier, 270–274  
 implant systems, 270–272. *See also* Implant systems  
 isolation amplifiers, 272–274, 274f
- machine learning, 267–270  
 overview of, 207
- signal acquisition. *See* Signal acquisition
- statistical techniques, 264–267  
 regression analysis, 264–267, 266f
- Signal processing and cochlear stimulation, 320
- band-pass filter bank, 320–321
- continuous interleaved sampling (CIS), 321  
 compression function, 323–324  
 design parameters, 323  
 envelope detection, 324–325  
 filter spacing, 325  
 stimulation rate, 323
- Signal processor, 372–373
- Silver oxide battery, 17. *See also* Batteries
- Simmons, F. Blair, 314
- Single-acting cylinder  
 hydraulic actuators, 138, 139f
- Single-coil DC motor, 100, 101f
- Single pole double throw (SPDT) devices, 10, 10f
- Single pole single throw (SPST) devices, 10, 10f
- Single-sided deafness (SSD), 301
- Sleep apnea, 519–520
- Slider, defined, 31
- SMA. *See* Shape memory alloys (SMA)
- Smeaton, John, 511
- Smith-Clarke “alligator” cabinet respirator, 505, 506
- SM5822 pressure sensor, 60, 60f
- Solenoids, 94–96, 94f, 95f, 96f  
 biomechatronic applications, 99  
 force–distance characteristics of, 96, 96f
- Solid-state relays (SSR), 12, 13f
- Sonar, input from, 352–353
- Sonar-based systems, 341–343  
 Binaural Sonic Aid (Sonicguide), 345  
 GuideCane, 348–349  
 issues with, 350  
 Mowat sensor and derivatives, 344  
 NavBelt, 345–348  
 pathsounder, 344  
 Sonic Pathfinder, 344–345
- Sonicguide, 352
- Sonic Pathfinder, 344–345
- Sound, 278  
 characteristic impedance and sound pressure, 278–279  
 intensity, 279  
 pressure, 60–61, 278–279
- Sound pressure level (SPL), 279
- Soundtec device, 310
- Soundtec direct system, 309–311
- SPDT. *See* Single pole double throw (SPDT) devices
- SPEAK strategies, 326
- Specific absorption rate (SAR), 454, 455
- Spectral estimation  
 digital filters, 260–264, 261f  
 Fourier transform, 261–264, 263f, 264f
- Speech threshold audiometry, 286
- Speed  
 BLDC motor selection and, 116  
 induction motors, 135, 135f
- Sphygmology, 396
- Sphygmomanometer  
 for blood pressure, 57
- Sphygmos, 396
- SPICE software, 171
- Spinning mass rate gyros, 46
- Spirometry, 488  
 oxygen uptake test, 490–492  
 volume fraction definition, 489  
 volume tests, 489–490
- Split hooks, 540
- Spontaneous ventilation, 517
- Spool valves  
 hydraulic actuators, 137–138, 138f
- SPR. *See* Surface plasmon resonance (SPR)
- SPST. *See* Single pole single throw (SPST) devices
- Spur gearheads, 150, 151–152  
 vs. planetary gearheads, 152t
- SSR. *See* Solid-state relays (SSR)
- Stanford University retinal implant, 379
- Stapedectomy, 312
- Starling law, 417
- State-variable filter topology, 233, 233f
- Static balanced mode, 24–26, 25f
- Static characteristics of stepper motors, 120, 121f
- Static position error of stepper motors, 123
- Static signals, 210
- Statistical techniques  
 signal processing, 264–267
- Steady-state error, 188, 188f
- Steffen, Simon, 562
- Stepper motors, 117–124  
 biomechatronic applications, 124  
 bipolar coils in, 118  
 characteristics of, 120–123  
 dynamic characteristics, 122, 122f  
 magnets, 123  
 static characteristics, 120, 121f  
 static position error, 123  
 step response, 122, 122f
- conventional, 120, 121f
- drivers, 118–120, 119f, 120t
- hybrid, 117, 118

- operational principle, 118, 118*f*  
 permanent magnet, 117, 118  
 types, 123–124, 123*f*, 124*f*  
 unipolar coils in, 118–119, 119*f*  
 variable reluctance, 117, 118
- Step response  
 digital filters, 259, 260*f*  
 of stepper motors, 122, 122*f*
- Stereocilia, 283
- Steuart, W., 498
- Stimulator of Tactile Receptors by Skin Stretch (STRoSS), 362
- Stimulus/actuation, in biomechatronic systems, 3
- Stimulus mapping, 330
- Stochastic signals, 211
- Strain gauges, 22–24  
 small displacements with, measuring, 57  
 wire, 57
- Submarine-launched ballistic missiles (SLBM), 46
- Subretinal implants, 375  
 Boston retinal implant project, 379–381  
 Optobionics, 375–378  
 research, in Germany, 378–379  
 Stanford University retinal implant, 379
- Summing amplifier, 234–235, 234*f*
- Sun gear, 150, 150*f*
- Sunshine Heart/C-Pulse, 444
- Superior colliculi, 338–339
- Supervised learning, 267
- Support vector machines (SVMs), 269–270, 269*f*
- Surface acousticwaves (SAW) chemical sensors, 79–80, 79*f*, 80*t*
- Surface plasmon resonance (SPR), 82
- Suspension methods, for prostheses, 583
- SVMs. *See* Support vector machines (SVMs)
- Swing-arm actuators, 96
- Switches, 9–13, 10*f*  
 analog, 12, 13*f*  
 hall, 11  
 limit, 10–11, 11*f*, 12, 161, 162*f*  
 optical, 11  
 push-button, 10, 10*f*, 11*f*  
 reed, 11  
 relays, 12, 12*f*  
 rotary, 11  
 toggle, 10, 10*f*
- Switching regulators, 15
- Switch-mode power supplies, 15–16, 15*f*, 16*t*. *See also* Power supplies
- Syncardia Systems, 412
- Syncardia Systems/CardioWest, 412–414
- System  
 closed-loop, 159, 162, 163  
 defined, 159  
 models of. *See* System models  
 open-loop, 159, 162, 163  
 representation of, 162–164, 162*f*  
 response. *See* System response  
 stability. *See* System stability
- System for wearable audio navigation (SWAN), 370
- System models, 164–174  
 comparison, 171, 171*f*, 172*t*  
 direct current motor model, 170–171  
 electrical elements in, 168–169  
 electrical model, 169–171, 169*f*, 170*t*  
 fluid flow elements in, 171–174, 172*f*, 173*f*, 174*t*  
 mechanical elements in, 164–166, 166*t*  
 mechanical model, 166–167, 166*f*, 167*f*
- System response, 174–181, 175*f*, 176*f*, 176*t*  
 analyzing complex models, 179–181, 180*f*, 181*f*  
 float controller reservoir, 175, 175*f*  
 Laplace transform in, 175–178, 176*f*, 176*t*, 177*t*  
 partial fraction expansion in, 178–179  
 pole position and, 183*f*  
 RC Circuit, response of, 178
- System stability, 181–186, 183*f*  
 poles in, 181, 182, 183*f*  
 root locus method in. *See* Root locus method  
 steady-state error in, 188, 188*f*  
 zeros in, 181–182, 182*f*
- Systole, 401
- T**
- Tachogenerators  
 measuring rate and angular rate, 43–44
- Tactile display technologies and interaction, 360
- Tactile sensing, 73–77, 74*f*
- Tactile sensors, 73–77, 74*f*. *See also* Sensors and transducers  
 capacitive, 75–76, 75*f*  
 carbon nanotubes, 76, 76*f*  
 CMC, 76–77, 76*f*, 77*f*  
 piezoelectric, 76  
 resistive, 74–75, 75*f*
- Tactile vision substitution (TVS) devices, 356
- Targeted muscle reinnervation (TMR), 570–573
- Target flowmeters, 65–66. *See also* Flow measurement
- Telemetry  
 back, 272, 273*f*  
 inductive, 271–272, 272*f*  
 optical, 271  
 and power interface, 373  
 radio frequency, 272
- Temperature flowmeters, 64. *See also* Flow measurement
- Temperature sensors, 67–73. *See also* Sensors and transducers  
 bimetallic thermostats, 68, 68*f*  
 LM35Z, 72  
 RTD, 68–69, 69*f*  
 semiconductor sensors, 70, 72  
 thermistors. *See* Thermistors  
 thermocouples, 72–73, 72*f*, 73*t*
- Terminal devices (TDs), 554
- Terminal swing  
 leg mechanisms, 575  
 walking dynamics, 536, 542
- Terumo Heart, Inc., 439
- Thermal imagers, 340
- Thermal noise, 212–213, 213*f*
- Thermistors, 69–70. *See also* Temperature sensors  
 linearization of  
 resistance-mode, 70, 71*f*  
 voltage-mode, 70, 71*f*  
 NTC, 69–70, 70*f*  
 PTC, 69
- Thermocouples, 72–73, 72*f*, 73*t*. *See also* Temperature sensors
- Thermotactile display, 370  
 outputs, 370
- Thoratec/Heartmate, 424, 450, 451
- Thoratec/Heartmate II, 433–435, 450, 451
- Three-terminal regulators, 14
- Tidal volume (TV), 489
- Tilt sensors, 55–56, 55*f*. *See also* Sensors and transducers
- Time-cycled ventilation, 518
- Time-domain response, digital filters, 258–259, 258*f*, 259*f*, 260*f*  
 impulse response, 259, 259*f*  
 step response, 259, 260*f*
- Time-of-flight ranging, 41–43, 42*f*
- Toggle switches, 10, 10*f*. *See also* Switches
- Torque, 456–457  
 induction motors, 135–136, 136*f*
- Total lung capacity (TLC), 489
- Totally artificial heart (TAH), 408
- Totally integrated cochlear amplifier (TICA), 304
- Touch Bionics, 554, 561

- Touch EMAS (Edinburgh Modular Arm System), 561
- Trachea and bronchi, 474–475
- Tracking filter  
digital filters, 251–254, 253<sup>f</sup>
- Transcutaneous energy transfer (TET), 452–454
- components of, 453
- coupling method of, 454
- Transducers and sensors. *See* Sensors and transducers
- Transfemoral amputations, 535
- Transfer functions  
digital filters, 248–249, 250<sup>f</sup>
- Transient signals, 210–211
- Translation screw devices, 153–154, 153<sup>f</sup>
- Transtibial amputations, 535
- Triangulation sensors, 39–41, 40<sup>f</sup>, 41<sup>f</sup>, 42<sup>f</sup>
- Turbine flowmeters, 66–67, 66<sup>f</sup>. *See also* Flow measurement
- Turbo pumps, 455
- Turbulent flow, in airway, 483, 484–485
- Two-point discrimination threshold (TPDT), 358
- Tympanometry, 287
- Tympanostomy, 288
- U**
- Ultralight hand, 556
- Ultrasound flowmeters, 64–65. *See also* Flow measurement
- Ultrasound transducers, 60–61
- Unipolar coils in stepper motors, 118–119, 119<sup>f</sup>
- University of Bologna (UB) hand, 558–559
- Unsupervised learning, 267
- Upper limb prosthesis  
actuation and control of, 539–541
- types of, 547
- Utah Arm, 548, 549, 568, 570, 571
- Utah electrode array (UEA), 389
- Utah slanted electrode array (USEA), 389
- V**
- Vacuum jacket, 495, 496
- Vacuum tubes, 290
- VADs. *See* Ventricular assist devices (VADs)
- Vagus nerve, 401
- Vancomycin, 286
- Vanderbilt arm, 552, 553
- Variable-area meters, 66, 66<sup>f</sup>
- Variable reluctance (VR) stepper motors, 117, 118
- Veins, 397
- VentraCo/VentrAssist, 435–437  
components of, 436
- Ventral cochlear nucleus (VCN), 328
- Ventricular assist devices (VADs), 99, 417–418
- CorAide, 439
- EvaHeart, 439
- extracorporeal ventricular assist devices, 420–421
- generation 1 LVADs, 421  
Arrow/LionHeart, 425–427
- Thoratec/Heartmate, 424
- WorldHeart/Novacor, 421–423
- WorldHeart/Novacor II, 427–428
- generation 2 VADs, 433  
Thoratec/Heartmate II, 433–435
- generation 3 VADs, 435  
Mohawk Innovative Technology/MiTHeart, 437–439
- VentraCo/VentrAssist, 435–437
- generation 4 VADs, 439  
AbioMed/Impella, 442–443
- HeartWare LVAD, 442
- Jarvik-2000, 440–441
- Micromed/DeBakey LVAD, 441–442
- history, 419–420
- ideal replacement heart, 443
- intracorporeal left ventricular assist devices, 421
- pulsatile pump technology, 429  
Linear Motor Driven LVADs and TAHs, 430–433
- Roller-Screw LVAD, 429–430
- pump types, 443  
pulsatile rotary pumps, 444
- shape memory alloy driven VAD, 444
- Sunshine Heart/C-Pulse, 444
- Terumo Heart, Inc., 439
- Venturi tube, differential pressure flowmeter, 63
- Verduyn, Pieter, 527
- Versalius, 472
- Very-large-scale integration (VLSI)  
chips, 374
- Very-large-scale integration (VLSI)  
processes, 297–298
- Vibrant Med-El, 307
- Vibrant soundbridge device, 307–309
- Vibrating ossicular prosthesis (VORP), 307
- Vibrotactile haptic device, 363
- Vibrotactile stimulation, 356
- Video encoder, 372
- Viscous friction, 481
- Visual cortex implants, 388  
electrode arrays, penetrating, 388–389
- Illinois Institute of Technology Research Group, 389–391
- Visual neuroprostheses  
alternative implants, 386–387
- components, 372  
neural interface, 374
- neural stimulator, 373–374
- signal processor, 372–373
- telemetry and power interface, 373  
video encoder, 372
- epiretinal implants, 381
- historical perspective, 371
- optic nerve stimulation, 387–388
- potential sites for, 371–372
- subretinal implants, 375
- visual cortex implants, 388
- worldwide research activity, 375
- Visual prostheses, 333  
blindness, causes of, 339
- common sites for, 375
- future, 391
- GPS-based systems, 370–371
- laser-based systems, 350
- optical prosthetics, 339–340
- sensory substitution and, 350–351
- auditory substitution, 351–352
- electrotactile and vibrotactile transducers, 356
- input from cameras, 353–355
- input from sonar, 352–353
- sonar-based systems, 341–343
- Binaural Sonic Aid (Sonicguide), 345
- GuideCane, 348–349
- issues with, 350
- Mowat sensor and derivatives, 344
- NavBelt, 345–348
- pathsounder, 344
- Sonic Pathfinder, 344–345
- visual pathway  
anatomy and physiology of, 335–339
- Vital capacity (VC), 489
- Vitreous humor, 336
- VOICE, 352–353, 355
- Voice coil actuators, 96–99, 97<sup>f</sup>, 97<sup>t</sup>, 98<sup>f</sup>, 99<sup>f</sup>  
biomechatronic applications, 99  
range of, 98, 99<sup>f</sup>
- Voltage-mode linearization, of thermistors, 70, 71<sup>f</sup>
- Volume-controlled mandatory ventilation, 514–516

Volume-cycled ventilation, 518  
VR. *See* Variable reluctance (VR) stepper motors

**W**

Walking, 534–535  
  dynamics, 535–536, 541–542  
  energy use for, 20  
Wang Shu-he, 396  
  blood circulation model of, 396

Watson, Thomas, 403  
Western Electric Model 134, 290  
Wet cell devices, 17  
Wheatstone bridge, measuring with,  
  24–27, 25*f*  
White cane, 334  
White noise, 213, 215  
Woillez, E.J., 495  
WorldHeart/Novacor, 421–423  
WorldHeart/Novacor II, 427–428

Worm gear drive actuator, 125, 125*f*  
Wrist, 529

**Z**

Zeros, defined, 181  
Zinc air battery, 17. *See also* Batteries  
Zinc carbon battery, 17. *See also* Batteries  
Zinc chloride battery, 17. *See also*  
  Batteries



## INTRODUCTION TO BIOMECHATRONICS

This text/reference provides fundamental knowledge of mechanical and electronic (mechatronic) components and systems and their interaction with human biology to assist or replace limbs, senses, and even organs damaged by trauma, birth defects, or disease.

The first half of the book provides the engineering background to understand all the components of a biomechatronic system: the human subject, stimulus or actuation, transducers and sensors, signal conditioning elements, recording and display, and feedback elements. It also includes the major functional systems of the body to which biomechatronics can be applied including:

- Biomechanical
- Nervous
- Cardiovascular
- Respiratory
- Musculoskeletal

The second half discusses five broadly based devices from a historical perspective and supported by the relevant technical detail and engineering analysis. These devices include hearing prostheses, sensory substitution and visual prostheses, artificial hearts, respiratory aids, and artificial limbs.

**Introduction to Biomechatronics** provides readers with the fundamental engineering (biomedical, mechanical, electronic) background to analyze and design biomechatronic devices and will inspire greater designs by discussing successful inventions that have done the most to improve our lives.

### ABOUT THE AUTHOR

Graham Brooker is a Senior Lecturer at the Australian Centre for Field Robotics at the University of Sydney. While completing his baccalaureate in electrical engineering, he developed a myoelectric controlled rehabilitative exercise device using an early microprocessor. During his two years of compulsory national service, his passion turned toward radar, which he continued in for 20 years until he left industry for academia in 1999. While completing his Ph.D. at the Centre for Field Robotics, he conducted research and lectured in sensors to reestablish his biomedical credentials. In 2007, he developed a course in Biomechatronics which has been offered as a final year elective for mechatronic and biomedical engineering students. This book has evolved from that course.

ISBN 978-1-891121-27-2



Raleigh, NC  
[www.scitechpub.com](http://www.scitechpub.com)

